ATTITUDE REPORTS AND RELATIVE BELIEF:

A NEW PERSPECTIVE ON THE SWAHILI DUAL-COMPLEMENTIZER SYSTEM

By

Aron David Finholt

M.A. Research Project in Linguistics

University of Kansas

February 2021

1. Introduction

Swahili (Bantu) is descriptively said to make use of two lexically distinct but functionally identical complementizers, *kwamba*, and *kuwa*, to introduce a finite indicative clause under a clause-embedding predicate (1).

(1) Hamisi a-li-ni-ambia kwamba/kuwa a-na-penda kusoma
Hamisi 1SM-PAST-1SG.OM-tell COMP/COMP 1sm-pres-like read.INF
"Hamisi told me that he likes to read"

(Mpiranya, 2015)

As such, Swahili is similar to so-called 'split-complementizer' (also 'dual-complementizer') languages – like Greek (Roussou, 2010) and certain dialects of Italian (Cruschina 2006; Vecchio, 2010; Ledgeway & D'Allesandro, 2010) – which are reported to make use of multiple distinct complementizers to introduce an embedded clause. In contrast to the superficially similar complementizer systems of Greek or Italian dialects like Francavilla (Vecchio, 2010), however, the Swahili complementizers *kwamba/kuwa* are not reported to coincide with any distributional or interpretive distinctions – properties that have been regularly attributed to split/dual-complementizer systems (Givon & Kimenyi, 1974; Cruschina 2006; Roussou, 2010; Vecchio, 2010; Angelopoulos, 2019). Instead, the two complementizers are reported to exist in free variation, with no interpretive differences arising from the use of one or the other (Ashton, 1944; Thompson & Schleicher, 2006).

In this paper, I investigate the distribution of the Swahili complementizers *kwamba* and *kuwa* under clause-embedding predicates. Specifically, I employ a regression-based analysis of (Tanzanian) Swahili corpus data to examine the claim that these two complementizers exist in free variation. Ultimately, the results of the regression analysis suggest that treating *kwamba/kuwa* as being in free variation is insufficient, as both complementizers are shown to correlate with distinct elements in the main clause that have been argued to influence complementizer choice cross-linguistically (Kiparsky & Kiparsky, 1971; Hooper & Thompson, 1973; Givon & Kimenyi, 1974; Roussou, 2010). Notably, *kwamba* is shown to correlate with first-person subjects, attitude predicates, and the presence of matrix negation, while *kuwa* is shown to correlate with third-person subjects and reportative predicates. On the basis of these results, I reject the notion that *kwamba/kuwa* are freely interchangeable, and instead lay out an

analysis of the Swahili complementizer system in which complementizer choice encodes relative belief. I argue that by using *kwamba*, the speaker conveys that someone, either the local subject or the speaker themselves, holds a specific belief about the embedded proposition, P; namely, that P is true. Conversely, I argue that *kuwa* is used to stay neutral about the attitude holder's commitment to the truth of P, likely to avoid overcommitment to a particular attitude report or to express speaker uncertainty. I further attempt to integrate this analysis into the existing theoretical landscape at the left-periphery, focusing specifically on whether *kwamba/kuwa* can be attributed to two distinct projections in the C-domain (Rizzi, 1997) – as has been proposed of similar (split) complementizer systems in Greek (Roussou, 2010) and various dialects of Italian (Cruschina 2006; Vecchio, 2010, Ledgeway & D'Allesandro, 2010). Ultimately, I find that the Swahili corpus data does not necessarily lend itself to such an analysis, though I leave open the possibility that the two complementizers may indeed be structurally distinct.

2. Background on Swahili

Swahili, or *Kiswahili*, as it is referred to by its speakers (*Kiswahili* being derived from the demonym, *Swahili*, Arabic for 'coast dwellers'), is a Northeast Coast Bantu language native to the coastal regions of Tanzania, Mozambique, and Kenya.[1] Though predominantly used as a lingua franca in the region with over 100 million L2 speakers throughout East Africa (Hinnebusch, 2003), Swahili is spoken natively by a population of roughly 2 million speakers, most of whom reside in Tanzania. Descriptively, Swahili exhibits a base word order of SVO and is highly agglutinative in the verbal domain. As such, Swahili has a rich system of inflectional and derivational morphology – one that is exemplified by its various verbal suffixes (called "extensions" in the Bantuist tradition), which serve to (minimally) encode subject agreement, object agreement, tense, aspect, mood. Broadly, the Swahili verb complex is comprised of (from left-to-right) the subject marker, tense marker, object marker, the verb root, valency related

[1]*Kiswahili* is derived from the demonym, *Swahili*, 'coast dwellers', through the affixation of the Bantu noun class 7 prefix, *ki-*. Though not universal, the noun class 7 marker is used pervasively throughout the Bantu languages as a 'language' marker; *ki-* often affixes to a general demonym to denote the language of a particular group. Swahili employs this derivational strategy quite readily, as can be illustrated in comparing the word for the English language, *ki-ingereza*, with the word for the English people, *wa-ingereza* (*wa-* being the noun class 2 marker used for plural persons).

affixes like applicatives, causatives, and passives, and the final vowel – which generally reflects mood. Example (2) below provides a concise illustration of Swahili's verbal morphology as well as its base (SVO) word order.[2]

(2) Juma a-li-wa-nunu-li-a watoto vitabu
Juma 1SM-PAST-2OM-buy-APPL-FV 2.child 8.book
"Juma bought the children books" (Ud Deen, 2002)

Like other Bantu languages, Swahili also makes use of an extensive noun class system, which broadly serves to group nouns based on (loose) semantic notions and/or certain lexical features (e.g. number) – though noun class assignment is largely arbitrary, much like the prototypical grammatical gender systems of various Indo-European languages (Ud Deen, 2002). For the purposes of this project, we need only describe the properties of noun class 1 and 2, as these two classes strictly refer to people (and as such, potential attitude holders). Critically, these two noun classes differ only with respect to number, with all singular [+human] nouns belonging to noun class 1, and all plural [+human] nouns belonging to noun class 2. Noun class identity (in addition to featural information pertaining to number) is encoded as a noun class marker that is prefixed to the lexical root (3) - (4).

(3) m-tu
1-person
'person'

(4) wa-tu
2-person
'people' (Meinhof, 1932)

As described in (2) above, subject agreement on the Swahili verb is realized as noun class agreement, with each noun class corresponding to a different subject prefix. Though the agreeing subject prefixes for noun class 1 and 2 are broadly used for [+human] subjects, first and second-

[2]Unless otherwise noted, all Swahili examples were taken from the Helsinki Corpus of Swahili 2.0 (CITE). Given the nature of the corpus, these examples were manually translated and glossed based on the information provided by the corpus. To ensure glossing uniformity, all examples – including those exterior to the corpus – were glossed using the following template:

SM- subject marker, OM- object marker, SG- singular, PL-plural, PAST- past tense, FUT- future tense, PRES- present tense, CONS- consecutive marker, PFV-perfective aspect, IRR- irrealis mood, SUBJ- subjunctive mood, APPL- applicative, NEG- negation, CAUS- causative suffix, PASS- passive suffix, GEN- genitive, POSS- possessive marker, FV- final vowel (mood), COMP- complementizer.

person subjects exhibit a distinct agreement pattern that varies by person and number, as shown in Table 1.[3]

Table 1. *Swahili Subject-Agreement Paradigm*

	singular	*plural*
first-person	ni-	tu-
second-person	u-	m-
third-person	a- (Noun Class 1)	wa- (Noun Class 2)

2.1 Clause Embedding

In complementation contexts involving clause-embedding predicates such as -*sema*, 'say', Swahili employs (at least) four different methods of introducing a finite embedded clause. In such contexts, embedded clauses may be introduced by one of three distinct lexical complementizers – *kuwa*, *kwamba,* or *eti* (5) - (7) – or by a null complementizer (8).

(5) Juma a-li-sema *kuwa* a-nge-kuja kesho yake
Juma 1SM-PAST-say COMP 1sm-IRR-come tomorrow 3SG.POSS
"Juma said that he would come the following day"

(6) Mwalimu a-li-sema *kwamba* a-ta-safiri wiki ijayo
professor 1SM-PAST-say COMP 1sm-FUT-travel week next
"The professor said that he will travel next week:

(7) Fatma a-me-sema *eti* h-a-ku-mw-ona mgeni
Fatma 1SM-PFV-say COMP NEG-1SM-PAST-1om-see stranger
"Fatma said that she didn't see the stranger"

[3]Given that first and second-person subject agreement does not follow the same agreement pattern as noun class 1/2, glossing will be coded slightly differently than a normal (noun class) subject markers, as is the Bantuist tradition. First and second person subject agreement will instead be glossed based on person and number (e.g. 1SG, 1PL, etc.).

(8) Daudi a-li-ni-ambia Ø a-nge-kuja kesho yake
Daudi 1SM-PAST-1SG.OM-tell COMP 1SM-IRR-come tomorrow poss
"Daudi told me that he would come the following day

(Massamba, 1986)

Though these four different embedding strategies may appear similar at first glance (e.g. they all introduce an embedded finite clause), it has been reported that *eti* (7) and the null complementizer (8) exhibit distinct selectional and/or interpretive properties. Massamba (1986) specifically reports that these two complementizers differ from *kwamba/kuwa* in that *eti* uniquely encodes pragmatic information pertaining to speaker attitude – something which has never been reported for *kwamba/kuwa* – while the null-complementizer contrastively exhibits distributional constraints that do not hold of the lexical complementizers (9a-b). We will return to the pragmatic function of *eti* in Section 7.2.

(9) a. Amina a-li-ni-faham-isha *kwamba* a-nge-kuja kesho yake
Amina 1SM-PAST-1SG.OM-know-CAUS COMP 1SM-IRR-come tomorrow POSS
"Amina informed me that she would come the following day

b. *Amina a-li-ni-fahami-sha Ø a-nge-kuja kesho yake

(Massamba, 1986)

Conversely, *kwamba* and *kuwa* have been descriptively and prescriptively described as freely-interchangeable (Mpiranya, 2015), with no distributional or interpretive distinctions between them. Notably, the fact that these complementizers are reported to be in free-variation is prima facie surprising, particularly given that they are both used productively outside of complementation contexts as distinct lexical verbs. This is perhaps most relevant with *kuwa*, which serves as the infinitive 'to be' (*ku-* being the infinitival marker); *kwamba* is largely limited to its applicative form, *kwambia*, 'to tell', in Modern Swahili. Nonetheless, it is generally accepted that their roles as complementizers are functionally synonymous, and further, that they do not coincide with any differences in interpretation.

3. Literature Review

In complementation constructions involving clause-embedding predicates such as *think* or *believe*, the role of a complementizer can be understood syntactically as a subordinating element that introduces the embedded finite clause, and semantically as a mediating element that relates the propositional content of the subordinate clause to the superordinate matrix clause. The syntactic and semantic functions of the complementizer can be illustrated by the following examples. Specifically, we see that in (10), the declarative complementizer *that* is possible following the clause-embedding verb *believe* but impossible following the inherently interrogative clause-embedding verb, *inquire*. Conversely, in (11), we see that the opposite is true; only the inherently interrogative predicate, *inquire*, is compatible with the interrogative complementizer, *whether*.

(10) We {believed/*inquired} that he was there.
(11) We {*believed/inquired} whether he was there. (Bresnan, 1970)

The fact that the distribution of the English complementizers *that* and *whether* correlate with the inherent semantic features of the matrix predicate (e.g. whether the predicate is interrogative or not) suggests that, at some level, the choice in complementizer must be sensitive to information relative to both its syntactic function as the interface between a matrix clause and its complement, and also its semantic function as the interface between the propositional content of the two clauses. Put differently, one could describe the ungrammaticality of the V-C pair *inquired-that* as the result of a semantic incompatibility between the selectional stipulations of the matrix predicate and the semantic content of the embedded clause; *inquire* selects for a non-propositional (interrogative) clause, and is therefore incompatible with the *that*-clause, [*that he was there*], which denotes a proposition that is evaluated as true or false. Likewise, one could attribute the incompatibility of the V-C pair, *believed-whether,* to the fact that unlike *inquire*, the predicate *believe* cannot select for an interrogative clause. Instead, it must select for a clause containing a proposition which the speaker/attitude holder can hold a belief about, meaning that it is incompatible with the interrogative complementizer, *whether*.

The observation that complementizers may encode both syntactic and semantic information has led many to treat the complementizer as a window into the syntax-semantics interface. As

such, a number of proposals have attempted to draw direct correlations between the semantic function of complementizers and the syntactic configuration in which they are ultimately realized (Rizzi, 1997; Vecchio, 2010; Ledgeway & D'Allesandro, 2010). One area of fruitful discussion on this topic concerns languages, unlike English, that exhibit multiple distinct complementizers to introduce an embedded clause. For languages of this type, e.g., Greek (Roussou, 2010; Kinyarwanda (Givon & Kimenyi, 1974), and certain dialects of Italian (Cruschina 2006; Vecchio, 2010), it has been observed that there exist strong correlations between predicate classes and the complementizer(s) they appear with. Specifically, it has been reported that, in number of these 'dual-complementizer' languages, complementizer distribution is sensitive to certain semantic properties of the embedding predicate; most notably whether the predicate presupposes the proposition of its complement (factivity), and/or whether the predicate asserts some sort of epistemic or doxastic evaluation by the speaker or attitude holder about the embedded proposition. In Greek, for example, the distribution of the complementizers *pu* and *oti* differ only in that the latter cannot co-occur with an emotive factive predicate, like *lipate*, 'be sad' (12).

(12)	I	Elena	lipate	pu/*oti	dhen	perase	tis	eksetasis
	the	Elena	be.sad.3Sg	C	not	passed.3Sg	the	exams

"Elena was sad that she didn't pass the exams." (Angelopoulos, 2019)

In addition to predicate class, other elements in the matrix clause, such as the person specification of the matrix subject (Givon & Kimenyi, 1974), or the presence of a matrix operator (Vecchio, 2010), among others, have been shown to influence complementizer choice cross-linguistically. Furthermore, such interactions between matrix elements and the individual complementizers making up these dual-complementizer systems have been shown to exhibit clear interpretive differences (Karttunen, 1971; Beaver, 2010; Angelopoulos, 2019; Djärv, 2019).

3.1 Predicate Class

One particular factor that has often been noted to influence complementizer choice is the lexical semantics of the embedding predicate (Kiparsky & Kiparsky, 1971; Hooper & Thompson, 1973). The relationship between embedding predicate and complementizer selection is most famously attested in the dual-complementizer system of Modern Greek, henceforth, MG (Roussou, 2010; Angelopoulos, 2019), a language that is reported to distinguish its complementizers based on factivity. Based on examples like (12) – repeated below as (13) – it has been argued that the two MG complementizers *oti* and *pu* differ with respect to the information they encode relative to the truth of the embedded clause. Specifically, it is argued that *pu*-clauses are factive, and necessarily presuppose the truth of the proposition of the clause they introduce, while *oti*-clauses do not. This distinction can be illustrated by the fact that the emotive factive predicate, *lipate,* 'be sad', strictly selects for *pu* and not *oti*.

(13)	I	Elena	lipate	pu/*oti	dhen	perase	tis	eksetasis
	the	Elena	be.sad.3Sg	C	not	passed.3Sg	the	exams
	"Elena was sad that she didn't pass the exams."							(Angelopoulos, 2019)

While this pattern of asymmetric selection holds for predicates that inherently presuppose the truth of their complement, other predicates seem to be able to select for either *pu* or *oti* (14).

(14)	I	Elena	thimotan	pu/oti	milise	s-ton	Jorgho
	the	Elena	remembered.3Sg	C	talked.3Sg	to-the	George
	"Elena remembered that she talked to George."						(Angelopoulos, 2019)

Even in such cases of apparent selectional alternation, there has been shown to be subtle interpretive differences between an embedded *pu*-clause and an embedded *oti*-clause that are derived from the fact that, as shown in (13), *pu*-clauses encode factivity. With respect to the predicate in (14), the effects of this presuppositional dichotomy result in the proposition *she talked to George* being necessarily true when 'remember' selects for a *pu*-clause, but not when selecting an *oti*-clause; only with *oti* can the proposition be directly falsified by a contrastive continuation, suggesting that the speaker is only committed to the truth of the proposition if it is contained within a *pu*-clause (15b), but not an *oti*-clause (15a).

(15) I Elena thimotan …
the Elena remembered.3Sg

a.) oti milise s-ton Jorgho …(but she didn't actually talk to him)
C talked.3Sg to-the George

b.) pu milise s-ton Jorgho …*(but she didn't actually talk to him)
C talked.3Sg to-the George

"Elena remembered that she talked to George." (Angelopoulos, 2019)

As such, complementizer selection in Modern Greek has been argued to be directly influenced by the semantic properties of the embedding predicate; in cases where the embedding predicate presupposes the propositional content of the embedded clause, only *pu* is available. Correspondingly, this selectional distinction leads to interpretive differences in cases where both *pu* and *oti* are available (e.g. with non-factive predicates), with *pu* yielding a reading in which the speaker commits themselves to the truth of the embedded proposition.

Though Modern Greek is an illustrative example of how predicate class can influence complementizer choice, it is important to remember that MG is but one of many different dual-complementizer systems. As such, we cannot assume that the interaction of predicate class and complementizer choice is necessarily the same in MG as it is in other dual-complementizer languages. Indeed, there are many different ways to classify embedding predicates, and it's not the case that all these classifications align – even in a single language. For instance, there have been a number of proposed predicate classifications concerned with the availability of so-called 'Main Clause Phenomena', or MCP, (e.g. adverbial preposing, V-to-C movement, tag questions, *wh*-extraction, etc.) in English embedded clauses (Emonds, 1970; Hooper & Thompson, 1973; Green, 1973). Much like the dual-complementizer system in Modern Greek, the availability of MCP in English has been argued to show sensitivity to predicate class, suggesting that this particular property is also linked to the clausal left-periphery. Interestingly, many of the proposed predicate classifications concerned with MCP in English rely on a similar distinction to what we have seen in Modern Greek; that is, a number of proposals distinguish predicates based on factivity (Kiparsky & Kiparsky, 1971), i.e. whether they presuppose the truth of the embedded clause. With that being said however, there has been considerable debate on the exact

classification of such clause embedding predicates in English, particularly with respect to the specific semantic properties that define the various predicate classes and/or license the availability of MCP. Take, for example, the subtle distinction between Hooper and Thompson's (1973) classification of English complement taking predicates, and Kastner's (2015) classification. Unlike the original classification of Hooper and Thompson (1973), which suggests that assertive predicates (e.g. *tell, see, believe,* etc.) – predicates that assert the truth of the embedded proposition – license the availability of MCP, the reformulated classification of Kastner (2015) instead suggests that presuppositional predicates (e.g. *be happy, remember,* etc.) – which assume the propositional content of the embedded clause to be in the common ground (i.e. presuppose the truth of the embedded proposition) – block MCP. Though a seemingly small distinction, the difference between the two classifications fundamentally changes the way in which Doxastic Factives – predicates, like *discover, see, notice, hear, remember,* etc., that presuppose the truth of the embedded proposition and assert that someone (an attitude holder) holds an epistemic attitude about that proposition – are analyzed (see Table 2).

Table 2. *Predicate Class Distinctions*

Proposal	Classes that allow Main Clause Phenomena	Classes that do not allow Main Clause Phenomena
Hooper and Thompson (1973)	Speech Act Predicates Doxastic Non-Factives *Doxastic Factives*	Response Predicates Emotive Factives
Kastner (2015)	Speech Act Predicates Doxastic Non-Factives	*Doxastic Factives* Response Predicates Emotive Factives

The difference between these two classifications can be summarized as follows. Under the assertion-based classification, Hooper and Thompson (1973) assume that Doxastic Factives (e.g. *remember*) allow MCP because they assert attitude holder belief towards the truth of the

embedded proposition, P. Conversely, under the presupposition-based classification, Kastner (2015) argues that Doxastic Factives do not allow MCP because they presuppose the truth of P.

Ultimately, the two classifications here represent only a small subset of the different predicate classifications proposed for English. Indeed, there are a number of other proposals that base predicate class assignments on different semantic properties (see stance/non-stance systems like in Hegarty, 1992 or evidence-based factivity systems like in Karttunen, 1971, Beaver, 2010, or Djärv, 2019). The point here is that, even if we find that predicate class has an effect on complementizer choice in Swahili, there are various ways of classifying predicates, and those classifications might be relevant for different phenomena.

3.2 Matrix Subject Person

Another salient element of the matrix clause that has been observed to have an effect on complementizer distribution is the subject itself; in certain languages, the choice of complementizer appears to be modulated by person. In the Bantu language Kinyarwanda,[4] for example, it has been shown that the distribution and interpretation of the hearsay complementizer, *kongo,* which is described as reflecting doubt on behalf of the speaker towards the source of the proposition encoded by the embedded clause, is sensitive to the person of the matrix subject (Givon & Kimenyi, 1974). Though this sensitivity only yields selective restrictions in the presence of a factive matrix predicate,[5] it still holds that *kongo* must track the person of the matrix subject, as it is only compatible with a third person subject (16), with both first and second person subjects yielding an ungrammatical judgement (17-18).

[4]Though both members of the Bantu language family, Kinyarwanda and Swahili are quite distinct from one another, with each being classified as a member of a different subfamily of languages; Swahili belonging to the Northeast Coast Bantu subfamily, and Kinywarwanda to the Great Lakes Bantu subfamily (Nurse & Philippson, 2003). Given this difference in classification – among other salient differences between the two (Kinywarwanda being a tonal language) – we do not necessarily expect a similar complementizer system.

[5]It is not trivial that the interaction between matrix subject person and the selection of the hearsay complementizer *kongo* may only occur in the presence of a factive verb, like *forgot*, which entails the truth of the proposition encoded by the complement clause. In fact, the incompatibility of said complementizer with a first- and second-person subject in such contexts can be straight-forwardly attributed to the fact that, given a factive matrix predicate, it is illogical for a speaker to cast doubt on the source of a presupposed proposition that they (1p) or another conversational participant (2p) uttered (see, Givon & Kimenyi, 1974; for a more in-depth discussion). The point illustrates that complementizer selection may be sensitive to multiple factors at once.

(16) y-iibagiwe kongo amazi yari mare-mare
he-forgot that water was deep
"He forgot that the water was deep" (and I doubt it, since my information was obtained via hearsay)

(17) *n-iibagiwe kongo amazi yari mare-mare
I-forgot that water was deep

(18) *w-iibagiwe kongo amazi yari mare-mare
you-forgot that water was deep

(Adapted from Givon & Kimenyi, 1974),

3.3 Matrix Operator

Yet another factor that has been shown to impact complementizer distribution is the presence of a matrix non-actual/non-factual operator; the presence or absence of either a question or negation operator can influence complementizer selection. Consider again the behavior of the non-factive declarative complementizer *oti* in Greek, this time with respect to the interrogative complementizer, *an* (English *if*). Comparing (19) and (20) below, it can be seen that the presence of a matrix operator in the later, in this case negation, influences complementizer selection in the sense that it allows for the selection of either *an* or *oti,* whereas the former, lacking a matrix operator, allows only for the selection of *oti*.

(19) O Janis kseri *oti/*an* efighan
the John know-3s that/if left-3p
"John knows that they left."

(20) O Janis dhen kseri *oti/an* efighan
the John not know-3s that/if left-3p
"John doesn't know that they left." (Roussou, 2010)

The fact that *an*, unlike *oti*, is only available in the presence of negation follows from the fact that *an*, as an interrogative complementizer, necessarily introduces a non-propositional clause; given that a declarative attitude verb like *kseri*, 'know', requires its subject to hold a belief about the truth of the embedded clause, *an* is expectedly unavailable in (19), as it crucially introduces a

clause that doesn't contain a proposition, and cannot be evaluated as true or false. Contrastively, *an* is expectedly available in the context of matrix negation (20), as this context presumably does not impose a belief on the matrix subject, therefore permitting the complement clause to be non-propositional.

3.4 Factors in the subordinate clause

As for factors relevant to the subordinate clause that can play a role in determining complementizer choice, the large majority of research on this topic has focused on the presence of a topicalized or focused element in the embedded clause (Rizzi, 1997; Cruschina, 2006; Vecchio, 2010). Take for example the distributional behavior of the complementizers /ka/ and /ku/ in the Southern Italian dialect Francavilla[6], which, while exhibiting distinct distributions in most embedding contexts, sees the underlying contrast between the two complementizers neutralized in cases where another element (e.g. a topicalized or focused constituent) is A' moved into the C-domain (Vecchio, 2010), as only /ka/ may be followed by a Foc/Top element (21) - (22).

(21)	Vogghiu	cu	(*CARLU)	veni	cu	nnui,	(Carlu)
	I-want	that	(Carlu)	comes	with	us	Carlu
	"I want Carlu to come with us"						

(22)	Vogghiu	ca	(CARLU)	veni	cu	nnui	(Carlu)
	I-want	that	Carlu	comes	with	us	
	"I want Carlu to come with us"						

(Adapted from Vecchio, 2010)

When considered alongside reports suggesting that choice of complementizer may also interact with/be modulated by the mood of the embedded clause (see discussion of Vecchio, 2010 in Section 3.5), data like that in (21) - (22) have been used to argue that the different complementizers in a language like Francavilla occupy distinct functional projections in the CP-domain (Rizzi, 1997).

[6]Following Vecchio (2010), the Francavilla complementizers in question are given in IPA (e.g. /ka/ and /ku/) when discussed in-text, and in standardized roman orthography when cited in data tokens (e.g. *ca* and *cu*).

In summary, cross-linguistic evidence suggests that complementizer choice in any one language may show sensitivity to certain classes of clause-embedding predicates, the presence of a matrix operator, the person of the matrix subject, movement of a topicalized/focalized element into the C-domain, the mood of the embedded clause, and moreover, may encode information about the semantic content of the embedded clause. Given these facts, it becomes an empirically relevant question as to whether such selectional and interpretational effects will hold for any language in which multiple possible complementizers are attested, particularly in cases where those complementizers appear to be in free variation.

3.5 The Extended Left-Periphery

Observations of such differences in distribution and interpretation, specifically in Italian and a number of related dialects, is what ultimately led to the postulation of an expanded, articulated C-system that remains, more or less, as the primary schematic of the C-system in the generative tradition today. In his seminal work on the left periphery, Rizzi (1997) describes this complex system as being composed of distinct functional projections that encode different information relative to a complementizer's role as the interface between the superordinate and subordinate material. Minimally, this system serves to split the interface in two, allowing for a distinct projection high in the system that encodes information relevant to the superordinate material, and another, lower projection that encodes information relative to the subordinate material.

The highest projection in the C-system, *Force*, is argued to encode information relevant to superordinate material, specifically with respect to clause type. As suggested by its name, the *Force* projection provides information pertaining to the illocutionary force of the embedded clause, i.e. whether the embedded clause exhibits declarative or interrogative force. Motivation for a projection that is sensitive to clause type/illocutionary force is well-captured by Roussou's (2010) argument that in languages like Greek and English, the distinction between a declarative complementizer (e.g. *that*) and an interrogative complementizer (e.g. *if*) is rooted in how they interact with particular elements in the matrix clause, as illustrated by the fact that an interrogative complementizer in English is only licensed by the presence of a negative/question operator in the matrix clause (23) - (25).

(23) John said [that/*if Bill went to the store].

(24) John didn't say [that/if Bill went to the store].

(25) Did John say [that/if Bill went to the store]? (Roussou, 2010)

Conversely, the lowest node of the articulated C-system, *Fin* (for "Finite"), is argued to encode information relative to the subordinate material of the embedded clause, specifically with respect to finiteness/modality. In her work on the distribution of the complementizers /ka/ and /ku/ in the Southern Italian dialect Francavilla, Vecchio (2010) clearly illustrates the need for a projection that encodes information relative to mood, and further, that this projection needs to be distinct from Force. The motivation for such a projection can be seen in the following data (26) - (27), which show that only the latter of these complementizers, /ku/ (written 'cu'), can introduce an embedded verb that is marked with what Vecchio refers to as 'deictic' tense, i.e. inflected verbs that may only be understood relative to the matrix clause (similar, if not identical, to the subjunctive).

(26) Creu ca/*cu teni raggioni iddu
I-believe that he-has reason he
"I think he's right"

(27) Vogghiu cu/*ca jjeni cu mme ala chiesa
I-want that you-come with me to-the church
"I want you to come to church with me"

(Adapted from Vecchio, 2010)

In addition, Vecchio suggests that, when considered alongside the fact that only /ku/ interacts with the mood of the embedded clause, the incompatibility of /ku/ with the presence of a topicalized/focused element in the left periphery (28) provides further evidence that /ku/ is syntactically lower than the alternative complementizer, /ka/. That is, given that the underlying contrast between the two complementizers appears to be neutralized if another element (e.g. a topicalized or focused constituent) is A' moved into the C-domain – as only /ka/ may be followed by a Top/Foc phrase (29) – /ku/ must crucially be lower than both /ka/ and the landing spot for a Top/Foc phrase.

(28) Vogghiu cu (*CARLU) veni cu nnui, (Carlu)
I-want that (Carlu) comes with us Carlu
"I want Carlu to come with us"

(29) Vogghiu ca (CARLU) veni cu nnui (Carlu)
I-want that Carlu comes with us
"I want Carlu to come with us"

(Adapted from Vecchio, 2010)

As suggested above, Topicalization/Focus plays a large role in the C-system, as the presence of A' movement in data like these appears to constrain the subset of possible complementizers available for selection. The relevant projections that house topicalized/focused constituents in the left periphery are widely argued to be found between ForceP and the lower FinP (see Figure 1). Particularly compelling data in favor of an articulate C-system analysis of this sort can be seen in Ledgeway & Alessandro (2010), which attempts to address the split-complementizer system in Abruzzese by suggesting that A' movement to C and the selectional requirements of the matrix predicate forces movement of a complementizer from a lower projection of the C system, Fin, to the highest projection, Force. Ultimately, they found that unlike regular A' movement to C, the movement of a phonologically heavy Top/Foc constituent to the left periphery can elicit the realization of both the highest copy of a complementizer at the left-edge in Force, as well as its lower copy at the right-edge in Fin.

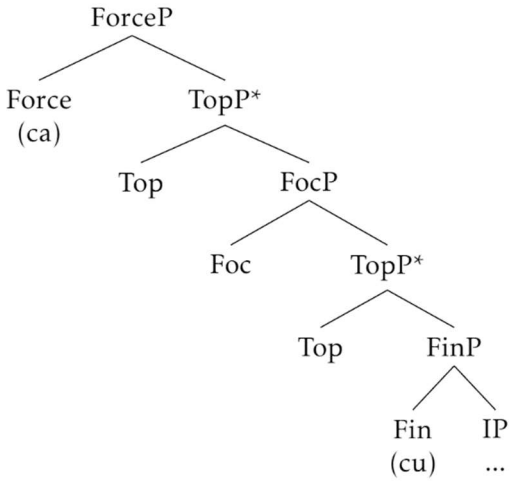

Figure 1. *Articulated C-System (Rizzi, 1997)*

With respect to how a Force/Fin-type analysis could play out in the Swahili dual-complementizer system, we would expect to see a difference in the specific factors that *kwamba* and *kuwa* each interact with. Specifically, if we expect a structural distinction to hold between the two complementizers, then the structurally superordinate complementizer (maps to Force),

whichever that may be, should interact with elements of the matrix clause (e.g. matrix predicate class, matrix negation, etc.), and be compatible with A' movement in the embedded clause, while the lower complementizer (maps to Finite) should instead interact with elements of the embedded clause (e.g. mood), and be incompatible with A' movement. Though it remains an open question as to whether the Swahili data can be explained by attributing *kwamba/kuwa* to distinct functional projections, it is important to note here that, from a syntactic perspective, this is the null hypothesis; given that both *kwamba* and *kuwa* are used to introduce an embedded finite clause, then, like other dual-complementizer systems, the two complementizers in Swahili must map to distinct projections in the C-domain.

4. Methodology

The data considered in this project were extracted from the Annotated Version of the Helsinki Corpus of Swahili 2.0 (henceforth, HCS 2.0), a restricted-access corpus of the Language Bank of Finland (Bartis & Hurskainen, 2016). The Annotated Version of HCS 2.0 was chosen as the primary corpus for this project because it is ideally situated for an in-depth corpus search addressing the probability of co-occurrence between the complementizers *kwamba* and *kuwa* and different elements of the matrix clause.

As one of the most recently published Swahili corpora available, HCS 2.0 is a repository of over 26 million individual tokens across four distinct sub-corpora, with each sub-corpus being defined by the specific medium from which its content was derived. As such, the four corpora differ slightly with respect to the types of texts they include, with the Bunge (parliament) corpus including official political documents from the Tanzanian Parliament (2004-2006), the Books corpus including complete or partial Swahili texts from various authors (prior to 2003), and the News (old) and News (new) corpora including transcribed interviews from prior to 2003 (News, old), and 2004-2015 (News, new) respectively. It is worth noting here that HCS 2.0 necessarily includes a range of Swahili dialects given the inclusion of transcribed news interviews and book/text excerpts. Nonetheless, it is likely that Tanzanian Swahili is the best represented and/or most dominant dialect present in the corpus, as the texts used in the Bunge and News (old and new) sub-corpora are largely of Tanzanian origin.[7] As such, our analysis and discussion will largely take Tanzanian Swahili to be the primary dialect of the corpus.

Arguably the single biggest reason as to why we elected to use HCS 2.0 in this project is that it has been carefully morphologically parsed and tagged; each individual word in the corpus is indexed according to the relevant features of its particular part of speech, with nouns being annotated with information relative to noun class, and verbs being annotated with information relative to subject marking, object marking, TAM marking, and negation, among other features, for example. Perhaps the most useful aspect of the corpus is the fact that it not only allows direct

[7]Though the nature of the data contained within the News (old) and News (new) sub-corpora necessarily precludes this corpus from being composed of *strictly* Tanzanian Swahili, we assume the data within HCS 2.0 to be predominantly of Tanzanian origin given that the Bunge sub-corpus is made up of Tanzanian Parliamentary documents, and the two News sub-corpora data taken from Tanzanian news channels.

access to the featural information of a particular token, but also permits tailored searches for specific feature combinations that apply to one or more items in a linear word string. Notably, this allows for complex searches of the nature required for this project, where, for example, one could search for only those items containing a negative element followed by a clause-embedding predicate, a complementizer, and a finite complement clause. With some manipulation of the regular expressions/query language template provided with the corpus, these searches can be carefully controlled to isolate individual features like negation, person, etc., so that the returned results consist of word strings that minimally contain all of the relevant features in a specified linear order, though not necessarily adjacent.

Indeed, this method was the only one used for data collection in this project. Only one search query was ultimately needed to return the relevant corpus data, which, given the specifications of the search string in question, consist only of tokens exhibiting word strings of the type [VFin + *kwamba/kuwa)* + (Noun) + VFin].[8] This exact search was customized to include other information relevant to the returned tokens, including raw data values, corresponding lemmas, and morphological glosses for each item within the target word string, as well as a broad translation, and the sub-corpus from which the token was extracted. The data were then extracted from the HCS 2.0 corpus, converted into a single external spreadsheet, and coded by subject person verbal morphology, the presence/absence of matrix negation, embedding predicate class, and complementizer choice/selection.

The external dataset, which I will henceforth refer to as the "complementizer dataset", consists of 26,066 individual tokens in total. Token distribution across the four sub-corpora can be seen in Table 3.

[8]To avoid any ambiguity with their infinitival verb forms, the syntactic function of *kwamba/kuwa* was marked as 'complementizer/conjunction' in the search string. By including the specification for an embedded finite verb 'VFIN', we consequently controlled for the appearance of the infinitival forms of *kwamba/kuwa* in the embedded clause (i.e. only tokens involving a finite embedded clause were included in the data).

Table 3. *Token Distribution by Sub-Corpora*

Sub-corpus	Total Tokens	% of Total Corpus
Bunge	9492	0.364153
News new	10162	0.389857
Books	5804	0.222666
News old	752	0.02885

Notably, of the 26,066 tokens in the complementizer dataset, roughly 60% of tokens involved the use of *kuwa*, while only 40% contained *kwamba*, as illustrated in Table 4. This imbalance was considered and accounted for by our regression model, as will be discussed in Section 4.4.

Table 4. *Token Distribution by Complementizer*

kwamba		*kuwa*	
Total Instances	Proportion of Corpus	Total Instances	Proportion of Corpus
10,365	.398	15,700	.602

4.1 Factors to Investigate

This project considers three distinct features argued to have an effect on complementizer choice, those being matrix predicate class, matrix subject person specification, and the presence of matrix negation (a matrix operator).

4.1.1 Predicate Classification

As discussed in Section 3.1, predicate class is argued to have an effect on complementizer choice (Kiparsky and Kiparsky, 1971; Hooper and Thompson, 1973), and as such, serves as one of the three possible predictive factors in our regression analysis. Although the annotations provided by HCS 2.0 allow for a number of elements/features to be broadly included in any given search query, matrix predicate class is (expectedly) not an annotated feature in the corpus.

In order to address the question(s) related to the correlation of predicate class and choice of complementizer, this investigation considers only a representative set of exemplars from each relevant predicate class. Once the general clause-embedding data had been extracted from the corpus, the data were manually sorted, and a select subset of the most pervasive predicates were coded according to matrix predicate class.

Though a number of different predicate classifications could have been used in our study, we have chosen to use an adapted version of the seminal clause-embedding predicate classification first laid out in Hooper & Thompson (1973), which groups predicate classes according to whether they exhibit Main Clause Phenomena (e.g. topicalization/focalization, verb fronting, or falsifying continuation). This classification lists each of the five clause-embedding predicate classes to be considered by this investigation alongside an exemplar predicate for both English and Swahili (see Table 5 and Table 6). It should be noted that this classification serves only to highlight/distinguish some predicate classes that may be relevant to this investigation; it remains an empirical question as to whether this classification, or any other, can account for the Swahili facts.

Table 5. *Predicate Classes that allow Main Clause Phenomena*

Predicate Class	*English*	*Swahili*
Speech Act Non-Factives	say	*sema*
Doxastic Non-Factives	assume/believe	*dhani*
Doxastic Factives/Semi-Factives	see/feel	*ona*

Table 6. *Predicate Classes that do not allow Main Clause Phenomena*

Predicate Class	*English*	*Swahili*
Emotive Factives	love/like	*penda*
Response Predicates	admit	*kiri*

(Adapted from Hooper & Thompson, 1973; Djärv, 2019)

To simplify the predicate class analysis, only a small number of predicates were chosen as representative exemplars for each of the predicate classes under consideration (see Table 7). The individual predicates chosen as class representatives were carefully selected to control for potential class ambiguity, using previous analyses (Hooper & Thompson, 1973; Kastner, 2015; Djärv, 2019) as a classificational baseline. This particular classification was chosen as a baseline because it allows us to analyze the behavior of the predicates with respect to Factivity, previous proposals pertaining to the availability Main Clause Phenomena (Hooper & Thompson, 1973; Kastner, 2015; Djärv, 2019) and also speaker/local subject attitude.

Table 7. *Predicate Classification*

Doxastic Factives	Doxastic Non-Factives	Doxastics (ambiguously Factive/Non-Factive)	Desideratives	Emotive Factives	Response	Speech Act
sikia (‘hear’)	*amini* (‘believe’)	**kuta* (‘find’)	*tumai* (‘hope’)	*furahi* (‘be happy’)	*kubali* (‘agree’)	*tamka* (‘pronounce’)
ona (‘see/feel’)	*fikiri* (‘think’)	-	*taraji* (‘hope/expect’)	*penda* (‘like’)	*kiri* (‘admit’)	*taja* (‘announce’)
gundua (‘discover’)	*dhani* (‘think/guess’)	-	-	*ogopa* (‘fear’)	*nakiri* (‘deny’)	*ambia* (‘tell’)
-	*zingatia* (‘consider’)	-	-	*lalama* (‘complain’)	*jibu* (‘answer’)	*sema* (‘say’)
-	-	-	-	*hofu* (‘fear’)	*ongeza* (‘add’)	*onya* (‘warn’)
-	-	-	-	-	-	*dai* (‘claim’)

**kuta*, ‘find’, is coded independently as “Doxastic”, given that it is ambiguously Factive/Non-Factive. As such, it will be presented as a separate class (e.g. ‘Doxastics’) in the results section. Ultimately, we will group *kuta* together with the Doxastic Factives in our analysis based on its behavior (i.e. we will collapse Doxastics/Doxastic Factives).

4.1.2 Matrix Subject Person

Matrix subject person has similarly been shown to modulate complementizer choice in dual-complementizer languages (Givon & Kimenyi, 1974). Much like predicate class, the

investigation of matrix subject person does not require its own unique search query, as the person specifications of the subject – as part of/in addition to subject noun class information encoded on the verb – are indexed by the corpus. However, once the data was downloaded from the corpus, matrix subject person, like predicate class, was similarly hand-coded specifically for person. Specifically, all tokens exhibiting 1P, 2P or 3P (Noun Class 1/2) subject agreement on the matrix verb were coded based on person, while all other examples were coded as '-' (non-person; used as the baseline).

4.1.3 Negation

The presence of an operator – whether a question or negation operator – has also been shown to affect complementizer choice (Roussou, 2010). In order to simplify our corpus search, this investigation considers only the matrix operator of negation, as opposed to both the negation and question operators. Given that negation (glossed as '*NEG*') is an indexed feature of HCS 2.0, searches of this type are relatively straightforward in that they only require a query that yields tokens containing a negative feature in the verb complex, or in some position higher than the embedded clause. Information relative to the presence of a negative operator was downloaded directly from the corpus in conjunction with the general search data, and then listed distinctly in the exterior "complementizer corpus" (i.e. the downloaded search data).

4.2 Statistical Analysis

Given that the primary objective of this investigation is to determine whether the aforementioned factors play a role in complementizer selection, any statistical analysis employed by this project must necessarily address the relative ability of these factors to predict the use of either *kwamba* or *kuwa.* That is to say that, in order to tease apart any relationship between the factors under consideration and complementizer choice, the question at the heart of this study must be framed as one of likelihood and predictability; given some subset of (predictive) factors, which of the two complementizers is more likely to occur? This question of relative predictability lends itself well to a regression-based statistical analysis, as regression models

crucially describe the relationship between the independent and dependent variables based on the ability of the independent (predictor) variables to predict the value/outcome of the dependent variable.

With respect to the present project, we employ the use of a binomial logistic regression model as a means through which to investigate the relationship between matrix negation, matrix subject person, and/or predicate class (independent predictor variables), and the use of *kwamba* or *kuwa* (binary dependent variable).

4.3 Data Transformations

Given the nature of corpus data, there is generally expected to be some degree of token variability and/or token imbalance that needs to be accounted for prior to data analysis. In the case of the complementizer data under investigation in this project, there are two salient categories of data imbalance that need be addressed, those being embedding predicate frequency, and the uneven overall distribution of *kwamba* and *kuwa*.

Unlike matrix negation and matrix subject person, whose relevant data consist only of a small subset of categorical values extracted directly from the morphological glosses in the primary corpus (e.g. NEG, 1Sg-SBJ, 2Sg-SBJ, etc.), coding the data for *predicate class* required an additional level of abstraction across the extracted data, as predicate class classification crucially groups individual predicates into larger sub-classes. Given that the complementizer data extracted from the corpus contains no information directly pertaining to predicate class, any effort to group predicates by class needed to be done by hand.

Predicate classification began with the identification of each unique predicate lemma used in the complementizer data. These predicate lemmas were extracted from the dataset, placed into a separate spreadsheet, and sorted by frequency of occurrence with each of the two complementizers (i.e. listed frequencies reflected the number of tokens including a predicate and *kwamba*, and a separate number of tokens including that same predicate and *kuwa)*.

Once sorted, the raw frequency values associated with each predicate lemma were subject to two distinct transformations to account for the expected effects of Zipf's Law (Zipf, 1949),

which states broadly that, for any set of linguistic frequency data (e.g. word frequency), the observed occurrences of a token and its rank in an ordered frequency will exist in an inverse relationship such that the frequency of a token, X, will be roughly twice that of the next most frequent token, Y.

First, the raw frequency count values were transformed using the Zipf scale (Van Heuven, Mandera, Keuleers, & Brysbaert, 2014), which assigns each token (e.g. predicate lemma) a "Zipf value" on a 1-7 scale based on log frequency per million words, with Zipf values of 1 being assigned to the lowest frequency tokens in the corpus, and 7 being assigned to the most frequent.

Given that the raw frequency values of each predicate lemma were split by complementizer (i.e. one value for each complementizer), the Zipf scale transformations were applied to both the predicate-*kwamba* frequency counts and the predicate-*kuwa* frequency counts. In order to make these frequency values comparable to one another, the derived Zipf values were then transformed to reflect the difference between the observed values of each predicate-complementizer combination, and their expected values. The formula(e) used to derive the expected values for each complementizer can be seen in Table 7 and Table 8 below[9]:

Table 8. *Expected Value Calculations – kwamba*

	Expected *kwamba* frequency values
General Calculation	((*kwamba* observed freq.) + (*kwamba* observed freq.)) * (total *kwamba* tokens) / (total corpus tokens)
Calculation with token counts	((*kwamba* observed freq.) + (*kwamba* observed freq.)) * (10,365) / (26,066)

Table 9. *Expected Value Calculations – kuwa*

	Expected *kuwa* frequency values
General Calculation	((*kwamba* observed freq.) + (*kuwa* observed freq.)) * (total *kuwa* tokens) / (total corpus tokens)
Calculation with token counts	((*kwamba* observed freq.) + (*kuwa* observed freq.)) * (15,700) / (26,066)

[9]The EXP/OBS values were computed for both raw frequency values and their corresponding Zipf values.

4.4 Data Training

As previously mentioned, the second issue related to data imbalance addressed prior/during statistical analysis was the uneven distribution of *kwamba* and *kuwa*. For ease of reference, the overall distribution of the two complementizers is repeated in the Table 10.

Table 10. *Complementizer Frequency*

kwamba		*kuwa*	
Total Instances	Proportion of Corpus	Total Instances	Proportion of Corpus
10,365	.398	15,700	.602

To ensure that such an imbalance in the dependent variable would not have any effect on the results of the multi-factor analysis, each of the candidate regression models were explicitly 'trained' prior to model testing. That is, in order to account for the 60-40 split in distribution, the candidate models needed to be trained to 'ignore' data imbalance and instead focus solely on the explanatory power of each predictor variable (i.e. the factors under consideration).

The following procedure outlines the training process. The complementizer data was first chunked into two distinct sample populations, with 14,510 tokens being allocated to a training dataset (i.e. the development sample), and 11,555 to a separate test dataset (i.e. the validation sample). The training data consisted of an equal distribution of *kwamba* and *kuwa* tokens, with 7,255 individual instances of each complementizer. The equal population sizes making up the training data is of crucial importance here; this distribution allows each candidate model to be trained/created using an unbiased dataset, meaning that any relationship found to hold between some factor(s) and complementizer selection could not simply be the result of a skewed distribution of the dependent variable (i.e. it can't be influenced by the fact that *kuwa* is statistically more prevalent). As such, each candidate regression model was trained using the evenly distributed training data, before being compared and ultimately ranked according to their ability to account for the distribution of the unbalanced data in the validation sample (e.g. test data).

5. Results

Given that this project considers only matrix predicate class, matrix subject person, and matrix negation as potential factors affecting complementizer choice, there are logically seven potential explanatory models for the data in question, with each model differing only in the specific combination of predictor variables it includes. Excluding discussion of factor interactions for the moment, Table 11 summarizes the possible explanatory models for the data.

Table 11. *Possible Model Outcomes by Factor Combination*

Number of Significantly Predictive Factors	Possible Model Outcome (factors predicting complementizer choice)
If one factor is predictive …	[Negation]
	[Predicate Class]
	[Person]
If two factors are predictive …	[Negation / Predicate Class]
	[Negation / Person]
	[Predicate Class / Person]
If three factors are predictive …	[Negation / Predicate Class / Person]

5.1 Model of Best Fit

Following data training, each of the above seven models were compared based on their ability to account for data in the test sample using simple ANOVA model comparisons. The results of these model comparisons found that for every addition of a predictor variable, the resultant model showed a statistically significant difference in predictive power relative to its predecessor, suggesting that each of the three factor variables under investigation do, to some extent, account for the distribution of *kwamba* and *kuwa* in the data. As such, it was ultimately

found that of the seven potential models listed above, the three-variable model including negation, matrix subject person, and predicate class as predictive factors was the model of best fit.

It should also be noted here that the ultimate model of best fit (i.e. the three-variable model) was compared with various models considering possible interactions between the predictor variables, and was still found to be the most powerful model with respect to its ability to predict complementizer choice given a particular combination of factors. Specifically, it was found that models including factor interactions offered no significant increase in predictive power as compared to the general three-factor model without interactions. Given this finding, no further discussion of factor interactions will be pursued in this paper.

5.2 Matrix Subject Person Morphology

Turning now to the results for each individual factor under consideration, let us first begin with matrix subject person morphology. Ultimately, it was found that, of the six person-number feature combinations under investigation, all were shown to be significantly predictive of complementizer choice, though not necessarily equal with respect to statistical significance. While each of the feature combinations is shown to be more predictive of complementizer choice than the null alternative (i.e. noun class specific matrix subject morphology, primarily of class 5/6, 7/8, and 9/10), they vary with respect to just how predictive they are. Specifically, it was found that the feature combinations of 1Sg, 1Pl, 2Pl, and 3Sg were the strongest predictors of complementizer choice, each having a highly significant p value of less than 0.001, while the remaining two combinations, 2Sg and 3Pl, each show a less significant p value of less than 0.05. That being said, this difference in statistical significance does not appear to show any bias towards one complementizer choice or the other, as both *kwamba* and *kuwa* are shown to correlate with feature combinations of either significance, with 1Sg, 1Pl, 2Pl, 2Sg correlating with *kwamba,* and 3Sg and 3Pl correlating with *kuwa*. Interestingly, this classification of feature combinations by complementizer correlation sheds light on what appears to be a bias for person, but not number; *kuwa* is shown to generally correlate with third-person subject morphology, while *kwamba* is shown to correlate with both first and second-person (see Table 12).

Table 12. *Matrix Subject Correlations and Significance*

Matrix Subject Morphology	Predicted Complementizer	Statistical Significance
1SG	*kwamba*	p < .001 ***
1PL		
2PL		
3SG	*kuwa*	
2SG	*kwamba*	p < .05 *
3PL	*kuwa*	

At this point it is worth noting that, while considered significant predictors based on our regression model, we will ultimately exclude second-person matrix subjects from our analysis, as there simply aren't enough second-person tokens to make any statistically legitimate claims about the relationship between second-person and complementizer choice. Table 13 provides frequency counts for co-occurrences of each complementizer with the different phi-feature combinations. As can be seen, second-person matrix subjects are significantly less frequent than both first and third-person matrix subjects, regardless of complementizer choice.

Table 13. *Complementizer Frequency by Matrix Subject Person/Number*

	kwamba		*kuwa*	
Person/Num.	Total Instances	Proportion of total *kwamba* Tokens	Total Instances	Proportion of total *kuwa* Tokens
1SG	1672	.161	598	.038
1PL	1424	.137	672	.043
First-person	**3096**	**.299**	**1270**	**.081**
2SG	111	.011	94	.006
2PL	131	.012	43	.003
Second-person	**242**	**.023**	**137**	**.009**
3SG	2517	.243	6330	.403
3PL	1432	.138	2824	.180
Third-person	**3949**	**.381**	**9154**	**.583**

As the predictive output of our regression model, the results of our analysis for matrix subject person (and the subsequent factors to come) are presented in terms of complementizer likelihood given the relevant factor. That is, the trained regression model assigns a predicted value to each person/number feature combination on a likelihood scale from 0 to 1, where 0 denotes a 100% likelihood of occurrence with *kuwa*, and 1 a 100% likelihood of occurrence with *kwamba*. The results of this analysis are illustrated in Figure 2, which provides a visualization of the dispersion of the predicted values for the individual data tokens in each predicate class, and Table 14, which lists median complementizer values by person. Note that Table 14 arranges the possible feature combinations by median likelihood value in order to clearly illustrate which predict *kwamba* (i.e. have a value > 0.5) and which predict *kuwa* (i.e. have a value < 0.5).

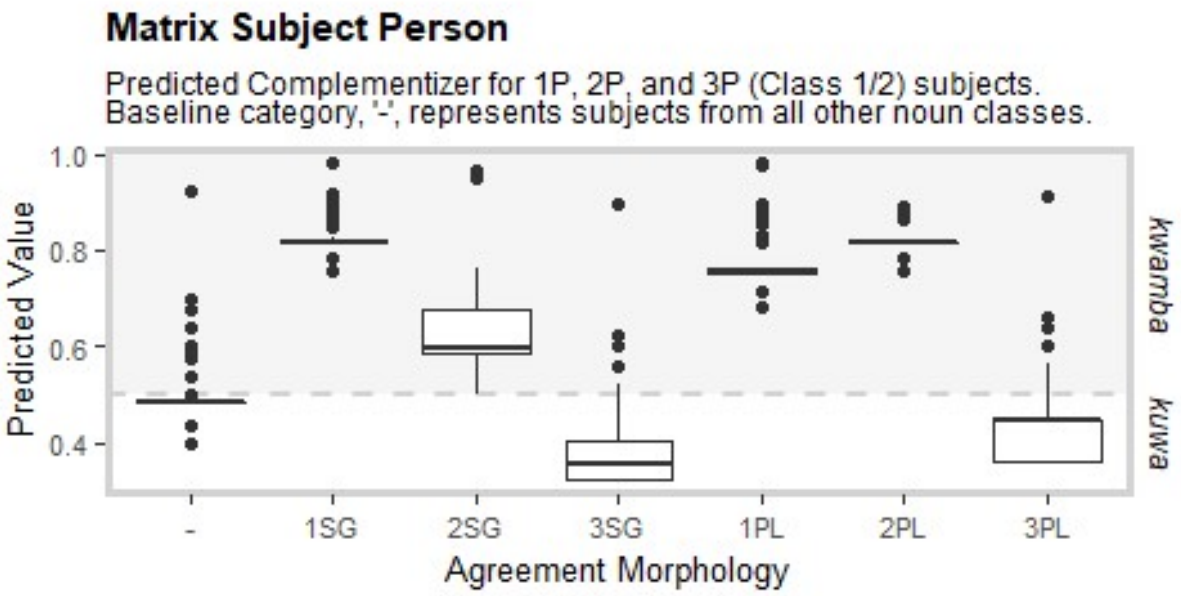

Figure 2. *Matrix Subject Person Boxplot*

Table 14. *Median Complementizer Value by Matrix Subject Person/Number*

Matrix Subject Morphology	Median Complementizer Likelihood Value (0 = 100% likelihood *kuwa*, 1 = 100% likelihood *kwamba*)	Predicted Complementizer
1SG	0.8273384	*kwamba*
2PL	0.7906945	*kwamba*
1PL	0.7588116	*kwamba*
2SG	0.5952662	*kwamba*
(baseline)	0.4794679	*kuwa*
3PL	0.4488631	*kuwa*
3SG	0.3699610	*kuwa*

5.3 Predicate Class

With respect to the second factor under consideration, namely, matrix predicate class, five total predicate classes were found to be significantly predictive of complementizer choice, with three classes correlating with *kwamba*, those being Doxastics, Doxastic Factives, and Emotive Factives, and two classes correlating with *kuwa*, specifically, Response Predicates, and Speech Act Predicates. Of these, Doxastics and Speech Act Predicates were found to be the strongest predictors of complementizer choice, with both classes having a p-value of less than 0.001. The remaining classes, namely, Doxastic Factives, Emotive Factives, and Response Predicates, can be seen classified by significance and complementizer correlation in Table 15.

Table 15. *Predicate Class Correlations and Significance*

Matrix Predicate Class	Predicted Complementizer	Statistical Significance
DOX	*kwamba*	p < .001 ***
SA	*kuwa*	
DOX-F	*kwamba*	p < .01 **
EF		
RSP	*kuwa*	p < .05 *
DOX-NF	*kuwa*	-
DSD	*kwamba*	-

Having identified the statistical significance of each factor level, in this case, the various predicate classes, each token was assigned a unique predicted complementizer value based on the predictive factors present (i.e. the expected complementizer value based on the combination of matrix subject person, matrix predicate class, and negation, on a 0-1 scale). Note that the Desiderative predicate class was ultimately omitted from this analysis, as there were simply not enough tokens within the class to yield any significant result. Moreover, though coded separately, we ultimately collapse Doxastic and Doxastic Factive classes, as the former only contains a single predicate *kuta,* 'find' (see Table 7 in Section 4.1.1).

Using the same likelihood scale as described with matrix subject person above, where 0 denotes a 100% likelihood of occurrence with *kuwa*, and 1 a 100% likelihood of occurrence with *kwamba*, Doxastic Factives are shown to predict a mean complementizer likelihood value of .684, Emotive Factives a value of .656, and Speech Acts a value of .393[10]. The results of the predicted values analysis can be illustrated by the following, with Figure 3 providing a visualization of the dispersion of the predicted values for the individual data tokens in each

[10]The mean complementizer value for Doxastic Factives is .684 when Doxastics (a class made up of a single predicate, *kuta*, 'find' that is ambiguously Factive/Non-Factive) is collapsed with Doxastic Factives. Without the addition of *kuta*, Doxastic Factives predict a slightly lower (though still significant) mean complementizer value of .668. Doxastics (e.g. *kuta*) predicts a mean value of .937. Though excluded from the analysis, Desideratives also exhibited a high mean value of .710.

predicate class, and Table 16 listing the median complementizer values – as opposed to the mean values – by class.

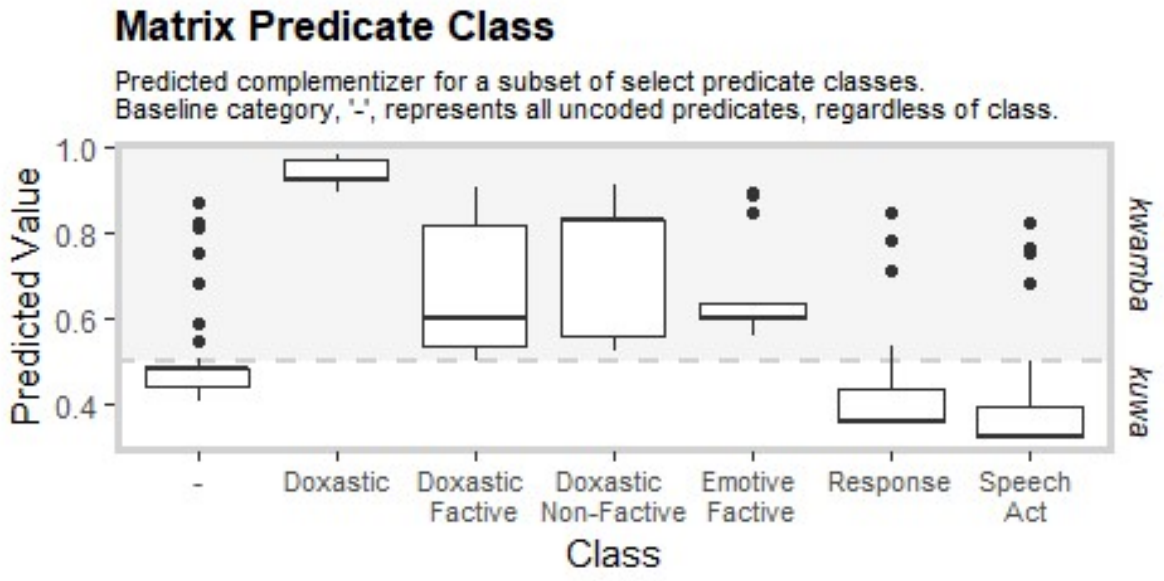

Figure 3. *Matrix Predicate Class Boxplot*

Table 16. *Median Complementizer Value by Matrix Predicate Class*

Matrix Predicate Class	Median Complementizer Likelihood Value (0 = 100% likelihood *kuwa*, 1 = 100% likelihood *kwamba*)	Predicted Complementizer
DOX	0.9203233	*kwamba*
DSD	0.8080274	*kwamba*
DOX-F	0.6341603	*kwamba*
EF	0.5779602	*kwamba*
(baseline)	0.4794679	*kuwa*
DOX-NF	0.4639719	*kuwa*
RSP	0.3699610	*kuwa*
SA	0.3236884	*kuwa*

5.4 Negation

Turning finally to negation, we find perhaps the simplest results of the three factors. Ultimately, it was found that, as compared to the null condition (e.g. no morphological presence of negation) the overt presence of negation in the matrix clause is significantly predictive of complementizer choice. Specifically, matrix negation is shown to strongly correlate with the use of *kwamba,* as illustrated in Figure 4 (Boxplot) and Table 17 (Median Predicted Values).

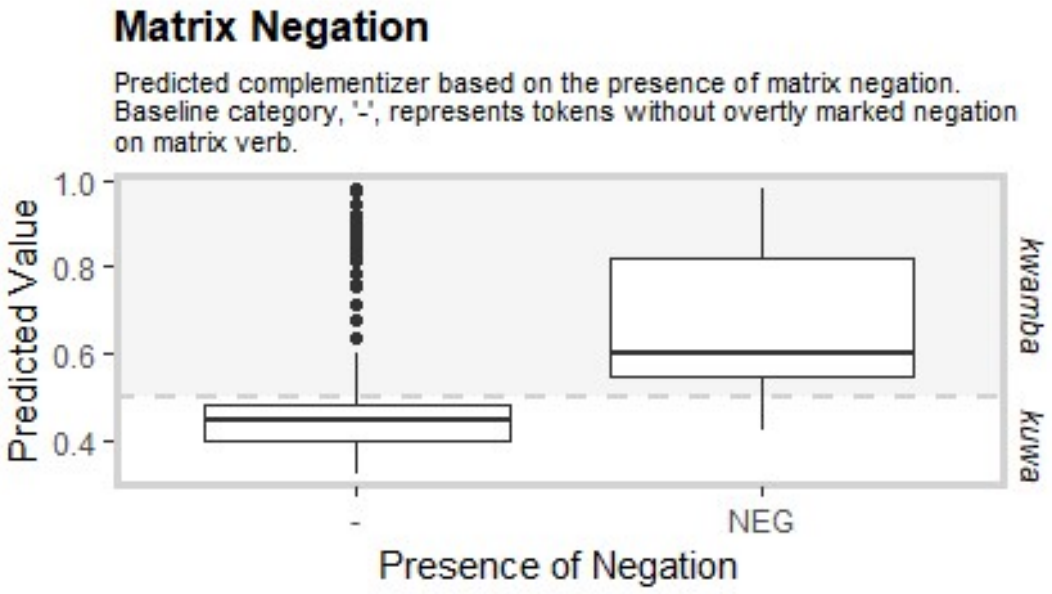

Figure 4. *Matrix Negation Boxplot*

Table 17. *Median Complementizer Value based on Presence of Negation*

Presence of Negation	Median Complementizer Likelihood Value (0 = 100% likelihood *kuwa*, 1 = 100% likelihood *kwamba*)	Predicted Complementizer
Negation	0.6341603	*kwamba*
-	0.4488631	*kuwa*

5.5 Individual Levels of Significance

Though matrix subject, matrix predicate and negation were all found to be significant predictors themselves, the individual levels (i.e. possible predictor values) making up those factors show significant variation with respect to their predictive power. That is, given that each of these variables consists of at least two possible values (e.g. matrix subject:1sg., 1pl, 2sg, etc.), one would expect that certain values correlate more strongly with complementizer choice, and therefore 'drive' the predictive power of the model, and indeed, this is what we have shown for each factor. Summarizing these findings, our model shows that 12 different predictor values across the three factors are statistically significant predictors of complementizer choice, with 7 of those being at a significance of $p < .001$, 2 at a significance of $p < .01$, and 3 at a significance of $p < .05$. Overall, 8 of these values were found to be significant predictors of *kwamba*, with the other 4 correlating instead with *kuwa*. Significance by factor level can be seen in Table 18.

Table 18. *Significance by Factor Level*

Complementizer	Predictor	Statistical Significance
kwamba	1SG	$p < .001$ ***
	1PL	
	2PL	
	Doxastics	
	+Negation	
	Emotive Factives	$p < .01$ **
	Doxastic Factives	
	2SG	$p < .05$ *
kuwa	3SG	$p < .001$ ***
	Speech Act Predicates	
	3PL	$p < .05$ *
	Response Predicates	

5.6 Dominance Analysis

Given that one of the primary goals of this project is to identify whether the factor variables under investigation play a role in complementizer selection, the logical next step once a regression model had been identified was to address the relative contribution of each factor to the model itself; that is, are each of the factors similar in their ability to predict which complementizer will appear, or do they vary with respect to predictive power?

In order to tease this question apart, the aforementioned model of best fit was subject to a dominance analysis (Azen & Traxel, 2009), a statistical analysis that is used to identify the relative contribution of each predictor variable to the overall predictive power of the full model. For this particular project, dominance analysis serves to determine the relative significance of each of the individual factors under consideration (e.g. matrix subject, matrix predicate class, and negation) to the regression model. Ultimately, the dominance analysis shows that of the three predictor variables under consideration, matrix subject person is by far the most dominant in that it yields the most significant change in R^2N (Nagelkerke) value when added to a smaller model[11]. Thus, of the three factors at play, the addition of matrix subject person as a predictor variable results in the most significant positive change in a model's overall predictive power regardless of model size; adding the factor of matrix subject improves every possible model significantly more than adding matrix predicate class or negation, meaning that matrix subject person is the strongest individual predictor of complementizer choice in the model. The results of the dominance analysis can be seen in Figure 5 below, which quantifies each factor's contribution to the overall regression model based on resultant change in R^2N.

[11]Though we employ the Nagelkerke pseudo r-squared value here, there are a number of other feasible alternatives (e.g. McFadden r-squared) that we could use to evaluate the predictive ability of our logistic regression model. Though these different measures yield slightly different r-squared values, such distinctions are largely irrelevant to the results of this study. We therefore accept the Nagelkerke measure as a sufficient general measure of model performance.

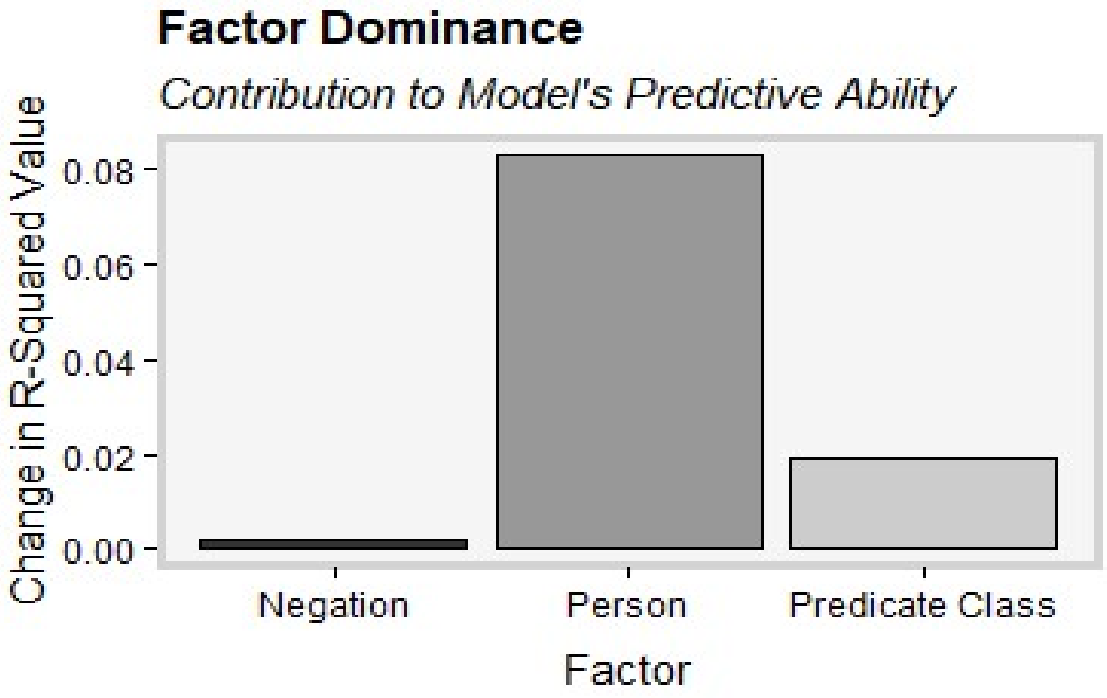

Figure 5. Dominance Analysis Results

As can be seen above, matrix subject person morphology clearly stands out as yielding the largest change in R^2N value, with matrix predicate class yielding the second largest change, and negation the smallest change. Given that this sort of 'rank' in relative R^2N value difference corresponds to the relative contribution of each predictor variable to the regression model, these results suggest that matrix subject person morphology is the strongest individual contributor to the regression model, with matrix predicate class and negation being the second and third strongest contributors, respectively. In sum, matrix subject person is shown to be the driving predictor of the model and is therefore the best individual predictor of complementizer choice of the three factors. It should be reiterated here, though, that matrix predicate class and negation do both significantly correlate with complementizer choice, and their respective additions do necessarily improve the regression model's predictive power; they just don't contribute as strongly as matrix subject person.

5.7 Model Explanatory Power

To ensure that the model in question is not just 'predictive' in the sense that its ability to account for expected complementizer choice is marginally above chance, we have included a brief statistical measure of a model's predictive power by plotting the Receiver Operating Characteristics Curve (ROC) and then calculating the area under its curve (AUC) – a value which essentially represents the likelihood that, given some subset of factors, the model's predicted output matches the true value of the data. Given the scale of Hosmer and Lemeshow (2000),[12] we found the AUC in ROC for the main model to be .674, a score which has generally been argued as a poor but nonetheless significant indicator of predicative power (i.e. the model does correlate with the data, but is not enough to accurately predict complementizer selection in and of itself). We interpret this to be consistent with our discussion earlier: there are many factors that influence complementizer choice cross-linguistically, meaning this model may simply not consider enough factors to explain (or predict) all of the data. In reality, there could be a number of other variables that could significantly improve the overall power of this model and therefore serve as a predictor of complementizer choice in Swahili, though this falls outside the scope of this particular project.

5.8 Sub Corpora Analysis

In order to determine whether the results of the regression analysis hold not just of the entire corpus, but also of its constituent sub-corpora, a post-hoc analysis of the Bunge sub-corpus – which is made up of government documents from the Tanzanian Parliament – was conducted using the same regression model (i.e. we tested whether complementizer choice in the Bunge corpus was also predictable based on matrix subject person, predicate class, and negation).

[12]Hosmer and Lemeshow (2000) assign what is essentially a letter grade scale to AUC in ROC values in order to provide a relative system of measuring model predictive power (or what they call 'model discrimination'). This scale can be broadly defined as the following: if the model's AUC value is ≥ 0.9, the model receives an (A) outstanding grade, if between 0.9 and 0.8, it receives an (B) excellent grade, if between 0.8 and 0.7, it receives an (C) acceptable grade, if between 0.7 and 0.6, it receives a (D) marginal/near passing grade; a value of 0.5 reflects chance level predictive power.

Ultimately, we found that the same general correlations attested in the larger corpus were shown to hold of the Bunge corpus as well. Specifically, the distinctions between first/third-person subjects, predicate class, and the presence of negation in the larger corpus mirrored the distinctions identified in the sub-corpus analysis. A full list of the predictive factors identified in the sub-corpus analysis can be seen in Table 19.

Table 19. *Predictive Factors by Complementizer (sub-corpus)*

Complementizer	Predictor	Statistical Significance
kwamba	1SG	p < .001 ***
	1PL	
	2PL	
	Emotive Factives	
	+Negation	
	2SG	p < .01 **
	Doxastic Factives	
	Doxastics *(Ambiguously Factive/NF*	p < .05 *
kuwa	Speech Act Predicates	p < .001 ***
	3SG	p < .01 **
	-	p < .05 *

6. Discussion

Having presented the results of our regression analysis, let us now turn to how these results reflect on the standard descriptive analysis of Swahili complementizer variation, and, more broadly, how they may fit into a complex theoretical picture of the left-periphery.

With respect to the descriptive generalization of *kwamba* and *kuwa* being in free-variation, the results of our regression analysis provide strong evidence to suggest that these two complementizers are not as freely interchangeable as has been argued (Ashton, 1944; Thompson & Schleicher, 2006). Instead, we find that the choice of complementizer is predictable, but depends on a variety of factors. This sort of predictability claim follows logically from the regression model presented here; of the predictor variables investigated in this study (i.e. matrix subject person, matrix predicate class, and presence of negation), all three were shown to be significantly predictive of complementizer choice, albeit with varying degrees of precision/association by factor sublevel.[13]

Though the notion of complementizer predictability is relatively surprising given the standard description of *kwamba/kuwa* as being interchangeable complementizers, it is unsurprising that we see such a distinction between the two lexical items given that they have distinct etymologies, with each having been lexicalized from different lexical sources. As briefly discussed in Section 1, although *kwamba* and *kuwa* appear to serve the same general syntactic function in clause embedding contexts, these two complementizers were grammaticalized from two distinct lexical verbs, with *kuwa* being the infinitival form (Bantu noun class 15) of the verb, 'to be', and *kwamba*, the infinitival form of the synchronically obsolete verb, 'to say' (Russell, 1992).[14] Given this (historical) distinction in meaning, we might hypothesize that *kwamba* was employed as a quotative marker, used after verbs of speech – consistent with what is found in many other Bantu languages (Guldemann, 2008).

[13]A 'sublevel' in this context refers to any sub-category within a given predictor variable. In this sense, 1Sg would be a sublevel of the predictor variable Matrix Subject Person, and Emotive Factive a sublevel

[14] Though the latter of these has been entirely reanalyzed as a complementizer survived only by its derivative form -*ambia*, 'to tell' (Russell, 1992) in modern Swahili, the infinitival form of *kuwa* is still available in the language, meaning that these two lexical items do convey different meanings synchronically.

However, such a hypothetical distinction would run into immediate problems with the results of our regression analysis, as crucially, *kuwa,* not *kwamba,* was found to correlate with speech act predicates like *sema*, 'say', or *eleza*, 'explain'. As such, I will discard this as a plausible synchronic analysis here. That being said, I will nonetheless argue that the predictability of *kwamba/kuwa* in the present corpus follows from the fact that the two complementizers exhibit distinct pathways of grammaticalization; though their synchronic meanings may not map directly to their lexical origins, it still holds that *kwamba/kuwa* represent distinct lexical items, and as such, may encode discrete meanings as part of their separate lexical entries. The question that remains is whether any difference in meaning caused by these different pathways persists synchronically, or, conversely, whether the meanings of these two complementizers have ultimately converged.

Before progressing further into our discussion of what *kwamba/kuwa* actually mean, it is important to note here that the relationship between these factors (predictor variables) and complementizer choice is in no way absolute. Though sometimes strongly predictive of complementizer choice, the presence of any one predictor variable (e.g. an Emotive Factive predicate, or a third-person matrix subject) provides only enough information to yield the model's 'best guess' as to which of the two complementizers is most likely to be realized. Put succinctly, the correlations between the predictor variables outlined by the model and the appearance of *kwamba* or *kuwa*, though often accurate/informative, merely provide information pertaining to the likelihood of one complementizer being realized instead of the other, and are therefore gradient and inexact. Nonetheless, it remains clear that we need to revisit the accepted distributional account of *kwamba* and *kuwa* as being in free variation, as our model is crucially shown to predict complementizer choice at above chance accuracy. In fact, given that our regression model is based on a limited number of factors (e.g. matrix subject person, matrix predicate class, and the presence of negation), the fact that we still see a significant degree of predictive ability strongly suggests that the descriptive, free-variation account is far too broad of a description for the distribution of *kwamba/kuwa*.[15] One would expect that, given the addition of other factors argued to influence complementizer choice (e.g. mood of the embedded clause, presence of a topicalized/focused element, matrix/subordinate subject match, etc.), the predictive

[15]Though this model is only near acceptable in terms of its predictive efficiency, the fact that it is capable of predicting complementizer choice at all is somewhat surprising given that it is driven only by three factors.

ability of this model would only increase in power. Though the investigation of such additions presents an interesting area of future research, this line of questioning is beyond the scope of the current project, and, as such, will be set aside for the time being. What is important here is that, given the current model, complementizer choice is predictable at an above chance rate, meaning we must reevaluate the accepted distribution of *kwamba/kuwa*. Furthermore, based on the results of the sub-corpus analysis (Bunge), the same patterns that were shown to hold for the broader corpus were also shown to hold for one of its sub-corpora, suggesting that these results may indeed reflect a pervasive correlation between the factors under investigation and complementizer choice.

Having identified that the results of our regression analysis do not fall in line with the expected (i.e. at-chance) distribution of *kwamba/kuwa*, we can now ask how this sort of probabilistic model of complementizer choice fits into a syntactic analysis of complementation in Swahili. Specifically, if we take the results of our regression model to be true, that is, that the use of either *kwamba* or *kuwa* can be predicted – but crucially not determined – by the presence of certain factors in the matrix clause, what does this say about the syntactic representation of selected clause-embedding in Swahili? More concretely, if *kwamba/kuwa* aren't as freely interchangeable as once thought, how do we go about marrying a traditional, selection-based syntactic analysis with a probabilistic model of complementizer likelihood?

Operating under the base assumption that the null alternative to the regression model (i.e. that the two complementizers in question are freely interchangeable and identical in meaning) is false, there are broadly two avenues one can take to approach the issue of integrating this probabilistic model into the syntax. One possible analysis, let us refer to this as Hypothesis 1, would be to interpret the results of the likelihood model as evidence in favor of a competition-based model of complementizer choice in Swahili. That is, given that our regression model functions on occurrence likelihoods, one could argue that *kwamba* and *kuwa* simply represent two allomorphs of the same syntactic head and, as such, are subject to slightly different distributional constraints. Under such an analysis, it is unclear as to whether the two complementizers in question need convey the same general meaning. Any variability in occurrence likelihood could simply arise from an arbitrary, associative relationship between a candidate complementizer and some element of the matrix clause, likely as a result of the lexical

semantics of the two complementizers' verbal forms (i.e. predictability would arise due to some association between their individual verbal meanings and the specific factors that predict each complementizer). In this sense, one could argue in favor of a sort of stochastic syntax, in which the selectional probability of either complementizer, is influenced – but not determined – by a subset of syntactic and/or discourse-level factors, likely including a speaker's subjective preference.

Under an alternative analysis, which we will refer to here as Hypothesis 2, one could instead interpret such a probabilistic model as evidence to the contrary; that is, if the appearance of either *kwamba* or *kuwa* is conditioned by the individual factors specific to each, then it may well be the case that these two complementizers, while identical in syntactic function, convey distinct discourse-pragmatic information. That is, one could argue that the reason why these two complementizers exhibit some degree of predictability based on matrix subject person, matrix predicate class, and the presence of negation is because they simply convey different meanings. If this were to be true, one would expect to find some degree of uniformity across the specific factors that predict a given complementizer, as such an analysis would crucially attribute the distributional associations between *kwamba/kuwa* and their respective predictors to subtle differences in their *meaning*.

In this case, the two complementizers in question would be analyzed as serving slightly different functions at the syntax-semantics/pragmatics interface, something that has been most readily accounted for under an articulated analysis of the C-domain (Rizzi, 1997), which crucially splits the information canonically encoded on C^0 into distinct functional projections that each convey different information/meanings.

Given these contrastive hypotheses (rewritten below), one may now ask which of the two is most strongly supported by the regression model used in this study. That is, do these results favor an analysis in which the factors predicting a given complementizer are arbitrary insomuch as they do not serve a unifiable semantic-pragmatic purpose (Hypothesis 1)? Or, conversely, do they support an analysis in which a complementizer's specific predictors share a 'common thread' that links the function of the complementizer with the shared meaning(s) of the factors that predict it (Hypothesis 2)?

Hypothesis 1.) *Kwamba/kuwa* are morphological allomorphs of the same head. The likelihood of their appearance is conditioned by a set of **arbitrary** factors which have arisen as a result of their contrastive diachronic pathways.

Hypothesis 2.) *Kwamba/kuwa* convey distinct meanings. The likelihood of their appearance is conditioned by a set of **specific** factors that can impact the meaning of the embedded clause.

In the remainder of this discussion, I will make the case that the distinct sets of predictive factors highlighted by the regression model provide strong evidence in favor of an analysis like that outlined in Hypothesis 2. Specifically, I will argue that the combined set of factors that predict the appearance of *kwamba*, in particular, suggest that *kwamba* is sensitive to the notion of an "evaluator" – someone who evaluates the truth of the embedded proposition. Specifically, the choice between *kwamba/kuwa* in Swahili reflects an attitude holder's confidence in the truth of the embedded clause; *kwamba* is used to convey that the (most) local attitude holder endorses the truth of the embedded clause, while *kuwa*, conversely, is used to remain neutral about the truth of the embedded clause. I support this analysis with a brief look at the additional Swahili complementizer *ati/eti*, which has been argued to encode doubt about the truth of the embedded proposition in reported speech (Massamba, 1986). I will conclude with an attempt to integrate this analysis into the existing theoretical landscape on the topic of the articulated left-periphery. As we will discuss, however, we will ultimately find that this particular complementizer system may not fit as nicely into a *Force-Fin* type model of the C-system as would be expected.

7. Analysis

In order to address the predictions of Hypothesis 2, namely, that factors can impact the meaning of the embedded clause and share a 'core' feature/meaning with the other factors that predict the same complementizer, let us now revisit the results of each individual factor in an effort to identify a) any possible relationships between the different factors that predict each complementizer, and b) whether such relationships, if found, can tell us anything about the discourse-pragmatic contributions (i.e. the meanings) of the two different complementizers.

7.1 Matrix Predicate Class

Let us begin with matrix predicate class. As discussed in the results section (Section 5.3), of the seven different predicate classes investigated, five were found to be significantly correlated with complementizer choice, albeit at varying levels of significance. When split by predicted complementizer, Speech Act and Response Predicates were shown to pattern together as significant predictors of *kuwa,* while Emotive Factives, Doxastic Factives, and Doxastics (i.e. ambiguously factive/non-factive Doxastic predicates) were shown to conversely predict *kwamba*.

The three predicate classes that correlate with *kwamba* (e.g. Emotive Factives, Doxastic Factives, and Doxastics) share the property of being attitudinal; each of these classes strictly contains predicates that describe some attitude or belief of a local attitude holder (Djärv, 2019). That is, the lexical semantics of these three classes – which we will henceforth refer to as Attitude Predicates – all share the property that the local (matrix) subject hold an attitude or belief about the propositional content of the embedded clause (i.e. about its proposition, P). As illustrated below, attitude predicates like English *know* (Doxastic) or *be happy* (Emotive Factive) attribute a belief about P to the local subject/attitude holder, in this case the subjects Mary and Anna (30).

(30) a. Mary believes that [$_P$ Bill is moving to Canada]

b. Anna is happy that [$_P$ Lisa got the job]

(Adapted from Djärv, 2019)

Given the fact that our model identifies only these attitude predicate classes as being significantly predictive of *kwamba*, we may tentatively reason that this particular complementizer encodes some information that is relevant to the presence of an attitude holder.

While such a classification does well to capture the identity of the predicates classes that correlate with *kwamba*, this sort of attitude-based dichotomy runs into immediate issues when attempting to account for the predicates that do not correlate with *kwamba* (e.g. Response Predicates and Speech Act Predicates). The primary issue here is that, if we assume the correlation between attitude predicates and *kwamba* to be evidence that this particular complementizer encodes information relative to the presence and/or beliefs of an attitude holder, we would necessarily expect the predicate classes that do not correlate with *kwamba* to also not be descriptors of attitudes. Put succinctly, if *kwamba* correlates with predicates that require an attitude holder, we would generally expect the predicates that correlate with *kuwa* to lack such a requirement. However, under the predicate classification employed in this study (Hooper & Thompson, 1973), the predicate classes that were shown to correlate more strongly with *kuwa* actually don't follow this expected pattern. Rather, a number of Response Predicates (e.g. *kubali*, 'agree') and even Speech Act Predicates (e.g. *tamka*, 'pronounce') appear to attribute to their subject a clear belief about P, despite the fact that, again, their broader predicate classes do not correlate with *kwamba*.

Given the variability among this type of predicate, I conducted a post-hoc analysis of these classes, looking at each individual predicate. As the data in Table 20 demonstrate, we do in fact find that attitudinal reportative predicates (e.g. *kubali*, 'agree', and *tamka*, 'pronounce') correlate with *kwamba,* despite the broader category of reportative predicates correlating with *kuwa*. Moreover, once constrained to contexts including a first/third person subject – an eligible attitude holder[16] – this correlation with *kwamba* grows even stronger for a number of attitudinal reportative predicates.

[16]We assume that a first/third-person subject guarantees that the local subject is an eligible attitude holder. When not constrained to first/third-person subjects, the set of possible matrix subjects could include arguments that simply cannot hold an attitude (e.g. a document, a law, or a book).

Table 20. *Select Predicates shown to correlate with kwamba*

Predicates that correlate with *kwamba*		% of overall tokens	% of tokens w/ *first/third*-person subject
kuta, 'find'	(Doxastic)	92 %	86 %
ona, 'see'	(Doxastic Factive)	62 %	62 %
amini, 'believe'[17]	(Doxastic Non-Factive)	45 %	46 %
fikiri, 'think'	(Doxastic Non-Factive)	65 %	67 %
kubaliana, 'agree'	(Response)	68 %	71 %
furahi, 'happy'	(Emotive Factive)	87 %	86 %
tamka, 'pronounce'	(Speech Act)	61 %	64 %

These results suggest that Hooper and Thompson's classification is not entirely relevant for Swahili's complementizer selection (though it may be for other phenomena). Instead, I distinguish between two broad classes of verbs: those that necessarily attribute a belief to the subject (Attitude Predicates), and those that do not (Reportative Predicates). Moreover, I further interpret the strong correlation between attitude predicates and *kwamba* as evidence that *kwamba* encodes some information relative to the local attitude holder. In particular, I suggest that *kwamba* is used to convey that the local attitude holder holds a belief about P; namely, that P is true**.** Conversely, I suggest that *kuwa* is used to remain neutral about the attitude holder's beliefs relative to the truth of P.

To tease apart the implications of this analysis, let us consider an example from the corpus involving an attitude predicate, specifically the Emotive Factive predicate, *furahi*, 'be happy'. As

[17]Though technically skewed toward *kuwa*, we interpret this predicate as generally predictive of *kwamba* given that the percentage at which it occurs with *kwamba* (45%) is higher than the overall percentage of *kwamba* tokens (40%) in the corpus. Moreover, when limited to contexts involving only *first*-person subjects, *kwamba* occurs with 56% of *amini* tokens. This relatively low percentage may be evidence that the use of *kwamba* to encode relative belief – which we will discuss in the coming sections – is redundant in the context of a predicate, like *amini*, that necessarily describes attitude holder commitment to a belief about P as part of its lexical semantics.

a factive predicate, *furahi* provides a nice illustration of how complementizer choice may reflect relative belief in the truth of P, as this predicate presupposes the truth of the embedded clause (i.e. it commits the *speaker* to the truth of P). Consider now the following example (31), which here occurs with the complementizer *kuwa*.

(31) a-na-furahi *kuwa* [$_P$ ndoa yake na Real i-li-yo-kumbwa na
1SM-PRES-happy COMP 9.marriage 1.POSS with Real 9SM-PAST-9REL-push.PASS with
matatizo … i-na-anza upata uhai]
6.problems 9SM-PRES-begin to find 14.life
"He is happy that his troubled marriage with Real (Madrid) is beginning to find life."

On the proposed analysis, by using *kuwa* in (31), the speaker opts to remain neutral about the local attitude holder's beliefs relative to truth of P when making the attitude report (i.e. the speaker doesn't convey that the matrix subject endorses the truth of P), potentially to avoid overcommitment to an attitude that the speaker can only report on based on real world knowledge and doxastic reasoning. Interestingly, this particular example (31) may also involve speaker-oriented meaning in the embedded clause itself, specifically related to the use of *matatizo*, 'troubles/problems'. That is, it may be the speaker – as opposed to the subject – that judges the player's 'marriage' with Real Madrid as 'troubled'. If so, the presence of such a speaker-oriented modifier in the embedded clause may further constrain complementizer choice (or rather, the choice of complementizer may constrain the orientation of the modifier); it is possible that, under an attitude predicate, embedded modifiers may only be subject-oriented if contained in a clause headed by *kuwa,* since *kuwa* is used to avoid directly commenting on the local subject's beliefs.

If, however, the speaker in (31) would have opted to use *kwamba* in place of *kuwa*, the statement would have instead been interpreted as making an explicit claim about the local attitude holder's beliefs; the use of *kwamba* would signal that the speaker is confident in the local subject's beliefs relative to the truth of P. The prediction here is that only with *kuwa* could a speaker follow an utterance like (31) with a statement of uncertainty (e.g. *lakini sina hakika*, "but I'm not sure") that directly refers to the speaker's confidence/knowledge of the local attitude holder's belief set (e.g. the speaker could hedge express uncertainty about the attitude holder's beliefs). The use of *kwamba* would be predicted to be incompatible with such a follow up statement, as its use would encode attitude holder commitment to the truth of P (i.e. it would be

odd for the speaker to admit uncertainty about the attitude holder's beliefs after directly reporting on the attitude holder's beliefs). This can be seen illustrated in (32) below, where the use of *kwamba* in the context of the attitude predicate, *furahi*, 'be happy', must reflect the beliefs of the matrix subject, not the speaker (i.e. *kwamba* encodes that Mramba believes P).

(32) Mramba a-na-furahi kwamba [*P* suala lake li-me-guswa halafu pia na Sumatra]
Mramba 1SM-PRES-happy COMP 5.issue 5.POSS.3SG 5SM-PFV-touch.PASS also by Sumatra
"Mramba is happy that his issue has also been touched/felt by Sumatra"

7.2 Matrix Subject Person

Moving now to matrix subject person, the results of the regression analysis for this factor can be summarized as follows. Ultimately, it was found that the model picked out first-person and third-person subject morphology as being significantly predictive of complementizer choice, albeit not in the same direction; first-person subject morphology was shown to correlate with the use of *kwamba*, while third-person subject morphology was shown to contrastively predict the use of *kuwa*. This sort of person-specific dichotomy seems to quite readily support an analysis whereby *kwamba* and *kuwa* convey distinct meanings. I interpret this data as evidence that *kwamba* may also encode information about the *speaker*'s beliefs about the embedded proposition. Specifically, I suggest that, in the absence of a local attitude holder (e.g. with attitude predicates), *kwamba* is speaker-oriented.

Crucially, this person distinction is driven by two factors: predicate class, and the ability of a speaker to make a confident report on the beliefs of an attitude holder. Of these, predicate class determines whether the matrix clause necessarily contains a local attitude holder (the local subject), and as such, determines whether *kwamba* reflects the beliefs of the matrix subject, or the speaker. With attitude predicates (ATTITUDEPRED) – which crucially attribute a belief to the matrix subject – *kwamba* encodes that the matrix subject (the most local attitude holder) endorses the truth of the embedded proposition (33). In the absence of an eligible attitude holder in the matrix clause, e.g. with *purely* reportative predicates (REPORTPRED) that do not attribute a belief to the local subject, *kwamba* encodes the speaker's commitment to the truth of P (34).

(33) [X ATTITUDEPRED [*kwamba*$_{X}$ P]]
→ X believes P is true.

(34) [X REPORTPRED [*kwamba*$_{Speaker}$ P]]
→ Speaker believes P is true

The difference becomes apparent when we look at the class of reportative predicates. To illustrate, consider the following examples involving the reportative predicate *sema*, 'say'. Unlike with attitude predicates, the use of *kwamba* in (35) cannot encode the relative belief of the matrix subject, as *sema* – a reportative predicate – presumably doesn't attribute an attitude to its local subject. Unable to anchor to the matrix subject, *kwamba* anchors to the next closest attitude holder, which we assume to be the speaker (see Pearson, 2013 for one analysis that syntactically represents the speaker in the left-edge of the matrix clause). As such, the use of *kwamba* in (35) encodes that the *speaker* endorses the truth of the embedded clause, even if the matrix subject and the speaker do not co-refer (e.g. with a non-first-person subject). To that end, the use of *kwamba* in (35) contributes the same meaning as the use of *kwamba* in (36) despite the two examples having saliently different matrix subjects; though the latter example includes a first-person matrix subject, both reflect *speaker* commitment to the truth of P.

(35) yeye a-li-sema kwamba [*P* ha-ku-sikia]
3SG 1SM-PAST-say COMP NEG.1SM-PAST-hear
"he said that he didn't hear"

→ Speaker believes P (that the local subject didn't hear) is true.

(36) na ni-li-sema kwamba [*P* katiba i-li-vunjwa]
and 1SG-PAST-say COMP constitution 9SM-PAST-break. PASS
"and I said that the constitution was broken."

→ Speaker believes P (that the constitution was broken) is true.

In contrast, *kuwa* is used in reportative contexts to reflect speaker neutrality, just as it is used to remain neutral about the local attitude holder's beliefs with attitude predicates. Under *sema*, 'say', and *kupiga kelele*, 'shout', for example, the speaker can use *kuwa* to avoid explicitly reporting on their beliefs relative to the truth of P (i.e. to report on speech without conveying any subjective attitude(s) relative to its content). As such, the use of *kuwa* in (37) - (39) can be seen as contributing no extra information to that included in the speech report, whereas the use of *kwamba* in (35) - (36) encodes information about the speaker's subjective interpretation of the embedded proposition, in addition to the reported speech.

(37) Bush wa Marekani a-li-sema kuwa [*P* Wapalestina wa-na-hitaji viongozi wepya]
Bush 2.GEN America 1SM-PAST-say COMP 2.Palestinians 2SM-PRES-need 8.leaders 8.new
"President Bush said that the Palestinians need new leaders."

(38) ni-li-sema kuwa [*P* sera yangu ni uwazi na ukweli]
1SG-PAST-say COMP policy 1SG.POSS be 14.space and 14.truth
"I said my policy is openness and truth."

(39) wao wa-na-kuja hapa … wanapiga tu kelele kuwa [*P* Dk. Slaa
3PL 2SM-PRES-come 16.LOC 2SM-PRES-hit only 9.noise COMP Dr. Slaa
a-li-iba fedha]
1SM-PAST-steal money
"they are coming here and shouting that Dr. Slaa stole money."

It is worth noting here that, as discussed previously, the distinction between attitude predicates and reportative predicates is often a difficult line to draw, as even the most "purely reportative predicates" (e.g. *sema*, 'say') sometimes exhibit attitude-oriented properties. Consider, for example, the use of *kwamba* in the following (40). Based on our analysis above, *kwamba* in this context would most likely be interpreted as denoting the speaker's subjective endorsement of P, as the reportative predicate, *sema*, 'say', does not necessarily attribute a belief to the matrix subject.

(40) mtu a-na-sema kwamba [*p* hii ha-i-nufaishi kabisa]
1.man 1SM-PRES-say COMP 9.DEM NEG-9SM-benefit absolutely
"the man says that this is absolutely useless"

However, one could just as easily argue that the local subject of (40) necessarily believes that P is true; assuming that the local subject is not intentionally obscuring the truth, we would expect the assertion, [X is useless], to reflect the local subject's beliefs, as the notion of 'usefulness' is by all accounts subjective, and therefore necessitates the existence of a judge (i.e. the local subject must believe [X is useless] if they judge X as useless). A more detailed study – and one that incorporates speaker judgement data – is needed to control for potentially ambiguous verbs like *sema* `say'. As such, I will leave this issue of ambiguity to future experimental work and/or extensions of this corpus study.

With that being said, the dichotomy between the two predicate classes stands as described above so long as we interpret 'purely' reportative predicates as those that crucially do not allow

the local subject to hold an explicit belief about P; if the local subject can hold a belief, then the predicate (at least in that particular context) will pattern with the broad class of attitude predicates. In the absence of a local attitude holder, *kwamba* falls into the general class of elements that convey speaker-oriented meaning (Papafragou, 2006; Ernst, 2009).[18] A speaker uses *kwamba* to indicate that they are confident in the truth of the embedded clause – even if nothing in the sentence asserts that the embedded clause is true. This is again illustrated in examples like (41) below, where the use of *kwamba* can be interpreted as an endorsement of speaker belief in the truth of P (i.e. the speaker believes that the police are indeed continuing their investigation).

(41) a-li- sema kwamba [$_p$ polisi wa-na-endelea na upelelezi]
1SM-PRES-say COMP 9.police 2SM-PRES-continue with 14.investigation
"he said that the police are continuing their investigation"

Shifting gears slightly, the fact that reportative predicates (e.g. *sema*, 'say') are shown to broadly correlate with *kuwa* is relatively surprising given that the most local attitude holder in these contexts is usually the speaker. Based on our discussion, we would conversely expect these predicates to be more predictive of *kwamba*, as the speaker in question need only account for their own subjective beliefs about P, seeing as there is (generally) no local attitude holder. With that being said, there are a couple of issues to be raised here. First, it is not clear that reportative predicates can be analyzed with respect to the impact of subject person in the same way that attitude predicates can. Given that reportative predicates crucially do not *require* anyone (including the speaker) to carry a belief about P, it is unclear whether a speaker would necessarily report their own beliefs about P – as we see in (41) – as often as they would with attitude predicates. Moreover, even if this lack of a required attitude holder does not influence the chances of a speaker making a direct report about their beliefs relative to P, one could still

[18]While reportative predicates work as a deterministic class in our predicate-based dichotomy, they largely serve as a catch all for all contexts without a local attitude holder. In this sense, we also include in the 'reportative predicates' class any other contexts in which there isn't an eligible attitude holder in the matrix clause (e.g. with expletive subjects, existential constructions, etc.). This can be seen in (23) below, where the absence of a local subject renders the speaker as the only possible attitudinal anchor for *kwamba.*

(a) lakini ni kweli kwamba watumishi ndani ya Sekta ya Afya ni wachache
but be true COMP 2.workers in 9.sector 9.GEN health be few
"But it is true that there are few workers in the health sector"

argue that it is unlikely that a speaker would commit to the truth of reported speech more often than they would stay neutral about that speech. That is, it is unlikely that a speaker would more often commit themselves to the truth of the propositional content of reported speech than stay neutral about it, as they presumably lack sufficient evidence to explicitly endorse the truth of the embedded proposition in most reported speech contexts.

To that end, I further argue that the use of *kwamba/kuwa* overall is necessarily contingent on the speaker's ability to confidently report on the beliefs of an attitude holder. That is, given that a speaker is intrinsically more aware of their own subjective beliefs than they are of a third-person subject's, they are significantly more likely to hedge a overt report about the beliefs of a third-person subject by using *kuwa*. I suggest that this subjectivity bias is at least partially responsible for the broad distinction between first and third-person subjects, though not necessarily more so than predicate class.

Before concluding this section, let us briefly consider some secondary evidence relevant to our analysis of *kwamba* as being (sometimes) speaker oriented. Specifically, let us consider the discourse-pragmatic properties of another Swahili complementizer, *eti*. In his work on indirect speech reports, Massamba (1986) argues that the use of the complementizer *eti* carries a general implication of surprise (on behalf of local subject or the reporter), and a stronger implication of doubt, specifically on behalf of the reporter (i.e. the speaker). This can be seen illustrated in (42) below, where the use of *eti* implies that the speaker (the reporter, not the subject Fatma) casts doubt on the truth of the original speech (i.e. the content of the embedded clause).

(42) Fatma a-me-sema ***eti*** [$_p$ h-a-ku-mw-ona mgeni]
Fatma 1.SM-PFV-say COMP NEG-1.SM-PST-1.OM-see stranger
"Fatma said that she didn't see the stranger"

(Adapted from Massamba, 1986)

In this example, the speaker of (42) uses *eti* to convey their doubt about the truth of what Fatma said; the speaker doesn't believe or isn't sure that Fatma didn't see the stranger (i.e. they doubt P).

Turning back to our discussion of matrix subject person, we can identify two salient properties of the complementizer *eti* that may apply to our analysis of *kwamba/kuwa*. First, it is apparent that *eti* can encode the speaker's beliefs about the propositional content of the

embedded clause – at least in reported speech contexts. The second potential application to our analysis, while not necessarily convincing in and of itself, is that the pragmatic function of *eti* – to cast doubt on the truth of P – finds a logical counterpart in our proposed description of *kwamba* as conveying commitment to the truth of P. More work is necessary to determine whether *eti* can be *subject*-oriented under an attitude predicate, as we propose for *kwamba*.

Taken together, I interpret the dichotomy between first and third-person subjects as evidence that complementizer choice in Swahili must not only reflect an attitude holder's relative beliefs about the truth of the propositional content of the embedded clause, but also the speaker's ability to confidently report about that attitude holder's beliefs. Specifically, I argue that, in the absence of a local attitude holder, *kwamba* is used to encode the beliefs of the speaker, rather than the beliefs of the local subject. I further argue that the correlation between subject person and complementizer choice reflects the fact that speakers are inherently more capable of making strong attitude reports (e.g. commitment to the truth of P) about their own beliefs than another person's beliefs. The predictions of this analysis can be seen summarized in Table 21 below.

Table 21. *Complementizer Meaning by Predicate Class*

	ATTITUDE PREDICATES (*subject*-oriented)	REPORTATIVE PREDICATES (*speaker*-oriented)
kwamba	used to convey that the *local subject* believes P.	used to convey that the *speaker* believes P.
kuwa	used to remain neutral about the *local subject's* beliefs toward P.	used to remains neutral about the *speaker's* beliefs toward P.

7.3 Negation

Turning now to the final factor investigated in this study, the regression model identified the presence of negative morphology on the matrix verb as being significantly predictive of *kwamba*, while the absence of negation was conversely shown to predict the appearance of *kuwa*.

Although negation appears to pattern with first-person subjects and attitude predicates based on this data (i.e. it correlates with *kwamba*), it is not immediately clear how this correlation relates to our proposed analysis of complementizer choice as encoding relative belief. Moreover, the fact that the presence of a negative operator in the matrix clause correlates with *kwamba* would seemingly suggest that the presence of other matrix operators, most notably, question operators, should also correlate with *kwamba,* as these two matrix operators have been shown to pattern together in other dual-complementizer systems (Roussou, 2010). Though the relationship between question operators and complementizer choice was not investigated as part of this project, a correlation – if found – between *kwamba* and the presence of a question operator could feasibly be explained through interrogative shift; by using *kwamba* in a question, the speaker casts the addressee as the salient attitude holder, and essentially requests that they make a self-oriented attitude report to confirm or deny their beliefs about the information introduced in the question.

With that being said, it is worth noting that a dominance analysis found negation to be (by far) the least powerful predictor in the regression model. Moreover, raw data counts seem to suggest that the predictive power of negation may simply be an effect of a more general correlation between negation and first-person subjects, as negation is significantly more likely to occur in the context of a first-person subject than a third-person subject given that the latter of these is much more pervasive across the complementizer corpus (see Table 22 and Table 23).

Table 22. Negation Raw Statistics

kwamba			*kuwa*		
Total Instances w/NEG	Proportion of Total Tokens w/NEG	Proportion of Total *kwamba* Tokens	Total Instances W/NEG	Proportion of Total Tokens w/NEG	Proportion of Total *kuwa* Tokens
493	.586	.048	349	.414	.022

Table 23. *Negation x Matrix Subject Statistics*

	kwamba			*kuwa*		
	Tokens with w/NEG	Total Token Count	Proportion of *kwamba* Tokens	Tokens with w/NEG	Total Token Count	Proportion of *kuwa* Tokens
1SG	113	1672	.068	45	598	.076
1PL	64	1424	.045	27	672	.040
***First*-person**	**177**	**3096**	**.057**	**72**	**1270**	**.057**
2SG	14	111	.126	14	94	.149
2PL	1	131	.007	0	43	-
***Second*-person**	**15**	**242**	**.062**	**14**	**137**	**.102**
3SG	88	2517	.035	73	6330	.012
3PL	47	1432	.033	65	2824	.023
***Third*-person**	**135**	**3949**	**.034**	**138**	**9154**	**.015**

Given this skew toward first-person subjects, I will suggest that the relationship between negation and complementizer choice is largely an artifact of the curious correlation between negation and first-person subjects, as matrix subject person is by far the strongest predictor of complementizer choice as outlined by the model. Further work will be needed to address the negation facts and their effect on the present analysis.

7.4 The Left-Periphery

With respect to how the current proposal fits into previous research on the left-periphery, the big question to address is whether the distinction between *kwamba/kuwa* can be mapped to distinct projections in the C-system (see Figure 6) – specifically the heads Force and Fin (Rizzi, 1997).

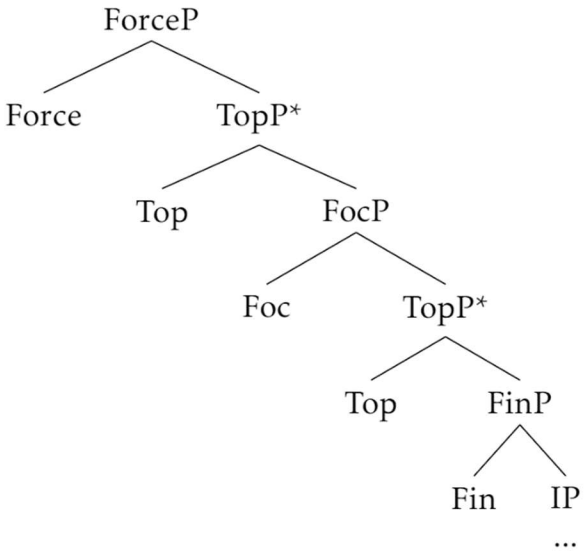

Figure 6.
Articulated C-System

Given the nature of the Force/Fin distinction (Rizzi, 1997; Roussou, 2010; Vecchio, 2010), we would expect the complementizer that maps to the higher projection (Force) to interact with elements of the matrix clause (e.g. matrix operators, matrix subject, matrix predicate), and the complementizer that maps to lower projection (Fin) to interact with elements of the embedded clause (e.g. mood). As such, the results of the regression analysis would seem to suggest that *kwamba*, given its correlation with first-person subjects, attitude predicates, and the presence of negation, maps to a higher position in the C-system than *kuwa*. Although the relative position of these two complementizers within the C-system is an interesting theoretical puzzle that merits further investigation, I largely reject the notion that *kwamba* maps onto a higher position than *kuwa* based solely on the data presented in this study.

Perhaps the strongest evidence against a structural distinction between *kwamba/kuwa* can be found in the fact that the two complementizers have been shown to co-occur in complementation contexts, with both linear orders (e.g. *kwamba kuwa,* and *kuwa kwamba*) attested in the HCS 2.0 corpus, as seen in (43) and (44). Though such examples are a statistical anomaly in the corpus, their existence strongly suggests that these two complementizer cannot map onto a Force-Fin type model of the left-periphery à la Rizzi (1997).

(43) ni-nge-penda kushauri kuwa kwamba Bajeti … i-ta-birike
1SG.SM-IRR-like advise.INF COMP COMP 9.budget 9SM-FUT-forecast
"I would like to advise that the budget be predictable/dependable"

(44) ukweli ni kwamba kuwa watu hawa sasa wa-me-gutuka
14.truth be COMP COMP 2.people 2.DEM now 2SM-PFV-be.alarmed
"the truth is that these people have now been frightened/alarmed"

In addition to the double complementizer facts, it is also not entirely clear whether *kwamba* and *kuwa* differ in their ability to interact with material in the matrix clause. Under the broadest interpretation of our analysis, only *kwamba* is anchored to an attitude holder in the matrix clause, not *kuwa*. Given that the predicate class dichotomy arises from the speaker's ability/likelihood to report on the beliefs of the relevant attitude holder (i.e. attitude reports with *kwamba* are more likely in the context of attitude predicates because there is a local attitude holder who must hold a belief, whereas with reportative predicates, the speaker must report on their own beliefs relative to the content of reported speech), and, moreover, that matrix negation correlates with the use of *kwamba*, one may interpret *kuwa* as the 'default', non-matrix interfacing complementizer. Though *kuwa* being the default complementizer is corroborated by the fact that it is more pervasive in the corpus overall (representing 60% of all *kuwa/kwamba* tokens), there is considerable evidence that it is not structurally subordinate to *kwamba*, nor used in the absence of some matrix element that can interact with the C-system (e.g. negation or a local attitude holder). Remember that the correlations between these matrix elements and *kwamba* are just that, *correlations* – their presence predicts the occurrence of *kwamba*, but does not guarantee it. As such, it would be difficult to conclusively discriminate *kwamba* from *kuwa* based solely on (apparent) interface properties.

Following that point, any attempt to structurally distinguish *kwamba/kuwa* would be contingent on the broad interpretation of our analysis in which only *kwamba* encodes relative belief. As brought up in previous sections, we leave open the possibility that *kwamba* and *kuwa* actually encode *degrees of belief* – much like the belief systems proposed of other Bantu (Luhya) languages (Gluckman & Bowler, 2016; Gluckman, 2021). Under such an analysis, *kwamba* would be interpreted as encoding a strong belief about the truth of P, while *kuwa* would be

interpreted as encoding a weaker/neutral belief; crucially, though, both *kwamba* and *kuwa* would be anchored to an attitude holder in the matrix clause.

One final piece of evidence against a Force-Fin analysis is that complementizer choice does not appear to be sensitive to the size of the embedded clause; though we argue that these two complementizers encode different meanings, they can still be found in the same syntactic context, seemingly without restriction. Further research – particularly on the interaction between complementizer choice and Topic/Focus movement in the embedded clause – is needed to rule out any effect of clause size.

In summary, it appears as though the results of our regression model do not elegantly map onto a classical split-CP analysis (Rizzi, 1997). Ultimately, a Force-Fin analysis cannot account for the overall variability *kwamba/kuwa*, nor can it account for the double complementizer facts. As such, I tentatively reject the notion that *kwamba* and *kuwa* correspond to two structurally distinct positions in the left-periphery, as it is not clear that we can distinguish *kwamba/kuwa* based on their interface properties, nor their syntactic distribution. I leave open the possibility that, though absent synchronically, these two complementizers may have mapped onto two distinct projections at some point in their grammatical development. Indeed, we would expect *kwamba* – which appears to be more readily influenced by matrix material than *kuwa* – to have once held a higher position in the syntax based on its origins as a quotative marker (Russell, 1992; Guldemann, 2008), before the two complementizers (presumably) converged on a single projection.

8. Conclusion

In this paper, I provide a novel regression analysis of Swahili corpus data to argue against the notion that the Swahili complementizers *kwamba* and *kuwa* are freely interchangeable under clause-embedding verbs like *fikiri*, 'think'. With respect to the three predictive factors considered in this study (e.g. matrix predicate class, matrix subject person, and matrix negation), the regression model identified all three factors as being significantly predictive of complementizer choice, with *matrix subject person* being the most significant predictive factor, followed by *matrix predicate class*.

On the basis of these results, I propose an analysis of the Swahili dual-complementizer system (*kwamba/kuwa*) in which complementizer choice reflects a speaker's relative knowledge of the beliefs of the local attitude holder. By using *kwamba*, the speaker conveys that the most local attitude holder (in the matrix clause) believes the embedded proposition to be true. As such, the use of *kwamba* under an attitude predicate encodes that the syntactically local subject believes P is true, while under a reportative predicate (which does not attribute an attitude to the local subject), it instead encodes *speaker* commitment to the truth of P. Conversely, *kuwa* is used to stay neutral about the beliefs of the most local attitude holder. With the broad class of attitude predicates, the use of *kuwa* results in a generic attitude report; the speaker presents the beliefs of an attitude holder, but makes no judgement as to whether those beliefs are genuine or correctly reported on. Similarly, *kuwa* is used with reportative predicates to make a basic speech report; by using *kuwa*, the speaker presents an already asserted proposition (asserted by the local subject of the speech report), but makes no claims as to whether this proposition is true.

After laying out this analysis, I discuss whether the proposed system, like other dual-complementizer systems, could be accounted for under an articulated analysis of the clausal left-periphery (Rizzi, 1997; Vecchio, 2010). Ultimately, I reject the idea that *kwamba/kuwa* map to distinct projections in the C-system, as such an analysis would crucially fail to account for the double-complementizer facts, and, more generally, overall variability in complementizer choice.

Although the analysis presented here seemingly accounts for the general correlations between complementizer choice and certain factors in the matrix clause, it is important to remember that this proposal serves only as a response to the corpus facts. To that end, given the

sheer absence of speaker judgment data, and, moreover, the limited predictive capacity of the model at hand, I reserve myself to a tentative endorsement of the proposed analysis of complementizer choice in Swahili. Nonetheless, I maintain that the results of this study necessarily suggest that the accepted characterization of *kwamba/kuwa* as being in free variation is far too broad; though future study incorporating speaker judgment data is needed to tease apart the exact semantic/pragmatic contributions of *kwamba/kuwa*, the fact that complementizer choice can be predicted at all – let alone by a model consisting of just three matrix factors – is cause to question whether these two complementizers are as freely-interchangeable as reported.

References

Angelopoulos, N. (2019). *Complementizers and Prepositions as Probes: Insights from Greek* (Doctoral Dissertation, University of Calfifornia, Los Angeles). Retrieved from https://escholarship.org/uc/item/83n8m5f7

Ashton, E. (1944). Outline Grammar of Bantu. In C. Doke (Ed.). *Africa*, 14(6), 355-356. doi:10.2307/1156565

Azen, R., & Traxel, N. (2009). Using Dominance Analysis to Determine Predictor Importance in Logistic Regression. *Journal of Educational and Behavioral Statistics*, 34(3), 319–347. https://doi.org/10.3102/1076998609332754

Beaver, (2010). Have you noticed that your belly button lint colour is related to the colour of your clothing? In R. Bäuerle, U. Reyle, and E. Zimmerman, editors, *Presuppositions and Discourse: Essays Offered to Hans Kamp*, 1–34. Oxford: Elsevier.

Bresnan, J. (1970). On Complementizers: Toward a Syntactic Theory of Complement Types. *Foundations of Language, 6*(3), 297-321. Retrieved from www.jstor.org/stable/25000462

Cruschina, S. (2006). *Informational focus in Sicilian and the left periphery*. 363–386. https://doi.org/10.1515/9783110197723.5.363

Djärv, K. (2019). *Factive And Assertive Attitude Reports* (Doctoral Dissertation, University of Pennsylvania). Publicly Accessible Penn Dissertations.

Emonds, J. (1970). *Root and Structure-Preserving Transformations*. (Doctoral Dissertation, Massachusetts Institute of Technology), Cambridge, Mass.

Ernst, T. (2009). Speaker-Oriented Adverbs. *Natural Language & Linguistic Theory*, 27(3), 497-544. Retrieved February 21, 2021, from http://www.jstor.org/stable/40270294

Givon, T., & Kimenyi, A. (1974). Truth, belief and doubt in Kinyarwanda. *Studies in African Linguistics*, (Supplement 5), 95–114.

Gluckman, J. (2021). Null Expletives and Embedded Clauses in Logoori. *Syntax*.

Gluckman, J., & Bowler, M. (2016). Expletive Agreement, Evidentiality, and Modality in Logooli. In *Proceedings of SALT 26.*

Green, G. (1976). Main clause phenomena in subordinate clauses. *Language*, 52, 382-397.

Güldemann, T. (2008). Quotative indexes in African languages: A synchronic and diachronic survey. *Empirical Approaches to Language Typology*, 34. Berlin: Mouton de Gruyter.

Hegarty, M. (1990). On adjunct extraction from complements. In Cheng, L.L.S., & Demirdash, H. (Eds.), *MIT Working Papers in Linguistics*, 13. 101-124.

Hinnebusch, T. (2003). Swahili. In William J. Frawley (ed.). *International Encyclopedia of Linguistics,* 2. Oxford: Oxford University Press.

Hooper, Joan B., & Thompson, S. (1973). *On the applicability of root transformations*. Linguistic Inquiry 4(4). 465–497.

Hosmer, D., & Lemeshow, S. (2000) *Applied Logistic Regression*. Wiley Publishing.

Bartis, I., & Hurskainen, A. J. (2016). *Helsinki Corpus of Swahili 2.0 (HCS 2.0) Annotated Version*. FIN-CLARIN-konsortio, Nykykielten laitos, Helsingin yliopisto.

Karttunen, L. (1971). Implicative Verbs. Language, 47(2), 340-358. doi:10.2307/412084

Kastner, I. (2015). Factivity mirrors interpretation : The selectional requirements of presuppositional verbs. *Lingua*, *164*, 156–188. https://doi.org/10.1016/j.lingua.2015.06.004

Kiparsky, P., & Kiparsky, C. (1971). Fact. *9*(628). In M. Bierwisch and K.E. Heidolph (eds), *Progress in Linguistics*, 143-73, The Hague, Mouton.

Ledgeway, A., & D'Alessandro, R.(2010). At the C-T boundary: Investigating Abruzzese complementation. *Lingua*, *120*(8), 2040–2060. https://doi.org/10.1016/j.lingua.2010.02.003

Massamba, D. (1986). Reported speech in Swahili. In Coulmas, Matejka, Ladislav and Krystyna Pomorska (eds.). 1971. *Readings in Russian poetics: formalist and structuralist views*, 99–119. Cambridge, Mass.: MIT Press.

Meinhof, C. (1932). *Introduction to the phonology of the Bantu languages*. Berlin: Verlag von Dietrich Reimer, under the auspices of the International Institute of African Languages and Cultures (IIALC), The Carnegie Corporation of New York, and the Witwatersrand Council of Education in Johannesburg.

Mpiranya, F. (2015). *Swahili Grammar and Workbook*. New York: Routledge.

Nurse, D., & Philippson, G. (2003). *The Bantu Languages.* London: Routledge.

Papafragou, A. (2006). Epistemic modality and truth conditions. *Lingua*, 116, 1688–1702.

Pearson, H. (2013). *The sense of self*. (Doctoral Dissertation, Harvard University). Cambridge, Mass.

Rizzi, L. (1997). *On the Fine Structure of the Left Periphery*. https://doi.org/10.1093/acprof:oso/9780190210687.003.0003

Roussou, A. (2010). Selecting complementizers. *Lingua*, *120*(3), 582–603. https://doi.org/10.1016/j.lingua.2008.08.006

Russell, J. (1992). From Reanalysis to Convergence: Swahili-Amba. In: Harlow, S. J. & Warner, A. R., (Eds.), *York Papers in Linguistics*, 16, 121-138.

Thompson, K., & Schleicher, A. (2006). *Swahili Learner's Reference Grammar* (2nd ed.). National African Language Resource Center

Ud Deen, K. (2002). *The Acquisition of Nairobi Swahili: the Morphosyntax of Inflectional Prefixes and Subjects (Kenya)*. (Doctoral Dissertation, University of California at Los Angeles). Ann Arbor: UMI.

Van Heuven, W. J., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *Quarterly journal of experimental psychology*, 67(6), 1176-1190.

Vecchio, P. (2010). The distribution of the complementizers/ka/and/ku/in the north Salentino dialect of Francavilla Fontana (Brindisi). In: D'Alessandro, R., Ledgeway, A., Roberts, I. (Eds.), *Syntactic Variation: The Dialects of Italy*, 312–322. Cambridge, Mass.: Cambridge University Press

Zipf, G. K. (1949). *Human behavior and the principle of least effort: an introduction to human ecology.* Cambridge, Mass.: Addison-Wesley Press