On complementizer choice in Swahili

SASAL 1
Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft
Universität Tübingen
February 24, 2022

Aron Finholt
University of Kansas
finhola@ku.edu, aronfinholt.com
This project investigates the distribution of the Swahili complementizers *kwamba* and *kuwa*, which are both used to introduce a finite indicative clause under clause embedding predicates like -*ambia*, ‘tell’ (1).

(1) *Hamisi a-li-ni-ambia kwamba/kuwa a-na-penda*

*Hamisi 1SM-PAST-1SG.OM-tell COMP/COMP 1SM-PRES-like kusoma*

*read.INF*

‘Hamisi told me that he likes to read.’ (Mpiranya, 2015:220)

*Kuwa* and *kwamba* are reported to be interchangeable, with no interpretive or distributional distinctions.
Distribution of *kwamba* and *kuwa*

Although such a distributional description of *kwamba/kuwa* is generally accepted, the fact that these two complementizers appear to exist in free variation is prima facie surprising.

- Each has distinct lexical origins as infinitival verb forms; *kwamba* being derived from ‘to tell’, and *kuwa* from ‘to be’.

- Interpretative differences have been reported in similar dual-complementizer systems (e.g. Greek, Italian Dialects) (Ledgeway, 2000; Angelopoulos, 2019)

- Many Bantu languages have multiple complementizers, including ‘say’-complementizers, many of which serve “evidential”-like functions (Botne, 1997; Güldemann, 2008; Diercks, 2013)
• We examine the question of complementizer choice through regression-based analysis of (Tanzanian) Swahili Corpus data, and follow-up elicitation.

• Ultimately, we find that complementizer choice in Swahili is (at least) partially predictable based on a subset of factors that have been shown to influence complementizer choice cross-linguistically.

• Based on these results, we propose a system in which complementizer choice in Swahili encodes *relative belief*.
**Methodology**

**Corpus:** Helsinki Corpus of Swahili 2.0
   → approx 25 million words
   → fully morphologically tagged

**Token Type:** Embedding Predicate+[\(_\text{CP}\) kwamba/kuwa...] 

**Total Token Count:** 26,064

The factors we investigated were chosen based on what was feasible to look at in a corpus.
Factor 1: Predicate class

Predicate class (or selection) is known to affect complementizer choice cross-linguistically (Kiparsky and Kiparsky, 1971; Hooper and Thompson, 1973; Noonan, 2007; Roussou, 2010), many others

- We initially divided up the predicates based on the classification in Hooper and Thompson (1973).
- Eventually, we collapsed these into just two categories
  - **Attitude predicates** (-fikiri ‘think’): those predicates which entail the existence of a belief-holder.
  - **Reportative predicates** (-sema ‘say’): those predicates which do not entail the existence of a belief-holder.
Factor 2: Person of subject

The person of the main-clause subject has also been shown to affect complementizer choice cross-linguistically (Kiparsky and Kiparsky, 1971; Givón and Kimenyi, 1974; Massamba, 1986)

• In Kinyarwanda, the complementizer *kongo is reported to not be possible under factives with 1st/2nd person subjects (Givón and Kimenyi, 1974).

(2) 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. * niibagiwe kongo amazi yari mare-mare
1SG.forgot COMP water was deep
[Intended: ’I forgot the water was deep (and I doubt it).’]
Factor 3: Mood

The mood of the embedded clause has also been shown to be a factor in complementizer selection (Ledgeway, 2000; Roussou, 2010), many others.

- Subjunctive mood in the embedded clause is often correlated with a particular complementizer as in, e.g., Greek.
Overview of results

Factors investigated

1. Predicate class
2. Person of subject
3. Mood

- All three factors are significant predictors of complementizer choice
- *However*, the person of the main clause subject is by far the most significant factor affecting complementizer selection.
Results: Matrix Subject Person

Key Findings

- Matrix Subject Person found to be the strongest individual predictor in the model.

- *First* and *third*-person subjects were shown to be significant predictors; *first*-person subjects correlate with *kwamba*, *third*-person subjects correlate with *kuwa*.

Our model assigned predicted values to each token in the corpus based on the likelihood of finding *kwamba* given the coded factors present. Predicted values closer to 1 indicate that *kwamba* is more likely. Predicted values closer to 0 indicate that *kuwa* is more likely.
Results: Matrix Predicate Class

Key Findings

- Matrix Predicate Class found to be second strongest predictor in the model.

- **Attitude** predicates (e.g. -fikiri, ‘think’) shown to correlate with *kwamba*, while **Reportative** predicates (e.g. -sema, ‘say’) instead correlate with *kuwa*.
Results: Mood of Embedded Clause

Our model assigned predicted values to each token in the corpus based on the likelihood of finding *kwamba* given the coded factors present. Predicted values closer to 1 indicate that *kwamba* is more likely. Predicted values closer to 0 indicate that *kuwa* is more likely.

Key Findings

- Mood of Embedded Clause found to be weakest predictor in the model.
- Presence of the subjunctive in the embedded clause (e.g. FV , -e) shown to correlate with *kwamba*.

Our model assigned predicted values to each token in the corpus based on the likelihood of finding *kwamba* given the coded factors present. Predicted values closer to 1 indicate that *kwamba* is more likely. Predicted values closer to 0 indicate that *kuwa* is more likely.
Discussion

There are two major takeaways from our corpus study.

- First, the distinct factor correlations demonstrate that *kwamba* and *kuwa* are not in free-variation.

- Second, despite these predicting factors, the choice between *kwamba* and *kuwa* is not categorical; either may appear with any of the aforementioned factors.

These takeaways suggest that *kwamba* and *kuwa* are not selected by an element in a higher clause (c.f. Roussou (2010) on complementizers in Greek); they simply interact with such elements indirectly.
Discussion

Based on the results of our corpus study, we propose that *kwamba* and *kuwa* make a semantic distinction:

- **Kwamba** anchors the embedded clause to an individual’s beliefs/thoughts/attitudes.
- **Kuwa** anchors the embedded clause to a pragmatically salient event/situation.

We assume that complementizers serve to identify propositions based on the information content of contentful “anchors”: e.g. events, individuals, mental states, etc. (Hacquard, 2006; Kratzer, 2006, 2013; Moulton, 2015).

Following Kratzer (2006), we further assume that complementizers house the function **content**, which identifies an event/individual with propositional content and returns the set of worlds modally accessible from that content (Hacquard, 2006; Moulton, 2015).
Meaning of *kwamba*

We define *kwamba* as a standard complementizer housing the function \( \text{CONTENT} \) (Hacquard, 2006; Moulton, 2015).

- With *kwamba*, \( \text{CONTENT}(e) \) is the set of worlds modally accessible from an event, \( e \), associated with the embedding verb (3).

\[
(3) \quad \left[\text{KWAMBA}\right] = \lambda P \lambda e. \forall w' \in \text{CONTENT}(e), P(w') = 1
\]

Put differently, \( \text{CONTENT}(e) \) simply provides an appropriate modal background given the embedding predicate (e.g. *believe*-type: DOX, *say*-type: RPG, Giorgi and Pianesi (1997); Mari and Portner (2021)).
Meaning of *kwamba*

The model shows a significant correlation between *kwamba* and *first*-person subjects, *ATTITUDE* predicates, and subjunctive mood. Each of these correlations follows straightforwardly from *kwamba* anchoring the embedded clause to the beliefs/hopes of an *individual*:

- **First-person subjects**: the speaker is intrinsically more aware of their own thoughts than of others’, making it easier to ‘check’ the modal base to ensure that $P=1$.
- **Attitude predicates**: such predicates are typically used to discuss the thoughts/beliefs/attitudes of local attitude holder.
- **Subjunctive mood**: crosslinguistically, subjunctive mood tends to trigger a layer of relativized belief that targets an attitude holder (Givón, 1994; Portner, 2018)
Meaning of *kuwa*

In contrast, it is a bit more difficult to determine whether *kuwa* contributes any meaning based on the results of our corpus study. Given the correlation with *REPORTATIVE* predicates, we may say that it minimally differs from *kwamba* in that it doesn’t appear to project the thoughts/beliefs/attitudes of an individual, but that is all.

To investigate whether *kuwa* contributes any additional meaning, we conducted a follow-up elicitation study with a native speaker of Tanzanian Swahili. Native speaker judgments reveal that:

- *kwamba* does appear to anchor the embedded clause to an *individual*; it sometimes results in ‘subjective’ belief reports.

- *kuwa* instead anchors the embedded clause to a pragmatically salient *event/situation*; it sometimes yields an evidential reading.
The contrast between *kuwa/kwamba* is best illustrated by modulating the available evidence in a circumstance of evaluation.

For example, when there is concrete evidence to support the truth of $P$, only *kuwa* is available. In the following context, the available evidence in the situation almost guarantees Tanzania’s winning the match.

(4) *We’re watching Tanzania play in a football [soccer] match. There is 5 minutes left to play, and Tanzania is up by 3.*

\[
i-na-onekan-a  \quad kuwa/\#kwamba  \quad Tanzania  \quad i-ta-shind-a
\]
\[
9\text{SM-PRES-seem-FV}  \quad \text{COMP/COMP}  \quad 9\text{Tanzania}  \quad 9\text{SM-FUT-win-FV}
\]
\[
\text{‘It seems like Tanzania will win.’}
\]

Here *kuwa* effectively expresses an evidential meaning; there is sufficient evidence in the situation to conclude that Tanzania will win.
Native Speaker Judgements

In contrast, only *kwamba* is available in a context where there is much less evidence that Tanzania will win:

(5)  *We’re watching Tanzania play in a football [soccer] match. It’s halftime, and Tanzania is up one to nil.*

\[i-na-onekan-a \quad kwamba/\#kuwa \quad Tanzania \quad i-ta-shind-a\]
\[9SM-PRES-seem-FV \quad COMP/COMP \quad 9Tanzania \quad 9SM-FUT-win-FV\]

‘It seems like Tanzania will win.’

The use of *kwamba* here ostensibly serves to project the speaker’s beliefs about the outcome of the match.
A similar contrast occurs with first-person subjects (6).

(6) ni-na-fikiri kuwa/kwamba Tanzania i-ta-shind-a
    1SG-PRES-think COMP/COMP Tanzania 9SM-FUT-win-FV
    ‘I think that Tanzania will win [the football/soccer match].’

The use of *kwamba* expresses what the speaker believes/hopes will occur; the speaker could simply be cheering for Tanzania.

With *kuwa*, however, the speaker is interpreted to have some extra insider-knowledge leading them to believe that P.
Discussion

The evidential properties of *kuwa* and the belief-projection properties of *kwamba* lend support to our proposed analysis:

- *kwamba* anchors the embedded clause to an individual.
- *kuwa* anchors the embedded clause to an event/situation.

With this distinction in mind, we may now present an analysis for *kuwa*. We invoke a direct evidential-type semantics for *kuwa* involving two events (de Haan, 1999; Faller, 2004; Speas, 2004; Garfield et al., 2010).

- *kuwa* asserts that there exists a (pragmatically relevant) event $e'$ that properly contains the event associated with the embedding verb, $e$, and that all accessible worlds from $e$ are $P$ worlds (7).

\[
\lambda P \lambda e. \exists e'[e \subset e'] \land \forall w' \in \text{CONTENT}(e), P(w') = 1
\]

where $e'$ is an event containing evidence pertaining to $P$. 

Discussion

*Kuwa* differs from *kwamba* in that the event content takes as its argument is contained within a larger event; *kuwa* situates the modal background associated with the embedding verb within a larger event containing evidence pertaining to the truth of P.

With this, we now have a straightforward analysis for the correlations between *kuwa* and reportative predicates/third-person subjects:

- **Reportative predicates**: make reference to a *situation* where something was uttered.
- **Third-person subjects**: the speaker could make reference to a situation to report about a *third*-person subject (e.g. the speaker heard them say P), or they could make reference to a situation that the subject themselves used to make a report.
In addition to the corpus data and elicited judgments, we find substantial diachronic and typological evidence to support the proposed analysis.

Analyzing *kwamba* as anchoring an embedded clause to an individual’s beliefs follows from its diachronic source as a *say*-verb.

- *say*-complementizers often project the beliefs of a local attitude holder in Bantu (Diercks, 2013) and other languages (Güldemann, 2008).

The evidential properties of *kuwa* may be similarly traced back to its diachronic origin as a copular verb.

- In this sense, *kuwa* is used to express that there is a relevant situation/event to support the truth of P; “the situation is that P”.
Conclusion

• Taken with speaker judgement data, the results of our corpus analysis suggest that *kwamba/kuwa* are not in free variation.

• Instead, we propose that *kwamba* and *kuwa* make a semantic distinction; *kwamba* anchors the embedded clause to an individual’s beliefs, while *kuwa* anchors the embedded clause to a pragmatically relevant event/situation.

• The evidence puts Swahili in line with other Bantu languages which make similar epistemic distinctions in the C domain, and, moreover, with languages that make a distinction between solipsistic belief (i.e. belief held by a single individual) and more general belief.
Special thanks to our consultant Chacha Wambura, my advisor and co-author John Gluckman, as well as the Bantu research/reading group (KUBantu) at the University of Kansas.


