

The Future of Prehistoric Comparative Linguistics

A Review of Guthrie's *Comparative Bantu* Methodology with
Insights from Recent Relevant Literature on Comparative Linguistics,
Language Change and Reconstructing Genetic Links and
Suggestions Toward Contemporary Applications

David Rowbory

TS 720 Special Topics in Translation Studies • 1 July 2008

— Contents: —

0. Introduction	1
1. Comparative Linguistics: Guthrie's Comparative Bantu Methodology	2
The process in brief	2
Comparative Bantu: An Overview	4
2. An Overview of Other Literature	5
Language Change (McMahon 1994)	5
African Languages: An Introduction (Heine & Nurse 2000)	6
Greenberg, Mathematical Models and Lexicostatistics (Fodor 1982)	8
3. Data Preparation	10
4. Simple Comparisons	12
Meaning as preferred connector	13
Underlying Assumptions	14
5. Systematising: Associated Comparisons and Comparative Series	15
Simple Computer-assistance	16
Assessment of the Computer-Assisted Comparison System	20
6. Inferring Relationships	22
Guthrie, Greenberg and Differing Interpretations	22
7. Possible Application Within and Beyond Bantu	23
The Point of Genetic and Other Classifications	24
Nigerian Languages	25
8. Conclusions and Further Study	26
9. Bibliography	28

0. Introduction

Twenty years after his monumental *Classification of the Bantu Languages* (Guthrie 1948) Malcolm Guthrie published in *Comparative Bantu* (Guthrie 1967) a substantial exposition of the comparative linguistic approach that lay behind his classification. In this time, the enduringly controversial Joseph Greenberg published probably his least controversial work *The Languages of Africa* (Greenberg 1963) using a methodology considered

dramatically different. Where Guthrie tentatively suggested possible links and worked towards a reconstruction of Common Bantu, Greenberg made bold, large-scale claims about the prehistory of African languages and their genetic relationships.

By the 21st Century the study, comparison and classification of African languages is still in its infancy compared to Indo-European study. One challenge is that without any significant historical records of Sub-Saharan African languages, any comparative linguistics is by nature pre-historical investigation. Are there insights we can draw from past work, more recent literature and contemporary techniques to advance work in this field? In this paper we examine and summarise Guthrie's methodology and the contributions of recent scholarship, then reflect on each stage of his methodology in the light of such contributions. Finally we attempt to apply this to African linguistic study beyond the Bantu languages.

1. Comparative Linguistics: Guthrie's Comparative Bantu Methodology

The process in brief

In the following section we give an overview of *Comparative Bantu*, but an introduction to the process and terminology is helpful first. Guthrie describes four stages of work which could be described as preparation, comparison, synthesis and inference. Discussions of the first three of these stages form the bulk of this paper. Guthrie's paragraphs are all numbered for reference in a four-digit format like 21.32 where the initial digit (2) signifies the chapter, the next a major division within the chapter (1), then after a period a minor section (3) and finally an incrementing paragraph number within a minor section (2). For the sake of convenient cross-reference we cite Guthrie's *Comparative Bantu* as 1967:<page>/<paragraph>.

First the data from a variety of languages must be prepared and organised to facilitate comparison. Today that would likely mean sourcing, collating, verifying and tidying word lists into a well-structured lexical database. Three elements are required for each word: a

unique language identifier, the form being compared (likely a lexeme or morpheme) and a gloss. Additional information may help as discussed below.

Then 'simple comparisons' (Guthrie 1967:15/21.23) are made, as a form from one language is compared with one from another language. Given two languages L and M, we link words in these languages which have an identical (or very similar) form or meaning:

$$\text{word}_L \leftarrow \textit{linked to} \rightarrow \text{word}_M$$

We must note that we are not claiming synonymy or equivalence for these words—we are merely linking them as being worth comparing. Both form and meaning must be similar to some extent, but one of these variables is chosen to be the major connector between the words.

Where several languages are being compared, we choose a 'connecting feature' and list the words in these languages which share this connecting feature. If we gather all words which look like **-pet-** then the 'connecting feature' of the simple comparison is the *form*. If we gather words with the English gloss 'buy', then *meaning* is the connecting feature.

The third stage takes these lists of simple comparisons between languages and seeks to systematise or synthesise them. The aim is to identify patterns of regular differences between languages by selecting and linking lists of simple comparisons into 'associated lists of comparisons' (Guthrie 1967:17/21.62). If enough evidence is found to demonstrate a pattern, Guthrie calls such an association of lists a 'Comparative Series'.

Finally, each Comparative Series contributes evidence about regular sound correspondences (similarities or shifts) between the languages in the series. From the evidence of many Comparative Series the comparative linguist then draws inferences about groups of sound changes and begins to reconstruct the likely pre-historic development of the languages.

Comparative Bantu: An Overview

In Part I (volumes 1 & 2) of *Comparative Bantu*, Guthrie explains his aims and methodology in some detail. Chapter 1 introduces the linguistic background to the study of over 300 agglutinative Bantu languages. Chapter 2 concentrates on the methodology itself: preparing data, comparing words, systematising comparisons, and beginning to make inferences. Chapter 3 takes this further to examine how the results of the comparative method can be used to reconstruct 'Common Bantu', the prehistoric common ancestor of all the Bantu languages. Further chapters focus on the processes of inferring and deducing the nature of the prehistoric changes from Common Bantu up to the various individual languages and clusters observed today. Several topograms show the geographical spread of relationships detected in the data: regular sound shifts and relationships between lists of related words (Comparative Series).

This paper focuses on the methodology and assumptions evidenced in the first two chapters of Part I. From chapter 3 Guthrie focuses on issues within Bantu languages, and analysis of the comparative series established, but any the outcome of any subsequent work depends crucially on the preparation, comparison and systematisation of the data. These aspects are also the most applicable to non-Bantu situations, such as studying the varied languages of Nigeria or the Nilo-Saharan languages.

Volume 2 begins with a synthesis of the tentative conclusions reached about the nature of Proto-Bantu (the common ancestor of all the Bantu languages), then a series of excurses, and various indices to the lists of Comparative Series which comprise part II (volumes 3 & 4). These indices allow one to investigate the data analysis beginning either with a language or group (according to Guthrie's 'areal' classification codes), or with a reconstructed morpheme, or with a common meaning (in English) lying behind a group of similar morphemes. The morphemes include stems, radicals and affixes. The abundance of indices vividly reveals the dependence of this study on good, patient analysis of very large sets of data. Nowadays the ability to present the data in a hyperlinked electronic form would probably dramatically increase its utility and verifiability.

Such is the vast quantity of analysis and cross-referencing between the evidence and conclusions that it is unsurprising this work took such a long time. Unfortunately, checking and improving the work, say with more data, would take equally long, following the same approach. As part of this assessment of Guthrie's method, we will therefore consider the extent to which data processing might be handled using computers to speed up analysis, allow more possible patterns to be detected and explored, verify and allow revisions to be made more easily, so that human intelligence can be spared the most mundane aspects of data analysis. Unfortunately the source data—the original word lists—are not published as part of *Comparative Bantu*. This makes it hard to compare Guthrie's original results with the results of someone repeating Guthrie's methodology with the help of electronic data processing.

2. An Overview of Other Literature

Having gained an overview of Guthrie's methodology we examine a selection of contributions from other scholars on related subjects. Principles of language change and other research on Bantu and comparative linguistics may complement Guthrie's method as well as mathematical and lexicostatistical suggestions and discussion of Greenberg's methods.

Language Change (McMahon 1994)

Giving an overview of the various ways in which language change has been studied, McMahon introduces Neogrammarian, Structuralist and Generative theories before examining some more recent thinking about lexical phonology and lexical diffusion. After the study of grammar being concerned with classics, Neogrammarians began to study mainly European languages, noting synchronic and diachronic links and differences. They saw languages changing as a result of largely logical phonetic simplifications; the chief motivation for change was economy and making words easier to say. Structuralists (following Saussure) held that phonemes change. Generativists focussed more on morphological elements of a language and saw language change in terms of morphological and lexical rules changing.

Neogrammarians (and to some extent Structuralists and Generativists) expected that sound changes would be phonetically gradual but affect the whole lexicon simultaneously. However some noticed that this is not always the case. Sound changes often seem to spread from a small group of words to others. Over a variable stretch of time most of the lexicon may be affected, but the impetus for the sound change may be removed before affecting the whole lexicon. Some words may resist change not for phonological reasons, but because they are either more commonly or more rarely used (depending on socio-cultural factors). Here a Neogrammarian approach would struggle since every exception must be explained with a rule.

It was also observed that not every sound change can be phonetically gradual. The constriction of a plosive can gradually become relaxed so that it becomes a fricative, and the openness and point of articulation of a vowel may move imperceptibly over time, but several sound changes must be abrupt. At some place in a sound changing from a bilabial to a dental point of articulation the lips would cease to become involved. A change of air mechanism cannot be a gradual process.

McMahon suggests that rather than having to choose between a Neogrammarian model of lexically-universal phonetically-gradual change and the lexical diffusion model of lexically-gradual, phonetically-abrupt change, both may be observed functioning at different times. This has implications for studying the prehistory of languages, since we cannot necessarily expect that all sound changes will have affected the entire lexicon of a language. Some sound changes might have affected only a certain set of words. Also since sound changes (as seen in Indo-European studies) seem to have been limited to particular geographical areas and times, different dialects of one language may experience some sound changes, but not others.

African Languages: An Introduction (Heine & Nurse 2000)

African Languages is probably the most recent comprehensive survey of African languages by topic and phylum. Descriptions of the four major language phyla present the fruit of comparative linguistic study and classification such as the work of Kay

Williamson and Roger Blench on Niger-Congo, but without much discussion of methodology. Only Afro-Asiatic languages (such as the Semitic languages) have any substantial historical records and it is the least controversial major language grouping ('phylum'). All other phyla and sub-groupings require much more work to refine current analyses before we can confidently chart the prehistory of these languages.

Seemingly then, much more investigation is still required to improve our understanding of relationships between African languages. Elsewhere, Nurse (1994) has drawn attention to the vast amount of work remaining to be done in Bantu studies. In the case of Bantu languages, Williamson & Blench (Heine & Nurse 2000:35) note that some have been trying to apply lexicostatistics, but that accepting Guthrie's arbitrary areal boundaries and a lack of good data may hinder discovering any improvements to his classification.

Paul Newman's chapter on comparative linguistics (Heine & Nurse 2000:259-271) focuses on the fourth stage as outlined above: drawing inferences about genetic relationships based on comparisons between languages. He shows several ways in which Greenberg's techniques of mass comparison can be constrained so as to achieve reliable results. Newman provides five distinctive principles underlying Greenberg's approach. Firstly, linguistic evidence alone (not racial) must be used. Secondly, specific (phonetic) resemblance is required (not typological similarities). Third, analysis based on vocabulary and morphology can be assumed to provide identical results. Fourth, classifications require no definite proof but simply to provide the best explanation of the data. Finally and most significantly, perhaps, transitivity applies so that if language A is related to B, and B to C, then we can take language A as being related to C.

One advantage of mass comparison is then, that even when explicit resemblances between A and C are scarce, other relationships can fill in gaps where perhaps a lack of data or some other 'noise' such as partial lexical diffusion or a series of complex sound changes obscures resemblances that otherwise might be found. However, we must note at this point that a posited relationship between A and C now rests on this one questionable assumption and two inferences. Inaccuracy in either of the relationships A-B, B-C would throw the relationship A-C into doubt.

Greenberg, Mathematical Models and Lexicostatistics (Fodor 1982)

István Fodor responded to Greenberg's claims with a withering attack on his classification method in the loquaciously-titled *A Fallacy of Contemporary Linguistics: J. H. Greenberg's Classification of the African Languages and His 'Comparative Method'*. In this brief book progressing through four editions from 1966-1982, Fodor introduces some welcome concrete (lexico-)statistical and mathematical methodology to the problem of assigning significance to correlations between languages. Clearly he is motivated by a strong disagreement with Greenberg, but his findings are nevertheless somewhat surprising. His key assertion is that depending on rate of change and time since separation, genetically-related languages will exhibit only 3-20% similarity due to their common origin. Furthermore, 68% of words will be considered similar due to chance, onomatopoeia or common borrowing. This would seem to cast significant doubt on all attempts to establish genetic links between any languages on the basis of similarity of forms where no historical data exists.

However, for all Fodor's attempts at mathematical rigor and objectivity, it must be noted that many of his source figures appear suspiciously arbitrary and unjustified. A relatively even-handed reviewer remarks "One may well suspect that Fodor is right in his conclusion that Greenberg has failed to prove his immense framework of classification, but his figures are open to grave suspicion...One feels very strongly that it would not be difficult to produce quite contradictory figures by looking elsewhere." (Palmer 1968:219) While Fodor's rejection of anything but absolutely regular phonetic changes may be unduly pessimistic, more inflexible than Neogrammarian requirements (Greenberg 1969:428) and lacking the insights of lexical diffusion, some aspects of the mathematical and statistical modelling may be profitable for establishing credible levels of significance. However the data used for the basis of this would need to be carefully sought to ensure that meaningful results were found. It is notable that in supposedly applying Greenberg's comparative methodology Fodor found many more similarities (over 68% in general) than Greenberg ever claimed. This suggests that a mismatch between Fodor's understanding of Greenberg's methodology and Greenberg's own understanding and implementation.

Fodor's discussion of experimental artificial languages (Fodor 1982:80ff) might usefully inform computer-based modelling and statistical investigation. Rosenfelder (1999a, 1999b) has some helpful programs to generate and compare artificial (unrelated) languages, and to calculate the probability of matches in unrelated artificial languages. This might facilitate experimentation with various phonological and lexical parameters to ascertain what level of chance resemblance might be expected. Rosenfelder's results unsurprisingly suggest that we should expect to see a bell curve (normal distribution) of chance resemblance between two unrelated languages. That is, depending on the degree of semantic and phonetic variation we can accept for a resemblance to be seen, the number of expected resemblances will cluster around a particular number.

Figure 1 shows such a distribution for a fairly wide range of phonetic and semantic variability. Here we would expect most likely 270 matches between the two 980-word lexicons. But significantly this shows that finding either fewer than 230 matches or more than 300 would be surprising.

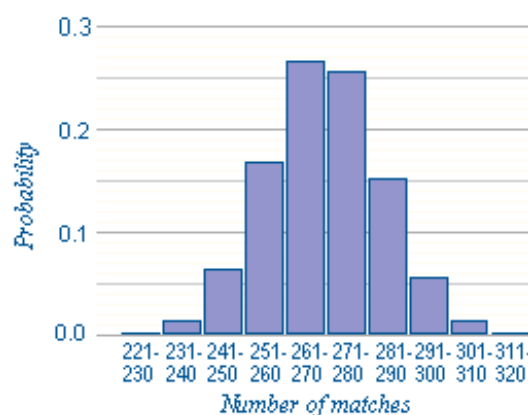


Figure 1: Probability of How Many Resemblances Might Be Expected For Two Random Languages

In the former case, if we found many fewer matches than the random expectation, it casts doubt on our method and implementation. We should suspect errors in the construction of the model of the two real languages we are comparing. Perhaps we have overlooked some resemblances which should have been included.

In the latter case, where we note many more resemblances than expected, the result may indicate a link between the languages. However, we must account for (that is remove) those resemblances due to universal onomatopoeia or recent borrowing from a common source before drawing any conclusions as to any genetic link. In removing onomatopoeic and loan-word links we must also adapt the model to account for the reduction of the lexicon, which is one aspect Fodor may have overlooked. The mathematical model becomes increasingly complex, but hopefully increasingly better reflects reality.

In personal correspondence Blench (2008) was somewhat dismissive of the value of lexicostatistics, presumably because ‘authoritative’ results can be so easily skewed by accident or deliberate misrepresentation. But is it necessary to dismiss mathematical modelling so summarily? Perhaps we should just work on doing the modelling and statistics better, rather than dismissing this field as useless because previous studies have failed to handle it properly. A helpful analogue might be the use of mathematical models in civil engineering where a useful model for predicting whether a bridge can bear a certain load must adequately reflect the reality of the construction of the bridge. Otherwise the model may prove less reliable and more dangerous than guesswork.

3. Data Preparation

Little controversy seems to surround the preparation of data. Guthrie (1967:15/21.21) explains that for comparison we require three elements about each ‘item’ (word, root or other type of morpheme): the language in which it occurs, its ‘shape’ (phonetic or phonemic form) and its meaning (generally an English gloss). Thus a simple word list for each language to be compared is sufficient for the comparative study.

Guthrie sees the main problem with these three elements as lying in assigning the meaning. Especially in an agglutinative language, such as a Bantu one, a word may have several constituent morphemes. By contrasting the meaning of words, phrases or sentences that differ in the smallest way (changing only one morpheme if possible), we derive a gloss by explaining the significance of the change. For example in English we might contrast “The dogs jumped” with “The dog jumped” and several other similar examples and see that the morpheme “-s” signifies number, but the constant morpheme “dog” signifies a different animal than “cat” in the sentence “The cat jumped”. Since there seem never to be direct equivalents between words in different languages across their whole semantic range, a gloss is necessarily an approximation. In later comparisons we must remember that the gloss is a very limited description of the function and meaning of the word it translates.

In addition, at the preparation stage we must decide what we can compare between languages (Guthrie 1967:13/14.01). Thus some rudimentary grammatical and morphological analysis of the languages concerned may help as words and morphemes are sorted into categories such as noun, verb, concord prefix, adjective etc. Then when comparing we can look carefully to check we are comparing like with like rather than even contemplating comparing verbs with concord prefixes, for example. Guthrie lists thirteen typical categories including a variety of prefix categories, nominals, verbals, radicals and bases

Only three elements are required but surely more information would help, especially when moving to the interpretive stage four, but also at earlier stages in the comparative process.

Although unmentioned in Guthrie's description of his methodology, an index by semantic domain at the end of Part I suggests that he saw the value of relaxing sole reliance on an English-based gloss system. A list of words sorted by English gloss would place 'eat', 'consume' and 'drink' quite far apart, hindering comparison, whereas in some cases semantic ranges of a word in one language may overlap significantly with other possible English glosses, so that it would be best to sort word lists not by gloss but by a semantic-based system which keeps potentially-related words together.

Some recent projects aimed mainly at rapidly and systematically building word-lists and dictionaries may well help here. Shore and van den Berg (2006) detail a workshop-based method of building dictionaries using Ron Moe's *Dictionary Development Program* (DDP). The numerically indexed list of questions and example words based on 1700 semantic domains comprising 9 major categories has ambitious (but necessary) aims "to be both exhaustive and universal" (Shore 2006:2). Word lists annotated with the relevant index would facilitate much better comparisons than those relying on glosses. For example we might use a semantic index of "3.4.1.2.1" for a word glossed "laugh" (that is, categorized under 3 Language and thought, 3.4 Emotions, 3.4.1 Good emotions). Later in the comparative stages it would then be easy to compare related words in other languages for potential resemblance based on the index. We could also vary the degrees of freedom

by choosing whether to match at the level of 3.4.1 or 3.4.1.2, for example. This is somewhat like the use of resource like Roget's (1979) *Thesaurus*. It allows us to specify more formally how tight is the match in meaning we are looking for, rather than informal definitions such as 'laugh' being considered similar to 'giggle' but not to 'grin'.

In Bantu studies, some recent word lists (such as from the Mara cluster) have databases annotated with "Bantu word list numbers", basically a unique three digit code which, independent of gloss, identifies a word according to questions used in the original elicitation of the word. This is weaker than the DDP method, but very useful for later comparison work. Without using such a numbering system, glosses might not always be consistent and comparable across different languages, even if word lists were compiled by the same person.

So while a simple word list is sufficient preparation according to Guthrie and others, supplementing it with more formal information will add benefits later.

Finally a summary phonological analysis of each language to be compared is required (at very least the inventory of phonemes and allophones) so that at later stages we can decide how to compare languages with differing phoneme systems. Obviously data must be represented in a consistent orthography which reflects all the distinctive phonemes and any lexical tone. The data orthography then may necessarily differ from the standard orthography adopted for language use.

4. Simple Comparisons

A simple comparison is essentially bringing together words from two potentially-related languages which are closely linked by shape, and which may resemble each other in meaning (or vice-versa). In terms of the three core elements for each item, element 1 (language) differs, and then one of elements 2 and 3 (shape and meaning) closely correlates while we test for some resemblance in the other:

Table 1: Simple comparison with ‘shape’ as connecting feature

Element 1 (language)	Yao (P.21)	Bemba (M.42)
Element 2 (shape)	-pet-	-pet-
Element 3 (meaning)	winnow	bend

In more mathematical terms then, this procedure when repeated aims to show a dependent relationship between sets which otherwise might be seen as being independent. Either ‘shape’ or ‘meaning’ may be the ‘connector’ or ‘connecting feature’ (Guthrie 1967:15/21.25) and choose an arbitrary item in an arbitrary reference language. If ‘shape’ is the connecting feature, then we search other languages’ word lists for identical or close matches to the reference shape then note their shape and meaning. In Table 1 above the shape ‘-pet-’ is the connector and Yao is the reference language. We find an identical shape in Bemba but the meaning of the Bemba shape ‘-pet-’ does not closely resemble that of ‘-pet-’ in Yao, so a link seems implausible.

Meaning as preferred connector

For a variety of reasons Guthrie prefers using ‘meaning’ as connector, which might yield the following simple comparisons (this time with three languages):

Table 2: Simple comparison with ‘meaning’ as connecting feature

Element 1 (language)	Yao (P.21)	Makua (P.31)	Bemba (M.42)
Element 2 (shape)	-pet-	-ver-	-el-
Element 3 (meaning)	winnow	winnow	winnow

Here the meaning is equivalent. If we had applied Ron Moe’s DDP4 list (Shore 2006:2) we could have searched for all words in Makua and Bemba with the semantic tag 6.2.6.1 ‘Winnow grain’, or possibly relaxed it to 6.2.6 ‘Process harvest’. Presumably Guthrie relied on word lists manually sorted by gloss. A computer database search would be much quicker and allow more flexibility. We will see some examples in this in our consideration of the systematising stage.

Guthrie’s (1967:15/21.22,21.32) preference for using meaning as the connector is due to a number of factors including the fact that when comparing languages with different phoneme systems, each sound must be considered within the context of that language’s sound system. So, for example, where one language lacks fricatives but the other has a

full set, it might be possible that **p** in the first language would be found where /f/ is found in the second. It is easier to discover these links if meaning is used as the connector. Similarly comparing a 5-vowel tonal language with an 11-vowel toneless language it would be very difficult to know which phonemes could be usefully equated.

However, in practical terms Guthrie (1967:16/21.42) admits that it might be worth gathering similar shapes from several languages, then using those lists to seek overlap in meaning. This is a largely a pragmatic consideration probably since the likelihood of a match for any arbitrary meaning may be quite low except in very closely-related languages. Here it may be that computer processing could bring some assistance in producing a 'work list' of items which have a similar meaning and whose form (shape) is very loosely similar.

Underlying Assumptions

Before we come to the stage of systematising the simple comparisons, we must consider the core assumptions underlying these two stages of making and systematising simple comparisons.

McMahon (1994:18ff) outlines how from Neogrammarian theory, language change was seen as a succession of regular, gradual sound shifts. This could be observed in the historical record of Indo-European language development. Neogrammarians saw these changes as affecting the whole lexicon simultaneously, a view challenged by lexical diffusionists who observed that often sound changes affect a small subsection of the lexicon before spreading gradually to the rest of the lexicon, but not necessarily in its entirety. Guthrie's methodology relies on the assumption that we can reconstruct the prehistoric state of languages with no recorded history by inferring a sequence of regular sound changes. We detect these regular sound changes by comparing languages that have similar forms for the same meaning, noticing patterns of differences between sounds and inferring the most plausible sequence of sound changes from a common ancestor. In some situations one language might have preserved the sounds of the common ancestor

language, and another language may have preserved other sounds. The simplest explanation is to be preferred (McMahon 1994:36).

Guthrie appears primarily to be governed by a Neogrammarian approach, but that has little impact on the first stages of his comparative methodology where we are merely observing the behaviour of the data. In languages that are genuinely related we can legitimately expect some consistent sound changes. It is the interpretation of those changes which depends the most on our theoretical assumptions and framework. So in the next section we will discuss the differences brought by a Neogrammarian, Structuralist, Generativist or other theory.

5. Systematising: Associated Comparisons and Comparative Series

While nothing useful can be done without well-prepared data and the mechanism for efficient comparisons between lexicons, it is at the point of systematising these comparisons that Guthrie and Greenberg diverge, and where much controversy rages. The synthesis of the data is the key step towards inference and drawing conclusions about links, but Guthrie (1967:16/21.41) insists that a good comparative linguist must avoid involving any (potential) theories in the data synthesis stage. This restriction guards one from the crippling circularity that would result if too quickly the researcher sought evidence of particular expected patterns. Other significant patterns might easily be overlooked, the opportunity to make fresh insights would be lost and we would only ever find what we already knew and were looking for.

The synthesis stage involves identifying every possible pattern in the data, and works towards producing lists which show good evidence for these patterns. Of course when beginning it will be unclear what simple comparisons might show accidental (random) similarity, and which resemblances are part of a wider pattern. It may also be difficult to know where to start. Ideally the systematising process must be as comprehensive as possible, to capture every regular sound change which can be observed in the languages under study.

Guthrie gives no explicit instructions or recommendations regarding choosing a place to start examining the data and determining when sufficient work has been done. But the existence of an apparently exhaustive list of comparative series in Part II show that he took comprehensive study seriously. His own study was made possible with the availability of a large resource of data (mostly word lists) at SOAS (School of Oriental and African Studies) in London. He reminds us that “it can never be claimed that the results are final, since when extra data become available from fresh languages it may well be possible to make use of some of the items that could not previously be incorporated into any associated list of the required type.” (Guthrie 1967:17/21.61)

Simple Computer-assistance

In order to establish some patterns of regular sound changes, a sensible place to start would appear to be a phoneme inventory for each language in question. We might begin with one language (fairly arbitrarily chosen) then for each phoneme, list words containing that phoneme (possibly differentiating between word-initial, medial and final occurrences). For each word we then want to find a semantic equivalent from the other languages we are comparing. In some ways this comparative study overlaps with stages of phonological analysis where we examine the behaviour of phones and phonemes in different environments.

We suspect some computer data processing may well help speed up the identification of potential comparative series, and help estimate how much of the data had been considered—either identifying possible resemblance or ruling it out. Unfortunately it appears difficult to achieve quick comparison using SIL’s software Fieldworks Language Explorer or Toolbox. *Paralex*, a very simple proof-of-concept demonstration (Rowbory 2008) takes data from the Mara language cluster in Tanzania (courtesy of Oliver Stegen, SIL UTB; idea courtesy Jeremy Lang). Several databases were transferred into *Paralex* via Flex ‘Lift’ XML export as lexeme with gloss (the core elements for comparison) but also indexed by Bantu Word List Number. This is a three digit number from a word list elicitation template and if present is used by the *Paralex* system as the connector between the data from each language in place of simply using the gloss. Additionally a

part of speech (grammatical category) is incorporated to help inform construction of comparative series. Languages were identified by a three-letter code: cwa, ikz, ngq, ntk, zak.

First, the system lets us analyse the constituent letters of each lexicon, to show which letters appear with what frequency in each language. Ignoring some spurious characters such as punctuation, the results are as follows:

Table 3: Distribution of characters in the four languages compared

	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	r	s	t	u	v	w	y	z	á	é	í	ó	ú	ĩ	ƙ
cwa	1741	498	10	96	1054	12	461	589	1219	127	1177	7	616	799	1414	20	1000	294	313	1469	1	341	373	1	31	19	29	18	31		
ikz	1807	519	53	100	849	1	547	318	1046	16	1200	2	628	699	963	18	1175	340	342	1000		361	305	145						487	956
ngq	1750	560	197	105	1634		691	462	1251		578		678	945	1685	2	1304	414	470	597		162	248	1							
zak	1611	500		78	1039		456	541	1072		1112		573	629	1377	1	1034	286	334	1203		345	392	117						112	229

Knowing nothing more about these languages, we might begin with Kabwa (coded as **cwa**), and examine words beginning with the bilabial voiceless plosive **p**:

Figure 2: Simple comparisons of p-initial words in Kabwa with other Mara language equivalents

Language	Search for p% in cwa • 2 Results	
cwa	<p>Choose Noun</p> <p>pilipili</p> <p>manga</p> <p>1222 pepper (green)</p>	<p>Choose Noun</p> <p>pilipili hoho</p> <p>1223 red pepper</p>
ikz	ikawarari	--
ngq	kawarare	--
ntk	--	--
zak	--	--

Notes

(eg search for ab)

- initial: **ab%**
- final: **%ab**
- med: **%ab%**
- any: **ab**

Set up

- Import xml databases from FLEX Full Lexicon export

(The figures in this section are screen captures from the web-based software to highlight relevant detail.) We see the first line shows only 2 words in the Kabwa lexicon that begin **p**- are clear anomalies (one word borrowed from Kiswahili). So we may conclude that this search is unproductive, and examine whether **p** occurs in any other more natural contexts. Incidentally though, we might notice a close correspondence between two of the other languages **ikz** / **ngq** though quite different from **cwa**. Such observations might be added to a work-list, perhaps to compare situations where **ngq** words end with **-e** and **ikz** words end **-i**. In the example above we see the Bantu word list number appear in grey under the 'master word', together with the gloss, which we assume should be the same for each equivalent in other languages. The words are listed first by part of speech

(adjectives, nouns, verbs) and then by word list number so as to gather the most comparable words as close to each other as possible with the data given. The search text is highlighted in red wherever it occurs in the language data to help the reader focus.

If we examine word medial -p- in cwa we find 15 matches:

Figure 3: Word medial -p- in Kabwa with equivalents for comparison

		Search for %p% in cwa • 15 Results										
Language	<<<<	1-10 • 11-15										>>>>
cwa	Choose Noun	Choose Noun	Choose Noun	Choose Noun	Choose Noun	Choose Noun	Choose Noun	Choose Noun	Choose Noun	Choose Noun	Choose Noun	
	omupango 0295 plan	ekipánde 0546 cloth	ekipúri 0555 earring	ekikapo 0626 basket	enchupá 0629 bottle	omuchipi 0651 strap	eripindo 0728 hem	omuchipi 0803 fishing line	empánga 0871 sword	eripaapayu 1213 papaya		
ikz	omoremo	egauni	ichogero	ikikaapo	ekeko	orogoye	umuringo	rirobo	risaba	ribabayu		
ngq	omorwecho	rigoti ryo goswara	rigoocho	ekehoncho	enchuba	embohero	omobindo	orosiri ryo kuroberi	risaba	ribabayu		
ntk	mihebo	--	kocho	sakÉ"Ī	nyerere	bara	ringo	--	--	--		
zak	--	ekitenge	engitawu	ekikapu	enzubha	omukanda	orukunyo	omukandaara	--	eribhabhaayo		

Unfortunately a list of 15 occurrences comprises only about 1% of the Kabwa word list, and the paucity of results suggests the phone(me) is quite rare. However, we can still observe several interesting comparisons, and click on the **Choose** button to set aside the (potentially interesting) columns headed with the words **ekikapo**, **omuchipi**, **eripaapayu**, **eripeera**, **omupiira**, **okukopa**, **okupakira**, **eripindo**.

Figure 4: Collecting Potential Comparative Series

		Search for %p% in cwa • 15 Results										
Language	<<<<	1-10 • 11-15										>>>>
cwa	Choose Noun	Choose Noun	Choose Noun	Choose Noun	Choose Noun	Choose Noun	Choose Noun	Choose Noun	Choose Verb	Choose Verb	Choose Verb	
	omuchipi 0651 strap	eripindo 0728 hem	omuchipi 0803 fishing line	empánga 0871 sword	eripaapayu 1213 papaya	eripeera 1215 guava	omupiira 1243 rubber	okukopa 0832 borrow	okukopesya 0833 lend	okupakira 0864 load		
ikz	orogoye	umuringo	rirobo	risaba	ribabayu	eripeera	umupiira	okokopa	ukuranda	--		
ngq	embohero	omobindo	orosiri ryo kuroberi	risaba	ribabayu	eripeera	omubiira	koba	saba	itweki		
ntk	bara	ringo	--	--	--	--	--	--	--	--		
zak	omukanda	orukunyo	omukandaara	--	eribhabhaayo	--	--	--	--	okusotera		

Comparative Series

cwa:0626,cwa:0803,cwa:1213,cwa:1215,cwa:1243,cwa:0832,cwa:0864,cwa:0728 <<< bookmark this if you want to save the list of comparative series

	Remove	Remove	Remove	Remove	Remove	Remove	Remove	Remove
cwa	ekikapo basket Noun	omuchipi fishing line Noun	eripaapayu papaya Noun	eripeera guava Noun	omupiira rubber Noun	okukopa borrow Verb	okupakira load Verb	eripindo hem Noun
ikz	ikikaapo Noun	rirobo Noun	ribabayu Noun	ripeera Noun	umupira Noun	okokopa Verb	--	umuringo Noun
ngq	ekehoncho Noun	orosiri ryo kuroberi Noun	ribabayu Noun	ripeera Noun	omubiira Noun	koba Verb	itweki Verb	omobindo Noun
ntk	sakÉ"Ī Noun	--	--	--	--	--	--	ringo Noun
zak	ekikapu Noun	omukandaara Noun	eribhabhaayo Noun	--	- Noun	--	okusotera Verb	orukunyo Noun

Choosing disconnected simple comparison lists on the top line moves them into a 'tray' of potential comparative series below. From here we can check to see if there are any

good patterns to observe. Here, the best that we could infer might be that **-p-** in **Kabwa** may be represented by a **-b-** in **ngq** or **-bh-** in **zak**. However the links are not particularly strong.

Searching for the voiced equivalent **b** yields 432 matching words in Kabwa, providing a much richer set of data to work with, so many words in fact that we would need to limit our search. Only 16 Kabwa words have word-initial **b**, but a number of patterns become apparent from 7 simple comparison lists (several of the other words were excluded because of lack of data):

Figure 5: Most Plausible Comparative Series from Kabwa word-initial *b*

cwa	bhoono now Adverb	bhwangu early Adverb	bhiringa how many interrog	bharina namesake Noun	bhasiirya day before yesterday Noun	bhenyu you-all pron	bhetu we pron
ikz	nangweno conj	--	iringa interrog	ring'ana Noun	basirya Noun	imwe pron	itwe pron
ngq	bonno Noun	mwecha Adverb	bareng interrog	bariina Noun	baseeronde Noun	inyu pron	itu pron
zak	bhoono Noun	bwangu Adverb	bhiringe interrog	--	bhasirya Noun	emwe pron	etwe pron

On the basis of the data in Figure 5, **b-** only occurs in the compound **bh-** word-initially, equivalent to **b-** in other languages such as **ngq** (and possibly **zak**). However the rightmost comparisons show **bh-** in Kabwa represented possibly by \emptyset in the other languages. Even this limited set of comparisons makes it plain that some comparisons may seem much more direct than others. Although we are concentrating on **b** here, we may identify signs of other possible patterns and again add these to a future work-list. Variation in length and vowel quality are evident (for example **bhoono** vs **bonno**)

More useful are some common collocations with **b**, such as searching for word-medial **amb**. Kabwa has 27 such occurrences, of which 19 may show evidence of patterns:

Figure 6: Comparative Series with word-medial *-amb-* in Kabwa.
(Shown without glosses and parts of speech, for clarity.)

	1	2	3	4	5	6	7	8	9	10
cwa	entambi	obhwina obhutambi	obhubhambare	eritambuka	ekigambó	ekitambaara	ekitambaara	obhwambo	erisambwa	nyawambwi
ikz	ndehu	--	kegare	ritambuka	ikigambo	ikisagi	ritambara	urwambo	--	nyawambwe
ngq	entambe	egetuko	embambare	ritamboka	ekigambo	ekimaga	ritasu	orwambo	emisambu	enyawambwe
zak	-tambi	-tambi	-bhambagare	eritarambuka	engamba	ekitambaara	eritasa	obwambo	erisambwa eribhi	nyawambwe
	11	12	13	14	15	16	17	18	19	
cwa	nyawambubhi	okwamba	okugamba omuntu	okusamba (na omuuro)	okuhamba	okwambuka	okuhambirira	okuhamba	okutambihya	
ikz	nyawambube	ubutange	kumugamba	ukukara	ukuhamba	kwamboka	ukuhambera	okomera	kongera oboreehu	
ngq	nyamse	yambere	gamba bobe	samba	emi	omboka	tochera	busura	engeri obutambe	
zak	nyawambubhi	hinga	okugambana	okwokya	okwemya	okwambuka	okwemererya	okwemya	okutambihya	

Assessment of the Computer-Assisted Comparison System

This system is very limited, but brings some immediate benefits: it should be possible to establish the easiest (and most robust) comparative series very quickly; we can easily ensure we cover all occurrences of a particular phoneme in one language thoroughly; we can easily switch the arbitrary ‘master’ language.

However, this basic system reveals a number of critical problems which limit the usefulness of this system. Firstly the standardisation of the data entered into such a database-driven system has a very significant impact on what can be found. The data here seem not sufficiently prepared for this comparative study. Supplied in their orthographic form, **b** in one language may represent the same phone as **bh** in another, so that direct correspondences between phonemes of the different languages are obscured. Some standardisation of phones or phonemes across the languages might be helpful. Also it is unclear whether morphemes have been analysed sufficiently here. Any class marking or compounding should probably be removed so that noun roots and verb radicals can be compared without the morphology confusing the picture. Sometimes the words are actually phrases in some languages, but single words in other languages, which hinders comparison.

Secondly, we are limited to only one possible link (via Bantu word list number) between items in different languages. It is quite possible that due to semantic shift another similar

word might be found as a match if we had a way to relax the requirements of matching. This is where the *Dictionary Development Program* indices could be helpful, since we could specify the degrees of freedom acceptable and for each word in the ‘master’ language, display a range of semantically-similar words in the other languages.

Thirdly, while we can search in this method for a sequence of letters it would be more powerful, especially in the early stages of analysis, to be able to use meta-characters (similar to variables in SIL Toolbox), to search for vowels, nasals, bilabials etc. seeking patterns. If we were to attempt to let the computer seek patterns automatically it would probably be advantageous to analyse the features of each phoneme, and in computing a match allow a certain number of degrees of freedom or fuzziness in terms of features matched. Thus **b** and **v** will match if allowing 2 degrees of freedom, but not **b** and **z**.

Finally, this process is still very laborious and some improvements should be sought to ensure that we cover the full range of data and identify as many patterns as possible. It might be that some semi-automation could run through all possible comparisons and look for the most plausible matches according to some algorithm. The current unoptimised system works almost instantly for the four 1500-word lexicons shown here (300ms per search), so this may well be feasible.

But, as a simple tool for speeding up comparing words, this shows some promise. It might be helpful to have a partner system which allows the user to postulate and test hypotheses about regular sound differences between languages. We might suggest that words ending **-i** in one language always end **-e** in another, then ask the system to show which words fit the pattern and which do not. This would be slightly more complex to develop, but if clearly specified would greatly assist the intelligent comparative linguist.

Any tools developed with a greater sophistication than the existing *Paralex* and any significant use of lexicostatistics begins leaning heavily on assumptions about language change. This is where the insights outlined in McMahon (1994) will be very significant.

6. Inferring Relationships

As already mentioned, Guthrie warns against letting any inferred hypotheses affect the systematisation stage. But when all the possible patterns have been found in the data, the patterns themselves must be examined to infer what changes they might show. This is where the assumptions about language change have the greatest impact. At very least the discerning comparative linguist must identify what kind of sound changes are to be observed, whether fitting a Neogrammarian or lexical diffusionist theory. The residue of simple comparisons which do not fit the major comparative series need to be explained in some way.

As mentioned earlier, Guthrie, Newman and others in the literature devote considerable time to the details of inferring relationships. This is beyond the scope of this paper, but the first result of the inference stage should be a collection of sound change rules explaining the relationship between the languages being compared. From these rules we may infer groupings of languages (which languages are more similar than others), and postulate the most likely original form from which each language is derived. There are several competing methodologies for determining the original form. If two or more related forms are observed, the one which is current predominant is not necessarily identical to the original form (or vice versa).

Guthrie, Greenberg and Differing Interpretations

Since Guthrie was concerned mainly about establishing the prehistoric 'Proto-Bantu', he was able to ignore some detail that would need to be considered to establish true genetic language families. His groupings (E.55, E.40 etc) were overtly based on area (despite some presuming genetic links). Many sub-patterns might be discovered within the data, but Guthrie was interested in the patterns that could be traced back to Proto-Bantu so does not appear to have explored these sub-patterns. He also warns against forming inferences on the basis of other inferences (Guthrie 1967:15/21.02) since that would risk multiplying erroneous judgments.

Greenberg, aware of that random change and borrowing may gradually conceal true genetic links, seems to have concentrated his efforts on identifying any patterns that could be seen rather than accounting for the residue of simple comparisons that fail to fit into any comparative series. Thus he was able to identify links which otherwise could have been hard to establish (Wikipedia 2008:MLC), but was susceptible inferring significance in forms which randomly coincided. Beginning with a hypothesis about probably language families, in constructing his comparative series he considered evidence from one language within a family to represent the whole family, so really compared families with families. Clearly this involves considerable selectivity and thus increases the possibility that spurious evidence could be given more significance than is justified. Guthrie, on the other hand, refused to give any comparative significance to his areal groupings.

Clearly there is great scope for variation in interpretation and vigorous debate in this final stage. It may well be possible to achieve 'correct' analysis using defective methodology. So, while increasing numbers have criticised Greenberg's methodology, many linguistics surprisingly still accept his classifications and are now seeking to find a better basis for them.

7. Possible Application Within and Beyond Bantu

A good deal of study remains to be done to improve on the work stimulated by Guthrie. Nurse (2001) sees an important rôle for SIL fieldworkers collecting more and better data to facilitate improved comparisons. In turn, Bible translators and literacy workers aware the large scale of the task remaining have become increasingly interested in 'cluster projects' where language development work proceeds in a group of related languages simultaneously. If the languages in a cluster are closely related then the aim is to enable insights from work in one language to feed into the work on related languages. The theory is that uninformed duplication can be reduced by language development workers working together, and consultants overseeing progress on the whole cluster. But two key questions arise: how are suitable clusters to be identified, and what specific linguistic advantages should be sought in running a cluster project?

To some extent cluster projects may be organised more on the basis of pragmatic and sociolinguistic factors than solely linguistic concerns. However as any cluster project is investigated and established, comparative linguistics could offer important information and frameworks to help the constituent language projects genuinely reinforce each other. Comparative study may offer suggestions for developing orthographies which function well across a region as well as suiting the immediate language. This may have been behind Kay Williamson's proposal of a Pan-Nigerian Alphabet (Wikipedia 2008:PNA). Some understanding of grammatical and lexical links and consistent differences between languages in a cluster may improve dictionaries, grammars and literature produced. Where staff are limited and languages seem very closely related some computer-aided translation between related languages may help produce quick first drafts or translation checking aids. But all this relies on good comparative linguistic study.

So practical and academic interests may well exist in some degree of synergy. Those doing the fieldwork required to produce good academic linguistic study of a language may later reap the benefits for their translation and literacy work, as the links between the languages are further explored.

The Point of Genetic and Other Classifications

However, we may also question the importance and usefulness of genetic classification and establishing a 'family tree' of languages. Nurse (2001) describes a variety of possible bases for classifications including area, typology, genetic/historical affiliation and reference. Most classifications (such as Guthrie) begin with area and reference, and classification by typology seems largely discredited as a solid basis for any useful classification. Guthrie offers only the barest skeleton of genetic classification, filling in the trunk (Proto-Bantu) and individual leaves (areally-classified languages) of the family tree. One motivation for tracing the historical development of one or more languages is to complement the wider prehistory of the people who speak that language. Where oral history is vague or apparently unreliable, we might hope to obtain some degree of objective evidence for the origins and prehistoric movements and contacts of a people through examining the prehistory of their language. However it must be admitted that

this leaves many imponderables and uncertainties, so that most comparative linguists would be reluctant to infer much sociological or anthropological information from the inferences of comparative linguistic study.

As Nurse (2003:132) suggests in considering the fairly cosmopolitan Zone F Bantu languages, genetic links in language show just one influence. Comparative linguistics may consider all other contacts or developments as random or irrelevant interference, but that assumes that the most important thing in understanding or classifying a language is identifying some 'pure' core that has survived various mutations. Perhaps a more well-rounded comparison would seek to understand the various different influences, from wherever they have come. So where lots of languages have had contact (as seems to be the case in several Bantu areas) we may see various different phonological changes, borrowings etc. occurring not only within one language but possibly across several languages, or within only some dialects of one language.

Nigerian Languages

Nigeria boasts over 500 living languages, primarily from Niger-Congo, but also significantly from Afro-Asiatic and Nilo-Saharan phyla (Ethnologue 2005). Compared to the Bantu languages much less study appears to have been carried out on these languages. Some within Niger-Congo are classed as 'Bantoid' or seem closely related to Bantu languages. Some have experienced significant influence from very different languages. Clearly classifying Nigerian languages is somewhat challenging, but a matter of considerable interest and importance. Considerable language development (including ultimately Bible translation) is required in many languages. Some 'cluster' projects could be possible, but more linguistic fieldwork is required to identify the sub-groupings and families within Benue-Congo and other groups, before it is clear which languages would benefit from a cluster approach.

Where an Afro-Asiatic language such as Hausa is concerned, we may see evidence of direct and indirect influence from Arabic, but in different ways and at different times. Both may be derived from a similar 'stock' or super-phylum, but Hausa seems to have

developed separately until a more recent time when Arabic words and constructions began to be incorporated afresh into Hausa. Unravelling the various timings of the influences may involve considerable study and care.

8. Conclusions and Further Study

Guthrie's comparative linguistic method is robust and appropriately cautious and can provide a helpful framework for some of the future research which seems greatly needed in African linguistics. However there are a number of areas in which it would benefit greatly in being extended and updated. Since later stages of systematisation and inference hang on what is found in earlier stages, it seems important to make the most of computerised data processing to prepare, pre-process and organise the data thoroughly. Much work is laborious and in such circumstances humans are prone to error, inconsistency, delay and exhaustion, requiring constant checking. Some improved automation of the most laborious parts, and tools to check the plausibility of hypotheses might speed up and improve the quality of comparative research.

Much maligned by some and much lauded by others, lexicostatistics seems best utilised as a checking mechanism. With very careful modelling (taking into account the unique aspects of the phonology and lexicon of each language) we should be able to produce a sanguine and relatively objective measure of the significance of lexicostatistical results. That is, we might be able to improve upon Guthrie's (1967:18/22.31) reasonable-sounding but unproven assumption that a pattern found in three or more simple comparisons is strong enough to be the basis for a comparative series.

Although most of the literature has a preoccupation with genetic-based classification, it seems unfortunate completely to disregard all comparative information which does not clearly yield evidence of genetic links. Languages exist as they are today for a variety of reasons. Producing a genetic classification is only one reason to study the prehistory of modern languages through comparative study.

Other legitimate motivations exist. We may want to understand in the best possible way why living languages today exist and function as they do. We may want to estimate how languages might change in the future due to internal and external influence. For these purposes we will need to examine the various contributions of genetic links, and shared or different influences from other languages. This includes learning how typology, grammar, phonology as well as the lexicon can be transferred from one language or group of languages to another.

Such multilateral study helps explain many of the vagaries of English, which has come under many varied influences over the last millennium. It would surely help deepen and broaden our understanding of African languages too with benefits for academia and practical language development.

9. Bibliography

- Blench, Roger (2008), Personal email correspondence 10 June 2008
- Crabb, David W. (1968) "Review of The Problems in the Classification of the African Languages: Methodological and Theoretical Conclusions Concerning the Classification System of Joseph H. Greenberg" by Istvan Fodor in *American Anthropologist*, New Series, Vol. 70, No. 4, (Aug., 1968), pp. 831-832, Oxford: Blackwell Publishing on behalf of the American Anthropological Association
- Fodor, István (1982) *A Fallacy of Contemporary Linguistics: J H Greenberg's Classification of the African Languages and His "Comparative Method"* Hamburg: Helmut Buske
- Gordon, Raymond G., Jr. (ed.), (2005) *Ethnologue: Languages of the World*, Fifteenth edition. Dallas, Tex.: SIL International. Online version: <http://www.ethnologue.com>
- Greenberg, Joseph (1963) *The Languages of Africa*. Bloomington: Indiana University Press / The Hague: Mouton & Co. (2e, 1966; 3e, 1970)
- (1969) "Review of The Problems in the Classification of the African Languages: Methodological and Theoretical Conclusions concerning the Classification System of Joseph H. Greenberg by István Fodor" in *Language*, Vol. 45, No. 2, Part 1, (Jun., 1969), pp. 427-432 Linguistic Society of America
- Guthrie, Malcolm (1948) *Classification of the Bantu Languages*, London: Oxford University Press for the International African Institute
summarised at <http://www.linguistics.berkeley.edu/CBOLD/Lgs/LgsbyGN.html>
- Guthrie, Malcolm (1967) *Comparative Bantu*, Farnborough: Gregg Press Ltd
- Heine, Bernd and Derek Nurse (2000 eds) *African Languages: An Introduction* Cambridge: Cambridge University Press
- Lang, Jeremy (2008) *A Comparative Study of E.40 Languages*, Unpublished paper, Nairobi Evangelical Graduate School of Theology
- McMahon, April M S (1994) *Understanding Language Change* Cambridge: Cambridge University Press
- Moe, Ron (2006) *Dictionary Development Program (version 4)* SIL
<http://www.sil.org/computing/ddp/>
- Nurse, Derek (1994) "'Historical' classifications of the Bantu languages" 65-81 in *Azania* 29/30: Nairobi: The British Institute in Eastern Africa
- (2002) A Survey Report for the Bantu Languages, SIL
<http://www.sil.org/silesr/2002/016/silesr2002-016.htm>
- (2003) *Stratigraphy and Prehistory: Bantu Zone F* in Andersen, Henning (ed) *Language Contacts in Prehistory*, Studies in Stratigraphy pp 115-134
- Palmer, F. R. (1968), "Review of The Problems in the Classification of the African Languages by István Fodor" in *Africa: Journal of the International African Institute*, Vol. 38, No. 2, (Apr., 1968), Edinburgh: Edinburgh University Press pp. 219-220
- Roget, Peter Mark (1979) *Thesaurus of English words and phrases : classified and arranged so as to facilitate the expression of ideas and to assist in literary composition*, New York: Avenel Books
- Rosenfelder, Mark (1999a) *Program for comparing artificially-generated languages*
<http://www.zompist.com/simm.html> accessed 27 June 2008
- (1999b) "How likely are chance resemblances between languages?"
<http://www.zompist.com/chance.htm>
- Rowbory, David (2008) *Paralex: Parallel lexicon searching for comparative linguistics*, Web-based software at <http://www.rowbory.co.uk/software/anna/paralex.php>

- Shore, Susan & René van den Berg (2006) *A New Mass Elicitation Technique: The Dictionary Development Program* Paper presented at Tenth International Conference on Austronesian Linguistics. 17-20 January 2006. Puerto Princesa City, Palawan, Philippines. <http://www.sil.org/asia/philippines/ical/papers.html>
- Wikipedia (2008:PNA) *Pan-Nigerian Alphabet*, http://en.wikipedia.org/wiki/Pan-Nigerian_Alphabet accessed 4 July 2008
- (2008:MLC) *Mass Lexical Comparison*, http://en.wikipedia.org/wiki/Mass_lexical_comparison accessed 4 July 2008