Chapter 16

A corpus study of Swahili's dual complementizer system

Aron Finholt[a] & John Gluckman[a]
[a]University of Kansas

We examine the distribution of the complementizers *kwamba* and *kuwa* in a corpus of Tanzanian Swahili. Our findings suggest that the complementizers are not in free variation, as is standardly assumed. Instead, their use is sensitive to a variety of factors known to affect complementizer choice crosslinguistically, specifically lexical class of the embedding predicate, person features of the main clause subject, and mood of the embedded clause. Given the distinct factors shown to predict the two complementizers, we suggest that *kwamba*/*kuwa* differ in terms of "relative belief," where *kuwa* is used to express a more general belief, while *kwamba* is used to express a privately held belief.

1 Swahili's two complementizers

Swahili[1] is reported to have two functionally interchangeable complementizers, *kwamba* and *kuwa*, shown in (1).

(1) a. Hamisi a-li-ni-ambia **kwamba** a-na-penda kusoma
Hamisi 1sm-past-1sg.om-tell comp 1sm-pres-like inf.read
'Hamisi told me that he likes to read.'

[1] Swahili is not monolithic; there are a number of dialects with significant differences between them (Maho 2009). Our claims here are about the pedagogical resources and descriptive grammars of Swahili, which uniformly report the ambiguity in (1). Ultimately, we restrict the findings of this paper to Tanzanian Swahili, as used in literature, news, and government reports.

Aron Finholt & John Gluckman. 2024. A corpus study of Swahili's dual complementizer system. In James Essegbey, Brent Henderson, Fiona McLaughlin & Michael Diercks (eds.), *Pushing the boundaries: Selected papers from the 51–52 Annual Conference on African Linguistics*, 341–357. Berlin: Language Science Press. DOI: 10.5281/zenodo.14038761

b. Hamisi a-li-ni-ambia **kuwa** a-na-penda kusoma
Hamisi 1sm-past-1sg.om-tell comp 1sm-pres-like inf.read
'Hamisi told me that he likes to read.' (Mpiranya 2015: 220)

Crosslinguistically, it is relatively common to find languages that have two (or more) lexical complementizers or complementation strategies. For instance, Greek has two complementizers, *oti* and *pu*, which are used to introduce finite, indicative subordinate clauses. Unlike Swahili, however, Greek's complementizers are well-known to correlate with distinct meaningful contributions. *Pu* gives rise to factive inferences, while *oti* does not. Indeed, the crosslinguistic pattern is clear: as a rule, when a language has more than one strategy for clausal embedding, the strategies have distinct distributions and/or meanings (Boye & Kehayov 2016).

The purpose of this paper is to ask whether Swahili also fits this pattern. Do the complementizers *kwamba* and *kuwa* have different functions in introducing an embedded clause? We investigate this question using a corpus of Swahili, probing whether known factors that influence complementizer choice crosslinguistically are present in Swahili as well. We focus on three factors: the effect of lexical class, the effect of the person features of the main clause subject, and the effect of mood in the embedded clause.

Ultimately, we find positive correlations for all the factors that we look at. *Kwamba* and *kuwa* are not in free variation, but in fact have distinct distributions, affected by well-known factors. Based on these preliminary findings, we suggest that *kuwa* and *kwamba* make a distinction between knowledge bases: *kwamba* expresses a "solipsistic" belief, while *kuwa* expresses a "general" belief. Our findings situate Swahili among more well-studied dual-complementizer systems.

2 Background: factors that affect complementizer choice

A variety of factors have been observed to have an effect on complementization strategy. Our study focuses on three factors: predicate class, person features of the matrix subject, and mood of the embedded clause. Our choice of these three factors was determined by feasibility in a corpus study.

2.1 Predicate class

The most widely documented factor that has been shown to have an effect on the complement clause is *predicate class*: different classes of predicates select different kinds of complement clauses (Kiparsky & Kiparsky 1971, Hooper & Thompson 1973, Noonan 2007), among many others. Moreover, such distinctions are crosslinguistically stable. We find that the same classes tend to pattern similarly across languages. Aspectual verbs (*start, stop*) tend to appear with "reduced" or nonfinite clauses, while doxastic verbs (*believe, think*) tend to appear with finite clauses.

(2) a. Mary started/stopped smoking.
    b. Mary believes/thinks that Sue left.

There are a number of proposed classifications of embedding predicates, depending on which factors are taken into consideration. In our study, we initially coded a subset of verbs reflecting the classification in Hooper & Thompson (1973), shown in (3).

(3) a. Speech act non-factives (*say*)
    b. Doxastic non-factives (*believe*)
    c. Doxastic factives (*know*)
    d. Emotive factives (*love*)
    e. Response predicates (*deny*)

In the end, our data suggests a broad two-way classification, which collapses the classes in (3) into (a) the predicates that comment on a mental state (*Attitude verbs*), and (b) the predicates that comment on (reported) speech (*Reportative verbs*) (c.f. Anand & Hacquard 2014 for the distinction between *private states* and *communicative acts*).

(4) a. *Attitude verbs*: Doxastic non-factives, doxastic factives, emotive factives
    b. *Reportative verbs*: Speech act non-factives, response predicates

This broad division crosscuts the fine-grained classifications of the above cited authors. Still, we also note that a more sophisticated corpus analysis might reveal distinct subclasses. The division in (4) is shown to influence the choice of complementizer in Swahili, but it is still expected that predicate classes may have other effects in Swahili, for instance, govern a finite/nonfinite distinction.

Theoretically, there are a number of ways to understand the effect of predicate class. The standard explanation is simply c-selection: some classes of verbs arbitrarily select for a particular complementizer/complementation strategy (as in, e.g., Roussou 2010). Ultimately, (lexico-)semantic factors are likely responsible for why a particular class correlates with a particular complementation strategy.

2.2 Person of subject

A second factor that has been observed to have an effect on complementizer choice is the person features of the main clause subject. For instance, in Kinyarwanda, the complementizer *kongo* is not possible with a 1st-person subject.

(5) a. yiibagiwe kongo amazi yari mare-mare
3sg.forgot comp water was deep
'He forgot that the water was deep (and I doubt it).'

b. * yiibagiwe kongo amazi yari mare-mare
1sg.forgot comp water was deep
'I forgot that the water was deep (and I doubt it).'
(Givón & Kimenyi 1974: 110)

The explanation for why *kongo* is possible in (5a) but not (5b) reduces to the meaning contribution of the complementizer. In (5a), *kongo* expresses doubt based on hearsay: the speaker expresses doubt toward the beliefs of the matrix subject because the source of the information was hearsay (Givón & Kimenyi 1974). The sentence in (5b) is ungrammatical because the complementizer *kongo*[2] expresses speaker doubt, but the verb *kwibagiwa*, a factive, commits the speaker to the truth of the embedded proposition. Such an explanation is predicated on the fact that complementizers are not simply functional linkers that connect clauses, but may bear meaningful content. Note moreover that there is an interaction between the person features of the subject and verb class in (5). The effect of subject person on the use of *kongo* is revealed with factive predicates because these commit the speaker to the truth of the embedded proposition and therefore do not allow the speaker to cast doubt on the beliefs of a 1st-person subject.[3] Other predicate classes, i.e., nonfactives, are compatible with *kongo* in the presence of 1st-person subject (Givón & Kimenyi 1974).

2.3 Mood in the embedded clause

A third factor that affects complementizer choice is the mood of the embedded clause. It is widely noted, particularly in Indo-European languages, that certain complementizers are correlated with certain moods (Ledgeway 2000, Roussou

[2] *Kongo* is arguably bimorphemic, composed of the independent complementizers *kó* and *ngo* (Botne 2020).

[3] Givón & Kimenyi 1974 show that *kongo* is similarly unavailable with 2nd-person matrix subjects, an effect they attribute to the addressee status of a 2nd-person subject; since the addressee is discourse present, they can supply the correct information.

2000, 2010, Giannakidou & Mari 2021). For instance, in Greek, subjunctive mood is strictly correlated with the complementizer *na*, thus found under desideratives in (6a). In contrast, under emotive factives, only the "indicative complementizer" *pu* (giving rise to a factive inference) is available.

(6) a. Thelo na kerdisi o Janis
want.1sg that.subj win.nonpast.3sg the John
'I want John to win.'

b. O Pavlos lipate pu efije i Roxani
the Paul be.sad.pres.3sg that.ind left.3sg the Roxanne
'Paul is sad that Roxanne left.' (Giannakidou & Mari 2021: 13)

Clearly, mood in the embedded clause is also affected by the lexical class of the embedding predicate: certain predicates require certain moods. However, crosslinguistically we also find variability. In Italian, for instance, doxastic verbs like *credere* 'believe' may take either indicative or subjunctive complements.

(7) Credo che sia/è bella
believe.1sg that be.subj/be.ind cute
'I believe that she is cute.' (Giannakidou & Mari 2021: 28)

The meaning difference between indicative and subjunctive embedded verbs will not be relevant below. What is important in these data is that (a) in some languages, the mood of the embedded clause is reflected in the choice of complementizer, and (b) in some languages, the mood of the embedded clause is not always predictable from the embedding verb. In Swahili, subjunctive mood is overtly expressed with final vowel *-e*.[4]

3 Methodology

This project employs a (logistic) regression[5] analysis of Swahili clause-embedding data to address the question of complementizer choice in Swahili. The data in this project were specifically extracted from the Annotated Version of the Helsinki Corpus of Swahili 2.0 (Bartis & Hurskainen 2016), a restricted-access corpus of the Language Bank of Finland. As one of the largest Swahili corpora available,

[4] Though we remain neutral as to the exact meaning contribution of the subjunctive in Swahili, our analysis of the corpus results in §5 aligns with the account in Portner 2018, which treats subjunctive mood as involving subjective or "solipsistic" belief on behalf of the speaker.

[5] A logistic regression is a statistical model that predicts the likelihood of an observation falling into one category of a dichotomous variable given a set of defined independent variables.

the Helsinki Corpus of Swahili 2.0 is a repository of over 26 million individual tokens across four distinct sub-corpora, with each sub-corpus containing data from a different source. As such, the four sub-corpora differ slightly in the types of tokens they include, with the *Bunge* (parliament) corpus including official political documents taken from the Tanzanian Parliament between 2004-2006, the *Books* corpus including complete or partial Swahili texts published 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.[6]

Importantly, the entirety of the Helsinki Corpus of Swahili 2.0 is morphologically tagged; each word in the corpus has been 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, TAM marking, negation, and mood, for example. With respect to our focus on the distribution of *kwamba*/*kuwa* under clause-embedding predicates, such featural information allowed us to isolate and extract only those tokens in which *kwamba*/*kuwa* introduces a selected finite clause under a finite matrix verb.[7]

In total, 26,065 such tokens were identified and extracted from the corpus for our analysis. Of these, roughly 60% involved the use of *kuwa*, while just under 40% involved the use of *kwamba*, shown in Table 1. This imbalance was considered and accounted for by our regression model, and will be discussed in §3.2.

3.1 Data coding

As discussed above, this project considers the effect of three factors on complementizer choice in Swahili: class of the matrix predicate, person of the matrix subject, and mood of the embedded clause. Upon extraction from the corpus, each token was tagged according to its relevant features for each of these three factors.

[6] Although the nature of the data within the *News (old)*/*News (new)* sub-corpora precludes this corpus from being composed of strictly Tanzanian Swahili, we assume the data within the Helsinki Corpus of Swahili 2.0 to be predominantly of Tanzanian origin given that the *Bunge* sub-corpus consists of Tanzanian Parliamentary documents, and the two News sub-corpora consist of data from Tanzanian news channels.

[7] Tokens were identified and extracted using the following search string (i), which filters tokens based on the linear adjacency of a finite matrix verb (V_{Fin}), *kwamba*/*kuwa*, an optional (subject) nominal, and a finite embedded verb (V_{Fin}).

(i) V_{Fin} + *kwamba*/*kuwa* + (Noun) + V_{Fin}

To avoid any ambiguity with their infinitival verb forms, the syntactic function of *kwamba*/*kuwa* was marked as 'complementizer/conjunction' in the search string.

Table 1: Token distribution by complementizer

kwamba		*kuwa*	
Total tokens	% of overall corpus	Total tokens	% of overall corpus
10,364	.398	15,700	.602

To investigate the effect of predicate class on the choice of complementizer, we initially coded a subset of the most pervasive predicates in the corpus based on the five embedding-predicate classes employed in Hooper & Thompson 1973 (e.g. Doxastic Factives, Doxastic Non-Factives, Emotive Factives and Response Predicates). It was necessary to select only those predicates for which there are substantial tokens to make accurate statistical inferences. As noted above, ultimately, this classification was collapsed into a broad distinction between *Attitude verbs*, or those predicates that attribute a mental state to their local subject, and *Reportative verbs*, or those predicates that introduce (reported) speech. When considered in our regression model, the factor 'Predicate Class' describes the class identity of the matrix predicate of a particular token (i.e. whether it as an *Attitude verb*, *Reportative verb*, or an unmarked baseline verb). For the sake of illustration, a few exemplars of these two classes are provided in Table 2.

Table 2: Predicate classification

Attitude		**Reportative**	
-amini	'believe'	*-ambia*	'tell'
-dhani	'guess'	*-jibu*	'answer'
-fikiri	'think'	*-ongeza*	'add'
-furahi	'be happy'	*-sema*	'say'
-tumai	'hope'	*-taja*	'announce'

The person feature of the matrix subject is the second factor considered in this project (§2.2). Using the morphological information provided in the corpus, each token was indexed by matrix subject person based on the subject morphology (e.g. the subject marker)[8] present on the matrix verb. In total, six person-number

[8] Swahili exhibits robust subject noun class agreement on the verb. The subject agreement pattern for noun class 1/2 — which includes human nouns — varies according to person and number: *ni-* (1sg), *u-* (2sg), *a-* (3sg), *tu-* (1pl), *m-* (2pl), *wa-* (3pl).

combinations were considered: 1sg/1pl, 2sg/2pl, and 3sg/3pl (equivalent to noun class 1/2). All other subject markers (e.g. noun classes other than NC 1/2) were marked as null and treated as the baseline by the model.

The final factor investigated in this project is the mood of the embedded clause (§2.3). Again using the featural information available in the corpus which encodes whether the verb is subjunctive or not, tokens were classified as either subjunctive or non-subjunctive, with the latter serving as the classificational baseline.

3.2 Data training

For each observation (i.e. instance of clause-embedding) in the data, our logistic regression model considers three independent variables (e.g. matrix predicate class, matrix subject person, and mood of the embedded clause), and predicts the likelihood of that observation, including the use of *kwamba*.[9]

As previously mentioned however, the overall distribution of *kwamba*/*kuwa* in our data set is imbalanced (see Table 1), as *kuwa* occurs in roughly 60% of the extracted tokens. 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 our candidate regression models was explicitly trained prior to model testing.

The following procedure outlines the training process. The complementizer data set was first chunked into two distinct sample populations, with 14,510 tokens being allocated to a training dataset (i.e. the development sample), and 11,554 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 being randomly selected from the data. 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 data set, 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 cannot 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.

[9] Since our data set consists exclusively of clause-embedding tokens involving either *kwamba*/*kuwa*, it is irrelevant which category (complementizer) is used as the ‘indicated’ dependent variable category — assuming that the sample size for each is equivalent. If the model instead predicted the likelihood of *kuwa* given the coded factors, the results would be the same.

4 Results

Following data training, potential models[10] 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 does, to some extent, account for the distribution of *kwamba* and *kuwa* in the data. As such, it was ultimately the maximal three-variable model — the model that includes matrix subject person, matrix predicate class and mood of the embedded clause as predictive factors — that was found to be the model of best fit. Using a basic predict function, we find that the trained model accurately predicts 72% of the test data (n=11,554 tokens). In the following sections, we walk through each factor in turn.

4.1 Predicate class

With respect to matrix predicate class, both *Attitude verbs* and *Reportative verbs* were found to be significantly predictive of complementizer choice (see Table 3). Moreover, the results of our regression analysis indicate a clear distinction between the two classes; *Attitude verbs* correlate with the use of *kwamba*, while *Reportative Verbs* instead correlate with the use of *kuwa*.

Table 3: Predicate class correlations and significance

Predicate class	**Predicted complementizer**	**Significance**
Attitude	*kwamba*	p < .001 ***
Reportative	*kuwa*	p < .001 ***

To illustrate the predictive output of our regression model, the results of our analysis for predicate class (and other subsequent factors) are presented in terms of complementizer likelihood – specifically the likelihood of *kwamba* appearing given the presence of some predictive factor. Complementizer likelihood can be understood as follows. For every token in the extracted data, the regression

[10] Potential models included all possible combinations of 1 or more predictor variables (i.e. factor) and all possible interactions between variables. Possible models therefore included one-factor models that consist of only one predictor variable (e.g. predicate class), two-factor models (e.g. predicate class + mood), the maximal three-factor model, as well as interaction models.

model considers the relevant factors at play (e.g. the class of the matrix predicate, the person of the main clause subject, etc.), and predicts which complementizer is most likely to appear given those factors. Specifically, the model assigns a predicted complementizer value based on a likelihood scale from 0 to 1, where 0 denotes a 100% likelihood of occurrence with *kuwa*, and 1 denotes a 100% likelihood of occurrence with *kwamba*. When considered at scale, this output allows us to analyze the dispersion of predicted complementizer values given specific, individual factors (e.g. predicate class) in order to get a broader view of the relationship between factor(s) and complementizer choice. The results of this analysis for predicate class are illustrated in Figure 1. When compared to the null baseline (i.e. tokens involving unclassified predicates), the dispersion of predicted complementizer values again distinguishes the two predicate classes based on the complementizer they predict; while tokens with *Attitude verbs* generally yield a predicted complementizer value closer to 1 (i.e. more indicative of *kwamba*), *Reportative verbs* yield a predicative value closer to zero (i.e. more indicative of *kuwa*).

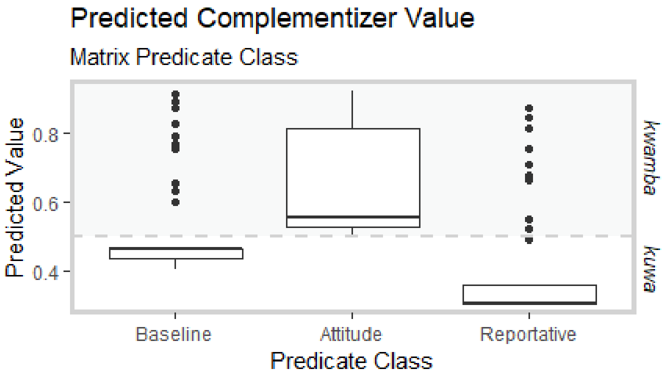

Figure 1: Dispersion of predicted complementizer values by embedding predicate class, as compared to baseline (i.e. tokens with unclassified matrix predicate)

4.2 Person of subject

Turning now to matrix subject person, the results of the regression analysis identify all six person-number feature combinations as being significantly predictive of complementizer choice (see Table 4). Though slightly variable in strength of significance, a clear pattern emerges across the six person-number combinations with respect to complementizer choice: 1st/2nd-person subject morphology correlates with the use of *kwamba*, while 3rd-person subject morphology correlates with *kuwa*. Ultimately, we focus our discussion of matrix subject person on the dichotomy between 1st/3rd-person in §5.[11]

Table 4: Person correlations and significance

Predicate class	**Predicted complementizer**	**Significance**
1SG	*kwamba*	p < .001 ***
1PL		p < .001 ***
2SG		p < .01 **
2PL		p < .001 ***
3SG	*kuwa*	p < .001 ***
3PL		p < .05 *

Using the same likelihood scale as described with predicate class in §4.1 — where 0 denotes a 100% likelihood of occurrence with *kuwa*, and 1 denotes a 100% likelihood of occurrence with *kwamba* — the dispersion of predicted complementizer values can be seen for each person-number combination in Figure 2 below. When compared to the classificational baseline (i.e. tokens involving non-1st/2nd/3rd-person subject morphology), the dispersion of the data is again indicative of a dichotomy between 1st/3rd-person subjects; 1st-person subjects correlate with *kwamba*, while 3rd-person subjects correlate with *kuwa*.

4.3 Mood of embedded clause

As for the third factor under consideration, the regression model identifies the presence of the subjunctive mood in the embedded clause as significantly predictive of complementizer choice. Specifically, it was found that subjunctive marking on the embedded verb correlates with the use of *kwamba* (see Table 5).

Considering again the same likelihood scale used in previous sections, we can compare the dispersion of predicted complementizer values for tokens that include the subjunctive in the embedded clause and those that do not. As can be seen in Figure 3, the presence of the subjunctive yields a higher predicted complementizer value (i.e. *kwamba* is more likely) than the absence of the subjunctive.

[11] We omit any discussion of 2nd-person due to lack of sufficient data. Compare the following token counts for each person/number combination: 1SG (n=1463), 1PL (n=1322), 2SG (n=129), 2PL (n=104), 3SG (n=4626), 3PL (n=2310).

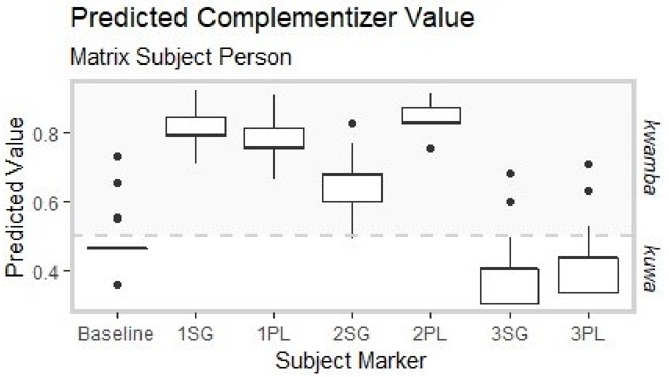

Figure 2: Dispersion of predicted complementizer values by matrix subject person morphology, as compared to baseline (i.e. tokens with any other subject marker)

Table 5: Mood correlations and significance

Mood	Predicted complementizer	Significance
Subjunctive	*kwamba*	p < .001 ***

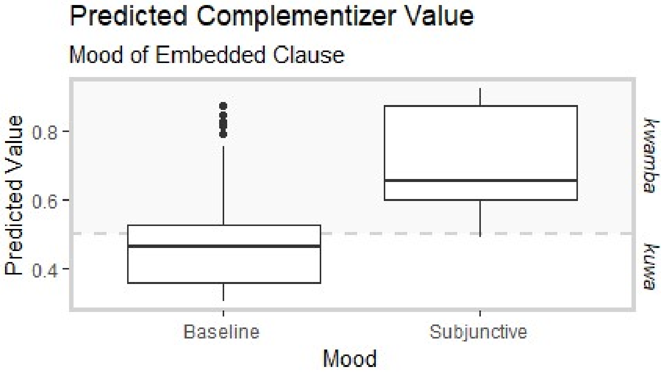

Figure 3: Dispersion of predicted complementizer values for tokens involving the subjunctive in the embedded clause, as compared to baseline (i.e. tokens without subjunctive)

4.4 Summary of results: factor strength

The results of our analysis are summarized in Table 6. Overall, we find a distinction in the specific factors that predict each complementizer, as 1st-person subjects, attitude predicates and subjunctive marking on the embedded verb correlate with the use of *kwamba*, while 3rd-person subjects and reportative predicates correlate with *kuwa*.

Table 6: Summary of correlations

Complementizer	Predictor
kwamba	1st-person Attitude predicates Subjunctive
kuwa	3rd-person Reportative predicates

Given that the results of our analysis found that all three coded factors are statistically significant predictors of complementizer choice, we conducted a follow-up dominance analysis (Budescu 1993, Azen & Budescu 2003) to identify the relative contribution of each factor to overall predictive power of the model. Using the McFadden index (McFadden 1993) as the measure for individual factor contribution,[12] our analysis found that the average contribution of matrix subject person (R^2M = 0.061) outweighs both predicate class (R^2M = 0.017) and mood (R^2M = 0.006), making it the most dominant individual factor in the model.

5 Discussion

There are two broad takeaways from our study. First, the complementizers *kwamba* and *kuwa* are not in free variation. Rather, their use is affected by a variety of factors. Second, even given the predicting factors, the choice of *kwamba* or *kuwa* is not categorical. For instance, while 1st-person strongly correlates with *kwamba*, we still find examples in which 1st-person subjects co-occur with *kuwa*. Similarly, we find that while attitude predicates occur more frequently with *kwamba*, *kuwa* still appears with such verbs.

[12] See Azen & Traxel 2009 for a more detailed discussion of measuring factor contributions in logistic regressions.

Both takeaways point to the conclusion that neither *kwamba* nor *kuwa* is directly selected by an element in the higher clause (as argued for Greek in Roussou 2010). Such a direct link would predict a categorical distinction between syntactic environments that require *kwamba* and those that require *kuwa.* Instead, it must be the case that the correlating factors we illustrate above only bear an "indirect" link to *kuwa* and *kwamba.*

Any analysis of these results should start from the observation that both *kuwa* and *kwamba* have non-complementizer functions. Synchronically, *kuwa* is the infinitival form of the copula. *Kwamba* is diachronically the infinitival form of the verb meaning 'say.' (It survives in its applicativized form *kwambia* 'tell,' shown in (1).) The informal analysis that we sketch here takes this lexical distinction as a starting place.

We suggest that the difference between *kwamba* and *kuwa* (as complementizers) is that *kwamba* situates the embedded clause from the perspective of a particular individual. This treats *kwamba* like other *say*-complementizers: it anchors the embedded clause to the local subject and projects the thoughts/beliefs/knowledge/etc. of that individual. *Kuwa*, on the other hand, situates the embedded proposition relative to a topical situation. When a speaker uses *kuwa*, they are indicating that there is a situation, in some cases a real-world situation, in which the embedded proposition is true.

The distinction between the complementizers is therefore not tied to any particular syntactic element. Rather, the complementizers play a role in the discourse; they are used to indicate something close to *relative belief*: *kwamba* indicates that the embedded clause is a "solipsistic" belief (Giannakidou & Mari 2021), while *kuwa* presents in some cases a more general belief, and in some cases remains neutral.

This characterization of the difference between *kwamba* and *kuwa* accounts for the corpus patterns above in the following way. The strong correlation between *kwamba* and 1st-person subjects in the main clause follows from the fact that speakers are self-aware: a speaker is able to confidently report her own thoughts, but may not be cognizant of the thoughts of others. This notion of self-awareness also explains the correlation between *kuwa* and 3rd-person: it is difficult for a speaker to confidently report on the thoughts of another attitude holder. On the other hand, using *kuwa*—even with a 1st-person subject—indicates potentially a more general or less "private" belief. "I believe *kuwa* P" can be employed to indicate something like "I believe (the situation is) P."

The weaker correlation between attitude predicates and *kwamba* also follows from this. Attitude predicates project the thoughts/beliefs/knowledges/etc. of an individual. A speaker will therefore use *kwamba* when they can confidently report what those thoughts are. This will of course be nearly all of the time with 1st person subjects, but may also reflect the thoughts of a third person subject.

Finally, the correlation between subjunctive and *kwamba* is accounted for in a similar manner. Assuming that subjunctive mood involves some kind of "subjectification" (Portner 2018), then the appearance of *kwamba* again is used to report a "self-centered" belief.

6 Conclusion

Our corpus study of Swahili's dual complementizer system demonstrates that native speakers use the two complementizers *kwamba* and *kuwa* distinctly. This puts Swahili in line with other more well-studied languages that have more than one complementizer or complementation strategy. Further investigation may shed light on more subtle distinctions that cannot be investigated in a corpus, like the influence of questions in the matrix clause, the affect of a topicalized/focused elements in the embedded clause, and the effect of a 2nd person subject in the main clause.

Abbreviations

Kinyamulenge has 20 noun classes. Following Bantuist convention, we mark noun classes via numerals at the beginning of nouns and verbs.

comp	Complementizer	past	Past tense
ind	Indicative mood	pres	Present tense
inf	Infinitive	sg	Singular
fv	Final vowel	sm	Subject marker
nonpast	Nonpast tense	subj	Subjunctive mood
om	Object marker		

Acknowledgements

Thanks to Chacha Wambura for help on a previous version of this work. Thanks also to the audiences at ACAL 51/52 at the University of Florida, at the 37th Annual Northwest Linguistics Conference (NWLC) at the University of British Columbia, and at the Research in Field and Formal Linguistics Seminar at the University of Kansas.

References

Anand, Pranav & Valentine Hacquard. 2014. Factivity, belief and discourse. In Luka Crnič & Uli Sauerland (eds.), *The art and craft of semantics: A Festschrift for Irene Heim* (MIT Working Papers in Linguistics 70), 69–90. Cambridge: MITWPL.

Azen, Razia & David V Budescu. 2003. The dominance analysis approach for comparing predictors in multiple regression. *Psychological Methods* 2(8). 129–148.

Azen, Razia & Nicole Traxel. 2009. Using dominance analysis to determine predictor importance in logistic regression. *Journal of Educational and Behavioral Statistics* 3(34). 319–347.

Bartis, Imre & Arvi J. Hurskainen. 2016. *Helsinki corpus of Swahili 2.0 (HCS 2.0) annotated version.* Chicago: FIN-CLARIN-konsortio, Nykykielten laitos, Helsingin yliopist.

Botne, Robert. 2020. Evidentiality in African languages. In Chungmin Lee & Jinho Park (eds.), *Between evidentials and modals*, 460–501. Leiden: Brill.

Boye, Kaspar & Petar Kehayov (eds.). 2016. *Complementizer semantics in European languages.* Berlin: De Gruyter.

Budescu, David V. 1993. Dominance analysis: A new approach to the problem of relative importance of predictors in multiple regression. *Psychological Bulletin* 3(114). 542–551.

Giannakidou, Anastasia & Alda Mari. 2021. *Truth and veridicality in grammar and thought.* Chicago: University of Chicago Press.

Givón, Talmy & Alexandre Kimenyi. 1974. Truth, belief and doubt in Kinyarwanda. In William Leben (ed.), *The papers from the fifth Annual Conference on African Linguistics*, 95–114.

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

Kiparsky, Paul & Carol Kiparsky. 1971. Fact. In Danny Steinberg & Leon Jakobovits (eds.), *Semantics: An interdisciplinary reader in philosophy, linguistics, and psychology*, 345–369. Cambridge: Cambridge.

Ledgeway, Adam. 2000. *A comparative syntax of the dialects of Southern Italy: A Minimalist approach.* Hoboken: Blackwell.

Maho, Jouni Filip. 2009. *NUGL online: The online version of the New Guthrie List, a referential classification of the Bantu languages.* Version dated March 25th, 2008. Available online at http://goto.glocalnet.net/maho/bantusurvey.html.

McFadden, Daniel. 1993. Conditional logit analysis of qualitative choice behavior. In Paul Zarembka (ed.), *Frontiers in econometrics*, vol. 4, chap. 4, 104–142. New York: Academic Press.

Mpiranya, Fidèle. 2015. *Swahili grammar and workbook*. London: Routledge.

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

Portner, Paul. 2018. *Mood*. Oxford: Oxford University Press.

Roussou, Anna. 2000. On the left periphery: Modal particles and complementisers. *Journal of Greek Linguistics* 1. 65–94.

Roussou, Anna. 2010. Selecting complementizers. *Lingua* 120. 582–603.