

UNIVERSITY OF DAR ES SALAAM  
FACULTY OF ARTS AND SOCIAL SCIENCES  
DEPARTMENT OF FOREIGN LANGUAGES AND LINGUISTICS

LANGUAGES OF TANZANIA PROJECT

## **WORKSHOP II REPORT**

REPORT OF THE WORKSHOP HELD AT BELINDA OCEAN RESORT  
FROM 8<sup>TH</sup> - 9<sup>TH</sup> MARCH 2002

*Editors:*

*Dr A.Y. Mreta*

*Dr. H.R.T. Muzale*

*Dr. J.M. Rugemalira*

## **OPENING REMARKS**

Prof. D. P. B. Massamba  
Institute of Kiswahili Research

### **Ladies and Gentlemen!**

It gives me much pleasure to welcome you all to this 2<sup>nd</sup> Languages of Tanzania (LOT) project workshop. As you are all aware, this workshop is a follow up of the first workshop that was held last year in preparation for the pilot study in data collection.

After the first workshop, a group of researchers and research assistants went to Kagera Region to test our research tools and methods. In that pilot study quite a number of experiences were gained. The main objective of this second workshop is to share those experiences and determine ways for moving forward to the main research assignment.

This workshop marks a very important stage in the project's development. It is this workshop that will give a clear direction as to how this whole exercise should be carried out in the light of what we have learned from the pilot study. It is my sincere hope, therefore, that we will make constructive comments and suggestions during our deliberations. We should bear in mind that the project is, after this workshop, entering a very crucial stage in terms of its implementation. We should therefore give it all the help we have got.

We have among us colleagues from other universities who have traveled all the way to join us in this noble endeavor. We have with us Prof. Herbert Chimundu from the University of Zimbabwe, Prof. Karsten Legere from Gottenborg University, and Prof. Herman Batibo from the University of

Botswana. We are also pleased to have with us colleagues from the Dodoma and Nairobi SIL offices, as well as participants from Pioneer Bible Translators in Morogoro; to them all we say *KARIBU sana!* Enjoy the workshop and enjoy the good atmosphere of the Haven of Peace.

With these few remarks I wish you all a wonderful workshop!

## EXPERIENCES FROM THE PILOT STUDY

Prof. K. Kahigi

Department of Kiswahili

### 1. Research Areas and Languages/Dialects Studied

The pilot study focused on Ruhaya and Runyambo as planned. The study areas included the following districts: Karagwe, Bukoba Urban and Rural and Muleba. Bukoba Rural was divided into two study areas: Kiziba/Nkenge and Kyamutwara.

Dialects studied include:

- "Ruhaya" Dialects:
  - (i) Ruziba, "Ruganda Kyaka" (Kiziba/Nkenge).
  - (ii) Ruhyoza, Ruhamba (Kyamutwara = Bukoba Urban & Rural)
  - (iii) Runyaihangiro, Ruhamba, Runyangote (Muleba)
- "Runyambo" dialects:
  - (i) Runyamabira (spoken in Karagwe West)
  - (ii) Runyamigongo (spoken in Karagwe East)

### 2. Participants

Participants in the pilot study included four researchers and eight research assistants (who were native speakers of the dialects) as noted below:

**Nkenge/Kiziba:** C. Rubagumya (R); Esta Felix and Deo Rutagwelera (RAs)

**Bukoba Rural & Urban (Kyamutwara):** H.R.T Muzale (R); Rosemary Mukandala and Stanslaus Gordian (RAs).

**Muleba:** K. Kahigi (R); Chrisant Bukuba and Safina Mashaka (RAs).

**Karagwe:** J. Rugemalira (R); Zeulia Jeremiah and Wilfred Kahumuza (RAs)

The other participants were the informants in the respective research areas.

### **3. Objectives**

The main objective of the pilot study was to test the instruments formulated on the basis of suggestions from the first workshop, i.e. the lexical and morphosyntactic questionnaires. The lexical questionnaire had 3400 items, while the morphosyntactic questionnaire had 256 sentences. The items on both questionnaires appeared in English and Kiswahili. These questionnaires had to be completed by the research assistants, under the supervision of the researchers.

In addition, the researchers and the research assistants were required to:

- (i) seek and obtain from the regional or district offices a district or divisional or ward map that would be used at a later stage in the drawing of a language map;
- (ii) seek a list of divisions, wards and villages to use in identifying the language/dialect spoken in each village/ward/division;
- (iii) list dialectal differences (using a notebook);
- (iv) tape oral texts: stories, songs, sayings;
- (v) collect documents written in Ruhaya and Runyambo.

### **4. Planned Activities**

The pilot study period was from 14<sup>th</sup> August to 15<sup>th</sup> September 2001. The planned activities were as follows:

- August 14<sup>th</sup> - 24<sup>th</sup>: Research Assistants, under the supervision of the researchers, filled out the lexical and morphosyntactic questionnaires. Items not filled in at this stage were to be filled out with the help of informants in the field.

- August 29<sup>th</sup> - August 31<sup>st</sup>: Research Assistants to seek letter of introduction and research permission from the regional office and the district offices. Also to seek the following from the same offices: information on language/dialects, regional/district/divisional/ward maps, lists of divisions, wards and villages. On arrival in research areas, the Research Assistants were supposed to identify key informants (at least one for each dialect) who were versed in oral literature and would be helpful in completing the questionnaires.
- August 31<sup>st</sup> - September 1<sup>st</sup>: Researchers to seek letter of introduction and research permission from regional and district offices. Also to seek the following from the same offices: information on language/dialects, regional/district/divisional/ward maps, lists of divisions, wards and villages. On arrival in research areas, the researchers were supposed to start tape-recording folk-tales and songs from the key informants, apart from coordinating and supervising the activities of the RAs.
- September 1<sup>st</sup> - September 15<sup>th</sup>: Researchers and research assistants to accomplish the following:
  - collect written documents (Haya bibles, prayerbooks, etc.) from seminaries, missions and individual Haya speakers.
  - list all divisions, wards and respective dialects spoken there.
  - list all dialectal differences between dialects in relation to pronunciation, lexical items and syntax.

All these activities were carried out more or less successfully (but see section 6).

## **5. Data and Materials Collected**

The following data and materials were collected during the pilot study:

## 5.1 Data

- The lexical data: the main list, additional lexical items, dialectal variants, plus specific vocabulary (from hunters, fishermen, etc.)
- Morphosyntactic data: the main list completed in Dar es Salaam.
- Atlas data: demographic data (1988 population census) for administrative villages in Bukoba districts.

## 5.2 Audio Tapes

- Taped folktales, riddles and songs.
- 1 set of Embandizo audio tapes
- 1 set of Ezaburi audio tapes

## 5.3 Books

- Ebigano bya Buhaya, (R.A. Mwombeki & G.B. Kamanzi), 1999.
- Ekitabo kya Embandizo, United Bible Societies, 1988.
- Ruti, The Bible Society of Tanzania, 1992.
- Okubonabona kw'omukama Yesu, 1981.
- Kitabo Ekilenga Okushoborora Ebigambo bya Katekismu, Bukoba White Fathers' Mission, 1925.
- Emigenzo y'aBahaya ab'eirai (S. Rweyemamu), Bukoba Diocese, 1994.
- 500 Haya Proverbs (H.B. Nestor), North Western Publishers, 1994.
- Katekisimu eya Buli Dominika, Vicariatus Apostolicus de Bukoba, 1952.
- Orodha ya Miti na Migomba Kagera, (P.P.I. Kanywa), Rumuli Press 1986.
- Emigani na Ebikoikyo (Leonidas Kalugila), North Western Publishers, Bukoba, 1992.
- Tusingize Omukama: Ekitabo ky'Enshala n'Empoya , Bukoba, 1981.
- Katekisimu y'Ebigambo Bikuru by'Edini, Diocese of Bukoba, 1984.
- Ikani-Ngambo: Oruhaya Dictionary, (Bona-Baisi) 1957.

**N.B.** District/divisional/ward maps for Bukoba Rural and Urban and Muleba districts were not available.

## **6. Problems**

### **6.1 Informants**

- Getting the ideal informants that we wanted, that is those that knew folktales and songs, was not easy, given the time limitation and transport problems.
- Some informants do not tell the truth about what they know/don't know in the language; for instance, they might say they are good at folktales, songs, etc. while they are not.
- Some informants took a long time to provide the data needed/requested.
- Some informants kept making promises without showing up, or could not be found.

### **6.2 Transport**

Transport from some areas to others was unreliable or unavailable. Thus some areas could not be visited.

### **6.3 Payments and Costs**

- Although all informants expected some payment for their information, most informants demanded more from the researchers than from assistants who are natives of the area.
- Actual payments and costs exceeded the amount budgeted.

### **6.4 Faint Recording**

This problem was noticed after one researcher played back some of the recorded tales and songs. The reason for this was not clear at the time.

### **6.5 Time**

Research time was not enough. Therefore:

- Researchers could not visit Kamachumu (which is central to the Ruhamba dialect) and Ikuza Island (which is central to Ruzinza dialect).
- Researchers could not get in touch with some people who are known to be rich sources of folkloristic and historical data.

## 6.6 Data Elicitation

Elicitation was the main method used in data collection. The following were some of the problems encountered:

- Given that the language of elicitation was mainly Swahili (although English equivalents were also available), some items on the lexical questionnaire were misinterpreted, e.g. in at least one case, *kamba* (lobster) was thought to be *omuguha* (rope);
- Some items had no equivalents in Ruhaya/Runyambo, especially those having to do with plants and culture specific items.

## 7.0 Conclusion

- (i) The main objective, that of testing the instruments, was in the main achieved. The instruments were on the whole found to be adequate for the purposes for which they were made, although they could be improved to eliminate the elicitation problems noted above.
- (ii) Concerning other objectives, i.e. collecting additional data and materials/documents, the following may be noted:
  - List of dialectal differences: this was not systematically carried out. Each researcher/research assistant made his or her own list.
  - Collection of written documents: these are useful sources of lexical, morphological, syntactic and semantic data; they included religious, cultural and literary texts.

- Audiotapes: these are useful sources of phonological, lexical and grammatical data.
- Maps: these were generally not available. A good map for Karagwe district was obtained, as well as a sketch map for Muleba district.

\* \* \*

## DISCUSSION

The discussion established the consensus that researchers should employ a more systematic approach for determining the dialectal differences. The dialectal divisions that are based on people's beliefs and other sentiments, as well as those based on geographical, political and traditional kingdom divisions should be avoided. In addition, it was suggested that, ideally, the researchers should start with a sociolinguistic survey in determining dialect divisions.

It was also proposed that the data should be transcribed in the field. By doing so, it will be possible to verify, crosscheck and clarify some unclear and misinterpreted items. The general feeling was that research assistants should have some basic linguistics knowledge that will enable them to make important decisions about the representation of phonetic information.

The participants expressed concern about the time constraint. Two weeks is too short for serious research work of this type. Researchers should not be in a hurry. It was also felt that the bulk of the written materials and audiotapes obtained had a religious bias. On the whole,

however, the participants were in agreement that the pilot study was a success, bearing in mind that there were a lot of new experiences gained in the field.

## PILOT STUDY EXPERIENCES AND LEXICAL DATA PROCESSING

Dr. Josephat M. Rugemalira

Department of Foreign Languages and Linguistics

### 1. General Observations on the Pilot Study

The pilot study was conducted over a period of three weeks in September 2001. Two languages – Ruhaya and Runyambo in Kagera region, west of Lake Victoria, were targeted. Four researchers and eight research assistants were engaged to collect the type of data described below:

- a) lexical data: completing 3400 items in the target language, and to expand the list using “type of” entries as leads;
- b) morphosyntactic data: completing 256 sentences in the target language designed to capture various forms of the noun and verb, including affixes and pronominals, and basic sentence structure;
- c) administrative divisions: a list of villages and wards in the district and a map of the district;
- d) information on the geographical distribution of the languages/dialects in the district and a list of linguistic features distinguishing one dialect from another;
- e) oral texts on tape: stories, songs, sayings;
- f) written materials in and about the respective languages: bible translations, books, grammars, dictionaries, newspapers, etc.

A few problems were encountered in the course of the pilot study. The lexical lists could easily be expanded, but very often a Ruhaya or Runyambo word would be obtained which could not readily be glossed in English or Swahili. This was especially the case with plant and animal names, as well as many

culture specific items and concepts. Take an example of a Runyambo word from the field of beer brewing: *orureeba* refers to a shallow basin made in the ground where ripe bananas are placed for subsequent crushing to extract the juice; and there is a whole host of vocabulary related to the materials and processes involved in preparing this site to hold the banana juice. Items of this kind require a long explanation in a phrase or sentence of considerable length. Yet for the purposes of the wordlist, there are no English or even Swahili lexical *entries* for such items.

Besides the absence of a ready English or Swahili gloss in cases of this type, a concern that arose was that, in order to compile a wordlist/dictionary that would seek to even approach double the number of items in the basic English list, the starting point would have to be the target ethnic community language itself. But in the absence of a body of written or even spoken samples of the language stored and retrievable for use, the compilation of a dictionary will depend very much on the elicitation efforts by knowledgeable researchers. In order to do justice to the language under study, a version based on *entries from the local language* itself must be produced.

The list of administrative units – villages and wards – was readily available. But the maps were not. Of the four districts covered, from only one, Karagwe, was an official map obtained.

Information on dialect distribution was rather scant. It may be that such information is difficult to capture, particularly due to the absence of a structured instrument for elicitation. What the researchers wanted to determine from the speech communities themselves was an indication of the perceived differences between two dialects/languages in the form of a list of

lexical, phonetic, and grammatical differences. A closer examination of the lexical and morphosyntactic data will bring out important differences, especially at a higher level (between Runyambo and Ruhaya), but internal dialect variations, within Runyambo and within Ruhaya, are likely to be lost. It will be noted that an attempt to capture this level of detail for the whole country will stretch the *financial* and *personnel* resources at the disposal of the project.

The national population census authorities would not accept a language question in the census questionnaire. A question on tribal affiliation was last asked in the 1967 census, which made use of a standard list of 126 tribes, but eliciting such information is now regarded as undesirable and subversive. For the purposes of making reasonable estimates of the number of speakers, it is assumed, in the Languages of Tanzania project, that each village would have only one language. So even where a different language comprises a *substantial* population of the village it would be statistically ignored. In addition, such information is impressionistic, the term "substantial" being subjectively determined since there are no records to rely on.

Oral materials were obtained in abundance. Each researcher was limited to a maximum of two cassette tapes (two hours). While the rationale for this material was to obtain data for phonological analysis and to get samples for preservation, it may be necessary to reconsider this aspect because of the demanding nature of such data at processing stage. The concern arises because tape data may not be easily stored and retrieved over a long period. To be of any use for future generations, oral data must be transcribed in normal orthography, entered into the computer (as written text), and also

digitalized on CD. This is extremely time consuming. It will be important for the project to collect such data very selectively.

Written materials were also collected, a large amount among them being religious literature. The preservation of such materials is relatively risk free, and they are readily available for use. To make them even more readily accessible, the creation of electronic files by typing or scanning and subsequent editing can be a formidable but worthwhile undertaking.

## **2. Lexical Data Processing**

The handling of the lexical and morphosyntactic data has posed a few challenges. Virtually all the data has been keyed into the computer – only a few of the lexical items that were added outside the basic list have not been entered. Although it was possible to engage research assistants to do most of the data entry, there has been a lot of time-consuming editorial work by the researchers involving

- a) making sure that each entry has been faithfully entered (correct spelling, correct word);
- b) correcting errors by the informant/assistant about the correct gloss for the Swahili and English items, and getting rid of literal translations and infelicitous expressions;
- c) establishing some consistency in the orthographical representations – vowel length, vowel elisions and coalescence, phonetic/phonological status determination (l/r; β), and choice of symbol (c/ch);
- d) tone marking on the lexical data.

A few examples from the data will clarify the nature of the task involved. Runyambo 1 (R1) and Runyambo 2 (R2) represent the data from the two research assistants engaged in the pilot study for Runyambo data:

ENGLISH	KISWAHILI	RUNYAMBO 1(R1)	RUNYAMBO 2 (R2)
lobster	kamba	omugoha	*
<i>omuguha = rope; this is the other sense of the Kiswahili gloss (note also correct spelling)</i>			
head of cattle	kichwa cha ng'ombe	engundu	echiraro che ente
<i>literal translation in Kiswahili; engundu = lead cow/bull; gloss R2 as "whole krall (of cattle)</i>			
herd of cattle	kundi la ng'ombe	amasyo ge ente	echiraaro
<i>orthographical issues here: ge ente = g'ente (of cattle); also above: che ente =ch'ente; note inconsistency in spelling echiraaro; preferred version: eciráaro (with tone mark)</i>			
plunder	kuteka (mji)	okunyaga	kutaha ameizi
<i>appropriate Kiswahili gloss is pora; R2 = fetch water! The mji in the Kiswahili gloss was read as maji (water)</i>			
roan	korongo	orusa	*
<i>Kiswahili gloss has three senses: (i) stork, crane; (ii) oryx (iii) gulley, ravine; R1 = gulley; Note that the English also has an adjective sense denoting an animal (esp. a horse) of mixed color, but the intended sense is probably oryx = choroa (Kiswahili). The Runyambo term should be enkórongo</i>			
sail	tanga	orufu/orumbe	*
<i>Kiswahili gloss has two senses: sail; and mourning period; R1 = mourning period.</i>			
shrew	kirukanjia	*	Omutingwa /malaya
<i>English senses: type of animal (mouse); bad tempered scolding woman.</i>			
<i>Kiswahili gloss senses: type of bird; prostitute; restless person. R2 senses: prostitute</i>			

TABLE 1: Editing the Runyambo Pilot Lexical Data

The notes in italics clarify the nature of the problem to be fixed by the data editor. The asterisk indicates a blank space in the returned questionnaire.

Three observations are in order. First, it will be important for the project to seek to establish a standard orthography for the languages of Tanzania and

train all researchers and research assistants to use that standard. Second, it may be more appropriate to give more prominence to word lists compiled from the local languages themselves. As the 'roan', 'sail' and 'shrew' examples show, working with three languages creates more problems of ambiguity and misinterpretation. The examples also show that the research assistants relied heavily on the Swahili column and ignored the English. Third, the marking of tone may become quite problematic as we approach the larger study because only a few researchers feel sufficiently confident to mark/capture tone. A possible solution may be to leave tone unmarked and only fill it in gradually for some of the languages as resources and personnel become available. It needs to be noted that no attempt was made to record the lexical and morphosyntactic data on tape. The decision was mainly based on a desire to collect a *large amount of data*, the audio recording and processing of which would pose extra challenges. It has subsequently been suggested that a researcher who is able to capture the relevant phonetic details should collect the morphosyntactic data; this requirement will further narrow the pool of personnel available to the project.

The lexical data for each language will be published in two parts: English - Ethnic Community Language (ECL)-Swahili and ECL-Swahili-English. A sample page from the first part of the Runyambo data (English-Runyambo-Swahili wordlist) is shown below.

<b>English</b>	<b>Prefix Runyambo</b>	<b>Kiswahili</b>
abandon	ku- nága/nájirana	ku-tupa
abdomen	ibondo	tumbo
abdomen, lower/below the navel	e- cinye, a-mayása	tumbo chini ya kitovu
abound	kw- íjura msera	ku-jaa tele
abstain	ku- yéima	ku-jinyima
abundant	msera	tele
abuse	e- cijúmi	tusi
abuse	ku- júma	ku-tukana
accompany	ku- jenda hámo, ku- jendana/sagarana	kw-enda pamoja
accompany (someone)	ku- séndecereza/hérecera	ku-sindikiza
accomplish	ku- mara	ku-maliza
accuse	ku- twéjera	ku-shitaki
acidity/sourness	o- busaarizi	ukali
across, lie	ku- cíkama	ku-kingama
across, put	ku- cínjira/síiciriza	ku-kinga
act	ku- kóra	ku-tenda
adam's apple	e- cimirônko, a-kasonda búro	koromeo
add	ku- téranisa	ku-jumlisha
add to	kw- onjera	ku-ongeza
add up	ku- téranisa	ku-jumlisha
adder, puff	e- mpíri	kifutu/ pili
admire	kw- égomba, ku-ríjira	ku-penda/husudu
admit guilt	kw- iciriza	ku-kiri kosa
adult	o- muntu mukúru	mtu mzima

TABLE 2: Sample Page from the English-Runyambo-Kiswahili Wordlist

The construction of the second part with ECL entries as headwords involves considerable back engineering and a heavy clean-up job. For a smooth result it requires that the ECL gloss supplied for the original English list be a lexical item, not a phrase or a sentence. If it is a phrase or a sentence, the editorial

task involves the choice of appropriate headword and form of entry. Also, this back engineering will result in a number of items appearing more than once as headwords. The decision to be made here concerns the appropriate representation for polysemous and homophonous items. Consider this portion on *gwa* as an example:

	<i>Runyambo</i>	<i>Kiswahili</i>	<i>English</i>
	[BEFORE EDITING]		
ku	gwa	ku-anguka	fall
ku	gwa	ku-anguka	tumble
ku	gwa (enjúra...)	mvua kunyesha	to rain
ku	gwa e-cihuumúra; ku-kába	ku-zimia	faint
ku	gwa empihi	ku-vimbiwa	overeat
ku	gwa iraro/raruka/neepa	ku-pata wazimu	crazy, become
	[AFTER EDITING]		
ku	gwa	ku-anguka; mvua ku-nyesha	fall, tumble; to rain
ku	gwa e-cihuumúra	ku-zimia	faint
ku	gwa empihi	ku-vimbiwa	over-eat
ku	gwa iraro	ku-pata wazimu	become crazy

TABLE 3: Back-engineering the Runyambo Wordlist

The first three rows have been collapsed into one row. The other related items, *ku-kába*, *ku-neepa*, and *ku-raruka* have been removed and will appear as separate headwords elsewhere in the list. And *crazy* need not appear in initial position any longer. A sample edited page from the second part of the Runyambo data (Runyambo-Swahili-English wordlist) is shown in Table 4 below.

	<i>Runyambo</i>	<i>Kiswahili</i>	<i>English</i>
ku-	bóneka	ku-onekana	be seen
o-	búbi	ubaya	badness
o-	bubóyi	ukali, ugomvi	fierceness, bullishness
o-	búce	udogo	smallness
o-	bucúreezi	ukimya	quietness
o-	bucúuya	mafuta juu ya mtindi	cream
o-	buféera	upumbavu	stupidity
o-	búfu	hali ya kosa	condition of impurity or danger because of violation
o-	bufúra	ukarimu	generosity, kindness
o-	bugáje	chakula kibaya	stale/stale food
ku-	búgana	ku-kutana njiani	meet on path
o-	buganga	baruti	gunpowder
o-	bugoro	ugolo	snuff
o-	bugúfu	ufupi	shortness
	bugoteka	kusini	south
o-	bugwa izóoba	magharibi	west
o-	bugweigo	kamba za katani ambazo hazijasukwa	sisal hemp
o-	buhângo	ukubwa	bigness
o-	buhéesi	uhunzi	smith's trade
o-	buhere	upele	skin rash, scabies
o-	buhûnga	unga	flour (any type)
o-	bujenyi	sherehe/tafrija, arusi	celebration, wedding
o-	bujubi	uvuvi	fishing
o-	bujúne	huzuni	grief/sorrow
o-	bukáma	utemi	chieftdom
o-	bukóko	mafuta juu ya mtindi	cream
o-	bukoma	uchoyo	stinginess
o-	bukúru	rika	age-group
o-	mwinganiro	rika	age-group
e-	saano	unga	flour (any type)

TABLE 4: Sample Page from the Runyambo-Kiswahili-English Wordlist

The sample in Table 4 shows a relatively neat result, giving word for word equivalents after the back engineering. But this does not give the full picture. The select portions in Table 5 below show some of the long definitions that are required for some Runyambo headword entries:

	<i>Runyambo</i>	<i>Kiswahili</i>	<i>English</i>
ku-	bûnja	ku-tembeza kitu ili kinunuliwe	show something around in order to sell it
o-	busya	mtego wa shimo ambamo mnyama hutumbukia	pit trap in the ground into which an animal falls
	caabwêra	aina ya mti wa <i>omutóoma</i> gome lake hutumika kutengenezea aina ya nguo	type of <i>ómutóoma</i> tree used in making barkcloth
ku-	cánkuza	ku-tafuna kwa kutoa sauti kubwa; tembea kwa kutoa sauti katika majani makavu ya migomba	eat noisely; walk noisely over dry banana leaves
e-	cicwamukágo	mali itolewayo kwa wakwe watarajiwa ili kuvunja undugu wa damu na kuruhusu taratibu za ndoa ziendele	wealth to sever blood relation so as to unblock marriage proceedings
o-	bukurura	aina ya magugu yenye mbegu zinazong'ang'ania kwenye nguo/mwili wa mtu	type of weed whose seeds stick onto clothing and body

TABLE 5: Examples of Long Definitions in the Runyambo Wordlist

It will be observed that these long definitions are not typical in a wordlist that usually seeks to supply one-word equivalents across languages. Also it would be difficult to do another back engineering and select headwords for Swahili or English for the items exemplified in Table 5.

The size of the wordlist for each language will depend on the amount of additional vocabulary obtained over and above the basic list of 3400 items (not all of which will be filled for every language). It is hoped that some of these vocabulary lists will be expanded into dictionaries. In this connection it

has been suggested that the project needs to take a bolder approach to vocabulary compilation by abandoning the English word list elicitation method. Moe (*in this report*) maintains that it is possible to elicit massive lexical data, up to 15,000 items, in a workshop setting over a period of five to ten days. The workshop participants make use of a list of *semantic domains* and quickly record every item that pops in the mind. After the list has been compiled, the glossing follows, partly facilitated by the semantic domain definitions.

The strengths of this method lie in the perceived universality of the semantic domain categories and the speed with which individual lexical items can be generated. Instead of struggling to obtain a native equivalent to a specific lexical item in a foreign language, the workshop participants search for native words for *concepts* that are shared across languages. But even in this method, the semantic domains will need to be translated from English into Swahili, rather than reinventing the wheel by trying to draw up a list of semantic domains for each language. And the glossing of culture specific items cannot be any easier – indeed the larger the vocabulary, the more demanding the task will be.

### **3. Concluding Remarks**

This limited experience of dealing with data from only two languages has demonstrated the critical role of researchers with a good knowledge of the languages to be studied. Ideally there should be at least one linguist researcher-in-charge for each language under study. The researcher-in-charge will attest to the accuracy/correctness of the data on that language and be in a position to detect problems in the data. Such a person needs to work with a pool of language consultants (informants) who even if not linguists, have a

literate person's understanding of the organization of their language. These would be able to transcribe oral data from tape recordings, recognize and mark tone on a list of words, comfortably recognize and write the language using a seven-vowel system (where the language has one) – it being noted that literacy is generally in Swahili with a five-vowel system and no tone.

As language teachers and learners we love bilingual dictionaries because they are such a convenient shortcut in deciphering the unknown. And yet in the history of any language, the publication of a monolingual dictionary and a grammar are milestones signaling the development of a rich meta-language as well as the presence of a large enough literate audience. In language teaching, a comparable development is the ability to teach a particular language using the language itself. All materials in the textbook are in the same language. There are, of course, several considerations in the choice of language teaching method, but all other things being equal, it is a mark of a well-developed language that it is used to teach in itself and talk about itself. It appears to me that this project has a noble contribution to make towards the development of a linguistic meta-language both in lexicography and in grammar. And so the need to compile large wordlists from the perspective of the ethnic community languages, rather than from English, needs to be pursued with greater determination.

#### **References:**

Moe, Ron 2002. Massive Data Collection in a Workshop Setting. (*in this report*)

#### **DISCUSSION**

Many participants agreed that there was need to have a **standard orthography** to cater for all languages in Tanzania. However, care needs to

be exercised to avoid the multiplication of orthographies, especially in the case of languages spoken across borders. Also appropriate modifications to the lexical list should be made in order to remove the many errors noted, especially with respect to the Swahili translations.

## SUMMARY OF GROUP DISCUSSION ON LEXICAL DATA

Prof. J. Mdee

The Institute of Kiswahili Research

### **1. Pilot Study**

The approach used in the pilot study was appropriate but needed some improvements. It was agreed that it is proper for some work to be done in Dar es Salaam first before going to the field. Improvements can best be effected by panel discussions.

#### **1.1 Compilation of the word list**

In order to ensure that no word is left out, the wordlist should be developed using the semantic domains approach first and later on it may be arranged alphabetically, when the data is processed for the eventual production of a glossary or dictionary.

#### **1.2 Swahili equivalents**

A panel should go over the definitions in the Kiswahili column and provide appropriate forms after thorough discussion.

#### **1.3 Community language definitions**

A panel of competent native speakers should provide the appropriate definitions in the relevant community language.

#### **1.4 Completion and verification of data**

The procedure whereby some work was done in Dar es Salaam and completed in the relevant research area by filling gaps and verifying the data was found to be convenient.

### **1.5 Questionnaire efficiency**

The questionnaires were relevant and useful but they were not as efficient as they should be. The cases of ambiguity and misinterpretation (e.g. kamba/lobster) require a tighter procedure to discern them. A semantic domains approach should be helpful in this regard. The quality of the data should be improved by means of audio equipment to capture accurate phonological information. Phonological information should not be ignored. As for quantity, the minimum target of 3400 is rather small. The project should attempt to collect as much information as possible.

### **1.6 Data entry**

Appropriate lexicographic software, such as shoebox, should be sought to process the data. Arrangement should be put in place to develop/adopt appropriate orthographic conventions for the project. Due attention should be paid to existing conventions in particular languages; but also the aim should be to develop conventions that would cut across individual languages and establish a standard for the languages of Tanzania.

## **2. Proposed Programme of Action**

The programme of action was considered good and implementable. It was noted that the Principal Researchers should be present in the field with their assistants and effectively supervise them. It is also desirable that data be collected from every dialect, as well as from urban and the most remote rural areas.

# MORPHOSYNTACTIC DATA COLLECTION AND ANALYSIS

Prof. Massamba D. P. B.

Institute of Kiswahili Research

## 1. Introduction

The first workshop of The Languages of Tanzania Project (LOT) was held from July 6-7<sup>th</sup>, 2000. The major theme of the workshop was “Getting Ready for Research”, the main objective of which was to make preparations for data collection in the field. The discussions focused on ways and means of formulating appropriate methodologies for collecting linguistic data (lexical, phonological, morphological, morphosyntactic and atlas data) from the different languages of Tanzania.

With regard to the collection of morphosyntactic data a number of issues were raised. These included:

- (a) identification of essential areas to be covered by a questionnaire for morphosyntactic data collection.
- (b) availing ourselves with what materials already exist in this area.
- (c) how much data should be collected in various forms.
- (d) whether or not an equal amount of data should be collected for all languages.
- (e) how to handle the collected data.
- (f) what are the most likely problems in the collection and analysis of morphosyntactic data, etc.

Taking into consideration what transpired during the discussion on this particular aspect, it was deemed more appropriate to focus on the questionnaire approach, in the collection of morphosyntactic data, following

closely the procedure of introspection (cf. Saville-Troike, 1982: 119-20; Kihore, 2000: 31-36).

After the workshop the principal researchers prepared a questionnaire for morphosyntactic data collection that was to be used during the Pilot Study. What I intend to do here is present to this workshop how the questionnaire was designed in order that we may make comments, if any, for improvement.

## **2. The questionnaire**

In the said workshop, one of the issues that were emphasized in designing the questionnaire was to make sure that it was elaborate enough to give as much information as possible. In order to achieve this goal, it was necessary to design a questionnaire that would require the respondents to give specific answers to questions on various aspects of language and culture. This consequently meant that the questionnaire had to be designed in the form of sentences. The questionnaire developed earlier by Prof. Batibo was adopted.

Since the Pilot Study was going to involve Bantu languages per se the questionnaire concentrated on morphological characteristics of Bantu languages. The main aspects that were included in the questionnaire were *nominal prefixes (viambishi awali vya nomino)*, *verb forms (maumbo ya vitenzi)* and *sentential constructions*. Let us examine these aspects very briefly using examples from Ruhaya, and making specific reference to Ruhyoza.

### **2.1 Nominal Prefixes**

In nominal prefixes the following categories were included:

### 2.1.1 Classes and agreements (Ngeli na upatanishi wa kisarufi)

The questionnaire had 42 constructions of different noun classes and their grammatical agreements. Here are a few examples:

- (a) This person is tall (mtu huyu ni mrefu)  
*Omuntu ogu muraimurai*
- (b) These person are tall (Watu hawa ni warefu)  
*Abantu aba baraibarai*
- (c) This egg is small (Yai hili ni dogo)  
*Eihuli eli like*
- (d) These eggs are small (Mayai haya ni madogo)  
*Amahuli aga gake*
- (e) This small goat is beautiful (Kijibuzi hiki ni kizuri)  
*Akabuzi aka karungi*
- (f) These small goats are beautiful (Vijibuzi hivi ni vizuri)  
*Obubuzi obu burungi.*
- (g) Drinking beer is good (Kunywa bia ni kuzuri)  
*Okunywa amaarwa kurungi*
- (h) On the wall there is a rope (Ukutani (yaani, juu ya ukuta) pana kamba)  
*A'rukuta aliwo omuguha*

### 2.1.2 Other Concordial Agreements and Demonstratives

There were 25 constructions in this category. These included the following examples:

- (a) The one child who is here is mine (Mtoto mmoja aliye hapa ni wangu)  
*Omwaana (omoi) ali aha / hanunju wange*
- (b) The two children who are here are mine (Watoto wawili walio hapa ni wangu) *Abaana babili / ababili abali aha bange.*

- (c) The one egg, which is yonder, is mine (Yai moja lililo kule ni langu)  
*Eihuli (limoi) elili kuliinya lyange.*
- (d) The two eggs, which are yonder, are mine (Mayai mawili yaliyoko kule ni yangu) *Amahuli gabili agali kuliinya gange.*
- (e) The two small babies that are there are mine (Vitoto viwili vilivyopo hapo ni vyangu) *Enkeremeke ibili / eibili ezili aliinya zange.*

### 2.1.3 Numeral agreements

There were 27 constructions in this category, the examples of which are as follows:

- (a) One person (Mtu mmoja)  
*Omuntu omoi*
- (b) Two persons (Watu wawili)  
*Abantu babili.*
- (c) Three persons (Watu watatu)  
*Abantu bashatu.*
- (d) Four persons (Watu wanne)  
*Abantu banai.*
- (e) Eleven persons (Watu kumi na mmoja)  
*Abantu ikumi na omoi.*

### 2.1.4 Personal Pronouns (Viwakilishi Nafsi)

Six constructions involving personal pronouns were provided as follows:

- (a) As for me, I am cultivating my farm (Mimi ninalima shamba langu)  
*Inye nindima ekibanja kyange / endimiro yange.*
- (b) As for you (sg), you are cultivating your farm (Wewe unalima shamba lako) *Iwe nolima endimiro yawe.*
- (c) As for him he is cultivating his farm (Yeye analima shamba lake)  
*Wenene nalima endimiro ya wenene.*
- (d) As for us, we are cultivating our farms (Sisi tunalima mashamba yetu)  
*Ichwe nitulima endimiro yaitu.*

- (e) As for you (pl.), you are cultivating your farms (Nyinyi mnalima mashamba yenu) *Inywe nimulima endimiro zanyu.*
- (f) As for them, they are cultivating their farms (Wao wanalima mashamba yao) *Bonene nibalima endimiro zabo.*

### 2.1.5 Verbal Prefixes and Genitive Forms

The questionnaire had 22 constructions of this category. Here are a few examples:

- (a) The child of the stranger is asleep (Mtoto wa mgeni amelala)  
*Omwana wa omugenyi anagiire.*
- (b) The tree of the stranger has fallen (Mti wa mgeni umeanguka)  
*Omuti gwa omugenyi gwakumba / gwagwa.*
- (c) The small baby of the stranger is crying (Kitoto cha mgeni kinalia) *Enkeremeke ya omugenyi nelila.*
- (d) The fingers of the stranger are broken (Vidole vya mgeni vimevunjika).  
*Ebyara bya omugenyi byahendeka.*
- (e) The inside of the house of the stranger is broken (Ndani ya nyumba ya mgeni pamebomoka) *Omunju ya omugenyi habomoka / hakumba.*

### 2.1.6 Object Substitute (Vitendwa Viwakilishi)

The questionnaire had 25 constructions in this category. Observe the following examples:

- (a) He found me there (Alinikuta huko)  
*Anshangile okwo / akashanga okwo / anshangileyo.*
- (b) He found you (sg) there (Alikukuta huko)  
*Akushangile okwo / akakushanga okwo*
- (c) He found them (people) there (Aliwakuta (watu) huko)

*Akabashanga okwo/abashangire okwo*

- (d) He found the (trees) there (Aliikuta (miti) huko)

*Akagishanga (emiti) okwo/ agishangire okwo*

- (e) He found it (the house) there (Aliikuta (nyumba) huko).

*Akagishanga (enju) okwo/agishangire okwo*

- (f) He found it (the inside of the house) good (Alipaona (ndani ya nyumba) pazuri) *Akaabona (omuunda ya enju) harungi.*

## **2.2 Verb Forms (Vipashio vya vitenzi)**

The following categories of verb forms were included:

### **2.2.1 Tense System (Mfumo wa Nyakati)**

This category had 36 constructions, the examples of which are:

- (a) I cultivate my farm everyday (Hulima shamba langu kila siku)

*Inye ndima ekibanja kyange buli kiro*

- (b) I do not cultivate my farm often (Silimi shamba langu mara nyingi)

*Tindima kibanja kyange buli kiro.*

- (c) I am cultivating my farm now (Ninalima shamba langu sasa).

*Nindima ekibanja kyange kaanya aka / mwaha.*

- (d) I have just cultivated my farm (Ndiyo tu nimelima shamba langu).

*Nalima ekibanja kyange.*

- (e) I cultivated my farm yesterday (Nimelima shamba langu jana)

*Ndimile ekibanja kyange nyeigoro.*

- (f) I cultivated my farm last week (Nimelima shamba langu wiki jana)

*Nkalima ekibanja kyange ilimansi elioire.*

- (g) I shall cultivate my farm tomorrow (Nitalima shamba langu kesho)

*Ninyija kulima/ndaalima ekibanja kyange nyenkyai.*

### 2.2.2 Verb Suffixes/Extensions (Viambishi Tamati)

This category had 21 constructions. The following are the examples:

- (a) He is cooking food for me (Ananipikia chakula) *Nanchumbira ekyakulya.*
- (b) He is cutting the rope with a knife (Anakatia kamba kisu) *Nashara omuguha na omuhyo/nashaza omuguha omuhyo.*
- (c) He is farming with the hoe (Analimia jembe) *Nalimisa enfuka.*
- (d) He is opening the door (Anafungua mlango) *Nakingura orwigi.*
- (e) He is closing the door (Anafungua mlango) *Nakinga orwigi.*
- (d) The children are walking together (Watoto wanatembea pamoja) *Abaana nibakyara/nibatambuka hamoi.*

### 2.2.3 Derivations/Deverbatives (Minyambuliko ya Nomino)

This category had 14 constructions. The examples are:

- (a) To cultivate (Kulima) *kulima (okulima?)*
- (b) Farmer (Mkulima) *Omulimi.*
- (c) Cultivated land (Shamba lililolimwa) *Obulime/endimiro*
- (d) Cultivating tool (Zana ya kulimia) *Ekikwaato kya eilima/ekilimiso.*

### 2.2.4 Derivative/Verbalization (Minyambuliko ya Vitenzi)

- (a) To become small (Kuwa mdogo) *Okukeeha*
- (b) To become tall (Kurefuka) *Okuraiha*
- (c) To become large (kuwa mkubwa) *Okuhanguha.*
- (d) To become dark (Kuwa mweusi) *Okwilagula*

### 2.2.5 Locative Enclitics (Vishikizi-Mahali)

This category too had 8 constructions, the examples of which are:

- (a) He is entering into the house (Anaingia nyumbani) *Nataaha omunju*
- (b) He is entering into it (Anaingiamo) *Nataahamu /nagitaahamu.*
- (c) He is going to the house (Anakwenda nyumbani) *Nagenda omunju.*
- (c) He is standing on it (Anasimamako) *Ayemereireho.*
- (d) He is passing through the farm (Anapitia shambani) *Naraba omu ekibanja/naguturana ekibanja.*

### 2.3 Sentence (Sentensi)

These were divided into two major categories, namely Phrase Distributions and Subordination.

#### 2.3.1 Phrase Distributions (Mpangilio wa Virai)

This category has 10 constructions. The examples are as follows:

- (a) The woman is cooking food for her children (Mwanamke anawapikia watoto wake chakula) *Omukazi nachumbira abaana ba wenene ekyakulya.*
- (b) The woman is cooking food in the kitchen (Mwanamke anapikia chakula jikoni) *Omukazi nachuumbira ekyakulya omu ichumbiro.*
- (c) The hoe is being used to cultivate the farm (Jembe linalimiwa shamba) *Enfuka nelimisibwa omundimiro.*
- (d) At home, there are visitors (Nyumbani kuna wageni) *Omuka aliyo abagenyi.*
- (e) Masanja has cut his finger with a razor (Masanja amejikata kidole kwa wembe) *Masanja yayeshara ekyara na akajilita.*

### 2.3.2 Subordination (Utegemezaji)

This category had a total of 10 constructions. The examples are:

- (a) If you come you will find me (Ukija utanikuta) *Kolaija oranshangao.*
- (b) If you had come, you would not have found me (Kama ungekuja usingenikuta) *Kuba waizire tiwakunshangireo.*
- (c) Since you are late, I will not talk to you (Kwa jinsi umechelewa sitaongea nawe) *Okwo wakeerererwa, tinfumoole na iwe/orwo wakeerererwa tindikwija kufumooru na iwe.*
- (c) Although she is beautiful, she is not married (Ingawa ni mzuri wa sura bado hajaolewa) *Araba ali murungu takashwerwaga/norwo ali murungi takashwerwaga.*

### 3. Observations and Recommendations

The questionnaire used had a total of 256 constructions divided into three major categories, namely *Nominal Prefixes*, *Verb Forms* and *Sentence Patterns*. The questionnaire has two major positive qualities. First, in the way it has been designed, it does give a good picture of Bantu morphological characteristics. The constructions provided, especially the *Nominal Prefixes* and *Verb Forms* give a lot of details with regard to Bantu morphology.

The second quality, which is very important, and which comes out quite clearly, is its cultural neutrality. By this I mean to say the materials contained in this questionnaire are not culturally bound; they can be used for collecting data in most linguistic communities in Tanzania. In my view, this questionnaire should form the core of all other questionnaires that are (linguistic) community-sensitive. In other words, the questionnaire should form the basis on which the culturally or geographically bound questionnaires should be developed.

The above major positive qualities notwithstanding, it seems to me that the questionnaire needs to be improved. As pointed out earlier, the questionnaire seems to give more weighting to morphological information. However, since it is the same collected data that will also be used in writing some grammars, we need to ask ourselves whether it would not buy us more by adding more syntactic information than is currently provided for by the questionnaire.

One of the additional issues we may wish to consider, for example, is the question of *Word Order*. It might prove profitable to have data that will provide us with information on word order, i.e. what is the nature of word order in the language we are studying? Is it *Free Word Order* or is it *Fixed Word Order*. If it is free how free is it? Another issue, which we also might wish to consider, is the inclusion of *Passive Constructions* and other sentence types. Of course these have to be simple constructions.

In languages that use pre-prefixes, it seems to me it would be more instructive to provide cases in which we find nouns with and without pre-prefixes. The present questionnaire does not provide room for cases of construction that have nouns without pre-prefixes. All the Ruhaya nouns given in this data have pre-prefixes. Since nouns can occur without pre-prefixes it would be important to see in what sorts of environments they do appear.

Finally, there is also the question of morphological parsing. The Ruhaya data provided has no morphological parsing. We know for sure that when it comes to data analysis morphological information plays a crucial role. It must, therefore, be provided. Of course this does not mean that this work should

necessarily be done during the fieldwork. Whatever the case may be, this requires both the native speaker of the language in question and the linguist.

#### **4. Conclusion**

The aim of this paper has been to present the morphosyntactic questionnaire as was used in the Pilot Study carried out in Kagera Region, to point out its merits and demerits and, finally, put forward suggestions for improving it. The suggestions put forward are, of course, open to discussion.

#### **DISCUSSION**

After a long discussion, a number of problems were noted in the questionnaire used for data collection and the methods employed:

- (a) Aspects not covered in the questionnaire include adverbs, conjunctions, interrogatives, onomatopoeic items, and idiophones.
- (b) Kiswahili was taken as a model. The issue here is, how far would Kiswahili reflect the complexity of the morphosyntactic structures of other Bantu and non-Bantu languages, e.g. cases like nouns, verbs, verbal extensions etc.
- (c) The research assistants used in the data collection process were not linguists.
- (d) The phonological aspect was not dealt with in the questionnaire; so it is a bit difficult to deal with morphophonological processes of the languages under research.
- (e) Intonation is not covered.
- (f) Some silly sentences need to be refined or dropped.

## **SUMMARY OF GROUP DISCUSSION ON MORPHOSYNTAX**

Prof. H.M. Batibo

University of Botswana

### **1. Introduction**

The Group on Morphosyntax was assigned the task of considering the different points in the Guideline Questions that were prepared by the Secretariat to guide the three Groups in their evaluation of the data obtained from the pilot survey conducted in September 2001. The Group on Morphosyntax deliberated the various points. Its first remarks were as follows:

- (a) that it was important that the objectives of the morphosyntactic study be made clear. Essentially the study should aim at collecting mainly morphological and syntactic data that would enable the team to prepare grammatical sketches for each of the Tanzanian languages. So as to have a complete grammatical study, aspects of phonology and morphophonology should be included in the description.
- (b) that the members of the Group should go through the Questionnaire and the different responses so as to make an appraisal of the quality and completeness of the outcome of the pilot study.

## **2. Results of the pilot study**

The various guiding questions were considered in relation to the quality and completeness of the pilot study results. The observations of the Group on each of the questions were as follows:

### **2.1 *Whether it was proper that the researchers should start in Dar es Salaam and only go to the field to complete the work and verify the accuracy of the data.***

The group found the procedure to be proper. In fact, the group went further to propose that the principal researchers should be the ones to collect the data from the field so as to ensure the accuracy of the data and the relevance of the data to the questions asked. The researchers should be linguists with training in research methods. In this case, they can start with data collection in Dar es Salaam, particularly where there were competent speakers of the relevant languages. The fieldwork in this case could concentrate on filling gaps, verifying doubtful cases and studying languages not found in Dar es Salaam. This would not only save time but also reduce costs.

### **2.2 *How usefully, relevantly or appropriately designed were the research tools***

It was observed that the research tools were generally well designed in that they captured most of the morphosyntactic aspects of the languages. However, some linguistic aspects were missing or could not be elicited by the questionnaires. These included:

- (a) Several of the grammatical categories like adverbs, onomatopoeia/ideophones and interjections. Moreover, there were only a few conjunctions which were captured by the questions in the questionnaire.
- (b) Some basic transformational rules such as passivisation, commands and topicalization were not included. Moreover, only a few interrogative cases were captured.
- (c) No allowance was made for any special peculiarities in each of the languages studied.

### 2.3 *The usefulness and relevance of the data collected*

It was agreed that generally the data collected from the pilot study was useful and relevant. However, the Group made the following observations:

- (a) There were some noticeable inconsistencies in the manner in which some of the questions were interpreted.
- (b) There were a number of errors and inaccuracies in some of the data.
- (c) Much of the phonological and morphophonological information was lost as the data were recorded through the conventional orthography.
- (d) Some of the responses were not relevant to the information intended, such as the case of Question No. 210 in the Questionnaire [To become tall – *Kurefuka*].

### 2.4 *The Difficulties of computing the data*

It was acknowledged that the work of data computing and processing should ensure that the end product was computer- readable. This was

an important aspect of the project. The following was therefore recommended:

- (a) that the computers to be used in the project should be powerful enough to handle all the data that will be collected in the project;
- (b) that the software should be appropriate for the various processing operations needed;
- (c) that the researchers should be trained in data handling and computing. In fact, MA. students and other language graduates should be involved in this work, NOT secretaries.
- (d) since the data collection exercise will be done by linguists, the data should be recorded phonetically (even broadly) and entered in the computer when the phonological and morphophonological analyses have been carried out. In this case, software for analyzing or recording tone, such as the one used in Helsinki, should be sought. Also a phonetician such as Prof. E. Elderkin or Prof. J. Maghway could be asked to organize special training seminars/workshops.

2.5 *Whether there is any method or principle that could have been included*

There is no other method or principle that could have been used, as all the data collection exercise was done systematically. However, as mentioned above, the data ought to include phonological and morphophonological aspects. Also, since the grammatical sketches are expected to be databased rather than theory-oriented, the analysis and interpretation of the data should be basically descriptive. The other aspects that ought to be considered are:

- (a) the idiomatic expressions in each language;

- (b) the major discourse markers and devices in the language;
- (c) transcribed oral texts (short stories), as an appendix to show how the language is used in practice.

**2.6 *Whether there was any principle or method, which should be avoided***

Generally, there was no principle or method that ought to be avoided. However, it was recommended that the following should be avoided:

- (a) any obscene or publicly offensive words, expressions or sentences;
- (b) any expressions or sentences which would interfere with certain articulated principles on gender, child abuse or human rights;
- (c) any expressions or sentences which would make the informants uneasy, uncomfortable or irritated.

It was observed that if the researchers wanted these words/expressions, they could get them by using less public-based strategies.

**3. Proposed Programme of Action**

**3.1 *The suitability of the team***

The group considered the general programme/plan of action that was presented in the plenary session in relation to the morphosyntactic study. It was generally agreed that the plan was proper in terms of timing and scheduling of activities. However, the following was observed:

- (a) It was important to know right from the outset the number of zones that will be established so as to know how to plan for the time and use of personnel and other resources.

- (b) There was a need for the researchers to make self-evaluation, in terms of time, availability and expertise, in order to make strategic work planning, time scheduling and quality assurance of output.
- (c) There was a need to emphasize the fact that only linguists should be involved in the data collection exercise so as to ensure the accuracy of the data collected.
- (d) There was a need to clarify the payment of money to research assistants and informants. This should be based on the general practice of the University.

### **3.2 *Any ways or strategies to improve the plan***

There are no other ways to improve the plan except that the following should be observed;

- (a) that linguists be responsible for the data collection;
- (b) that literature review on all the languages be carried out to know about the existence of documentation on each of the languages. The work that is going on at the University of Gothenburg (Sweden) should be consulted as a starting point.
- (c) the processing of the data should not wait until all the fieldwork is completed. It should be an on-going activity.
- (d) every effort should be made to publish the data that will have been analyzed into grammars and dictionaries.

### **3.3 *Any alternatives?***

No alternatives were suggested, as the work-plan was found to be appropriate.

#### **4. The Project in General**

##### **4.1 *Anticipated Problems***

It was observed that, generally, there would not be any problems. The only caution was in the planning of *time* to ensure that the work was done according to schedule and within reasonable time limits. Also the commitment of the researchers was important, particularly in the areas of data collection, analysis, computing, writing of the grammars and compilation of the dictionaries. The researchers were therefore, urged to be highly committed and dedicated to the project.

##### **4.2 *Recommendation on the methodological approach and project set-up in general.***

There was no other recommendation made, except that it was, once again, emphasized that the researchers should be committed, dedicated and persevering. It was also stressed that they should use the results of the pilot data to make some publications, such as academic articles.

# THE LANGUAGE ATLAS

Prof. C.M. Rubagumya

Foreign Languages & Linguistics

## 1. What kind of Atlas?

At the last workshop several levels of maps were identified:

- A countrywide map of all languages in Tanzania.
- Regional maps showing major language groups (i.e. Bantu, Nilotic, Cushitic, and Khoisan) and their sub-divisions.
- Regional/district maps showing specific languages and their appropriate locations.
- Maps showing languages and their respective dialects.

## 2. Data already available

- Maps produced by The Summer Institute of Linguistics (SIL) will be an important input, but they have to be appraised for accuracy by comparing them to data collected during our fieldwork.

*(Some errors have already been spotted: e.g. Kinyanwanda in Western Karagwe; Luganda in Northern Bukoba District.*

- Maps of administrative Districts (where available) would also be useful for this exercise.
- Census data (1988) are already available. Estimations of population based on these data can be made. Thus, if we know that in a certain district language 'X' is spoken, we can estimate how many speakers that language has.

### **3. Getting new data**

- A questionnaire for mapping data will be designed.
- Information for mapping data can be obtained in Dar es Salaam, then at district level, and then at division/ward levels.

### **4. Naming system**

- Languages will be named according to what speakers of those languages call them. Where there are several names, an explanation will be necessary. Offensive names should be avoided.

### **5. Drawing Maps**

- Maps will be drawn as data from each linguistic zone are made available from fieldwork. So the process will amount to a step-by-step building of the national atlas.
- Local facilities (e.g. the cartographic units of IRA and Department of Geography at UDSM) will be used as far as possible.
- Where local facilities are not available, foreign assistance/collaboration will be sought. Collaboration with SIL will be given due consideration.

**APPENDIX**  
**HALMASHAURI YA WILAYA - BUKOBA**  
**ORODHA YA TARAFU, KATA, VIJJI NA LUGHA**

<b>TARAFU</b>	<b>KATA</b>	<b>KIJI</b>	<b>LUGHA</b>
<b>K A T E R E R O</b>	Ibwera	Ibwera	Ruhamba
		Kibona	"
		Karonge	"
		Itongo	"
	Katerero	Kanazi	"
		Rwagati	"
		Mutahya	"
		Kyema	"
	Bujugo	Katoju	Ruhyoza + Ruhamba
		Buganguzi	" "
		Minazi	" "
	Kasharu	Kashule	Ruhamba
		Kasharu	"
		Ntoija	"
		Kabajuga	"
		Butainanwa	"
	Kaibanja	Kaibanja	"
		Nyakigando	"
		Kiinjogo	"
		Kazinga	"
	Katoro	Katoro	"
		Musira	"
		Ngarama	"
	Mikoni	Mikoni	"
		Kahyoro	"
		Rutete	"

		Kagondo	“
	Kyamula	Kyamulaile	“
	-ile	Mashule	“
		Omukihisi	“
		Nyakibimbili	“
		Kitahya	“
		Bugengere	“
		Bundaza	“

## 6. DISCUSSION

In the discussion session, the participants suggested that the researchers should try their best to be precise on the kind of atlas they intend to draw at different levels.

In addition to that, the participants proposed that the maps should include some basic information like names of languages/dialects, origins, vitality, current population figures, and mark some physical features: i.e., rivers etc. However, one pre-caution was that the addition of physical features into the map should not bring in the problem of readability. On the problem of cross border languages, the participants suggested that the researchers should not go beyond the borders i.e., the case of Ganda/Haya.

Lastly, the discussants expressed their concern on the failure to get a language question into the coming population census. The participants noted that the official position contradicts the spirit of the official cultural policy (Tanzania Government, 1997, *Sera ya Utamaduni*. Government Printer).

## SUMMARY OF GROUP DISCUSSION ON THE LANGUAGE ATLAS

Prof. C. M. Rubagumya

Department of Foreign Languages and Linguistics

During the group and plenary discussions, a number of ideas were put forward.

### 1. **Sociolinguistic surveys**

SIL experience shows that sociolinguistic surveys can be used to elicit information about where different languages are used in a given area. Questions can be asked about the heartland of the language, about which village speaks a different language, etc. It was suggested that traditional leaders and 'outsiders' who have lived in an area for a long time could be good sources of information on which language is spoken where. Once enough information has been obtained, the Global Positioning System (GPS) can be used to determine the exact location where the language in question is spoken.

A list of villages and languages spoken in those villages can be fairly accurate if we collect information systematically and accurately. However, as language boundaries are no neat and clear-cut, some kind of arbitrariness and approximation is inevitable in determining them. This is also true of dialect boundaries. It was emphasized that the question of language/dialect distinction is to be addressed at the data analysis stage, not before.

## **2. Number of speakers**

With available census data (1967, 1988), it is possible to make reasonable estimates about the number of speakers for each language. If we know that in village 'A' language 'X' is predominant and village 'A' has 2000 people, we can say that this is the approximate number of speakers of language 'X' in this village. If this is done for all villages where 'X' is predominant, then we get an approximate number of all the speakers of this language.

It was noted that these figures would not account for about 25-30% of the population of Tanzania, who live in linguistically mixed urban areas. This means the estimates will be based only on geographical 'heartland' areas of the different languages.

## **3. Language vitality**

It was suggested that a questionnaire be constructed to capture the linguistic vitality (or lack of it) of the different languages. The kind of information expected from this questionnaire will include:

- a. Geographical position (District, Ward)
- b. Ethnic identity and mother tongue
- c. Whether the respondent can read and write in the mother-tongue (e.g. books, newspapers, etc.).
- d. Whether the respondents' children can speak the mother-tongue.
- e. In which domains the mother-tongue is used.
- f. How many members of the respondent's family live in town permanently.
- g. The main economic activity of the respondent's village.

This information can be used to determine whether there is a trend of language shift or language maintenance for each of the languages surveyed and the extent of language shift. Information on language shift and the number of speakers may be a good indicator of whether a given language is in danger of dying or not.

#### **4. Output**

It was agreed that the output of the mapping information will be:

- (a) language maps for each linguistic zone (e.g. the lake zone for phase one of the research project).
- (b) the language atlas for the whole country at the end of the project.
- (c) approximate number of speakers for each language.
- (d) information on the linguistic vitality of each language (e.g. where they came from, if they spoke a different language before, etc.)

## **1 INTRODUCTION**

We are glad that the first year of the project has successfully come to the end. Virtually all of our objectives set out for Year One have been achieved, despite the delay in starting the activities. We are three months behind schedule because funds could not be released in time.

A pilot study was conducted in Kagera region in Year One (2001) and data were collected and entered into the computer as reported by Kahigi, Rugemalira, Massamba, and Rubagumya, in this report. As outlined in those papers, the pilot study is a springboard for the coming phase of data collection. After considering the methods used, the results achieved, and the problems encountered in the pilot study in relation to the project proposal and recommendations given in the first workshop, some modifications have been made on how the first phase of data collection should be conducted.

In principle, there will not be major changes on the work plan of the project proposal (as outlined in §4.2 and 8.0) in terms of the geographical area to be dealt with. Thus, there will be three main goals for the second year: (i) to study the languages around Lake Victoria, (ii) to collect as much data as possible, and (iii) to use the method that will be most cost effective without undermining the validity of the study.

## **2 MAJOR APPROACH**

## 2.1 Geographical zones

In order to collect data in a systematic way, the major approach of the project will be to divide the country into area zones. The formation of the zones will be mainly based on the geographical location of the languages as well as their genetic affiliation. A tentative list of the proposed zones (with their respective regions or part thereof indicated in brackets) includes:

- |    |               |   |
|----|---------------|---|
| a. | Lake Zone     | Kagera, Kigoma, Mara, and Mwanza;             |
| b. | Northern Zone | Arusha, Manyara, Kilimanjaro, Tanga, Singida; |
| c. | Western Zone  | Shinyanga, Tabora, Mbeya, Rukwa;              |
| d. | Eastern Zone  | Dodoma, Morogoro, Coast, Dar;                 |
| e. | Southern Zone | Iringa, Mtwara, Ruvuma, Lindi.                |

Each zone will then be organised into language groups. The creation of the language groups will take into consideration the existing genetic classification of the languages/dialects known.

## 2.2 Phase One

The first three years of the project (i.e. Phase One: 2001–2003) focus on languages around Lake Victoria. Thus, Year Two (2002) will deal with data collection from the Lake Zone, namely the regions of Kagera, Mara, Mwanza, and Kigoma.

The zone will be analysed into seven language groups, each of which will be coordinated by a linguist referred to as *Principal Researcher* (PR).

- |    |                    |   |
|----|--------------------|---|
| a. | Rutara I:          | Runyambo, Rubumbiro, Ruzinza, etc.      |
| b. | Rutara II:         | Ruhaya, Kisubi, Kikerebe, etc.          |
| c. | Western Highlands: | Kihangaza, Kishubi, Kiha, Kivinza, etc. |
| d. | Suguti:            | Kijita, Kikwaya, Ciruri, Kiregi, etc.   |

- e. South Mara: Kishashi, Kiikizu, Kizanaki, Kinata, Kingorimi, etc.
- f. North Mara: Kikuria, Kihacha, Kirieri, Kisuba, etc.
- g. Western Tanzania: Kisukuma, Sisumbwa, Kinyamwezi, etc.
- h. Nilotic: Luo

The principal researchers will work under the coordinatorship of the heads of the three sections, namely Lexicon, Morphosyntax, and Atlas/Mapping. Each section will thus be under *Section Head* (SH) (see attached managerial structure chart)

### **2.3 Advantages of this approach**

The proposed approach has a number of advantages including the following. One, it ensures that all the three components (lexicon, morphosyntax, and atlas/mapping) of the project are covered at each stage. Two, it is systematic in that it covers one geographical area after another, rather than selecting patches here and there. Three, it concentrates efforts and resources on one area, hence increasing the rate of reliability on the data collected. Four, it deals with languages that are most likely closely related, hence making it easy in terms of data processing and analysis.

However, there are also a few disadvantages of this approach. First, some areas are dominated by languages that are not closely related, and sometimes not even genetically related. This could make it difficult in terms of supervision, data processing and data analysis. Second, some researchers may not be comfortable to work on a certain area, and thus keep waiting for other phases.

### **3 METHODOLOGICAL STRATEGIES**

A number of strategies are envisaged to ensure that large data is collected in time and at reasonable cost. The following are some of the steps and strategies that will be applied.

First, the project coordinators (PCs) will generate a database of all prospective language informants available in Dar es Salaam, especially at UDSM. Some of these informants will later serve as research assistants. In principle, these should be individuals who are well informed on the language, its geographical location and speakers.

Second, the PCs, in consultation with the Project Management Committee (PMC) will identify three Section Heads (SHs) and seven Principal Researchers (PRs) who will coordinate the three language sections and seven language groups, respectively. Their selection will be mainly based on the information collected from the forms distributed and returned to LOT (November 2001). The forms sought to identify linguistics researchers who are readily available for the project.

The roles of the Section Heads, who will report directly to PCs, will include the following:

- a. Ensuring that all necessary tools, relevant to their sections, are ready before starting data collection.
- b. Advising the PCs on identifying appropriate PRs.
- c. Checking, proofreading, reviewing, and editing research tools to ensure that they are appropriately prepared and in good order before they are used to collect data.

- d. Collaborating with the PRs across all languages under study in data collection exercises.
- e. Writing a report on the section under their supervision.

Beside these roles, each SH will have to select a language on which they will produce a grammar, dictionary or lexical list. The aim is to ensure that we have individuals committed to project outputs that will enable us to achieve the project goals as stipulated in the proposal.

The roles of the Principal Researchers will include:

- a. Selecting appropriate RAs for the language(s) under their supervision.
- b. Supervising RAs in the data collection workshops.
- c. Suggest any appropriate strategy or methodology for improvement of data collection, analysis, and editing.
- d. Writing a report on the language(s) under their supervision.

Each PR will report matters relevant to lexicon, morphosyntax, and atlas to respective SH, and general matters to the PCs. Apart from these roles, each PR will have to select a language or dialect on which they will produce a grammar, a dictionary, or lexical list.

Third, the PCs, in collaboration with the SHs and PRs, will organise a series of workshops in Dar es Salaam for data collection. Informants from different languages will be invited to data collection workshops where they will provide data according to the tools and methodology that will be adopted by the research team. The research team will deal with one language group at a time. The major activities of the workshops will include:

- a. Listing all the languages and dialects in the Lake Zone;
- b. Completing relevant questionnaire(s);
- c. Brainstorming on any other relevant matters concerning the language, such as other names of the language, dialects, key resource persons in rural areas, etc.

Fourth, once the data have been collected, a few research assistants and data entry clerks, under close supervision of the Section Heads, Principal Researchers, and Project Coordinators, will enter them into the computer. The SHs, PRs, and PCs will be responsible for the good results that include proofreading, correcting, and editing the data, as well as crosschecking (with other sources) in order to verify, modify, and revise the data.

After the data have been verified and modified accordingly, they will be stored in the project computers, and made ready for data analysis.

#### **4 OTHER METHODS AND PRINCIPLES**

It is likely that there will be researchers who would like to conduct research on a language outside the selected area, and who would like to get project support. The project will look into ways of supporting such individuals in their research provided that, first, they agree to publish their results under LOT-Project and, second, they agree to abide by LOT-Project format and conventions.

Although the pilot study was conducted in Karagwe, Bukoba and Muleba districts (Kagera region) and much data was collected, more data will

necessarily be collected from the same area in the coming phase, in order to expand, improve, and verify existing data.

## **DISCUSSION**

Careful consideration is needed in the process of setting up zones since a combination of linguistic (related languages) and political (administrative regions) criteria appears to be in operation. It will not be easy to draw boundaries; languages and dialects do not have discreet edges.



## GUIDING QUESTIONS FOR GROUP DISCUSSION

### Note:

- (i) The questions are the same for all three groups. Please discuss them in relation to the relevant topic of our group (i.e. lexical data, morphosyntactic data, and atlas data).
- (ii) You are also allowed and encouraged to add any other relevant question(s).

### 1. Pilot Study

- 1.1 There were 4 researchers, each supervising two RAs.\* The RAs completed the questionnaires in Dar es Salaam first, and then took them to the field for further completion, consultation, verification, etc. How suitable is this approach?
- 1.2 How useful, relevant, or appropriately designed were the research tools?
- 1.3 How relevant or useful/suitable is the data collected in terms of its quality, quantity, format, reliability, objectives of the study, etc?
- 1.4 Entering data into the computer, proofreading/verifying/editing the data, and marking lexical tones are posing one of the major problems in data processing (especially in terms of getting suitable personnel to do it in time). What are your views on the issue?
- 1.5 Is there anything else, such as method, principle, theory, aspect, etc, that you think (was overlooked/ignored/forgotten/etc and) should have been done/applied/include?
- 1.6 Is there any method, principle, step, etc, that you think should have been avoided or ignored?

## **2.0 Proposed Programme of Action**

- 2.1 How suitable is the proposed approach in terms time, resources, financial implications reliability of the data to be collected, and the like?
- 2.2 Are there any other ways, means, strategies, etc (whether basic or additional) that can be used/added to improve the plan and thus ensure better results?
- 2.3 What would be other reliable/better alternatives if the proposed approach was to be abandoned?

## **3. The Project in General**

- 3.1 What problems should be anticipated and what are their respective solutions?
- 3.2 What recommendations would you make on the following aspects?
  - 3.2.1 Methodological approach
  - 3.2.2 Project in general

---

\* RA(s) = Research Assistant(s).

## **EXPERIENCES FROM ALLEX/ALRI**

Dr Herbert Chimhundu

African Languages Research Institute,  
University of Zimbabwe

### **1. Introduction**

The project started in 1992 by the name of ALLEX (African Languages Lexical Project). The project has now been institutionalized as the African Languages Research Institute (ALRI). The endeavour has been a collaborative effort involving researchers from the Universities of Zimbabwe, Oslo (Norway), and Gothenborg (Sweden).

### **2. Objectives:**

- To develop monolingual dictionaries
- To create and maintain a corpus for lexicographical work
- To establish formal computational linguistics training of staff and students
- To further cooperation between North and South in the field of lexicography

### **3. Achievements**

- Monolingual Shona and Ndebele dictionaries have been produced.
- The institute maintains a large electronic corpus of Shona and Ndebele materials, which form the backbone of the lexicographical work. Corpus building work is still in progress.

- Plans for a children's dictionary, a dictionary of literary terms, and an advanced Ndebele dictionary are at various stages of implementation.

#### **4. Other Plans**

- Bilingual dictionaries from Ndebele and Shona and other African languages are being considered.
- A dictionary of Zimbabwe Sign Language is being planned.
- Using experience from the two main languages, lexicography units will be set up for the other fourteen languages of Zimbabwe.

#### **5. Challenges Ahead**

- The institute is still very young (set up in 2000); consolidation of the programmes is essential.
- Corpus building needs to be a continuous process for the languages under study.

### **DISCUSSION**

The responses to the discussion questions highlighted the fact that the initial indifference of the public has been replaced by a strong appreciation of the work of ALLEX/ALRI. The project has disseminated its products widely and the researchers in the institute have become vital resource persons and consultants on language related issues.

# Lexicography and mass production

Ronald Moe

Linguistics Consultant, SIL Uganda-Tanzania Branch

## Abstract

Lexicography is a fruitful area for mass production techniques. Consultants can provide templates (standardized guidelines) to enable people who do not have a specialty in lexicography to produce extensive, quality dictionaries. Using an exhaustive list of semantic domains, it is possible to collect over 10,000 words within a week or two. Doing so at the beginning of a project provides a substantial dictionary, which can be used throughout the project for a variety of tasks. Collecting words by semantic domain results in a dictionary that is classified by domain. It is far more efficient to expand the dictionary field by field than word by word. It is far more insightful to investigate semantics domain by domain than word by word.

## 1. Mass production

Mass production techniques are highly effective whenever large numbers are involved. With thousands of languages in the world and tens of thousands of words in each, we need to collect and describe something on the order of 100,000,000 words. Efficiency is not a luxury but a necessity. Large publishing houses can afford to hire scores of professional lexicographers to work on a single language. SIL cannot. However, mass production techniques, such as the use of templates and task specialization, can multiply our efficiency and productivity, enabling mother tongue speakers to produce massive dictionaries quickly and with relatively little training or consultant help.

Within SIL most dictionaries are produced one word at a time over the course of a project. Words are collected as they are encountered. A word may be researched and described when it is added to the dictionary, or it may be entered with very little description. The result of such a hit or miss approach is usually a small dictionary that is very uneven in its breadth and depth of

coverage. We have policy documents that recommend that a team should collect 1,500 words before beginning translation. Considering that the Greek New Testament contains 12,000 words, a database of 1,500 is going to be insufficient to suggest translation options.

Many teams do not do serious work on the dictionary until after the translation is completed. This is tragic, because the dictionary can be a tremendous aid to translation and many other aspects of a program, including language learning. Virtually every computer program SIL uses for language work depends on a dictionary database, including those for interlinearization and phonemic analysis. I recommend that at least 12,000 words be collected in a two-week workshop at the very beginning of the project, so that the dictionary can serve as a tool throughout the project.

## 2. Templates

The Bantu Initiative<sup>1</sup> has recommended the production of templates to facilitate various tasks. A template is a mold or pattern that can be used repeatedly to produce similar objects. A ruler is a simple template for drawing straight lines. An example of a linguistic template would be a guide to producing a phonology statement. Much of the work in SIL could be facilitated by the use of templates. Whether producing a dictionary, orthography guide, or grammar, consultants can write guides and outlines of various sorts to standardize and facilitate the procedure and product. Consultants can design linguistically universal templates or modify a universal template for use within a particular language family.

---

<sup>1</sup> The Bantu Initiative is a consortium of linguists seeking to leverage the linguistic similarities among the 500 Bantu languages to facilitate research and language development.

SIL's training tends to focus on abstract principles to guide the researcher. The standard works on lexicography are no exception, giving principles along with examples to illustrate how the principle is to be applied. But it takes a great deal of study, imitation, practice, and corrections to learn how to apply abstract principles to real life. A template can shortcut much of this process, enabling a person with much less training to produce a quality product, because he is guided and restricted. A lack of constraints in lexicography results in very messy databases. Constraints may hinder a professional craftsman, but even craftsmen use templates, and they are incredibly helpful to someone who does not know all the intricacies of the job. With a ruler, even a child can draw a straight line.

A template also makes the process accessible to more people. As long as the procedure remains a complicated technique in the mind of one person, it is limited to the time he has available to devote to the task. A dictionary template enables people who do not have a specialty in lexicography to produce extensive, quality dictionaries. It enables the speakers of a language to do the bulk of the work and places the development of their language in their hands.

### 3. Lexical relations and semantic domains

The words of a language are organized in the mind in a multi-dimensional network of relationships of various sorts. Words are linked by patterns of syntactic distribution (part of speech), phonology (e.g., rhyme), semantics, and pragmatics. Lexicographers have catalogued many types of semantic links, called lexical relations (or lexical functions). Lexical relations tend to cluster around topics of conversation. These clusters form what are called

semantic domains. Semantic domains tend to be dominated by the “Generic-Specific” lexical relation but can center around a single semantic notion or area of life and include many lexical relations.

Many lexicographers have recommended that we utilize semantic domains and lexical relations to elicit and investigate words. What has been lacking is an exhaustive and universal list of domains. For instance, the *Outline of Cultural Materials* (Murdock et al, 1987) presents a list of anthropological domains, but is missing many lexical domains. *Roget's Thesaurus* (Roget, 1958) has 1000 domains, but due to its purpose it also omits many domains. Louw and Nida (1989: xix) admit that their list is uneven due to the subject matter of the New Testament. To fill this gap, I have been developing a list of semantic domains and related materials to facilitate the production of dictionaries. The list can be used to elicit, classify, and investigate the words of a language. Once a dictionary is classified by semantic domain, it can be sorted by domain for research purposes, to produce a semantic index, or to publish a semantically organized dictionary such as Louw and Nida's Greek-English lexicon.

As I have compared lists of semantic domains from around the world, it has become clear that almost all of the domains are universal. The differences come from minor differences of culture and the necessity to squash the multi-dimensional network into a two-dimensional list. Even organizing the list hierarchically fails to maintain all the semantic links. Some links can be maintained, while others must be lost. I have also tried to attain a level of detail such that each domain would contain ten to twenty words. At this level of detail the list contains approximately 1500 domains. So the organization of

the list is etic, somewhat arbitrary, and based on the commonalities of the lists available to me.

#### 4. Collecting words

Eliciting vocabulary has been a topic of interest for some time, and the literature contains a wealth of practical suggestions, such as using semantic domains and concordancing a text corpus (Beekman, 1968, Ballard, 1968, Pallesen, 1970). However, the use of semantic domains is by far the most effective, efficient, and productive. The combination of semantic domains and lexical relations is particularly powerful. Since lexical relations tie the entire lexicon together, the mind can jump rapidly from word to word, especially within a domain.

An exhaustive list of semantic domains filled out by lexical relations makes it possible to efficiently elicit the entire vocabulary of a language. This sort of systematic approach will ensure that the dictionary covers all domains of the language and does so to a relatively uniform depth.

However, the number of lexical relations is quite large, and it is far too inefficient to test each word against the entire list of relations. Rather than requiring each lexicographer to reinvent the wheel, I have thought through each domain, identifying the lexical relations applicable to that domain. It is more efficient for a single consultant to think through the theoretical issues than for each end user.

Lexical relations are easy to use, but hard to grasp in the abstract. So I have worded each relation in the form of a simple question. For example, the domain “Wind” has the following productive lexical relations:

What words describe a wind that lasts for a short time? *breath of air, puff of wind, gust*

What words describe a light wind? *draft, breeze*

What words describe a strong wind? *gale, howling (wind)*

What does the wind do? *blow, freshen, rise, fan (flames)*

What words describe the direction of the wind? *north wind, northeaster, updraft*

What sounds does the wind make? *sigh, moan, whistle, howl, shriek*

Answering these questions elicits a wealth of lexical material which might otherwise be overlooked. The example words following each question are merely meant to be illustrative. It takes very little mental effort to think of other words.

## 4.1 Example of a domain template

### 2.7.3 Lose consciousness

Use this domain for words related to losing consciousness, including: fainting, being knocked out, and anesthesia. For visions, hallucinations, and spiritually induced trances use “Vision, hallucination.”

What words refer to losing consciousness? *lose consciousness, go unconscious, faint, swoon, pass out, black out, be knocked out, go into a coma*

What words refer to the state of being unconscious? *be unconscious, be in a coma, be out, fainting spell*

What words refer to something causing someone to lose consciousness? *knock (someone) out, put under (anesthesia)*

What words refer to regaining consciousness? *regain consciousness, come to, come out of (the coma)*

What causes someone to lose consciousness? *hit on head, be sick, pain, shock, anesthesia*

What happens or what symptoms occur when someone begins to lose consciousness? *feel faint, feel dizzy, stagger, become incoherent*

## 4.2 Elicitation workshop

The procedure was tested using a beta version of the semantic domains list in a workshop for the Lugwere language of Uganda.<sup>2</sup> In ten days, fifteen participants collected over 10,000 words and 1000 example sentences.<sup>3</sup> One participant said, “The words are falling out of my head.” With an exhaustive list of domains and a little practice, a person can collect words almost as fast as he can write. I am currently revising the materials and method, so that even better results should be possible.

The procedure consists of the following.

- a. Organize a workshop with twelve to twenty participants.
- b. Print the list of domains, one domain per page. Divide the domains into sections of about twenty domains, and place each section in a folder.
- c. Give instructions to the participants, explaining the concept of a domain, the procedure, and working through some domains as a

---

<sup>2</sup> Thanks are due the Bantu Initiative for funding this workshop and initial research for the semantic domain list.

<sup>3</sup> By comparison the published Swahili dictionary on my shelf only contains around 5000 entries.

group to give them some practice. Decide on citation forms for the major parts of speech. Try to minimize orthographic issues.

- d. Divide into small groups. Allow each group to choose a folder of interest to them. Try to get people to work on domains that they have special knowledge of (pastors work on “Religion,” farmers work on “Agriculture”). Have each group review the domains in the folder then work on one domain at a time. Have them read the instructions then think of as many words as they can.
- e. If time permits, have them go back and give a simple gloss in the national language. If there are unpredictable word classes, have them give the plural of nouns and affixed forms of verbs that will enable identification of the class. However, the primary goal is to collect as many words as possible. These other tasks can be done later.

The workshop facilitator should closely monitor each group’s work, especially at the beginning. Incorrect work will result in a huge cleanup job later.

In the Lugwere workshop the participants were highly motivated and did not get bored, partly because the topic is constantly changing. They finished going through all the domains in about five and a half days. The participants improved over the course of the workshop, so that by the end they were coming up with twice as many words per domain. I asked them to review the earlier domains that had few words and see if they could add more words.

## 5. Expand the dictionary field by field

Once the words of a language have been collected, the dictionary can be filled out field by field. It is far more efficient to expand the dictionary field by field than word by word. Some fields can be added using macros. Speakers of the language can be trained to do things like add the part of speech or write example sentences. Some tasks, such as correctly identifying all the parts of speech, require a trained linguist, but non-linguists can do the vast majority of the work.

## 6. Investigate semantics domain by domain

One benefit of using semantic domains to collect words is that the resulting dictionary is automatically classified by domain. Lexicographers recommend that words be investigated in semantic sets. It is far more insightful to investigate words domain by domain than in isolation. Definitions can be standardized within a domain. It is also better to write example sentences for all the words of a domain at one time. Many pragmatic issues, such as connotation and register, are also better investigated within the confines of a domain. Anthropological issues also tend to be domain specific.

For all these reasons I am expanding each domain template to include instructions for investigating the words of the domain. Each template will note the semantic features that are likely to distinguish the members of the domain and give sample definitions. It will note anthropological, pragmatic, and other potential issues.

## 7. Bibliography

- Beekman, John. 1968. Eliciting vocabulary, meaning, and collocations. NT 29.
- Ballard, Jr., D. Lee. 1968. Studying the receptor language lexicon. NT 29.
- Louw, Johannes P. and Eugene A. Nida. 1989. Greek-English lexicon of the New Testament: based on semantic domains, Vol. 1. New York: United Bible Societies.
- Morehead, Albert H., ed. 1985. The new American Roget's college thesaurus in dictionary form. New York: Signet.
- Murdock, George P., et al. 1987. Outline of cultural materials. 5<sup>th</sup> ed. New Haven: Human Relations Area Files.
- Pallesen, Kemp. 1970. More on elicitation. NT 36.
- Roget, Peter Mark. 1958. Roget's Thesaurus. Harmondsworth, Middlesex: Penguin Books.

## DISCUSSION

During the discussion, a number of suggestions and ideas were put forward. Many participants agreed that the method used for massive data collection, i.e. semantic domains and lexical relations, is quite handy, and could be used to capture as much data as one would like in a very short time. It is possible to collect 10-15 words per minute, if no immediate attempt to provide definitions is made. In this approach it is easier to obtain cultural vocabulary that is specific to a particular language. It was also noted that the final dictionary need not be semantically organized but could easily be presented alphabetically.

## GENERAL OVERVIEW AND RECOMMENDATIONS

Dr. A. F. Lwaitama

Department of Foreign Languages and Linguistics

### 1. General overview

- 1.1 The papers presented on Day One of the Workshop and the discussion that followed their presentation allowed the participants to be appraised on how well the pilot study had been implemented.
  - 1.1.1 First, Prof. Kahigi gave a report on the general experiences gained from the pilot study. Though noting the various problems encountered due to constraints of time, financial resources and human capabilities, it was agreed that, in general, pilot study experiences were positive.
  - 1.1.2 Second, Dr. Rugemalira gave a report on lexical data collection and analysis. The difficulties encountered were noted and experiences were shared on how these may be resolved. The novel analytical procedures introduced to capture the unique lexical features of the lexicon of the languages/dialects under study were noted and critically scrutinized. The use of the notion of semantic domains in generating further lexicon items in addition to items generated by the use of standard English/Kiswahili lexicon prompts was noted.
  - 1.1.3 Third, Prof. Massamba presented a report on morphosyntactic data collection and analysis. The presentation generated a lot of discussion. Suggestions for improvements on the procedures adopted in the pilot study were made, but on the whole the approach adopted in the pilot study was validated.

- 1.1.4 Fourth, Prof. Rubagumya presented a report on the atlas data collection and analysis. The difficulties encountered were noted and suggestions for improvements on the research methodology adopted were made. On the whole it was agreed that the experiences gained from the pilot study could now be used in extending the study to the whole, of the Lake Zone and beyond.
- 1.1.5 Fifth, Dr. Muzale provided the participants with a description of the Year 2 Programme of Action. Experiences were shared and consensus sought on the way forward in resolving some of the anticipated problems.
- 1.1.6 Sixth, Dr. Chimhundu made a spirited presentation on the experiences from ALLEX/ALRI in Zimbabwe. The presentation was inspirational not least in showing how an African University could best use Foreign Donor Support to build capacity in the study of African languages.
- 1.1.7 Finally, the Day One deliberations were concluded by a presentation by Prof. Mlacha. This was a presentation whose purpose was to offer an overview and some critical observations on what had been presented and discussed on Day One of the Workshop. Some criticism was made of the overall approach of the LOT Project. Reservations were expressed about whether it was the best use of the available resources to implement the project by starting with a pilot study and then proceeding with a focus on the Rutara and North Nyanza, Suguti and South Nyanza, and Western Highlands languages/dialects in Phase 1 of the LOT Project, which was to be concluded by 2003. All the

same, it was agreed that the LOT Project ought to be allowed to proceed as originally conceived, at least in Phase One, not least in order to allow for the research outputs of that very phase to be of such good quality as to attract further and perhaps even higher levels of funding for subsequent phases of the project.

- 1.2 On Day Two of the Workshop participants heard a presentation by Dr. Ron Moe of SIL on massive data collection. It was suggested that one could use the notion of semantic domains in generating massive lexical data on many languages in the shortest possible time. Ronald Moe offered some illustrations of how his semantic domain technique could work. Some of the drawbacks of using the technique were also touched on in the course of the general discussion of the said technique. The LOT Project was encouraged to liaise with SIL and find out how at least some aspects of the technique could be adopted or adapted by the LOT Project.
  
- 1.3 Group Discussion on Lexical Data, Morphosyntactic Data, and Atlas/Mapping Data took up the rest of Day Two of the Workshop. The various Group Discussion outputs were summarized in terms of recommendations on how best to proceed with the collection of lexical, morphosyntactic, and atlas/mapping data in the subsequent implementation of the Action Plan of the LOT Project which had been discussed earlier on Day One of the Workshop.

## **2. General Observations**

2.1 The discussions on Day One indicated that all agreed that on the whole the pilot study had gone well and that therefore the study could proceed as originally envisaged.

2.2 The overall research approach adopted by the LOT Project has been seen to work. The Project Coordinators saw it fit to organize the First LOT Workshop in July 2001. At this Workshop the relevant experiences of colleagues drawn from language departments and institutes in Tanzania, in Southern Africa in general and in Europe were taped even at that early stage of articulating the details of the Research Work Plan. The LOT Project Coordinators have also seen it fit to organize this Second LOT Workshop in March 2002. At this Workshop the pilot study, which constitutes an important first step in the implementation of Phase One of the LOT Action Plan, has been subjected to intense criticism. By allowing their research endeavours to be constantly held up to the magnifying glass of close scrutiny by fellow linguists, the LOT Project Coordinators are adhering to professional standards which can only be commendable. The quality of the research endeavours subjected to such constant criticism can only be better.

## **2.3 Recommendations**

2.3.1 It is recommended that the LOT Project continue to organize workshops like this one every after a year of research work not least to share experiences with colleagues on the way forward in resolving

difficulties encountered in the implementation of the given Research Action Plan.

2.3.2 The Project Coordinators must at this stage begin to articulate the division of the whole of Tanzania into appropriate research zones as well as to tentatively indicate the time frame appropriate for the extension of the LOT study into the relevant zones.

2.3.3 The Project Coordinators should be allowed to proceed to implement their original Research Action Plan provided they took into account some of the key recommendations offered to them in the summaries presented after the Group Discussions on Day Two of the current Workshop. The detailed recommendations made by Prof. Batibo are particularly relevant in this regard. Let all linguists at University of Dar es Salaam who are willing and able be involved in the study of the languages of the Lake Zone irrespective of whether they do or do not speak any of the relevant languages to be studied. Let all linguistics at the University of Dar es Salaam feel a sense of ownership of the LOT Project. Let all the linguists at the UDSM, work as a team and together cover one zone after the other. They should produce an atlas for the relevant zone, a description of at least one of the languages of the given zone as well as producing classified word lists for at least two of the languages of the given zone. Let the LOT Project be guided by the desire to produce quality outputs that will capture the unique features of African languages, in this way avoiding some of the shortcomings of research outputs undertaken in previous centuries by foreign missionaries and explorers.