Nordic Journal of African Studies 22(3): 129–157 (2013)

Grammatical and Lexical Comparison of the Greater Ruvu Bantu Languages

Malin PETZELL
University of Gothenburg, Sweden
Harald HAMMARSTRÖM
Max Planck Institute for Psycholinguistics, The Netherlands

ABSTRACT

This article discusses lexical and grammatical comparison and sub-grouping in a set of closely related Bantu language varieties in the Morogoro region, Tanzania. The Greater Ruvu Bantu language varieties include Kagulu [G12], Zigua [G31], Kwere [G32], Zalamo [G33], Nguu [G34], Luguru [G35], Kami [G36] and Kutu [G37]. The comparison is based on 27 morphophonological and morphosyntactic parameters, supplemented by a lexicon of 500 items. In order to determine the relationships and boundaries between the varieties, grammatical phenomena constitute a valuable complement to counting the number of identical words or cognates. We have used automated cognate judgment methods, as well as manual cognate judgments based on older sources, in order to compare lexical data. Finally, we have included speaker attitudes (i.e. self-assessment of linguistic similarity) in an attempt to map whether the languages that are perceived by speakers as being linguistically similar really are closely related.

***Keywords**: Bantu languages, lexical comparison, morphosyntactic comparison, automated cognate judgement, sub-grouping.*

This article discusses lexical and grammatical comparison and subgrouping in a set of closely related Bantu language varieties in the Morogoro region, Tanzania[1]. The comparison is based on 27 morphophonological and morphosyntactic parameters, a lexicon of 500 items and the speakers' self-assessment of linguistic similarity. The language varieties[2] in the region include Kagulu [G12], Zigua [G31], Kwere [G32], Zalamo [G33], Nguu [G34], Luguru [G35], Kami [G36] and Kutu [G37][3]. These language varieties are poorly described, as are many of Tanzania's languages (Maho and Sands 2003).

The present study makes use of a set of parameters to investigate the structural relationships between the Greater Ruvu Bantu language varieties (cf.

[1] We would like to thank Rebecca Grollemund, Birgit Ricquier, Lutz Marten, Bernard Comrie and two anonymous reviewers for their constructive comments and Mary Chambers for the proof reading. Any remaining mistakes are of course our own.

[2] By the term *language variety*, we mean variations of a language used by particular groups of people, including regional dialects.

[3] The languages are labelled according to Maho's (2009) updated version of Guthrie's (1971) divisions.

Marten et al. 2007). In order to determine the relationships and boundaries between the varieties, grammatical phenomena constitute a valuable complement to counting the number of identical words or cognates. Consequently, the focus of this comparison is on grammatical (i.e. structural) features such as morphophonological processes, noun class marking, negation and verbal morphology (tense, aspect and mood markers).

We have used automated cognate judgment methods (to be described in the following) as well as manual cognate judgments based on older sources (Nurse and Philippson 1975 and 1980, Gonzales 2002) in order to compare lexical data. Finally, we have added speaker attitudes (i.e. self-assessment) in an attempt to map whether the languages that are perceived by speakers as being linguistically similar really are closely related.

All language data stem from the first author's field work in the area (unless otherwise stated).

1. THE GREATER RUVU LANGUAGES

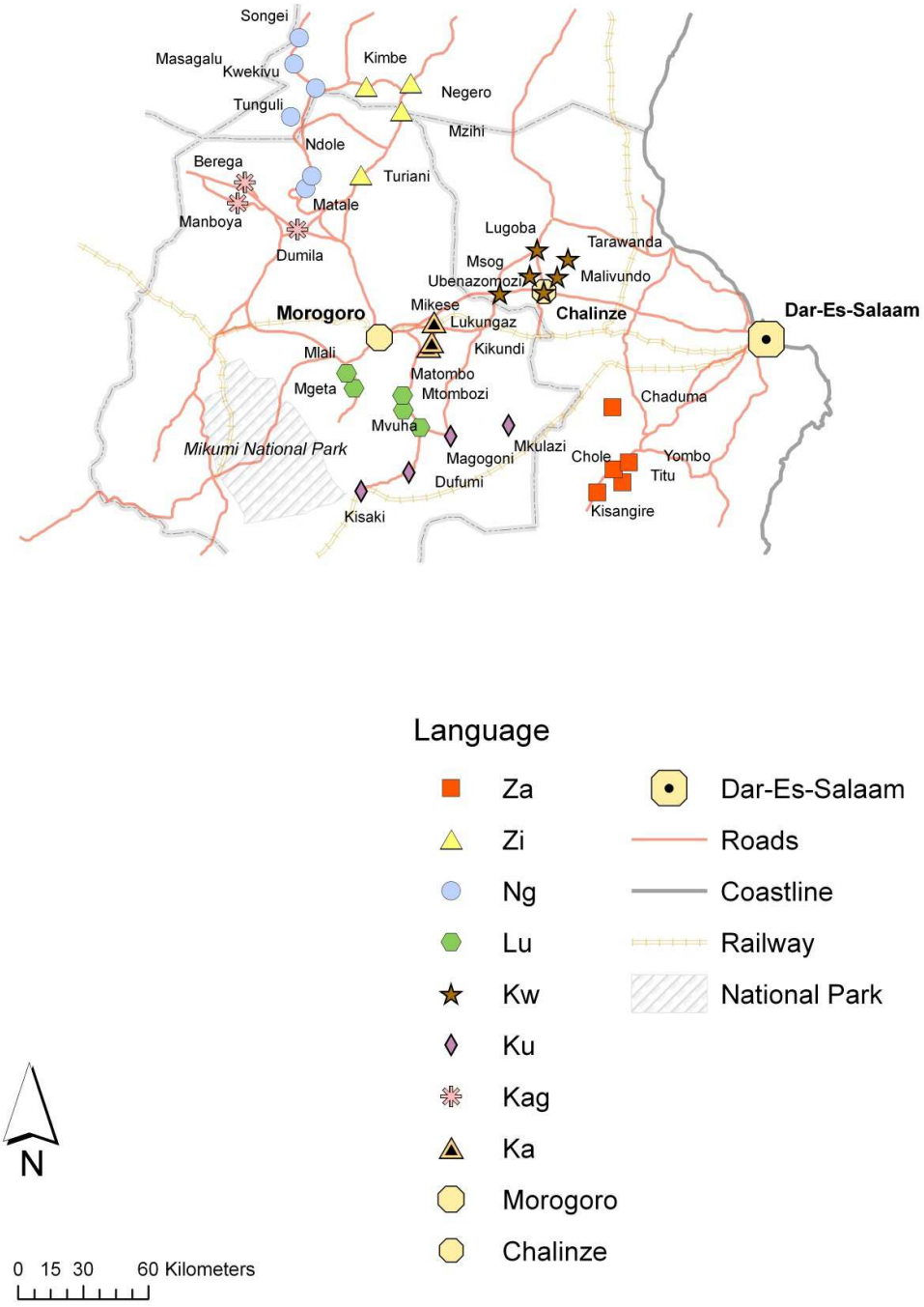

Figure 1. *Map of the linguistic centres of the language varieties*[4].
Za = Zalamo, Zi = Zigua, Ng = Nguu, Lu = Luguru, Kw = Kwere, Ku = Kutu, Kag = Kagulu and Ka = Kami

4 This geo-referenced map showing the answers to the question 'If I want to learn the "pure" version of your language, where shall I go?' was created by a cartographer, Ulf Ernstson (from the Department of Human and Economic Geography, University of Gothenburg) by correlating GPS points collected in the field by the first author.

From the earliest times, the Greater Ruvu languages have been recognised as Bantu and various broad remarks have been made with respect to their internal subgrouping (see Polomé (1975: 23–44) for an excellent survey of the literature on the classification of Tanzanian Bantu languages prior to 1975). Detailed previous classifications involving these languages are the lexicostatistically based classifications of Nurse and Philippson (1975, 1980, 2003)[5] and of Gonzales (2002). Nurse and Philippson (1980) group languages based on rates of shared cognates (aiming to exclude loans) among 400 words (Nurse and Philippson 1980: 27–28). Once lexicostatistical percentages for each pair of languages have been calculated, the languages are broadly classified into groups within which the average percentage of similarity is higher within the group than in comparison with the most similar language outside the group ("strong groups") or almost so ("weak groups"). According to Nurse and Philippson (1980: 27–28, 31, 46–47), choices in "the proper context"[6] resolve borderline cases. The relevant part of Nurse and Philippson's (1980: 50) classification is shown in Figure 2. Gonzales (2002: 29–42) uses only 100 words and includes potential borrowings, but otherwise uses a similar methodology. Gonzales also goes on to interpret elevated similarities among languages of different groups as borrowings, and displays them in the tree, as shown in Figure 3.[7]

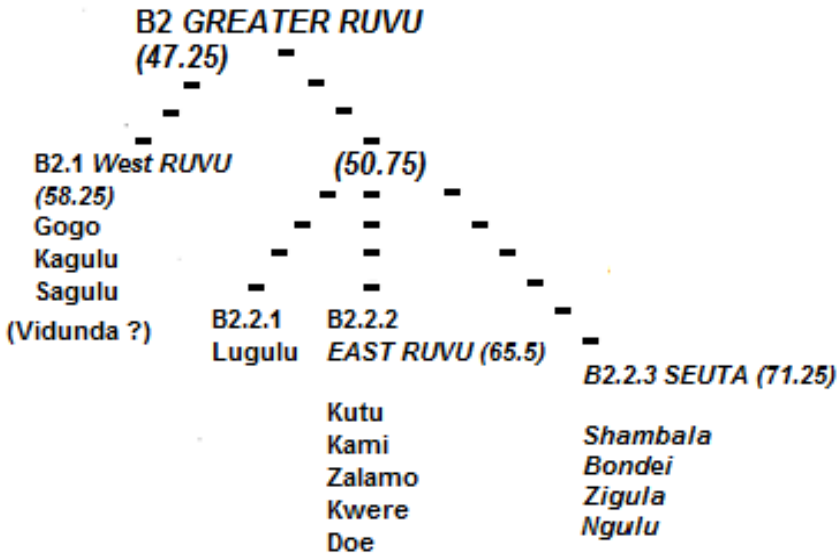

Figure 2. *Nurse and Philippson's (1980: 50) classification of Morogoro languages.*

5 The subsequent study by Hinnebusch (1981: 103–113) notes the conflicting signals between phonological isoglosses and lexicostatistical percentages and leaves it unresolved, concluding that "further study will be necessary to settle the matter" (Hinnebusch 1981: 113).

6 The phrasing is Nurse and Philippson's (1980) own and does not seem to be reducible to objective criteria. Judging from the outcome, the actual choices may involve any linguistic or non-linguistic clue.

7 Gonzales also interprets cognacy rates as reflecting time depth of separation, and links the branches in the tree to non-linguistic information. Such considerations fall outside the scope of the present paper, which is concerned only with the linguistic evidence for subgrouping.

Nurse and Philippson (1980), for reasons unclear, do not fully resolve the internal classification of the smallest groups, and the Seuta[8] languages are not included in Gonzales' (2002) study. On the issues which are included in both studies, the different authors agree,[9] except for the position of Kagulu viz-a-viz non-Greater-Ruvu languages (which fall outside the scope of the present study).

In addition, Nurse and Philippson (2003) propose a new classification of Bantu languages (80 languages) based on lexical evidence, but also on phonological and morphological features. The results of this classification have divided the Ruvu languages into three groups: (i) the G10 languages, (ii) a group composed of G23-4 (Shambala-Bondei), G31 (Zigula), G34 (Ngulu), (iii) a third group composed of G32-3 (Ngh'wele-Zaramo), G35-9 (Luguru, Kami, Kutu, Vidunda and Sagala). This classification is based on different features (lexicon and phonological and morphological criteria), and shows major agreement with the classification established in 1980 as well as in our current study.

We may note, however, that Hinnebusch's (1981: 103–113) subgrouping, based on shared phonological innovations, consistently keeps the Seuta group as a separate branch from the Luguru-East Ruvu group.

8 Bondei, Nguu, Shambala and Zigua.

9 Both Nurse and Philippson (1980: 26, 39) and Gonzales (2002: 206–209) have worked out sound correspondences in order to assess cognacy. Since Nurse and Philippson did not publish their actual correspondences, we cannot compare them with those of Gonzales (2002).

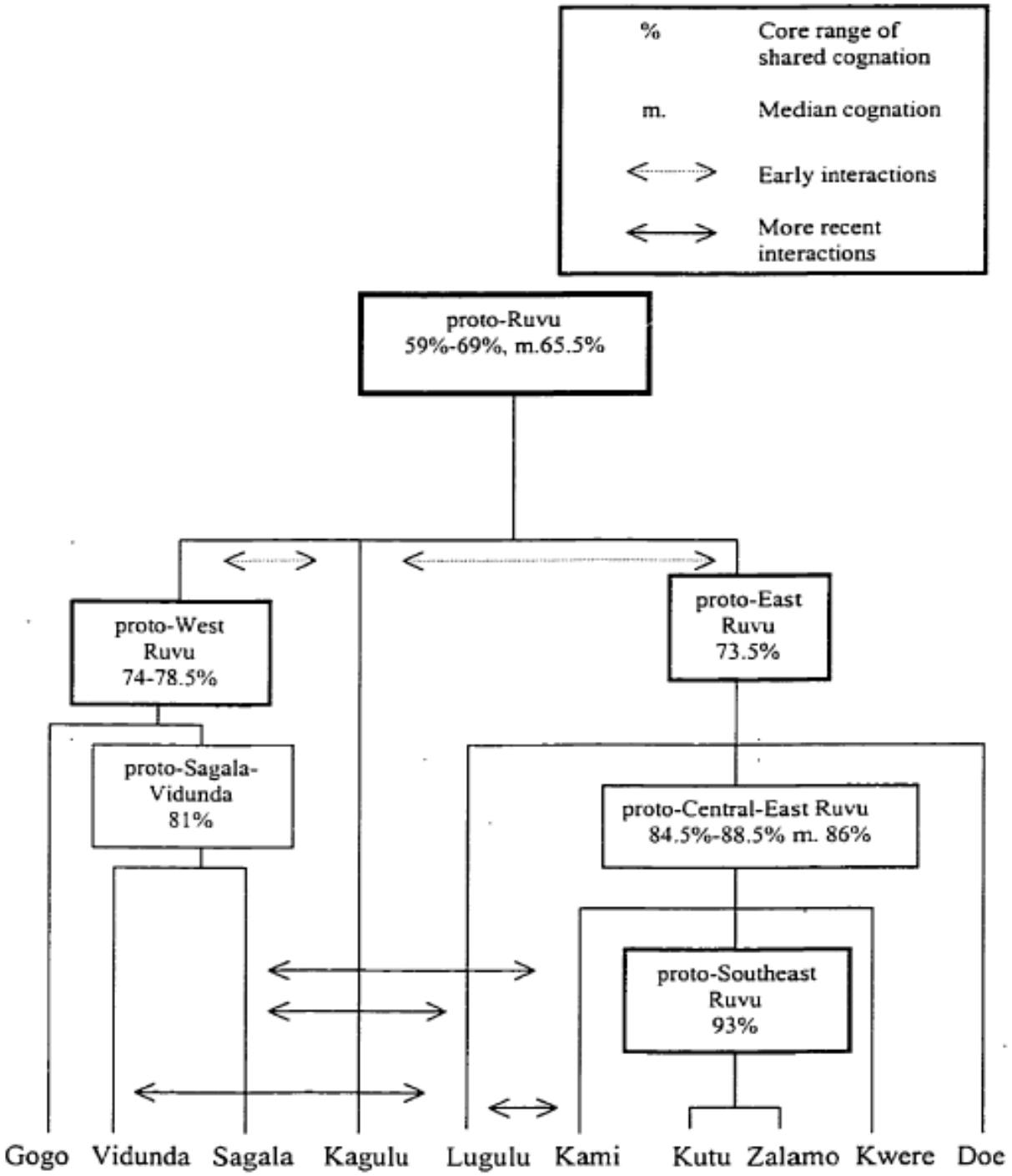

Figure 3. *Gonzales' (2002: 34) classification of the Ruvu languages.*

2. INTRODUCTION

2.1 PURPOSE OF STUDY

This article sets out to compare a number of closely related Bantu languages. The comparison is based on the study of lexical as well as structural (grammatical) parameters. We have also added the speakers' self-assessment, which is naturally subjective. Of the 27 structural parameters used in this study,

7 have been chosen as examples, and these will be discussed in detail. The first comprehensive, systematic study of Bantu morphosyntactic micro-variation is Marten et al. (2007), and we will be using a similar parametric approach in our study. As Marten et al. (2007: 253) observe, while Bantu languages are quite similar with regard to general typological parameters, there is a high degree variation in detail. The term 'parameter' should be taken at face value here, i.e. a variable or any factor that defines a system, and does not represent a small finite set of universal parameters as used in e.g. generative approaches to syntactic variation.

While we follow the parametric approach of Marten et al. (2007), only some parameters proposed in that study proved to be useful for our data. The majority of the parameters used had to be developed for this particular study. The reason for this is that the languages are so closely related that the parameters used for languages in general, or even for Bantu languages in particular, are not fine-grained enough. For instance, the parameter from Marten et al.'s study comparing single or multiple pre-verbal object markers is not suitable, since none of the eight languages in this survey allow multiple markers. The same is true for other parameters. All language varieties allow lexical object NPs to co-occur with the OM in the verb, they all have three distinct locative subject agreement markers (cl. 16, 17, 18) and locative object agreement, and they all lack a systematic distinction between conjoint and disjoint verb forms.

Closely situated language varieties may differ both in phonology and morphology, and the findings of these two areas can be contradictory, due to language contact. It is difficult to say what is in fact resemblance due to a common history, and what is influence due to language contact. There is a high resemblance between all of the Eastern Bantu languages, which can be explained by "a common ancestor and relatively recent splitting up, or to a period of physical contact, or, presumably to both" (Nurse and Philippson 1975: 3). Similarly, Holden and Gray (2006: 23, 28), in their study of tree-model-compliant classification of 85 Bantu languages find that "Some of the most complex relationships in East Bantu appear among the East African languages of zones E (excluding E5 and E6), F and G."

This is addressed in this study by systematically comparing the language varieties involved, both structurally and also lexically. "Syntactic change is different from lexical change, and hence results combining both lexical and morpho-syntactic data can lead to a more complex picture of language relationship" (Marten et al. 2007: 28). Moreover, speakers are generally more aware of the lexicon and of attitudes, while they are less conscious about language structure and thus less likely to be biased by their own preconceptions.

In addition, we wanted to compare the relationship between the languages as illustrated by grammatical parameters and by their lexical proximity. Is the subgrouping induced by grammatical similarities the same as that produced by the lexicon? For example, since speakers are conscious of the lexicon, one may hypothesize that the lexicon is more easily borrowed than grammar

(Hinnebusch 1999). This approach is supported by Nurse and Philippson, who say that "while there is today widespread agreement that almost any linguistic feature or system can be transferred, vocabulary is the component of language that is most readily and quickly transferred" (2003: 166). If this hypothesis is correct, similarity based on grammatical features should reflect fewer relationships induced by borrowing. On the other hand, structural borrowing is becoming better understood (Matras & Sakel 2007) and its efficacy may be conditioned more by the type of contact situation (Muysken 2010) than by some universal (dis)preference. In this paper, we therefore opt to explore the difference between grammatical and lexical borrowing, rather than to assume its existence.

Given that there have already been lexicostatistic studies carried out on these language varieties (cf. Nurse and Philippson 1975, 1980), we decided to make use of a new method that allowed us to produce automated cognacy judgments, namely the Levenshtein distance. The results obtained raise an epistemological question: is human cognacy judgment more reliable that automated cognacy judgment?

Lastly, apart from the purely grammatical and computationally generated lexical comparison, we wanted to compare speakers' opinions on how close the neighbouring languages are. Self-assessment is a fairly reliable means of measuring bilingualism (Skutnabb-Kangas 1981: 198), and given our results here, a fairly reliable way of measuring lexical linguistic resemblance.

2.2 METHOD, SPEAKERS AND DATA COLLECTION

The study is based on empirical fieldwork, and all language data were collected by the first author in the Morogoro region during 2008–2009. The main method was elicitation through questionnaires and interviews, as well as recordings of words, sentences and stories. The speakers are all mother tongue speakers of these languages and were born in the area where the language in question is spoken. Elicitation sessions and recordings were conducted with at least two speakers from each language. When possible, more speakers were consulted, but due to time restrictions, two speakers were set as the minimum requirement. Given that the study required us to use informants from an area where it is impracticable to carry out a random sample, snowball sampling[10], which automatically entails judgment sampling as well, proved to be the most appropriate and the most feasible method.

Two sentence questionnaires were used in the elicitation, plus a word list consisting of 500 lexical items, and several additional interview questions. The first questionnaire is a modification of the one used in the Languages of Tanzania (LoT) project, run by the Department of Foreign Languages and

10 Also called *chain sampling* or *referral sampling*.

Linguistics at the University of Dar es Salaam together with the Department of Languages and Literatures at the University of Gothenburg. The questionnaire was constructed specifically to compare the Tanzanian Bantu language varieties and constitutes a significant tool in mapping these closely related language varieties.

The second and most significant questionnaire was especially created for this project. It is loosely based on the comparative study of morphosyntactic variation in Bantu languages (Marten et al. 2007), which focuses on Bantu-specific morphological structures. This includes parameters such as symmetric/asymmetric object marking and locative inversion, which is typical for a number of Bantu languages. Nonetheless, instead of including, for instance, Marten et al.'s parameter of anaphoric relative marking, we added parameters relating to the devoicing of nasals and the internal ordering of verbal extensions. All parameters from Marten et al.'s study can be found in Appendix 1. Out of the 40 parameters that we set out with, 27 turned out to be viable for comparing this particular group of languages. In other words, only parameters that differentiated between these particular languages were included. Nevertheless, a few parameters had to be taken out due to the conflicting data they produced. 'Is there relative marking in copulas?' is one example of a too complex parameter where the answers were inconclusive, and thus, the parameter was removed.

In addition, toward the end of the second questionnaire, we created a story to translate, the ending of which has been taken out. This was done in order to get spontaneous speech/writing and to let the speakers use their own words as well as imagination. The first part of the story that the speakers were asked to translate focuses on the use of the pre-prefix. This morpheme is highly contextual and, unlike other nominal prefixes (such as the noun class prefixes), it is dependent on discourse. That is why, for instance, sentences in isolation are not sufficient to capture this phenomenon.

The word list, containing 500 semantically categorized words (which constitute the basis for our automated cognate judgments), stems from Aunio's (Institute for Asian and African Studies 1992) list from the University of Helsinki, which in turn is adapted from Heine-Möhlig's wordlist from the University of Nairobi, and was translated into Swahili by the Institute of Kiswahili Research at the University of Dar es Salaam.

Parallel to this comparative study, the speakers were interviewed with open ended questions. One of the questions was 'What language/s is/are the most similar to yours?' This was done in order to map the speakers' perceptions (i.e. self-assessment) of how similar the languages are and how they group. The results can be seen in Appendix 2. The thicker arrows indicate that more than two informants stated the resemblance, while the dotted lines indicate that only one or two informants mentioned that resemblance. For instance, Kwere and Kutu speakers mutually agree that their languages resemble each other, while Kami speakers consider Kwere to be closer to Kami than the Kwere speakers do.

Zigua speakers consider Nguu to be the most closely related language, while the Nguu speakers state that their language is most similar to Zigua, but also resembles Kagulu.

2.3 COMPUTATIONAL TOOLS

Cognacy judgments require human effort and are known to be somewhat subjective (Blench 2006: 4). A simple, even simplistic, automated procedure to gauge cognacy between two words of the same meaning is to calculate the Levenshtein distance between the two surface strings, i.e., to count the number of deletions/substitutions/insertions required in order to transform one of the strings into the other (see, for instance, Kondrak (2002) for details). Dividing the number of deletions/substitutions/insertions by the length of the longer string gives a score between 0.0 (complete identity) and 1.0 (complete difference). For example, Table 1 shows the surface strings for the meaning 'head'. The forms *litwi* and *ditwi* differ only in one character substitution (*l* to *d*), that is to say one out of five characters, and thus have a distance of 1/5 = 0.2. At the other end of the scale, *pala* and *ditwi* require four substitutions and one insertion/deletion to match, so they have a distance of 5/5=1.0. The prefixes are included with the stems since we want to track any changes in the prefixes. In any case, the test calculations when the prefixes were not included generated very similar results.

Table 1. *Cognates of the word 'head'.*

Laguages and speakers	**'head' (Proto-Bantu** *-túe*)
Kami 1	*di-twi*
Kami 2	*di-twi*
Kutu 1	*pala*
Kutu 2	*pala*
Kwere 1	*di-twi*
Kwere 2	*di-twi*
Luguru 1	*di-twi*
Luguru 2	*li-twi*
Nguu 1	*m-twi*
Nguu 2	*m-twi*
Zalamo 1	*pala*
Zalamo 2	*di-twi*
Zalamo 3	*di-twi*
Zigua 1	*m-twi*

In theory, cognacy is a strict yes/no distinction, but even in human assessments one often needs to relax this requirement somewhat – as evidenced by the halves and quarters in Gonzales' and Nurse and Philippson's tables for the same

languages (see Tables 6–7 below). As we shall see, a score between 0 and 1, as opposed to a strict yes/no decision, has some advantages and can, if necessary, be turned into a yes/no decision by imposing a threshold. To calculate the similarity between two languages, we take the average Levenshtein distance for all pairs of words with the same meaning, and then take 1 minus this score to transform distance into similarity. If one language has two (or more) words for one meaning, the average of these is used. If the data for one of the languages lacks a word for a meaning, that meaning is skipped in the calculation.

As with the lexicon, a list of languages and their grammatical parameter values can be turned into a similarity matrix that shows how similar the languages are on basis of their parameter values. The similarity of two languages is obtained simply by adding up the amount of agreement for each parameter, then dividing by the total number of parameters (for which both languages have a defined value). A comparison of two parameter values yields a full point if the values are the same, half a point if the values are 1 vs. 0 or 0 vs. -1, and zero if the values are 1 vs. -1.

The parameters in this study are logically independent. There are potential functional dependencies between parameters, but they are highly unlikely to have a significant influence on the resulting similarity matrix (Hammarström and O'Connor 2013).

3. RESULTS

3.1 THE PARAMETERS

The following paragraphs exemplify and discuss some of the parameters used in the study. This section is included for two reasons: first, we wanted to provide an example of what our parameters look like and the different areas they span; and second: we have so much qualitative data in addition to the binary parameters that is too interesting not to be put on display. First is a list showing all the parameters used in this study (Figure 4). As mentioned earlier, the parameters that did not differentiate between these 8 language varieties are not included here.

Are the tenses marked in more than one slot?
Can either object become the subject under passivisation?
Can the infinitive take the pre-prefix?
Do the languages display reflexes of Meeussen's *-*nóo*, *-*día* and *-*o* demonstratives?
Does class 5 commonly pair with class 4?
Do indefinite pronouns take the ACP?

Does noun class 1a take the inherent agreement?
Does the language have an intensive extension?
Does the language use a falsetto voice to portray distance in demonstratives?
Does the language use an object marker for the reciprocal?
Does the locative possessive prefer the inherent ACP?
Does the verb take the locative subject marker in locative inversion?
Is an object marker optional in object relatives?
Is partial agreement with conjoined NPs possible?
Is the general negative marker the regional 'default' *ha*-?
Is the negative imperative marked with an auxiliary?
Does the negative subjunctive use an auxiliary (or is it marked morphologically)?
Is the object marker placed before the subject marker?
Is the pre-prefix used in everyday speech?
Does the reciprocal marker have a CV structure?
Is the subjunctive used for giving negative commands?
Is there a designated relative marker in copular phrases?
Is there a diminutive class 12?
Is there a morphological past marker distinct from the perfective?
Is there a 'non-past' tense?
Is there a relative marker in the subject relative?
Is there general animacy concord (GAC) on the verb?

Figure 4. *The parameters*.

The first two parameters that will be discussed in more detail have to do with the noun class prefixes. One of the most prominent features in Bantu languages is the noun class system. The noun classes go back to an original Proto-Bantu system. It is a canonical system – meaning that these languages have "about six classes paired for singular and plural, plus about the same number of classes that are not paired (e.g. infinitive and locative classes)" (Katamba 2003: 108). The nouns comprise a stem and one or two prefixes. A formula for the morphological structure of the noun is given below:

(pre-prefix) + nominal class prefix + noun stem

The pre-prefix, which may also be called *augment* or *initial vowel*, is optional, while the other two components (the noun class prefix and the noun stem) are integral constituents of any noun. There is a third set of morphemes, namely the agreement class prefixes, which show agreement with other constituents in the

clause. The agreement class prefix is predominantly used on determiners and possessives, but is also used in the verb phrase both as a subject marker and an object marker (except in class 1).

3.1.1 Does noun class 1a take the inherent agreement?

In some Bantu languages, the animate class 1 has a subclass (referred to as 1a), in which the nouns do not display any noun class prefix, but take the agreement class prefix of class 1 rather than their inherent class agreement. This is sometimes referred to as *general animacy concord*. Class 1a is a minor class among the languages in this region. In other Bantu languages, this class usually contains animals, while the animals in the languages under study usually take the agreement of their inherent class. Kagulu and Nguu and Zigua always follow the inherent noun class for all animals, while Kutu and Kwere sometimes use general animacy concord. Compare the following data from Kutu and Kwere (the forms are in this case identical): *dibwa dinogile* 'good dog' *mbagile dihile* 'the bad hyena'. Unlike the other languages, Luguru seems very liberal: animals can take either their inherent classes OR the animate classes (therefore they were given the value half a point). Compare Luguru *dibwa diha* 'bad dog' (agreement from class 5) or *yumbwa keha* 'bad dog' (agreement from class 1), where both forms of agreement are accepted.

3.1.2 Does class 5 commonly pair with class 4?

Commonly, class 3 pairs with class 4 and class 5 with class 6, but in some of these languages, the pairing 5/4 occurs as well. This pairing is common in Kami, Kutu and Zalamo, as seen in the Kami *tsoka* (cl. 5) *mitsoka* (cl. 4) 'snake, snakes'. Naturally, class 6 can also be the plural of class 5, as seen in Zalamo: *nanasi dino* (cl. 5), *mananasi gano* (cl. 6) 'this pineapple, these pineapples'. What the semantic difference between the two classes is, is unclear. The pairing may occur in the other languages as well, but in that case, it usually carries an augmentative meaning.[11] This augmentative derivation expresses not only the size of the noun, but also the speaker's attitude to the noun. Augmented nouns may be used derogatorily. In Kagulu, *matamu* 'diseases' (cl. 6) are considered more dangerous than *nhamu* 'diseases' (cl. 10). In Nguu, we find *(d)ikuli* (5), *mikuli* (4) 'bad dog, bad dogs' in classes 5/4 which is, as mentioned, a not altogether uncommon pairing in these language varieties. What is interesting is that when one wants to say simply 'big dog, big dogs' with no derogatory meaning, classes 5/6 are used instead in Nguu, as in *(d)ikuli* (5), *makuli* (6) 'big dogs'. Hence, classes 5/6 are used to create an augmented noun and classes 5/4 are used derogatorily. This is in contrast to Kutu and Kwere, where the

[11] Classes 5 and 6 are the common augmentative classes in these languages.

augmentative, and not necessarily the derogatory, is displayed by using classes 5/4, as in *dibwa, mibwa* ‘big dog, big dogs’.

3.1.3 Does the locative possessive prefer the inherent agreement class prefix?

The possessive takes the agreement class prefix. For most languages, both the inherent noun class and the locative class prefixes can be used in a locative phrase. However, Kagulu prefers the locative agreement class prefix while Kutu, Kwere, Nguu and Zigua prefer the inherent agreement class prefix. Kami, Luguru and Zalamo display no apparent preference in our data.

Table 2. *Locative expressions ’in/at my house’.*

	locative noun ‘house’	possessive (locative class)	possessive (inherent class)
Kagulu	*ha/u/mu-kaya*	*ha/ukw/mwangu*[12]	*(yangu)*
Kami	*m-ng’anda*	*kw/mwangu*[13]	*yangu*
Kutu	*m-ng’anda*	*(mwangu)*[14]	*yangu*
Kwere	*u-kaye*	–	*yangu*
Luguru	*m-ng’anda*[15]	*mwangu*	*hakaye yangu*[16]
Nguu	*he/kwe/mwe-kaya*[17]	*(mwanu)*[18]	*yangu*
Zalamo	*m-ng’anda*	*mmwangu*	*yangu*[19]
Zigua	*he/kwe/mwe-nyumba*	–	*yangu*

Parentheses around a form mean it is less common. A dash means that the form does not occur in that language.

12 This is more common than the use of the inherent class. All three locatives are accepted. Note that class 17 behaves a bit differently as it makes use of the pre-prefix instead of the noun class prefix on the noun, and both the pre-prefix and the agreement class prefix occur on the possessive.

13 For class 16, another word for ‘house/home’ must be used, as in *hakae hangu* ‘at my home’.

14 This form only occurs once when the Swahili equivalent in class 18 is given. The other speaker gives *ikae yangu* ‘at my place’, also contracted to *ukayangu* ‘at my place’, and says that the locative agreement class prefixes can never be used for possessives.

15 This refers to the actual building and is used here since the locative carries the meaning of ‘inside’, while *kaye* translates as ‘homestead’.

16 This is usually contracted into *ukaiyangu* ‘by my house’.

17 *Kaya* is ‘homestead’ or ‘compound’ and therefore class 18 *mw-* is unusual since it usually means ‘inside’. If that is to be portrayed, the phrase *kundani kwenyumba yangu* ‘inside my house’ is preferred.

18 This is much less common than the usage of agreement class prefix of the inherent class.

19 This is used for classes 16 and 17 which cannot take the locative agreement class prefix on the regular possessives. If the locatives are used together with the possessives, they are used alone without the head noun.

3.1.4 Do the languages display reflexes of Meeussen's 3 **-nóo*, **-día* and **-o* demonstratives?

Bantu languages usually have three types of demonstratives. All in all, four demonstratives have been reconstructed for Proto Bantu (Meeussen 1967: 107) The languages under study show reflexes of three of Meeussen's demonstratives, namely **-nóo*, **-día* and **-o*,[20] although these follow Meeussen's matrix to varying degrees. The first demonstrative denotes proximity, the second distance and the third anaphoricity or emphasis. There is a fourth demonstrative consisting of a vowel and the agreement class prefix in the Proto-Bantu reconstructions, but this appears to have merged with the proximal demonstrative in these languages. Some languages allow the use of the initial segment *a-*, the status of which is unclear.

Table 3. *Demonstratives for noun class 1*.

	Near	Far	Referential[21]
Kagulu	*yuno*	*(a)yuya*	*yuyo[22] or ayo[23]*
Kami	*ino, (a)yuno, ayu[24]*	*(a)ija, (a)yuja, yula[25]*	*iyo, (ayo)*
Kutu	*ino*	*ija*	*iyo, (ayo[26])*
Kwere	*ino*	*ija*	*iyo, (ayo[27])*
Luguru	*(a)yuno*	*(a)yula ((a)yuwa)[28]*	*ayo[29]*
Nguu	*uyu[30]*	*yudya*	*uyo[31]*
Zalamo	*yuno[32]*	*yuja*	*ayo*
Zigua	*yuno*	*yudya*	*uyo*

Parentheses around a letter mean the segment is optional. Parentheses around a form indicate that it is less common.

---

20 The asterisk means that the word is a Proto-Bantu reconstruction that is not directly attested in any sources.

21 It marks "a referent that was previously mentioned in discourse" (Güldemann 2002: 275) or something that is of common knowledge.

22 No pre-prefix is possible on this form of the demonstrative.

23 This abbreviated form is more specific since it includes the pre-prefix.

24 This form is not attested in the written sentence, it is only elicited orally. It could be short for *ayuno*.

25 This form is not attested in the written sentence, it is only elicited orally.

26 The speaker gave this form in an interview, but it never occurs in the stories or sentences. The speaker explained that it is an older form.

27 The speaker gave this form in an interview, but it never occurs in the stories or sentences. The speaker explained that it is an older form.

28 This form is only given by one speaker and only for noun class 1.

29 The full form *imunu ayo* 'this person' is often contracted to *imunuyo* 'this person'.

30 The plural form of this is *awa*.

31 The referential demonstrative in class 1 never occurs spontaneously in the data. This form was elicited.

32 One speaker gives *ino* in the sentences.

The three-order demonstrative system is displayed in Table 3 (above). The demonstratives in the first column denote proximity, the ones in the second column distance, and those in the third column refer to something not within the range of visibility but previously mentioned. In Basaá [A43], for instance, this demonstrative is used referentially for 'the one in question' (Hyman 2003: 267), which goes for these languages as well.

The Kami demonstratives vary the most. Since the Kami area is quite small and homogenous, the more likely reason for the variations in the forms is influence from neighbouring languages (in this case Luguru).

3.1.5 Is the negative imperative marked with an auxiliary?

Kagulu, Nguu and Zigua are the only languages in the corpus that use morphological marking for giving a negative imperative. Nguu and Zigua use the morpheme *se-*, while Kagulu uses *ng'ha-*. All the other languages in the corpus use an auxiliary, which has the form of *seke* or *leka* 'leave' in all languages except Zalamo, where it is *samba*; see Table 4 below.

Table 4. *Giving a command.*

	Imperative (sg) 'Do that!'	Negative imperative (sg) 'Do not do that!'
Kagulu	*Golosa nheifo!*	*Ung'ha golose nheifo!*
Kami	*Tenda (p)fino!*	*Leka/seke kutenda/utende fino/ivo*!
Kutu	*Tenda vino!*	*Seke utende vino!*
Kwere	*Tenda vino!*	*Seke utende vino!*
Luguru	*Tenda(pfi)*!	*Uleke kutenda!*
Nguu	*Danmanya ivi!*	*Usekudamanya ivi!*
Zalamo	*Tenda vino!*	*Sambi utende vino!*
Zigua	*Tenda vino!*	*Usi/ekutenda vino!*

3.1.6 Does the language have an intensive extension?

In many Bantu languages, new verbs can be generated by adding suffixes to the existing verb root. These suffixes are often referred to as *extensions*. The verbal extensions can be valence-increasing, -decreasing or -maintaining. The valence-maintaining operations create another change in the verb, such as intensifying the action or reversing it. One such valance-maintaining extension is the *intensive*. It intensifies the action of the verb, as seen in the Kagulu *kugolos**es**a milimo* 'to work hard', from *kugolosa milimo* 'to work'; and in the Nguu *agel**es**a langi* 's/he paints a lot' from *kugela* 'to put'. There is no Proto-Bantu counterpart to this extension (Schadeberg 2003: 72). The intensive is quite common in Nguu and Kagulu, but is not found in Kami, Kutu, Kwere and Zalamo. It also exists in Zigua and Luguru, although it appears to be slightly less common in Luguru.

3.1.7 Does the verb take the locative subject marker in locative inversion?

Affirmative existential constructions are introduced by the locative prefixes. Existentials introduce participants and have a presentative function. There is an extended function of the existential that is used when there is no need to mention the actors. In Bantu it is mostly referred to as *locative inversion* (Demuth and Mmusi 1997, Marten 2006). It resembles existentials in that the locative introduces the predicate, while, in locative inversion, other verbs can be used, not only 'to be'. In these types of constructions, the locative appears to be the subject since the verb agrees with the locative, which is in the subject position, while the inverted, or logical, subject is in the object position after the verb.

For many of the languages in this survey, both the inverted subject marker (pertaining to the noun class of the subject) and the locative subject markers (from classes 16, 17 or 18) can be used without any apparent change in meaning. Compare, for instance the Kami examples *Mmibiki **ga**kala gamanyani*. 'In the tree sit baboons' (inherent noun class: 6) and *Mmibiki **mu**kala gamanyani*. 'In the tree sit baboons' (locative noun class: 18).

Table 5. *Locative inversion.*

	locative noun	locative subject markers	inherent subject markers	
Kagulu	*mu-ma-biki*	*ha/ku/mwi-kala*[33]	*ge-kala*[34]	*manyani*
Kami	*m-mi-biki*	*mu-kala*	*wo-kala*	*gamanyani*
Kutu	*m-mi-biki*	*mo-kala*	*wo-kala*	*nyabu*
Luguru	*m-chanya m-ne mi-ti*	*ha/ku/m-kala*	*wo-kala*	*wanyani*
Kwere	*mu-na i-mi-biki*	*ku/mu-kala*	*wo-kala*	*nyani*
Nguu	*mwe-ma-ziti*	*mwe-kala*	*ye-kala*	*manyani*
Zalamo	*m-mi-biki*	–	*wo-kala*	*nyabu*
Zigua	*mwe-mi-ti*	*mwe-kala*	*ye-kala*[35]	*manyani*

A dash means that the form does not occur in that language.

Nevertheless, in some languages, if the inherent subject marker is used instead of the locative, the meaning can change slightly. In Luguru, one informant claims that the meaning becomes more habitual with the inherent subject marker. The same goes for Nguu: two informants claim that if the locative subject marker is used, they sit every day, while if the inherent subject marker is used, they sit only once. Zalamo is the only language that never allows locative subject markers, at least not in this sample.

33 In some languages, the locative subject markers are more versatile and all three locative noun classes can be used regardless of the noun class of the locative NP.

34 This form is preferred over the locative subject markers.

35 This form was elicited and the locative form came more naturally.

3.2 How Closely Related are the Languages?

Tables 6 and 7 show the paired cognacy percentages calculated by Nurse and Philippson (1980: 56, without borrowings) and Gonzales (2002: 32, with borrowings) for the languages discussed in the present paper. Nurse and Philippson (1980) show only parts of the entire cognate matrix that formed the basis for their classification, so we do not know the actual cognate rates that they found for Kagulu, Luguru, Zigua and Nguu.[36] As already noted, Gonzales does not include Zigua and Nguu.

Table 6. *Nurse and Philippson 1980: 56.*

Kutu	Zalamo	Kwere	
69.25	65.25	69.25	Kami
	68.5	63.75	Kutu
		61.50	Zalamo

Table 7. *Gonzales 2002: 32.*

Kutu	Zalamo	Kwere	Kagulu	Luguru	
86	88.5	88.5	72.5	85.5	Kami
	93	84.5	70	80	Kutu
		86	74.5	84	Zalamo
			75	79.5	Kwere
				69	Kagulu

Greenhill (2011) finds that Levenshtein distances fail to capture what human cognacy judgments capture on a certain set of cognates (200 words from a well-studied subset of Austronesian languages). Not surprisingly, this effect is larger the more distantly related the languages are, where cognacy is less detectable in the surface strings. The languages used in this study, however, are so closely related that human cognacy judgment is sufficiently reproducible by Levenshtein distances. Although the numbers are on somewhat different scales, the automatically generated similarity matrix in Table 8 agrees – in the internal ranking of pairs – as much with the two human-derived matrices as those two do with each other.[37] A Neighbour-Joining tree (Felsenstein 2004: 166–170), based on a 500-word list for the Greater Ruvu languages including Swahili as a reference point (i.e. data from Table 8), is found in Figure 5. A Neighbour-Joining tree is the simplest principled way to derive a tree with branch-lengths from a distance/similarity matrix. First, note that any given tree defines a

[36] The cognate matrix attributed to Nurse, quoted in Polomé (1975: 223), spans the full set of languages relevant for this paper, but appears to be different – it possibly includes borrowings – from the one used in the Nurse and Philippson (1980) study.

[37] We omit a formal measure of this as it is complicated to derive with the missing data points, and shows the same thing as we aim to show with the table.

distance matrix as the distance along the branches between any two leaves. Neighbour-Joining is designed to find the tree with branch-lengths whose distance matrix is as close as possible to the input matrix. The sought after tree is built iteratively by "joining neighbours" (for details see Felsenstein 2004: 166–170), but it is not necessarily the closest pair of leaves (i.e. the pair with the highest cell value in the matrix) that are joined in each step.

Table 8. *Similarity matrix based on Levenshtein distances in a 500-word list for the Greater Ruvu languages including Swahili as a reference point.*

	Kagulu	Kami	Kutu	Kwere	Luguru	Nguu	Swahili	Zalamo
Kami	53							
Kutu	54	67						
Kwere	57	67	70					
Luguru	54	65	62	62				
Nguu	57	58	56	62	57			
Swahili	47	55	50	57	55	60		
Zalamo	54	67	76	69	62	54	48	
Zigua	51	54	52	58	51	68	51	53

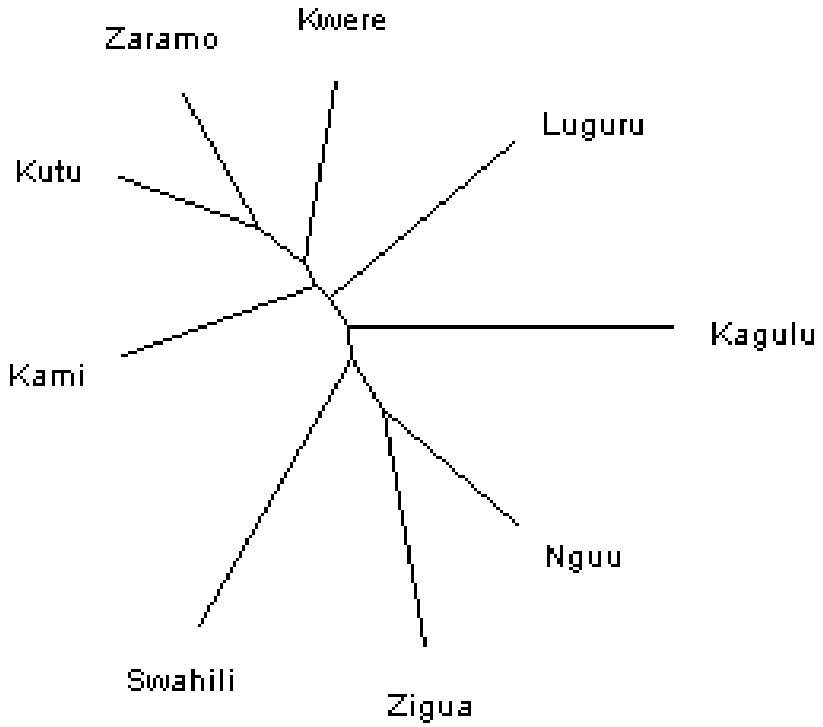

Figure 5. *A neighbour-joining tree based on the lexical distance matrix of Table 8.*

3.2.1 Subdialectal Varieties

At which point does idiolectal/elicitational variation interfere with dialectal classification? For example, two elicited lists from different speakers of one and the same dialect might well differ more than two lists containing words from different dialects. This is possibly the reason for Nurse and Philippson's (1980) reluctance to resolve their classification at the deepest level, i.e. to subclassify all the languages within a group. With an automated cognate judgment and classification method, we may quickly illustrate the boundaries. The outcome is shown as a similarity matrix in Table 9.

Table 9. *Similarity matrix based on Levenshtein distances in a 500-word list for each speaker[38] of the Greater Ruvu languages.*

	Kagulu	Kami 1	Kami 2	Kutu 1	Kutu 2	Kwere 1	Kwere 2	Luguru 1	Luguru 2	Nguu 1	Nguu 2	Zalamo 1	Zalamo 2	Zalamo 3
Kami 1	54													
Kami 2	53	64												
Kutu 1	55	68	67											
Kutu 2	52	72	61	80										
Kwere 1	56	71	60	67	71									
Kwere 2	58	71	67	74	67	74								
Luguru 1	55	68	67	65	63	63	66							
Luguru 2	52	61	65	63	57	57	63	69						
Nguu 1	57	55	57	59	51	57	65	56	55					
Nguu 2	58	59	60	62	52	60	68	58	57	86				
Zalamo 1	52	66	57	69	73	67	63	59	55	50	52			
Zalamo 2	55	72	70	78	82	70	72	67	60	53	59	82		
Zalamo 3	55	74	66	77	80	71	72	67	61	54	57	79	96	
Zigua 1	51	56	53	54	51	57	60	55	47	67	69	51	55	54

Although the similarity between different speakers of the same variety is far from 100%, all subvarieties can be subgrouped with their respective language variety partners, except for the outlying Kami 2, which looks more like a Luguru variety than Kami 1. Kami and Luguru are neighbours and have borrowed from each other, as well as being phylogenetically close (Gonzales 2002). Indeed, in

[38] Since there was no speaker of Swahili (which was the working language), it is not included in this matrix.

this case, the particular Kami speaker, represented as Kami 2, lives closer to the Luguru area than the Kami 1 speaker.

Table 10 below is a similarity matrix and Figure 6 is a neighbour-joining tree (Felsenstein 2004: 166–170) based on the 27 grammatical parameters. The matrix shows the percentage of similarity between the languages, meaning that for instance Kagulu and Kami are 46% similar in this study. The most similar languages are Kutu and Kwere, where the similarity is approximately 92%, and the most distant languages in the cluster are Nguu and Zalamo, where the similarity is only approximately 26%.

Table 10. *Similarity matrix based on 27 grammatical parameters*.

	Kagulu	Kami	Kutu	Kwere	Luguru	Nguu	Zalamo
Kami	46						
Kutu	42	81					
Kwere	42	73	92				
Luguru	70	58	66	66			
Nguu	76	37	36	44	54		
Zalamo	44	83	74	66	60	26	
Zigua	50	46	40	48	42	67	38

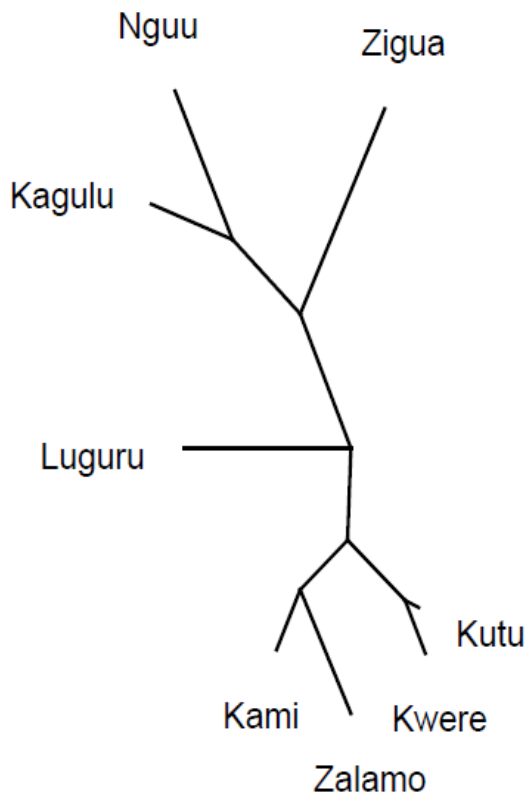

Figure 6. *A neighbour-joining tree based on the grammatical distance matrix of Table 10.*

3.2.2 Discussion

Our structural parameters, the automated cognate judgements, the previous studies by Nurse and Philippson (1980) and Gonzales (2002), and the self-assessment made by the speakers, point to the following groupings:

Nguu and Zigua form a clade the lexical tree, although grammatically, Nguu is closer to Kagulu. Nguu and Zigua show the highest figure in Nurse's investigation as well (95% lexical similarity) (Nurse 1970s: 45), compared with our figure (68% for lexicon). Assuming a tree model, and that chance and universal tendencies can be ruled out, this lexical versus grammatical mismatch could logically come about in four different ways. There could be structural diffusion between Kagulu and Nguu, or lexical diffusion between Zigua and Nguu, or there could be accelerated lexical change in Kagulu, or accelerated structural change in Zigua.

There are principled ways to gauge what the most likely reason is in such three-participant situations. Suppose A and B have a unique common ancestor, and a language C borrows from B. In such a scenario, the similarity of the pair B and C increases. The similarity between C and A, and indeed C and all other languages, does not increase as much, because some of the borrowings are likely to be items unique to B. Now suppose A and B have a unique common ancestor and their next-of-kin is C, and then, sometime after the break-up of A and B, A innovates much more than B does. This should cause the similarity between A and all other languages to drop proportionately, because there is no reason for the innovations in A to be concentrated among items retained by both A and any other language.

Let us now consider the idea that Nguu and Zigua form a subgroup (their structural similarity is 67%) and that there has been structural diffusion between Kagulu and Nguu (structural similarity 76%). Thus, the question is whether the 9+ percentage points between Kagulu and Nguu are localized to exactly that pair, or whether the Kagulu and Nguu similarities to the remaining languages are similarly higher (than between Nguu and Zigua)? The relevant figures, drawn from Table 10, are as follows:

Kagulu	46	42	42	70	44
Nguu	37	36	44	54	26
Zigua	46	40	48	42	38
	Kami	Kutu	Kwere	Luguru	Zalamo

We see that Zigua and Nguu consistently have more similar percentages to the other languages than do Kagulu and Nguu. This is indeed evidence that there has been structural diffusion between Kagulu and Nguu. At the same time, we can reject the idea that Zigua has had a period of unique accelerated aberrancy

because then it should have had lower similarity to all other languages than Nguu.

Similarly, let us now consider the idea that Nguu and Kagulu form a subgroup (their lexical similarity is 57%) and that there has been lexical diffusion between Zigua and Nguu (lexical similarity 68%). Thus the question is whether the 11+ percentage points between Zigua and Nguu are localized to exactly that pair, or whether the Zigua and Nguu similarities to the remaining languages are similarly higher (than between Nguu and Kagulu)? The relevant figures, drawn from Table 8, are as follows:

Kagulu	53	54	57	54	47	54
Nguu	58	56	62	57	60	54
Zigua	54	52	58	51	51	53
	Kami	Kutu	Kwere	Luguru	Swahili	Zalamo

Here there is about equal resemblance between the figures for Zigua and Kagulu to those for Nguu. To the extent that any small differences like these are meaningful, we do not observe that Kagulu and Nguu consistently have more similar percentages to the other languages, and thus there is no strong evidence for lexical diffusion between Zigua and Nguu. At the same time, we can reject the idea that Kagulu has had a period of unique accelerated aberrancy because then it should have had less similarity to all other languages than Zigua.

To summarize, regarding the lexical and grammatical subgrouping mismatch involving Kagulu, Nguu and Zigua, our numbers favour a scenario where Nguu and Zigua form a historical subgroup and where there has been structural diffusion between Nguu and Kagulu. Such diffusion is not necessarily the result of direct borrowing, but may be evidence of language shift/substrate effects. Should we find more cases like Nguu-Zigua-Kagulu, this would constitute evidence that lexicon is more stable than grammar.

As for the other groupings, Kami, Kutu, Kwere and Zalamo range between 62% and 76% lexically, and 58% and 92% structurally. The highest percentage parameter-wise is between Kutu and Kwere (92%) and the lowest is between Kami and Luguru (58%). Lexically, Luguru and Zalamo have the lowest similarity (62%) and the highest is between Kutu and Zalamo (76%).

Luguru is in a group of its own, although is closest to Kagulu structurally (70%) and Kami lexically (65%), and is furthest away from Zigua, both lexically (51%) and structurally (42%). This is in accordance with Gonzales' figures: 86% similarity with Kami and only 69% with Kagulu. According to the speakers' self-assessment, Luguru belongs with Kami, Kutu, Kwere and Zalamo.

Our findings are also corroborated by Gonzales' second study (2008). In this study, her shared cognate percentages are 69% for Kagulu and Luguru (our

figure is 70% for the structural parameters and 54% for lexicon), while within the group containing Kami, Kutu, Kwere, Luguru, Zalamo the numbers range between 79.5% and 93%, the average being 85.5%. Her highest percentage is between Kutu and Zalamo (where we reach the number 74% for the grammatical parameters and 76% for lexicon), while the lowest (in this group, that is) is between Kwere and Luguru (where we have 66% for the parameters and 62% for lexicon).

Our results likewise reproduce the subgroups of the trees in Nurse and Philippson (1980) and Gonzales (2002), except that ours lift up the Seuta group (Zigua, Nguu, Shambala and Bondei) to a higher node in the tree from its placement with Luguru-East Ruvu in Nurse and Philippson (1980). This may actually be a positive difference, since the subgrouping of Seuta with Luguru-East Ruvu is a "weak subgroup" according to Nurse and Philippson's (1980: 31) lexicostatistical subgrouping criteria, as well as being unsupported by sound shifts in Hinnebusch (1981: 103–113).

4. CONCLUSION

To only use lexicostatistics when mapping and comparing languages is too blunt a tool. There are other methods of differentiating languages – in this case grammatical parameters, automated cognate judgement and self-assessment. We have developed methods of differentiating language varieties that on the surface appear to be very similar. When we compare our three ways of measuring, we see some interesting results. Our automated cognate judgements and Levenshtein distance actually match the manual/human cognate assessments made by previous researchers. Even though these language varieties show quite extensive lexical variation (possibly even idiolectal) between different speakers, that variation is still smaller than the variation across language varieties. Speakers' self-assessments of similarity, in particular, correspond chiefly to lexical similarity, which is to be expected. As mentioned, speakers are more conscious of their lexicon and less aware of grammatical structures.

In our empirical study, we have shown how these languages can be grouped, based on different features and methods. We do not speculate on the wider implications of these groupings since the material is too limited. The descriptive data we have presented stand on their own. However, as in every study there are issues that deserve further analysis. It would have been interesting to investigate what a grammatical classification can tell us about the history of these languages, and what an investigation of language contact might reveal, but this is, unfortunately, not within the scope of this study.

According to our findings the four groupings are: Kami, Kutu, Kwere, and Zalamo together, and Nguu and Zigua in a second group, while both Kagulu and Luguru stand alone.

REFERENCES

Blench, R. M. 2006.
The Niger-Saharan macrophylum. Ms.

Demuth, K. A. and S. O. Mmusi. 1997.
Presentational focus and thematic structure in comparative Bantu. **Journal of African languages and linguistics** 18: 1–19.

Felsenstein, J. 2004.
Inferring phylogenies. Sunderland, Massachusetts: Sinauer.

Gonzales, R. 2002.
Continuity and change: thought, belief, and practice in the history of the Ruvu peoples of Central East Tanzania, C. 200 B.C. To A.D. 1800. University of California at Los Angeles (UCLA).

2008 *Societies, religion, and history: Central East Tanzanians and the world they created, c. 200 BCE to 1800 CE*. New York, Columbia University Press.

Guthrie, M. 1971.
Comparative Bantu: an introduction to the comparative linguistics and prehistory of the Bantu languages. Vol. 2: Bantu prehistory, inventory and indexes. London: Gregg International.

Güldemann, T. 2002.
When "say" is not say: the functional versatility of the Bantu quotative marker *ti* with special reference to Shona. In: T. Güldemann and M. v. Roncador (eds.), *Reported discourse: a meeting ground for different linguistic domains*, pp. 253–288. Amsterdam: John Benjamins.

Hammarström, H. and L. O'Connor. 2013.
Dependency sensitive typological distance. In: L. Borin and A. Saxena (eds.), *Linguistic distances*. Berlin: Mouton.

Hinnebusch, T. J. 1981.
Northeast coastal Bantu. In: T. J. Hinnebusch, D. Nurse and M. J. Mould (eds.), *Studies in the classification of Eastern Bantu languages*, pp. 21–125. Hamburg: Helmut Buske Verlag.

1999 Contact and lexicostatistics in comparative Bantu studies. In: J.-M. Hombert and L. M. Hyman (eds.), *Bantu historical linguistics: theoretical and empirical perspectives (Lecture notes 99)*, pp. 173–205. Stanford: CSLI (Center for the Study of Language and Information) Publishers. (Lecture notes 99).

Holden, C. J. and R. D. Gray. 2006.
Rapid radiation, borrowing and dialect continua in the Bantu languages. In: P. Forster and C. Renfrew (eds.), *Phylogenetic methods and the prehistory of languages*, pp. 19–31. Cambridge, UK: McDonald Institute for Archaeological Research.

Hyman, L. M. 2003.
Basaa (A43). In: D. Nurse and G. Philippson (eds.), *The Bantu languages*, pp. 257–282. London & New York: Routledge.
Institute for Asian and African Studies. 1992.
500 word list. Ms. University of Helsinki.
Katamba, F. 2003.
Bantu nominal morphology. In: D. Nurse and G. Philippson (eds.), *The Bantu languages*, pp. 103–120. London & New York: Routledge.
Maho, J. 2009.
NUGL Online. The online version of the New Updated Guthrie List, a referential classification of the Bantu languages. Retrieved 18 September, 2013, from http://goto.glocalnet.net/mahopapers/nuglonline.pdf.
Maho, J. and B. Sands. 2003.
The languages of Tanzania: a bibliography. Göteborg: Acta Universitatis Gothoburgensis.
Marten, L. 2006.
Locative inversion in Otjiherero: more on morphosyntactic variation in Bantu. *ZAS papers in linguistics: Papers in Bantu Theory and Description* volume 43. Retrieved 2006-10-03, from http://www.zas.gwz-berlin.de/index.html?publications_zaspil.
Marten, L., N. C. Kula and N. Thwala. 2007.
Parameters of morphosyntactic variation in Bantu. **Transactions of the Philological Society** 105(3): 253–338.
Matras, Yaron & Jeanette Sakel. 2007.
Introduction. In: Yaron Matras & Jeanette Sakel (eds.), *Grammatical Borrowing in Cross-Linguistic Perspective (Empirical approaches to language typology 38)*, pp 1–14. Berlin: Mouton de Gruyter.
Meeussen, A. E. 1967.
Bantu grammatical reconstructions. **Africana linguistica III**. Tervuren. 79–121.
Muysken, Pieter. 2010.
Scenarios for Language Contact. In: Raymond Hickey (ed.), *The Handbook of Language Contact*, pp. 265–281. Oxford: Wiley-Blackwell.
Nurse, D. (ed.). 1970s.
A phonological and morphological sketch of 15 of the principal languages of Tanzania. Dar es Salaam, Institute of Kiswahili Research (IKR), University of Dar es Salaam.
Nurse, D. and G. Philippson. 1975.
The north-eastern Bantu languages of Tanzania and Kenya: a tentative classification. **Kiswahili** 45(2): 1–28.
1980 The Bantu languages of East Africa: a lexicostatistical survey. In: E. C. Polomé and C. P. Hill (eds.), *Language in Tanzania*, pp. 26–67.

London: Oxford University Press for the International African Institute (IAI).

2003 Towards a historical classification of the Bantu languages. In: D. Nurse and G. Philippson (eds.), *The Bantu languages*, pp. 164–181. Richmond: Curzon.

Polomé, A. R. 1975.
The classification of the Bantu languages of Tanzania. Austin (Texas), University of Texas. Pp. 279.

Schadeberg, T. C. 2003.
Derivation. In: D. Nurse and G. Philippson (eds.), *The Bantu languages*, pp. 71–89. London & New York: Routledge.

Skutnabb-Kangas, T. 1981.
Bilingualism or not: the education of minorities. Clevedon: Multilingual Matters.

About the authors: *Malin Petzell*, PhD, currently holds a position at the Department of Languages and Literatures, University of Gothenburg. Her research interests include Bantu languages, language description (documentation and analysis), nominal and verbal morphosyntax, language endangerment, and field methods. *Harald Hammarström* currently holds a position at Max Planck Institute for Psycholinguistics, Nijmegen. His research interests include language typology, language description, Papuan languages, and quantitative methods in historical linguistics.

APPENDIX 1
THE PARAMETERS FROM MARTEN ET AL. 2007

Object markers	
1 OM – obj NP	Can the object marker and the lexical object NP co-occur?
2 OM obligatory	Is co-occurrence required in some contexts?
3 OM loc	Are there locative objects markers?
4a One OM	Is object marking restricted to one object marker per verb?
4b Restr 2 OM	Are two object markers possible in restricted contexts?
4c Mult OM	Are two or more object markers freely available?
4d Free order	Is the order of multiple object markers structurally free?
Double objects	
5 Sym word-order	Can either object be adjacent to the verb?
6 Sym passive	Can either object become subject under passivisation?
7 Sym OM	Can either object be expressed by an object marker?
Relatives	
8 Agr Rel mark	Does the relative marker agree with the head noun?
9a Res OM obl	Is an object marker required in object relatives?
9b Res OM barred	Is an object marker disallowed in object relatives?
9c Res OM optional	Is an object marker optional in object relatives?
Locative inversion	
10 LI restr	Is locative inversion thematically restricted to intransitives?
11 Full loc SM	Are there three different locative subject markers?
Conjunct agreement	
12 Partial Agr	Is partial agreement with conjoined NPs possible?
Conjoint/disjoint	
13 Conj/disj	Is there a (tonal) distinction between conjoint/disjoint forms?
14 Tone case	Is there a (tonal) distinction of nominal 'cases'?

APPENDIX 2
LINGUISTIC SIMILARITY, SELF-ASSESSED