

1. OPENING REMARKS

By Prof. C.S.L. Chachage, Associate Dean,
Research and Publications, Faculty of Arts and Social Sciences.

It gives me pleasure to make these opening remarks at the first workshop of the Languages of Tanzania Project. It is a pleasure because I realize that we are launching a project of great significance and potential in the study of the languages of Tanzania.

As the person in charge of coordinating the research programmes in the Faculty of Arts and Social Sciences, I am particularly encouraged by the launching of this project because the humanities disciplines in general, and the languages in particular, have suffered considerable under-funding and neglect during the so-called development decades. These disciplines have often been regarded as non-developmental luxuries that deserve no resource allocation in a poor country's budget. Worse still, the study of Tanzania's ethnic community languages has at times been considered retrogressive and inimical to national unity.

Consider two manifestations of this attitude. First, within the structure of the University of Dar-es-Salaam (UDSM) there exist the Department of Kiswahili, the Department of Foreign Languages and Linguistics, and the Institute of Kiswahili Research. As the names clearly indicate, none of these units is directly charged with the study or promotion of Tanzanian languages other than Kiswahili. Beyond the University of Dar-es-Salaam there is the National Kiswahili Council (BAKITA) established by an act of parliament. But there is nothing anywhere close to these institutions for the study and promotion of the other languages of Tanzania.

Second, in connection with the various laws and policies on broadcasting, elections, and education, the consensus seems to be that the ethnic community languages must be ignored at best, or directly suppressed at worst. Broadcasting in a language other than Kiswahili or English is virtually prohibited; registering a non-Kiswahili or non-English newspaper or magazine is impossible; addressing an election campaign rally in an ethnic community language constitutes an election irregularity; and teaching any subject in any educational institution in Tanzania in a language other than English or Kiswahili is against the official policy on the medium of instruction.

The university structures, of course, reflect the hopes and fears of the society of which the university is part. But I am glad to note that the proposals for restructuring the Faculty of Arts and Social Sciences include the setting up of a Department of African Languages. It is not yet clear to what extent such a department would copy the structures of the existing teaching departments. It appears to me though, that the strength and rationale of an African Languages Department at the University of Dar-es-Salaam would reside in the kind of enterprise that this project is embarking upon - namely, research and documentation of the languages of Tanzania.

Beyond the university, again there is a ray of hope in the form of the national cultural policy (*Sera ya Utamaduni*). The policy proclaims that the

ethnic community languages are a major cultural treasure. It makes a public commitment to the need:

- a) to promote research, preservation, and translation of the ethnic community languages;
- b) to produce dictionaries and grammars of the ethnic community languages;
- c) to publish various materials in the ethnic community languages.

This policy is a significant step in the right direction. But it should be noted that, in the implementation of the policy, it will be important to take special measures to allay the fears and prejudices of a sizeable and influential section of the population. We need to persuade the doubters that in order to promote national unity we do not have to reject and suppress the rich linguistic diversity of our ethnic communities. We need to assure the young and restless that modernity need not imply the denigration of the past.

Another aspect of the project's significance concerns the creation and consolidation of research capacity in the university. This workshop is the first of a series of workshops planned under the project and designed to enhance participants' skills in research planning, field methods, and reporting of results. In addition, the project will be sponsoring several candidates for post-graduate study in linguistics. Above all, the project will provide tremendous opportunity for linguists at the University of Dar-es-Salaam to do basic fieldwork. The experience to be gained through this kind of work will be invaluable.

I note, with pleasure, the presence of prominent linguists who are not strangers to the University of Dar-es-Salaam. Prof. Karsten Legere, Prof. Herman Batibo, and Prof. Derek Elderkin, all taught at this university in the 1970s, 1980s and 1990s. We also have representatives from the Summer Institute of Linguistics in Dodoma, and Pioneer Bible Translators in Morogoro. The presence of all these researchers is an important reminder of the need for cooperation between the project team and other people engaged in similar work. The project will need to tap the experience and expertise of various people and be prepared to learn from and share experiences and materials with other researchers.

In this connection, one important consideration comes to mind. There exists a large amount of material on Tanzanian languages produced by various religious organizations during past one hundred years. I hope that this project will make a point of collecting samples of the various written materials in these languages in the form of Bible translations, hymn books, catechism manuals, newspapers, and the like. I believe that the university can and should be able to establish a languages archive, which would preserve these materials for posterity.

Thank you for your attention, and good luck in your deliberations.

2. THE PROJECT PROPOSAL AND WORK PLAN

By Dr. H.R.T. Muzale & Dr. J.M. Rugemalira

2.1 Project Summary

The Languages of Tanzania Project is a research endeavour conceived in the Department of Foreign Languages and Linguistics (FLL) of the University of Dar-es-Salaam (UDSM). The project has two major objectives. First it seeks to produce a language atlas showing the geographical location of the languages of Tanzania, number of speakers for each language, and the genetic classification. Second, the project will produce a series of descriptive studies which will document the grammar and vocabulary of each of the languages spoken in Tanzania. The materials put together under this project will form the base of a Tanzanian languages archive.

The documentation of the languages of Tanzania will fill an acknowledged gap. The project is expected to provide more reliable and up to date information on these languages. This information will be made readily accessible to scholars; it is such information that is made use of in the study of the nature of language (linguistic theory), language change, and language preservation. In a broader perspective, the documentation of language is a major contribution to the preservation of a people's cultural heritage. Nowhere is this more urgent than in the sphere of ethnic community languages which are threatened with extinction at a much faster pace (cf. Mekacha 1993).

The research team will be led by Dr. H.R.T. Muzale and Dr. J.M. Rugemalira of FLL, in collaboration with the African Linguistics group at Göteborg University (Sweden), headed by Prof. Karsten Legere. The project will rely on the linguistic research capacity within FLL, the Department of Kiswahili, and the Institute of Kiswahili Research at the University of Dar-es-Salaam, and the Department of Oriental and African Studies at Göteborg University.

At least three phases are envisaged for the project as a whole, each covering three years. The first phase (2001–2003) will be a planning, capacity building, and testing phase. It will thus lay the foundation for the whole project. The first year of this phase will consist of workshops for discussing the proposal and training in the methods and procedures, the setting up of the project office and equipment, as well as a pilot study focusing on two languages. The second year will consist of fieldwork to collect data on ten languages. The third year will be for data analysis and production of maps, at least two classified vocabulary lists, and one descriptive grammar. The second phase (2004–2006) will deal with data collection from all of the remaining languages, analysis of the data, and production of the language map by the sixth year. The third phase (2007–2009) will deal with further analysis of the languages studied for the production of at least twenty classified vocabulary lists and ten descriptive grammars.

2.2 Project Description (2001–2009)

2.2.1 Introduction

This research project has been conceived as a long-term project with two main objectives:

- i) to produce a language atlas showing the geographical location of the languages of Tanzania, number of speakers, and genetic classification,
- ii) to produce a series of descriptive studies which will document the grammar and vocabulary of each of the languages spoken in Tanzania.

The project seeks to produce a body of information that will form the backbone of a Tanzanian languages archive. It will constitute a formidable database and springboard for further research in linguistic theory.

Documenting the languages of Tanzania is of critical importance. As in all other fields of study, documentation places the relevant material on the scholastic map and draws the attention of current and future scholars to its existence. Documentation preserves the material for future study and reference. At the moment, very little is known about the history and development of the languages spoken in Tanzania today primarily because there are no written records. For instance, the immigrants who established the Hima empire in the Interlacustrine region (North-Western Tanzania, Western Uganda, Rwanda, and Burundi) around the 15th century are thought to have been speakers of a non-Bantu language. It is believed that these conquerors later adopted the language of their subjects and lost theirs. This raises some linguistic as well as historical questions, such as the following. What was this lost language? Did it leave any linguistic traces in the Bantu language that survived? How could conquerors so easily lose their language? What were the socio-cultural and socio-economic consequences of the change and/or loss of a language? If we had written records, such questions would find answers readily considering that the historical events lie only five hundred years back.

Given the level of technology at the disposal of humanity, it should be possible to preserve a spoken and written record of these languages. Such a record would give future generations tremendous advantages as they try to understand the phenomenon of language change. It would also help them to understand the cultures and identities associated with each language group before and after any linguistic changes. The forces at work in the Tanzanian sociolinguistic arena today promise to alter the linguistic map in a relatively short span (perhaps a century). The ethnic community languages are threatened with extinction at a much faster pace (cf. Mekacha 1993).

Committing Tanzanian languages to writing makes them more readily accessible to scholars interested in historical and comparative linguistics so that information from these languages can enrich our understanding of language change. Also of equal significance is the potential contribution of

African (Tanzanian) language structures in the development of linguistic theory. Whether it is in the generativist search for Universal Grammar or in the Greenbergian search for cross-linguistic generalisations, the importance of data from a variety of languages cannot be overestimated. The development of Lexical Functional Grammar, for instance, has benefited greatly from the study of African language structures (cf. Mchombo 1997).

The ethnic community languages of Tanzania survive mainly in the rural homes where they are spoken. But in the public domain, they are effectively ostracised; their use is excluded in education (e.g. all functional literacy materials in the 1970s were in Kiswahili), at political rallies and religious functions; they cannot legally be used by any radio or television station, and there are no newspapers published in any one such language. The general political context discourages the use of the ethnic community languages in favour of Kiswahili. Their use is seen as a retrogressive step undermining national unity. In the national population census there are no questions regarding languages spoken or ethnic affiliation. In such a context of official neglect and outright hostility, detailed information regarding these languages needs to be obtained by careful linguistic study. Also the preservation of samples of linguistic material as part of the people's cultural heritage must be deliberately planned and effected since no such samples will be preserved as a matter of course. And while there exists an Institute for Kiswahili Research, there does not exist even a token fund/organ for research in the ethnic community languages.

2.3 Background to the Study of Tanzanian Languages

There does not exist a reliable language atlas for Tanzania. The one hundred plus figure of the languages spoken in Tanzania is accepted by force of tradition, but there is no reliable source for it. It is worth noting that even a well-documented language count would not be accepted by everybody since the task of delineating languages and dialects is not entirely, or even primarily, linguistic, but is always subject to social and political pressures.

The production of grammars and dictionaries of ethnic community languages in Tanzania goes back to the early Christian missionaries and adventurers. Among the earliest available works are Madan (1905) on Gogo and Werner (1859-1935) on Hehe. Among the latest in this category is Felberg's (1996) dictionary of Nyakyusa. Most of these works were produced as aids for learning and teaching the relevant language for furthering missionary and colonial administration tasks.

A second category of works involves those produced by linguists as part of their normal scholarly pursuits. A representative sample would include Batibo's (1985) work on Sukuma, Whiteley's (1966) Study of Yao Sentences, and Mous's (1993) Grammar of Iraqw.

A critical look at the bibliography will reveal that only a few of the languages have so far been documented. Moreover, these studies vary in quality and coverage. The earlier ones are obviously rather dated. And yet

they are quite valuable if only because a particular piece of work may be the only written source on the language concerned.

There has not been a concerted effort to study a large number of languages with a similar format like the one proposed by Comrie (1977). The few studies that exist are rather like isolated chance occurrences in the study of Tanzanian languages. The language survey of the 1960/70s took a socio-linguistic focus, particularly in the case of Tanzania (cf. Polomé and Hill 1980). The Institute for the Study of the Languages and Cultures of Asia and Africa (ILCAA) of Tokyo University of Foreign Studies funded studies of several Tanzania languages beginning in 1987–1988. These studies have added some valuable published information on these languages. The ILCAA (1989) publication consists of a collection of papers on different topics in various languages. Perhaps more significant for our purposes are the classified vocabularies that have been published on Pare (Kagaya 1989), Nilamba (Yukawa 1989), Sandawe (Kagaya 1993), Shambala (Besha 1993), and Haya (Kaji 2000). But even this series has not produced a full description of any of the languages studied, and has refrained from attempts to compile dictionaries, preferring classified vocabularies instead.

2.4 Organisational Strategy

2.4.1 Planning and Organisation

This project requires a substantial amount of resources particularly in terms of finance, personnel, and time. We plan to implement the project in three phases:

2.4.1.1 Phase One: 2001–2003

This phase will lay the foundation for the whole project. It will be the capacity building phase that will consist of the following activities. First, a number of reliable researchers who can work both independently and in collaboration with others will be selected. Second, the two Project Coordinators (PC) will receive necessary training in computational and corpus linguistics in order to gain the necessary skills for managing this project. This will also give them an opportunity to study the way related studies in other countries (e.g. Kenya, Botswana, and Mozambique) were carried out. They will also obtain the relevant literature associated with such projects. Third, a group of up to eight students will be selected to embark on MA (linguistics) studies so as to ensure sustainability of the project. Among these, a few will be expected to pursue further studies at PhD level. Fourth, there will be a series of workshops that will enhance the research skills of linguistics staff at the UDSM, as well as selected students who will serve as research assistants. Fifth, through a pilot study covering two languages, it will be possible to get the researchers and assistants to try out the research instruments and skills before embarking on the larger study. Sixth, the research team will then cautiously approach the larger project piecemeal; they will do field work to collect data on ten languages from the following linguistic groups/zones: Rutara, North Nyanza, Suguti, East Nyanza, and

Western Highlands, that is, Kagera and Kigoma regions, and pockets of Mwanza and Mara regions.

The analysis of the data collected in this survey is expected to provide useful lessons to the research team. The team will seek to produce a language map for the area, a description of one of the languages, as well as classified word lists for at least two of the languages.

2.4.1.2 Phase Two: 2004-2006

It is envisaged that experience gained from the first phase will enable the team to embark on fieldwork covering the whole country. Accordingly, the second phase will collect data from all of the languages that were not studied in the first phase. The data will be analysed and, by the sixth year, the project should be able to produce a language map for the whole country.

2.4.1.3 Phase Three: 2007-2009

The third phase will deal with the remaining part of the study. That is, it will complete the analysis of the data collected in the second phase for all the languages, and then use that data to produce at least twenty classified vocabulary lists and ten descriptive grammars.

It is hoped that funds will be available beyond the ninth year to facilitate the production of more grammars and, perhaps, dictionaries. It is hoped that the data that will be collected in this study will be made use of by scholars to produce other scholarly works during and after the third phase.

2.4.2 **Project Management**

The Project Management Committee consists of the following members:

- | | | |
|------|-----------------------|---|
| i) | Prof. C.S.L. Chachage | Chair (Associate Dean, Research & Publications) |
| ii) | Dr. J.M. Rugemalira | Project Coordinator |
| iii) | Dr. H.R.T. Muzale | Project Coordinator |
| iv) | Dr. C.M. Rubagumya | Member (FLL) |
| v) | Prof. K.K. Kahigi | Member (Kiswahili Dept.) |
| vi) | Prof. D.B. Massamba | Member (IKR) |

The functions of the committee will be as follows:

- a. To provide day-to-day management and guidance for the project, such as
 - i) planning for field research;
 - ii) selecting researchers and research assistants for all the languages to be studied;
 - iii) receiving data and reports from researchers;
 - iv) reviewing and editing the data and reports from researchers;

- v) handing in the reports to the Project Coordinators who will be answerable to the Faculty on behalf of the Department.
- b. To publicise the project to interested people in order to win their support in the form of
 - i) intellectual commitment by doing research;
 - ii) moral support by encouraging their friends, students, and others to participate in the project;
- c. To seek financial resources from a variety of sources in order to sustain the project beyond the proposed period.

2.5 Data Collection and Analysis

Data collection needs to be in a form that lends itself to easy transfer, storage and retrieval. In particular questionnaires for eliciting vocabulary items should be in such a form that the responses can be entered into a central data base that is amenable to comparisons and use in the compilation of dictionaries. Data on the morphology, syntax, and phonology of the relevant language will be sufficiently detailed to provide the basis for their description and classification. Two main types of linguistic data will be collected, namely lexical and syntactic data. The former will be used for the genetic classification, the compilation of classified lexical lists, and determining the geographical distribution of the languages; the latter will be used for producing a language description.

The data collected from the field will be fed into the computer. A computer Database programme (to be determined later) will be used to organise, arrange, codify, and process the data. The choice of the programme to be used will largely depend on the availability of the programme itself and also its compatibility with the computer that we will get, that is, in terms of its size, the kind and version of the operating system, and our ability to use it on the data available.

Part of the planning work will give appropriate consideration to the place of a standard questionnaire/format for the language map and descriptive grammars. If the descriptive grammars adhere to a similar format, it will be possible for all of them to provide the same breadth and depth of coverage desired. This is important because a particular grammar of a language may be the only one that will ever be written for that language.

It should be pointed out that the amount of data to be collected will vary from one language to another. This is based on the reason that, whereas all of the languages will be classified and also included in the map, only a few grammars and dictionaries/lexical lists will be compiled on selected languages. The following chart presents a tentative plan for implementing the data collection during the first and second phases.

Phase	Linguistic/Referential area	Languages/dialects to be studied
I	Year 2001-2003	

Phase	Linguistic/Referential area	Languages/dialects to be studied
	Rutara & North Nyanza	Ruhaya/Runyambo, Ruzinza, Kikerebe, Rubumbiro, etc.
	Suguti & East Nyanza	Jita/Kwaya/Ruli, Zanaki/Nata, Kuria, etc.
	Western Highlands	Hangaza/Shubi, Ha/Vinza, etc.
II	Year 2004-2006	
	Western Tanzania	Sukuma, Nyamwezi, Sumbwa, etc.
	Central Tanzania I	Sagala, Gogo, Kagulu, etc;
	Central Tanzania II	Nyilamba, Nyaturu, Langi, Mbugwe, Kimbu, etc.
	Eastern Tanzania	Ng'wele, Doe, Zaramo, Kami, Rugulu, Vidunda, etc.
	Northern Tanzania	Meru, Chagga, Chasu (Pare), etc.
	North-Eastern Tanzania	Zigua, Shambala, Bondei, Digo, etc.
	Southern Highlands	Hehe, Bena, Kinga, Pangwa, Pogolo, etc.
	South-Western Tanzania	Nyiha, Safwa, Pimbwe, Mambwe, Nyakyusa, Ndali, etc.
	Southern Tanzania I	Ngoni, Ndendeule, Matengo, Manda, etc.
	Southern Tanzania II	Ngindo, Ndamba, Matumbi, Ndengereko, etc.
	Southern Tanzania III	Yao, Mwera, Makonde, Makua, etc.
	Cushitic	Iraqw, Gorowa, Burunge, Alagwa, etc.
	Nilotic	Maasai, Tatoga (Barbaig, Mang'ati), Luo, etc.
	Khoisan	Sandawe, Hadza (Tindiga), etc.
	Sign Language	Tanzanian Sign Language (TSL)

With regard to lexical and grammatical data, the researchers will be paying attention to the following aspects:

- i) Lexical equivalents in the target language and noting any dialectal differences;
- ii) Pronunciation of the lexemes in the target language/dialect (i.e. tone, phonetic variations, etc);
- iii) Basic morphological variations (e.g. Nominal classes, verbal extensions, tense inflections, etc);
- iv) Variations in meaning - in terms of dialects, use, collocation, etc).

The researchers will make use of the tape recorder judiciously, and they will also determine the orthographic conventions to be used.

The lexical and grammatical comparative method will be used in establishing the relationships between the languages/dialects. One of the methods to be used, apart from collecting lexical data, will be to test the rate of mutual intercomprehension between users of two or more languages.

In order to determine the number of language/dialect speakers, the researchers intend to make use of the national census information. The number of speakers of each language will be determined on the basis of The number of households/persons in each village. We assume that there will normally be one language per village. Therefore, there will be a questionnaire seeking to determine the language spoken in each village and the characteristic features distinguishing the variety in village X from the variety in village Y. We expect that most of this information can be collected at district headquarters and, therefore, the researchers will not need to go to each village. The pilot study will address this issue to see how workable the proposed procedure is.

The research will also make a point of collecting sample texts, both oral and written. Sample narratives will be tape-recorded for preservation and for use in further linguistic analysis. Available written materials will also be collected – translations of the Bible, prayer books, hymn books, newspapers, teaching manuals, and any available manuscripts in the targeted languages.

2.6 Project Objectives for Phase One (2001–2003)

The first phase of the project, 2001–2003, will focus on the following objectives:

- a. to organise and conduct workshops on how to collect, analyse, and store linguistic data, and then use the data to produce linguistic material;
- b. to train two principal researchers who will be the co-ordinators of the project (i.e., in corpus/computational linguistics, lexicography, and other consultations);
- c. to train linguistics students and enable them to apply their theoretical knowledge in linguistic field research;
- d. to establish an office that will deal with the data collected;
- e. to produce a corpus of teaching/learning materials in linguistics, especially in the areas of syntax, lexicology, etymology, lexicography, morphosyntax, dialectology, historical and comparative linguistics, and corpus linguistics;
- f. to produce a language atlas that will show:
 - i) the geographical location of the ten languages studied,
 - ii) the number of speakers for each of the ten languages/dialects;
- g. to produce a genetic classification of the ten languages/dialects studied;
- h. to produce descriptive studies which will document:
 - i) a grammar of at least one language,
 - ii) classified lexical lists for at least two languages;
- i. to produce a database and other forms of linguistic information/data (e.g. written and oral) for further research in linguistic theory;

- j. to publish papers, journal articles, manuals, and books based on the project findings.

2.7 Work Plan: 2001–2003

Sub-activities	Duration in months	Time Frame
** Four (4) M.A. Scholarships	(18)	Oct.2000–Mar.2002
a. Setting up the office, b. Acquiring/setting up and organising the relevant equipment/instruments	2	Jan.–Feb. 2001
a. Getting ready to start, and b. Preparing for the first workshop	1	March 2001
First workshop: a. Presenting and discussing the proposal b. Brainstorming on methods, procedures, organisation, scope, problems and their solutions, etc. c. Setting up the Project Management Committee (PMC)	1	April 2001
a. Training i) for the two (2) Project Coordinators, i.e., a trip to Sweden to meet and discuss with colleagues, study and acquire related literature, learn computational linguistics, and prepare questionnaires and maps; ii) for assistants, and b. Obtaining relevant literature associated with such a project	3	May–July 2001
Pilot Study (field data collection)	1	Aug. 2001
**Four (4) M.A. scholarships	(18)	Oct.2001–Mar.2003
a. Analysis of data from the pilot study b. Second workshop: i) presenting and discussing research instruments and findings from the pilot study, ii) drawing up a work plan and strategies for the project, iii) training on methods and procedures, iv) scheduling of field work vi) setting up the research team.	4	Sept.–Dec. 2001
Fieldwork (for ten languages from 4 regions) to collect, organise, and store the data: Kagera, Kigoma, Mwanza, and Mara regions.	10	Jan.–Oct. 2002
Third Workshop: a. to review fieldwork experience, b. to devise data analysis/handling formats, c. to plan/schedule activities for year 2003.	2	Nov.–Dec 2002

Analysing the data: to establish the number of languages/ dialects, their geographical locations, language map(s), (genetic) classification, and classified word lists.	5	Jan.–May 2003
Trip to Sweden by five (5) researchers to consolidate data handling skills and procedures, etc.	-	April–May 2003
Fourth Workshop: To review progress of data analysis	-	May 2003
Analysing the data: to establish the number of languages/ dialects, etc. continues.	5	June–Oct. 2003
Fifth Workshop: a. presentation and discussion of research findings, b. final project report, c. planning for Phase Two (2004–2006).	2	Nov.–Dec. 2003

References

- Batibo, H. 1985. *Le Kisukuma (Langue Bantu de Tanzanie): Phonologie et Morphologie*. Paris: Editions Recherche sur la Civilisations.
- Besha, R. 1993. *A Classified Vocabulary of the Shambala Language with Outline Grammar*. Tokyo: ILCAA.
- Comrie, B. 1977. *Lingua descriptive studies questionnaire*. *Lingua*, 42: 1–71.
- Felberg, K. 1996. *Nyakyusa–English–Swahili and English–Nyakyusa Dictionary*. Dar-es-Salaam: Mkuki na Nyota Publishers.
- Institute for the Study of Languages and Cultures of Asia and Africa (ILCAA). 1989. *Studies in Tanzanian Languages*. Tokyo: University of Foreign Studies.
- Kagaya, R. 1993. *A Classified Vocabulary of the Sandawe Language*. Tokyo: Institute for the Study of Languages and Cultures of Asia and Africa (ILCAA).
- Kagaya R. 1989. *A Classified Vocabulary of the Pare Language*. Tokyo: Institute for the Study of Languages and Cultures of Asia and Africa (ILCAA).
- Kaji, Shigeki 2000. *A Haya Vocabulary*. (Asian and African Lexicon, No. 37). Tokyo University of Foreign Studies: Institute for the Study of Languages and Cultures of Asia and Africa (ILCAA).
- Madan, A.C. 1905. *An Outline Dictionary Intended as an Aid in the Study of the Languages of the Bantu (African) and Other Civilised Races*. Microfilm.
- Mchombo, S. 1997. *Contributions of African Languages to Generative Grammar*. In R. Herbert, *African Linguistics at the Crossroads*. Cologne: Rudiger Koppe Verlag.
- Mekacha R. 1993. *The Sociolinguistic Impact of Kiswahili on Ethnic Community Languages in Tanzania: A Case Study of Ekinata*. Bayreuth: Bayreuth African Studies Series.
- Mous, Maarten 1993. *A Grammar of Iraqw*. University of Leiden.
- Polomé, Edgar and C. P. Hill. 1980. *Language in Tanzania*. London: Oxford University Press.
- Werner, A. 1859-1935. *Deutsch–Kihehe, Kihehe–Deutsch (und) syntax*.
- Whiteley W. 1966. *A Study of Yao Sentences*. London: Oxford University Press.
- Yukawa, Y. 1989. *A Classified Vocabulary of the Nilamba Language*. Tokyo: ILCAA.

2.8 Discussion

It was pointed out that although SIDA/SAREC are committed to supporting the project, it should be born in mind that the support for Phases Two and Three will depend mainly on the success of Phase One. So all efforts should be focussed at producing something worthwhile during the first phase.

It was also suggested that modifications to the original proposal could be made in order to make use of existing capacity within the university much faster than allowed for in the schedule. This could be in the form of deploying university students to collect data countrywide. In addition, the project should seek ways of getting other interested parties on board; these include institutions that can use their own resources and contribute to the project (e.g. the Summer Institute of Linguistics (SIL) and Pioneer Bible Translators), as well as individual researchers.

For the long-term sustainability of the project, besides the capacity building in the form of M.A. scholarships, the project should also seek to tap other sources of funding. It was noted that even after the first nine years, there would still be plenty of work to do on the languages of Tanzania.

3. BACKGROUND TO THE LANGUAGES OF TANZANIA PROJECT

By Prof. H.M. Batibo, University of Botswana

3.1 Background to the Project

In the early 1980's, the Linguistics Section of the Department of Foreign Languages and Linguistics (FLL) of UDSM put forward a proposal to make a coordinated and rigorous description of the languages of Tanzania. The proposal was motivated by the following realities at that time:

- a. Tanzania had one of the highest concentrations of languages in Africa, with more than 120 languages within only about one million square kilometres. This made Tanzania rank as the 5th country in Africa in terms of number of languages. The number and the diversity of the characteristics of these languages constituted an enormous linguistic and cultural wealth for Tanzania, which required proper preservation.
- b. Many of these languages were threatened by extinction, particularly as a result of the expansion of Kiswahili, the national language and lingua franca, as well as the fast rate of urbanization in the country. Hence the disappearance of these languages would leave no records for national heritage or linguistic and cultural preservation. The only way was to make a rapid description of these languages.
- c. The descriptive work that was going on was rather too fragmentary and uncoordinated. The local linguists were in most cases overburdened with teaching and other duties. Moreover, there were no funds for them to conduct extensive studies on the minority languages, particularly as most focus was on Kiswahili, the national language. As for the external researchers, many of them came with personal interests, either to conduct specific studies leading to theoretical discoveries or undertake descriptive work related to their academic requirements. Such studies often lacked a holistic description or relevance to the Tanzanian needs. Moreover, most foreign researchers left the country without leaving any of their findings behind, and often did not care to send back any of the publications they made out of their studies.

3.2 Justification for the Project

If no concrete steps were taken, not only most of the minority languages (also known as community languages) would disappear without any preserved records within the next few generations, but also the descriptions made by both local and foreign scholars would remain fragmentary, uncoordinated, and many of them would not be available for use in the country. The following steps were therefore necessary:

- a. to accelerate the research undertakings in the country by mobilising all the available human and material resources;
- b. to make the Linguistics Section of the Department of FLL the coordinating body of the project. Its main role would be to monitor all the research undertakings on the Tanzanian languages, to advise local and foreign researchers on the languages or aspects in languages which needed to be studied, and to make a periodic evaluation of the levels of achievement in the various undertakings;
- c. to make a rough survey of the languages of Tanzania by using the available information. The survey would determine the level of descriptive work already carried out on the Tanzanian languages and the work still to be done;
- d. to encourage foreign scholars to conduct research in Tanzania, particularly in the area of language description. The Department would arrange with the relevant authorities at the Commission for Science and Technology to make it easy for language researchers to obtain research permits to conduct research in Tanzania;
- e. to put forward a rigorous work plan spelling out clearly the objectives of the project, the priorities in the research undertakings, the strategies for the mobilisation of resources, the mode of operation, and a tentative time frame.

3.3 The Type of Data Needed in the Project

Although the original idea was to conduct a Linguistic Atlas of Tanzania indicating the number, distribution, and characteristics of all the languages in the country, such an undertaking was found to be of little importance given the mobility of communities, the disappearing tendencies of most minority languages and the complex sociolinguistic patterns in the country. Moreover, a linguistic atlas tends to concentrate on the distribution of a few lexical or phonological features and the drawing of maps and isoglosses without paying much attention to the linguistic systems or the comprehensiveness of the data collected on each language. The data to be obtained from the project would therefore be of the following type:

- a. a general idea of the location of the various languages in the country without too much concern for the exact location and distribution. Here the practical method of Bernd Heine used for Kenya, was to be applied in this study. According to this method, the speakers themselves provide the information about the location and distribution of their languages. This method is quicker, and more economical than the dialectometrical method of Wilhelm Moehlig;

- b. once the various languages have been identified and the representative dialects for each language determined, questionnaires would be used to collect both linguistic and cultural data for each of the languages.

3.4 Problems Encountered in the 80's and 90's

Although the Project was well planned and the aims, objectives, and form of activities were well defined, it could not take off vigorously for the following reasons:

- a. lack of resources to carry out the project systematically. The Department could not attract significant sponsorship for the project;
- b. understaffing and other pressing duties in the Linguistics Section made it difficult to concentrate on the project;
- c. lack of cooperation from some of the foreign scholars who promised to send their research findings or published work but did not;
- c. absence of storage a system in the Department to preserve the data and other materials in such a way as to easily retrieve them whenever needed.

In spite of these problems, the Department succeeded in the following:

- a. it was able to attract many foreign researchers to conduct research in Tanzania, hence many languages were described;
- b. it managed to arrange with the officials of the Commission for Science and Technology to make it easy for language researchers to get clearance to conduct research in Tanzania;
- c. it prepared sets of questionnaires which have been very useful in many research undertakings, even outside the Department, as they contain both English and Kiswahili equivalents;
- d. it helped to create cooperation and linkages with other institutions in Europe and the USA;
- e. it attracted donations of documents from individuals and institutions on specific and general topics.

3.5 Lessons for the Planned Project

The planned project can learn from the earlier project in the following ways:

- a. the undertaking should aim at the most desired data, namely linguistic and cultural data. Any effort to make a linguistic atlas should be considered as of secondary importance. Also given the dynamics of language change and language use patterns in the country and the impact of such dynamics to communication and development, it might be useful to include some dimensions on the current patterns of language use and language competence. Hence the focus of the project could be **language preservation** and determining the **language dynamics**;
- b. an inventory should be made of all the languages of Tanzania as a starting point, preferably using the information already available;
- c. appropriate questionnaires should be worked out so as to collect both

linguistic and cultural information for each of the languages. The questionnaires could be based on the ones used in the old project, one for morphosyntactic structures and the other for linguistic/cultural vocabulary;

- d. a plan of action should be made to ensure that the most threatened languages are described first, since these are the ones in urgent need;
- e. a location should be identified in the Department or University which would be the custody of all findings and a fast and easy system of retrieval should be arranged. If possible an on-line website system should be established for networking and access;
- f. foreign researchers should be encouraged to participate and even be helped to obtain research clearance so as to hasten the operation;
- g. the Project Management Committee should keep a close record of the levels of research undertakings for each language and encourage researchers to work on any understudied languages. Ideally, the project should collect data for all the languages in a systematic way.

3.6 Discussion

The discussion established a consensus that the production of a language atlas, with carefully considered levels of detail, was a worthwhile objective, and that the study of language dynamics as suggested in the presentation would constitute a separate study. Relevant details in an atlas would include the number of speakers, approximate geographical location, and boundaries with neighbouring languages.

Many participants agreed that it was a good idea for the project to start from data/information that is already available, and to make use of the lexical and morphosyntactic questionnaires inherited from the old project. Although the caution about collecting too much data is in order, it was pointed out that good and carefully structured data should be collected in abundance and be available for future researchers to refer to and analyse.

4. COLLECTING LEXICAL AND PHONOLOGICAL DATA

By Prof. James S. Mdee, Institute of Kiswahili Research, UDSM

4.1 Introduction

The following are the reasons for conducting linguistic research:

- a. A linguistic research is a way of obtaining linguistic data and studying linguistic phenomena (Samarin 1967).
- b. The linguistic data can shed some light towards the understanding of language universals.
- c. The results of the research can be applied practically to human affairs, such as in language planning, whereby languages of a society are assigned roles that each can play: national language, official language, language of education, or just a language for social interaction.
- d. The research results can be stored in different forms for future uses such as teaching the languages or developing them to be teaching languages at various levels.

Research on ethnic languages, which until now were left on their own, will eventually contribute significantly to their development because they are the languages in which the culture of its speakers is preserved and, through them, it is passed down from one generation to the next. Documenting lexical and phonological data of the ethnic languages is not only important in preserving them and saving them from extinction, but also because they play a vital role in the transmission of a society's heritage.

It is in this context that the project of studying the languages of Tanzania has evolved, and one of the aspects to be investigated for documentation is lexical and phonological data. In this presentation we shall do the following:

- a. examine the objectives for which lexical and phonological data are collected;
- b. examine the various word lists used in lexical data collection, and analyse their strengths and weaknesses;
- c. discuss methods of lexical and phonological data collection.

4.2 Objectives for Lexical and Phonological Data Collection

Lexical and phonological data can be collected for three linguistic objectives, as outlined in the following subsections:

4.2.1 Making a comparative study of two or more languages

In a comparative study, a researcher's objective is to compare some linguistic aspects across the languages or dialects under study. For example, Ngonyani (1988) made a comparative study of Kindendeule, Kingoni, and Kimatengo when he was investigating the relationship between three languages. Mreta (1990) compared Chasu, Kigweno, Kikahe, and Kirombo in his linguistic investigation of the said languages, and Lukula (1991) investigated the influence of Kikerewe on Insular Jita. The data needed for a

comparative study of languages is usually a representative sample. This explains why Ngonyani used a wordlist of 300 lexemes, Mreta needed a wordlist of 400 lexemes, while Lukula had about 711 lexical items for his task.

4.2.2 Describing specific aspects of a language

In a descriptive study, a field linguist collects data for the purpose of describing a linguistic aspect in order to discover the language's basic structure, such as phonology, morphology, and syntax. For example, Byarushengo (1975) examined the "Segmental Phonology of Haya". He used 409 lexical items, while Mkomagu (1991) collected 100 lexical items for his study in which he examined the tonal patterns in Kisongwe. Mshindo (1989) examined the morpheme *ka-* in the Pemba dialect of Kiswahili. He collected his data from 25 texts covering various topics such as autobiography of an informant, history of one's village, school or family, and narration of an incident. Kagaya (1989) conducted two different linguistic studies of the tonal systems in Gonja and Mbagha dialects of the Southern Pare language. Linguistic research for the purpose of describing an aspect of a language requires also a representative sample to accomplish the required objective. This explains why Mkomagu collected only 100 items for his research.

4.2.3 Documenting the vocabulary of a language

Documenting a language's vocabulary entails the recording of the words of a language. It requires hunting every word ever spoken or written. Lexical data that is collected for this purpose will eventually end up in a Vocabulary or Dictionary of a language. The classified Bantu vocabulary series by the *Institute for the Study of Languages and Cultures of Asia and Africa* were compiled on these lines although they did not endeavour to record many of the common and cultural words of the languages they studied. The wordlist for each language was between 1000-2500. Documenting the vocabulary of a language is the basis for this presentation. The objective of this presentation is to discuss how we can collect lexical and phonological data of an unwritten language so that eventually the lexicon of the language can be documented.

4.3 State of the Art

The conventional way of collecting lexical data is the lexicostatistical method. This technique assumes that languages have a basic/core vocabulary, which is less subject to change. The core vocabulary constitutes specific semantic groups that can be used to compare languages or dialects of one language, or to compile wordlists of unwritten languages.

Different scholars have proposed wordlists of varying lengths for collecting lexical data. These include Swadesh (1955), Samarin (1967), Batibo (1989), Lukula (1991), Yukawa (1979) and others. Swadesh (1955) prepared a basic/core vocabulary of 100 lexical items arranged in semantic groups; the list has been reproduced or expanded by other researchers.

Samarin (1967) has a wordlist of 218 basic words whose semantic groupings are as follows:

- i) Parts of human and animal anatomy;
- ii) Clothing and personal adornment
- iii) Artifacts, tools, utensils, weapons, machines, furniture, conveyance;
- iv) Occupational and personal terminology;
- v) Geographical and astronomical items, winds, types of weather, phases of the moon, season of the year, constellations;
- vi) Flora and fauna;
- vii) Food and methods of food preparation;
- viii) Measurement of time, space, volume, weight, quantity;
- ix) Disease and Medicines;
- x) Games, amusement;
- xi) Religious objects, beings;
- xii) Etiquette and taboo words, insults;
- xiii) Colours, textures and shapes;
- xiv) Systems of enumeration, counting money, telling time, etc
- xv) Classifiers: bunch, flock, handful, piece of.

Batibo (1989) has two wordlists, one list is a basic vocabulary of 232 words and the second list of cultural vocabulary has 1563 items. The basic vocabulary lacks many common words such as cry, mouth, waist, open, fill, etc. The cultural vocabulary is also arranged in semantic fields:

- i) Agriculture/*Kilimo* (1-254)
- ii) Animal husbandry/*ufugaji* (255-486)
- iii) Hunting/*uwindaji* (487-716)
- iv) Fishing/*uvuvi* (717-762)
- v) Beekeeping/*ufugaji wa nyuki* (763-785)
- vi) Collecting insects/*kuokota wadudu* (786-845)
- vii) Food and drinks/*vyakula na vinywaji* (846-926)
- viii) Tools and instruments *zana na vifaa* (927-1010)
- ix) Containers/*viwekeo* (1011-1081)
- x) Settlements/*makao* (1082-1163)
- xi) Social systems and activities/*mfumo wa jamii na shughuli zake* (1164-1306)
- xii) Cultural practices/*mila na desturi* (1307-1498)
- xiii) Attire and Adornments/*mavazi na mapambo* (1499-1563)

Lukula (1991) like Batibo has two lists: A list of basic and common vocabulary of 208 words, and a second list of cultural or localized vocabulary of 503 words. The basic vocabulary, partly adopted from Swadesh's wordlist and partly from Batibo, is categorized as follows:

- A. Parts of the body and secretions

- B. Social life
- C. Domestic life
- D. Human activities
- E. Eating and clothing
- F. Natural phenomena
- G. Quantities and Qualities

The cultural vocabulary consists the following subgroups:

- A. Physical environment
- B. Animal husbandry
- C. Wild animals, reptiles, birds and hunting activities
- D. Boat/canoe building and fishing activities
- E. Types of food, beverages, and their preparation
- F. Tools and instruments
- G. Metallurgy, archery, and other weaponry
- H. Containers
- I. Social systems
- J. Supernatural beliefs
- K. Social practices
- L. Settlements
- M. Attire
- N. Adornment
- O. Diseases.

Studies on Bantu vocabulary conducted by different scholars from the *Institute for the Study of Languages and Cultures of Asia and Africa (ILCAA)* in Japan use a wordlist prepared by Yukawa in 1979: "A Tentative Questionnaire for the Words of Bantu Languages" which has words arranged in 22 main categories each of which is classified further into a total sum of 82 subcategories, as follows:

- a. Human body: head, body, arm, leg, inside of the body, physiological phenomena, senses;
- b. Illness and injury: illness, mental disorder, injury, skin disease, symptoms, physical handicap, treatment;
- c. Clothing and dressing: clothing, sewing, grooming, ornaments.
- d. Eating: food, cooking, tableware, eating, condition of food;
- e. Dwelling: house, furniture, cleaning;
- f. Kinship: family, marriage, birth and rearing;
- g. Human being: human being, death;
- h. Animals: domestic animals, wild animals, hunting, reptiles, small creatures, fish, birds, insects and worms;
- i. Plants: plants, parts of a plant, life of a plant;
- j. Body actions: sleeping, body actions, movements;
- k. Daily life: work, fire, water;

- l. Social life: language, playing, fighting, giving and receiving, society, politics, war;
- m. Mental life: feelings, mental activity, number and counting, religion;
- n. Action towards things: movement of things, connecting and disconnecting, transformation, cutting and breaking, other kinds of actions;
- o. Things: parts of things, relation to things, colours and shapes.
- p. Natural phenomena: nature, light and sound;
- q. The earth: geographical features, rivers, earth and minerals;
- r. Time: time;
- s. Nature of things: nature of things;
- t. General (action): beginning and ending, other verbs;
- u. General (things): general things;
- v. Miscellaneous: pronouns, interrogatives, adverbs and conjunctions, greetings.

Counting the lexical items in the classified Bantu vocabulary series we note that although the researchers used the same questionnaire with the same number of semantic groups the lexical data collected for every language differed in number. For example, Kagaya (1989) listed 2301 items in his classified vocabulary of the Pare; Yukawa (1992) collected 2126 lexical items in his classified vocabulary of the Luba language in which he also indicated phonological data for every lexical item. Kaji (2000) listed only 1136 in his Haya Vocabulary. This difference simply shows how speakers of every language divide their own world in different ways. The difference can also be explained by the field researcher's skill of collecting data. Samarin (1967) notes that more and useful vocabulary can be collected if the field worker will take time to ask for other ways that an object is designated. However, a critical examination of Yukawa's questionnaire will reveal that the list incorporates Swahili words which have not really been adopted in the language. For example, in the Pare classified vocabulary, Swahili words are entered as Pare words although Pare has its own words for those concepts and those particular lexical items have not been adopted into the said (language). Compare the following Swahili-Pare equivalents:

- | | | |
|-----|------------------|------------------------------|
| (1) | kifua = mbafu | ngozi = kitanga/mkota/ikonde |
| | ubavu = msavaju | ndui = kiseghera |
| | harufu = mphungo | kigoda = kichumbi |

It should be emphasized here that although code switching is very common in our languages, this should not imply that such words have been adopted into the language.

4.4 Comments on the Wordlists in General

Examining the different wordlists we can make the following observations:

First, the Swadesh and Samarin wordlists were designed for a limited scope, that is, for a comparative study of languages. This is why they have a wordlist of 100 or 200.

Second, the wordlist of Swadesh and that of Batibo show cultural differences of their authors, reflecting the cultures of their respective speech communities. Swadesh and Samarin include machines, geographical and astronomical items, counting money, and telling time, but lacks hunting, collecting insects, beekeeping and animal husbandry which Batibo includes. Batibo does not include machines, counting money or telling time.

Third, Lukula's list though borrowed from Swadesh and Batibo and compiled by a Tanzanian from Ukerewe Island, also shows some cultural differences with that of Batibo, also a Tanzanian from Sukumaland, which is not an island. Lukula adds in his cultural list the category of boat/canoe building and fishing activities. For example, Lukula has words for different types of boats/canoe, sections of a boat, materials for building a boat, different methods and instruments of fishing etc. Batibo's list lacks boat building but includes a fishing category under which he lists fishing in rivers, ponds, lake and sea, and different types of fish, which is lacking in Lukula's list. Cultural differences can only be the explanation for these gaps. The islanders do not fish in ponds or rivers. The Sukuma may have become fishmongers too; hence the need to be specific with fish names. Kagaya has limited vocabulary on fishing. He has four types of fish and a word for fishing, skin and bone, as well as two more Swahili loan words: *ndoana* 'hook' and *kuvua* 'to fish'.

Fourth, there are no cultural words common to all speech communities regardless of their historical or geographical affinity. Therefore, compilation of a cultural wordlist shall require a researcher to explore all elements of the speech community's culture in every aspect of life. For example, although all human beings eat food, every culture has distinct dishes even if the raw materials are the same. Compare the following dishes of the Chagga, the Pare and the Swahili:

- (2) *Chagga dishes* are based on bananas and have the following types: shiro, machallari, kyena, mapoko, mtori, ngande, and kyumbo;
- (3) *Pare dishes* are based on maize and have the following types: mphure, nkokoro, ntunyula, mhawe, kibulu, and magolo. Although the pare eat bananas too, they have only two dishes made from them: 'kishumba' and 'mnyanyanto';
- (4) *Swahili dishes* are based on rice and have the following types: wali, ubwabwa, pilau, and biriani.

4.5 Some Observations on Lexical and Phonological Data

4.5.1 Lexical data

We have surveyed wordlists compiled by different scholars and noted their scope, their weak points, and their strengths. We noted that cultural words are very elusive because of the diversity of culture itself. Let it be said here that cultural vocabulary is a stock of words, which reflect the society's inherited assemblage of practices and beliefs determining the texture of our lives (Sapir 1949:207). Because of this, it is obvious that there are no two languages which have a common cultural vocabulary. Contrasted with this, basic core vocabulary is the sector of the lexicon which deals with those elements of universal human experience which exist irrespective of the speaker's culture. In basic/core words you will find:

- a. biological activities: eat, sleep, give birth, die, etc.
- b. major divisions of the body: head, arm, leg, etc.
- c. natural physical phenomena: fire, water, sun, moon, wind, rain, cloud, etc.
- d. general relational concepts: personal nouns, demonstratives, negations, size etc. (Bynon 1983).

This means that whereas we can use basic core vocabulary of one language to compile equivalent words in other languages, it is not possible to do the same for cultural words. Such a list will have to be compiled for every language. In order to be able to compile this list one has to acquaint oneself with the culture of its speakers, that is, to learn and study the culture of the people and then determine semantic groups for cultural vocabulary.

4.5.2 Phonological Data

Phonological data is a record of the sounds of an utterance by broad or narrow phonetic transcription. In a layman's language, it is the pronunciation of words and sentences. It shows the phonological form of a lexical item (including supra-segmentals). The supra-segmentals include:

- a. stress (primary or secondary),
- b. tone (extra high, high, mid, and extra low, etc.)
- c. long/short vowel,
- d. aspiration, etc.

To be able to show these linguistic features, a linguist should be versed in phonetics. Kagaya (1989) marked tone in his Pare classified vocabulary. It is important to recall that every language has its own basic sounds taken from the pool of sounds that human beings can produce. Although there are some common sounds that all languages share, almost every language has some

specific sounds. Now let us examine how a wordlist for a vocabulary of a language can be documented.

4.6 Documenting the Lexicon

We have seen that Batibo and Yukawa have each proposed a detailed vocabulary that can be used to elicit words from any speech community in order to build a lexicon for a language. Both wordlists have common and cultural words and both lists are arranged in semantic groups. I propose that, for the basic/core vocabulary, we take Yukawa's list because, with this list, one can develop the vocabulary systematically taking one semantic group after another and then filling any gaps with Batibo's list and from other sources. Yukawa's list starts with the human body and examines first the main parts of the body: head, body or the trunk, arm, leg, inside of the body, physiological phenomena, and senses. He then takes each item under this category and sorts out every separable part, as given in the following example.

(5) head

brain (ability = intelligence), occiput, hair, white hair, baldness, face, forehead, wrinkles, eye, eyelash, pupil, eye brow, squint eye, nose, nostril, mouth, lip, tongue, tooth, canine tooth, molar, moustache, beard, cheek, ear, ear lobe, chin, neck, and throat.

The list as it stands is not complete. Elicitation may add a few more words which are lacking in Yukawa's list such as Adams apple, nape, goatee, crown. The same could be done for the rest of the items under human body and for the other semantic groups. Looking at Batibo's list of common words we note also that his list lacks mouth, nostril, jaw, molar, canine, tooth, occiput, crown, goatee, moustache, and nape in the semantic subcategory *head*.

Although Yukawa's list includes also cultural items such as food, tableware, house, family, marriage, domestic animals, wild animals, hunting, birds, fish, insects etc., Batibo's list is more detailed and specific to the animal husbandry category. For example, under *Animal Husbandry* he provides seven subcategories:

(6) Animal husbandry

cattle; other domestic animals; herding; grazing grass; livestock description; livestock colours, and use of livestock.

The *cattle* subcategory has 22 items about cattle alone:

(7) Cattle

cow, bull, steer, heifer, calf, herd of cattle, old cow, head of cattle, barren cow, lead cow, unweaned bull/calf, pack-ox, etc.

Yukawa on the other hand has only two: cattle and bullock under subcategory *domestic animals*. Despite all these, there are some gaps in this list too. The following words in Yukawa's list are missing in Batibo's list: to bark (as made by a dog), castrated male sheep, and young female sheep.

From the aforesaid it is clear that there is no list that is complete, and indeed there cannot be a list that can cover every aspect of every culture. For example, while Pare has words for a male sheep (*ndorome*), a castrated male sheep (*ngulata*), a male goat (*nzenge*), and a castrated male goat (*ngirazio*), other languages do not have these fine distinctions. Therefore, it is important that field linguists collecting lexical data should familiarize themselves with the culture of the language so that they look for such fine distinctive lexical items.

4.7 Data Collection

Linguistic researchers collect data that will provide the information they need. In order to achieve this goal, they organize their investigation in such a way that they can discover it most easily. It is thus important to briefly look at the various aspects and variables involved in the collection of data, especially research tools, informants, and methodology in general, all of which have direct effect on the reliability of the data.

4.7.1 Research Tools for Collecting Lexical and Phonological Data

At least two types of tools will be considered. The first type is the wordlist from both the basic/core vocabulary and speech community's cultural vocabulary, which will normally be collected from the various semantic fields of a language.

The second type of tool is a tape-recorder for recording lexical items provided by the informants. Informants should be native speakers of the language under study and, in addition to that,

- a. they should know the language well, be masters of pronunciation and not code-switchers; and
- b. they should have lived in the area where the language is spoken long enough to master the vocabulary in the various semantic domains.

4.7.2 Research Methods

Data can be collected by elicitation or direct interview, and by filling in a questionnaire.

4.7.2.1 Elicitation

This method is used to solicit information from the informants in the form of interview whereby the subject is requested or asked to provide specific information. Analytical elicitation techniques involve the use of open-ended questions to seek more and more information. The reverse translation technique involves use of the contact language. The following are the procedures of eliciting information:

- a. Read the text to the informants;
- b. Ask them to produce corresponding lexemes found in their language;

- c. Record/tape the oral responses and read or play the recorded responses for the informants to check their responses;
- d. Transcribe the phonological realizations of the sounds in phonetic symbols.

4.7.2.2 Questionnaires

Provide a questionnaire consisting of a list of vocabulary to the informants and ask them to provide equivalent words in their language for every word provided in the wordlist. The lexical data will then be administered to another informant to read aloud for phonological data that will be transcribed.

4.7.3 **Reliability of the Data**

A reliable and consistent set of data can be obtained if the informants understand the research objective. In order to avoid incorrect data, it is advised that data should be collected from more than one informant and verified by other informants. It is thus recommended that data collection be done in two phases, as follows.

Phase One

- a. Administer a wordlist to the informants and (tape) record or transcribe the phonological form of the item in phonetic symbols;
- b. Verify the data by having another informant go through it and comment on its validity and reliability;

Phase Two

- a. Administer the same wordlist to another informant and repeat the same process of recording and transcribing;
- b. Verify the data;
- c. Compare the data and make corrections if any.

References

- Batibo, H.M. (1989) Linguistic and Cultural History of Tanzania Questionnaire on Cultural Vocabulary. ms. University of Dar-es-Salaam.
- Byarushengo, E.R. (1975). An Examination of the Segmental Phonology of Haya. M.A. Thesis. University of Dar es Salaam.
- Bynon, T. 1983. *Historical Linguistics*. Cambridge: CUP.
- Crystal, D.1997: *A Dictionary of Linguistics and Phonetics*. Oxford: Blackwell.
- Kagaya R. 1989. *A Classified Vocabulary of the Pare Language*. Tokyo: Institute for the Study of Languages and Cultures of Asia and Africa (ILCAA).
- Kaji S. (2000) A Haya Vocabulary. Tokyo University of Foreign Studies. Tokyo: ILCAA.
- Lukula, A.M. (1991) The Nature and Extent of Linguistic Change Among Insular Jita in Ukerewe. M.A. Dissertation. University of Dar es Salaam.
- Mkomagu, J.D. 1991. Tone and Accent in Kisongwe. Unpublished M.A. Dissertation. University of Dar es Salaam.
- Mreta, A. (1990) The Problem of Bantu Linguistic Affiliation: The Case Study of Chasu, Kigweno, Kikahe and Kirombo. M.A. dissertation. University of Dar es Salaam.

- Mshindo, H.B. (1989) The Uses of ka- in Pemba Swahili Variety. M.A. Dissertation. University of Dar es Salaam.
- Samarin, W. J. (1967) *Field Linguistics; A Guide to Linguistic Field Work*. New York: Holt, Rinehart and Winston, Inc.
- Sapir, E. 1949. *Language: an Introduction to the Study of Speech*. New York: Harcourt Brace.
- Swadesh, M. (1955) Towards greater accuracy in lexicostatistic dating. IJAL 21:121-137.
- Yukawa, Y. (1992) *A Classified Vocabulary of the Luba Language*. Tokyo: ILCAA.

4.8 Discussion

In its discussion, the group on "Collecting Lexical and Phonological Data" was guided by eight questions. The following subsections present the reactions of the group members in relation to respective questions.

4.8.1 How many lexical items are needed?

After a long discussion on the number of lexical items required for each language, it was agreed that since the idea is to document and preserve the languages, and since the items will reflect different cultural and social aspects we should aim at collecting as many lexical items as possible. It was suggested further that the minimum number for each language should be at least 3000 words.

During the plenary session it was noted, however, that the number of words to be collected would differ from language to language and from field to field. Therefore, it may not always be possible to get the same number of words in each language. Moreover, it was recommended that in collecting our data, since some languages already have some written texts and vocabularies, we should begin with what is available.

One of the issues that provoked a prolonged discussion had to do with the treatment of borrowings. Should these be included or should they be excluded since they are foreign? It was the opinion of many members of the workshop that loan words that have fossilized and become part of the language ought to be included. Even in cases where you have two versions of loans, one representing an earlier form and the other representing a later form (e.g. *iwindo* and *idirisha* in Ci-ndali), both forms should be recorded.

4.8.2 What consequences would the distinction between a classified vocabulary and a dictionary have on the amount of data to be collected?

A clear distinction was made between a dictionary and a classified vocabulary list. As a general rule, a classified vocabulary list is something that is limited, not open-ended. It is selective in nature. The selection of the words to be included is done according to certain interests, such as semantic domains and/or what the researcher is looking for. In other words, a

classified vocabulary list depends entirely on the preferences of the individual researcher.

A dictionary, on the other hand, is not limited; it is open-ended, taking as many words as possible, as many semantic domains as possible, and many different fields as possible. That is why there are many different types of dictionaries.

As to the consequences of this distinction, it is clear that it is easier to deal with classified vocabularies than dealing with dictionaries. Moreover, since classified vocabularies are limiting, they can help one achieve one's goal within a shorter period of time. Of course this does not in any way mean that one cannot enlarge one's classified vocabulary list.

4.8.3 What form should the data collection/recording take?

Through the discussions, it became evident that as far as collecting phonological data is concerned, questionnaires have certain limitations and have to be used with caution. This is mainly because not every one can read and write. Secondly, in data collection that has to do with sounds of any particular language, the aspect of speaking is of vital importance. It is, therefore, necessary to make sure there are as many tape recordings as possible and, where possible, video cameras should be used. The questionnaires should be used only for eliciting answers through speaking, not writing by informants.

4.8.4 What other forms of collection?

As far as the aspect of other forms of collection is concerned it was recommended that efforts must be made to use existing written texts, poetry (both written and oral), and stories. In addition, any other means that can become handy could be used, since the idea is to capture as much data as possible.

4.8.5 How many languages should the glossing have?

With regard to this, the opinion of the majority was that the glossing should be done in Kiswahili and English. In other words, the language being studied will be on the left-hand side, and the glossing languages should be on the right-hand side.

4.8.6 What should be the age limit of the informants?

The researcher must know exactly what he/she has in mind and look for the most suitable informants. It was suggested that in order to get a wide variety of data informants should be classified in three groupings: from 13 - 35 years; from 36 - 50 years and from 51 years and above. It is very important to make sure that there is a gender balance, because men and women tend to use language differently.

It was suggested that at least 15 people for each one of the proposed groups should be interviewed. Since the sources will of necessity be different, any source should have a minimum of five people.

4.8.7 How should the data be verified to ascertain authenticity?

It was recommended that, as for data verification, there was no reason why the researcher had to go to the field twice; rather data verification should be done in the field at the same time when the research is being conducted. The approach would be for the researcher to check the information he/she gets by asking and eliciting from other informants in the field.

4.8.8 Where should informants come from?

In the field it is recommended that we use people from locations within the area of study. Moreover, one should try to see which places seem to have significant dialectal variations.

5. COLLECTING MORPHOSYNTACTIC DATA

By Dr. Yared Kihore, Institute of Kiswahili Research, UDSM

5.1 Introduction

The term *morphosyntax* is defined as “grammatical categories or properties for whose definition criteria of morphology and syntax both apply, as in describing the characteristics of words” (Crystal (1995:226)). Crystal considers the distinctions under the heading of “number” in nouns to constitute a morphosyntactic category and “perfect”, “indicative”, “passive”, “accusative” as some of its properties. Number, for example, is considered a morphosyntactic category because its contrasts affect syntax (e.g. a singular subject requires a singular verb) and because it requires a morphological definition (e.g. add *-s* for plural in English language). This means that the morphosyntactic data in question are mainly word classes in their various patterns of usage, especially where inflectional and derivational processes are involved through affix attachments. For the African language families present in Tanzania, the word classes in question are mainly nouns, adjectives and verbs all of which in their patterns of usage involve category of number and other inflectional and derivational processes that constitute the morphosyntactic properties listed above (i.e. perfect, indicative, passive, accusative, etc.). Examples of patterns of usage of nouns, adjectives and verbs displaying morphosyntactic properties in Bantu and Nilotic families are as shown in (1) below.

- | | | | |
|-----|---------------------|----------------------|---------------------------------------|
| (1) | Kiswahili | | Luo |
| | (a) M-toto m-zuri | a-me-ku-j-a | Nyathi ma-b ϵ r o-bi-ro |
| | child pref-nice | 3P-pperf-inf-come-md | child pref-nice pref-come-tns/md |
| | | | ‘A nice child has come’ |
| | (b) Wa-toto wa-zuri | wa-me-ku-j-a | Nyithindo ma-b ϵ -yo o-bi-ro |
| | | | ‘The nice children have come’ |

In these examples, the word order is that of a subject noun followed by an adjective and verb predicate for both languages belonging to different families. Also in both languages affixes bearing morphosyntactic categories and properties are attached to word classes in question. As we shall see later, such a process is not very common with other word classes in both these languages.

For many years, data on the elements such as those mentioned above have been collected by using word lists of various types and also obtained from some written or oral texts. The word lists were produced mostly in a language known to the researcher and used to collect corpora of equivalents in other languages for various linguistic research purposes. Early scholars such as L. Krapf (Griepenow-Mewis 1996:161-172) and M. Guthrie (1970) relied heavily on such lists in their research work. Among such lists is the one popularly known as “100 word list”. This list contained basic vocabulary considered as culturally core items in any language. These are items such as

references to parts of the body as well as a variety of cultural activities and artefacts. The list was used mainly to collect core vocabulary equivalents from different languages the availability of which signified language authenticity, that is to say, its sustenance of original development and characteristics. For this reason, the outcome based on this list (i.e. the corpora of equivalents in other languages) was widely used for comparative and contrastive purposes as well as in the historical reconstruction of linguistic forms of related languages. Greenberg's (1963) classification of African languages, for example, is seen to have involved "mass comparison of language vocabularies..." (Welmers 1973:1) obtained from both written and unwritten sources.

The main criticism against the word list approach is its being 'corpus-restricted'. This criticism has been advanced by Generative Grammarians who think it is important to project to the language as a whole in data collection. Another problem is that by basing the list on one particular language (often a foreign one), the collector runs the risk of missing relevant forms that are language specific (e.g. languages of livestock keepers having more general and refined vocabulary on livestock than languages of non-livestock keepers). Of late, lists have been broadened with a wide range of items to improve data collection of equivalent items in other languages (see Yukawa 1992). Although this approach can be applied to both written and unwritten sources of data, it still remains only one type of approach for morphosyntactic data collection. The other main question then is, are there any other approaches besides this one? In the following sections we shall consider various morphosyntactic data collection approaches and provide relevant examples of such data from a number of Tanzanian languages. Before that, however, we need to say a word or two on the sources of such data.

In the preceding section, we have mentioned that the sources of morphosyntactic data can be written or unwritten. With respect to written sources, we need to mention here that they are still quite meagre in quantity, content, and context. Of the quantity available, much of it is considered unreliable because a good portion of such sources is based on "travellers' reports" (Saville-Troike 1982:121). As such, much of the data is confined to certain categories and does not cover all languages. This leaves many languages uncovered or untouched. The unwritten source, on the other hand, is still mainly untouched.

5.2 Approaches to Morphosyntactic Data Collection

The only morphosyntactic data collection approach mentioned in the introduction above is that of the use of word lists. These, as we have noted, have always been corpus-restricted. That is to say, they cover only certain areas considered by the researcher or collector. There have been efforts to broaden these lists to include a wide range of items. This, however, does not seem to ascertain the coverage of all areas of grammar as envisaged by Generative Grammarians. And since many such lists are still based on some 'foreign' language, it is not known how much data in the target language has not been covered.

In spite of such shortcomings, the wordlist approach has remained the backbone of linguistic data collection. We note, for example, that some of the lists (like those referred to above) also tend to include patterns of noun-adjective agreements instead of listing the vocabulary only. Other wordlists, such as Batibo's (1989) also include brief questionnaires on basic constructions involving nouns, adjectives and verbs. Thus the use of such questionnaires could also be considered when doing morphosyntactic data collection.

With respect to questionnaires, we observe that besides those that appear as short attachments to wordlists (e.g. Batibo's 1989 list), which are used to seek equivalents in other languages, the others have mainly been those used to obtain morphosyntactic data from one particular language. The examples of the latter are the Institute of Kiswahili Research's (IKR) questionnaire for *Mradi wa Utafiti wa Sarufi Miundo ya Kiswahili Sanifu* prepared in 1993 for research on Kiswahili grammatical aspects taught in Tanzanian secondary schools and Kihore's (1989) questionnaire for research on morphosyntactic aspects of the Kiswahili verb. Both these questionnaires mainly sought information on the acceptability of various Kiswahili syntactic and morphosyntactic patterns. Again, like some of the word lists referred to above, both were also corpus-restricted.

It seems to us that what is needed for this exercise is an elaborate questionnaire to be used to seek equivalents in other languages. Such a questionnaire may need to focus only on noun-adjective patterns and verbal inflectional and derivational patterns. These, as we have indicated in section 5.1 above, appear to be the bearers of various morphosyntactic elements, at least, insofar as the examples from the two language families (Bantu and Nilotic) in (1) show. Sketches of Cushitic and Khoisan grammatical patterns may also have to be prepared to see if there is need to focus on other word classes too. The other problem concerns the language on which the questionnaire is to be based. We showed earlier that basing word lists or questionnaires on some 'foreign' language may lead to failure to capture some relevant concepts or patterns in other languages whose equivalents are being sought. However, since it is just the beginning, general details may just suffice. Some of these other fine details can come at later stages. So far we are not certain to what extent the use of one member of the language family as a questionnaire base may help resolve the issue of capturing all the fine details.

We also need to consider, briefly, the question of data collection procedures. With respect to this, what we think we need to say is that, since we strongly recommend the questionnaire approach, such procedures should focus on questionnaire preparation and questionnaire administration. With regard to questionnaire preparation, we think it is important to consider the procedure known to discourse analysts as "introspection". This according to Saville-Troike (1982:119-20) is a means of data collection only about one's own speech community. In this procedure, members of the speech community answer questions about various aspects of language and culture and may be asked to formulate very specific answers from their own experience. In our case this would mean involving members of a speech community in

preparation of details that go into the questionnaire. Using this procedure may help in obtaining the fine language specific details referred to in the preceding paragraphs which normally tend to escape the outsiders. Questionnaire administration, on the other hand, could just be carried out by any competent researcher/collector. Other data collection procedures such as “participant-observation” and “observation”, do not, in our opinion, seem to be necessary. However, since almost all African languages have phonological characteristics that cannot be handled adequately by the existing orthographies, questionnaire filling should be accompanied by recorded sound backup. The question of inadequate orthography is also a problem with respect to written sources.

5.3 Examples of Data

Below we give a few examples of the morphosyntactic data we expect from the field. Here we assume that our questionnaire, based on Kiswahili, is used to collect equivalents from Kihacha (a Bantu language like Kiswahili) and from Luo (Nilotic family):

(2) Nominal patterns

Kiswahili	Kihacha	Luo	Gloss
mtu/watu	umuntu/aβantu	dhano/ji	person/persons
mguu/miguu	ukuguru/amaguru	tielo/tiende	leg/legs
jino/meno	eriino/amino	lak/leke	Tooth/teeth
ndege/ndege	inyonyi/iβinyonyi	winyo/winy	bird/birds
mti/miti	umuti/εmiti	yath/yien	tree/trees, etc.

(3) Verbal patterns

Kiswahili	Kihacha	Luo	Gloss
-piga	-tema	go	beat/hit
(u)nipige	(u)nteme	goya	beat/hit me
(u)mpige	(u)mteme	go(y)e	beat/hit him/her
(u)tupige	(u)tuteme	gowa	beat/hit us
(u)wapige	(u)βateme	gogi	beat/hit them
amenipiga	yantema	ogoya	he/she has beaten me
nimepigwa	ntemiwi	ogoya	I have been beaten
anani-piga	arantema	ogoya	He/she is beating me

The examples in (2) and (3) above indicate nominal and verbal patterns of usage for three languages, two of which belong to the Bantu group and one to the Nilotic family. The examples in (2) show singular/plural alternation, while those in (3) show verbal inflectional and derivational patterns. Take

note that the Kihacha /t/ is dental. Take note also of the differences between the Luo singular and plural forms in (2) and that its last three verbal forms in (3) have different tone patterns (though not indicated). These are examples of the elements that cannot be taken care of adequately by the existing orthographies.

The following noun phrase patterns also show further examples of morphosyntactic data involving the adjective category:

(4) Noun phrase patterns

	Kiswahili	Kihacha	Luo	Gloss
a.	kitabu kizuri vitabu vizuri	ekitaβu kiiya eβitaβu βiiya	buk maber buge mabeyo	a nice book nice books
b.	mtu mnene watu wanene	umuntu munene aβantu βanene	ng'at mchwe ji machwe	a fat person fat people

In (4) we see examples of noun-adjective patterns in the three languages. These examples also display morphosyntactic properties/categories that form the needed data. An exercise of the sort could be extended to other language families present in Tanzania to help researchers know the type of linguistic elements that constitute morphosyntactic data.

Finally, we note that, at least for the language families surveyed here, other word classes such as conjunctions and adverbs, do not involve affixation in their patterns of usage. The examples in (5) below show the patterns involving the conjunction “and” in the three languages:

(5)	Kiswahili	Kihacha	Luo	Gloss
	mimi na wewe	uni na uwe	an gi in	I and you

In (5), the conjunction forms for “and”, appearing in boldface characters for all the three languages, do not change form, which is an indication that their patterns of usage do not involve any morphosyntactic properties or processes.

5.4 Conclusion

We have attempted in the preceding sections to define what is meant by morphosyntactic data. We further described the various approaches and procedures that are used or could be used in collecting such data. We have also provided examples of patterns of usage of nominal, adjectival, and verbal word classes for three languages, which are composed of various morphosyntactic categories/properties. It is unfortunate that we could not provide examples from Cushitic and Khoisan language families. But it is our hope that the discussion and the examples can form the basis for further work on the collection of morphosyntactic data for languages belonging to different families.

References

- Batibo, H. (1989), Unpublished Word list.
- Crystal, D. (1995), *A Dictionary of Linguistics and Phonetics*. Oxford: Blackwell.
- Greenberg, J. (1963), *The Languages of Africa*. Indiana University Research Centre, No. 25.
- Griekenow-Mewis, C. (1996), "J. L. Krapf and his role in researching and describing East African Languages", *Swahili Forum* AAP No. 47.
- Guthrie, M. (1970), *Collected Papers on Bantu Linguistics*, London: Gregg International Publishers Ltd
- Kihore, Y. (1989), Unpublished Questionnaire on morphosyntactic aspects of Kiswahili verb.
- Saville-Troike, M. (1982), *The Ethnography of Communication*, Oxford: Basil Blackwell.
- Welmers, Wm. (1973), *African Language Structures*, Berkeley: University of California Press.
- Yukawa, Y. (1992), *A Classified Vocabulary of the Luba Language*, Tokyo: Institute for the Study of Languages and Cultures of Asia and Africa (ILCAA).

5.5 Discussion

5.5.1 General Discussion

The first participant to contribute to the discussion of this presentation posed what he called expected questions. These were, according to him, do we need a questionnaire for morphosyntactic data? Do we need one questionnaire for the whole project? Some answers to these questions were suggested as follows. First, it is important for us to acknowledge the fact that African languages are so diverse that we cannot rely on one questionnaire across the board. Languages are different and as such we may need different sets of questionnaires. We might need help from experts in the various language families in Africa who have done a lot of work on those language families or family groups.

It was also pointed out that research has been commercialised, which casts some doubt on the reliability of data sources. Informants are increasingly becoming difficult to part with necessary information for free, and researchers are increasingly finding it difficult to get data because in many cases money for the informants is not budgeted for. It was also observed that data reliability is affected by the difficulty in deciding who ascertains that a certain individual is a reliable and appropriate informant. All in all however, it was stressed that, as researchers, we must strive to identify reliable informants, and we must have some insight into what constitutes appropriate data. It was also stressed that it is better to rely on a group of people rather than individual informants, and data should be collected in the source area for ease of crosschecking.

It was agreed that for the Bantu languages one questionnaire would suffice, but there is need to work out appropriate questionnaires for the other language families represented in Tanzania. Thus, contact with colleagues who have worked on these languages will be essential.

5.5.2 Discussion of specific questions

Question 1: *What essential areas must be covered by a questionnaire for morphosyntactic data?*

A questionnaire for morphosyntactic data is expected to cover markers of syntactic aspects that are readily visible in the syntax of a language. These may include affixes and single words. In order to know the type of form, structure, or position, of these elements, we need a model. Some work has been done on Kiswahili, but for the other languages we need to develop a model. We also need to find out where or in which areas these morphosyntactic elements are found. Affixes, for example, are found in noun categories and their satellites and in verb categories and their satellites.

With noun categories, these elements are seen broadly as class elements before the radical and derivational elements after the radical. With verb categories, derivation markers (e.g. causative, benefactive, mood, etc) tend to appear after the radical, while elements marking the subject, object, tense, negative, etc appear before the radical. All these have to be captured and recorded, and researchers must be aware of these and be sensitive towards maximizing their recording. It was also pointed out that it may be important to find out the characteristics of a language on the basis of word order (e.g. verb-initial, subject-initial, etc.), and this is where the issue of examining sentence structure comes in.

Question 2: *What is already available (questionnaires) for the various language families represented in Tanzania?*

Library work and website visits will have to be carried out to find out what and how much has been done in this area. Batibo's questionnaire is readily available.

Question 3: *How can morphosyntactic data be handled/stored for efficient retrieval/analysis?*

This depends on the availability of funds. Modern technology should be used for storage and easy retrieval. It should also make it possible to store data for a long time (e.g. on CDs).

Question 4: *What are the most likely problems in the collection and analysis of the data?*

The most likely problems include the following:

- Transcription might be a problem because it is not standardized, and, as a result, different people code data differently.
- Failure to identify reliable informants may be a problem: who speaks a certain language, and in which area is a language spoken?

- Sending researchers in areas where they are familiar/insiders: this is only important for inside information, but for the atlas it is not necessary.
- Information has been commercialised and as a result more resources for informants are needed.
- At the analysis stage there may be problems in identifying elements like allomorphs, morphemes, and morpheme boundaries.
- In many cases, there are no appropriate translations of the languages that will be studied. This may affect the way questionnaires are set, interpreted, and understood by the informants and the researchers themselves.

Question 5: *How much data should be collected in various forms (written questionnaires, spoken texts, etc).*

We should try as much as possible to collect both written and spoken data, but emphasis should be on capturing spoken discourse.

Question 6: *Should an equal amount of data be collected for all languages or should we consider detailed data for only those languages for which we plan to write a grammar?*

We need to collect detailed information for the languages we wish to write a grammar. For the languages that we need only for the atlas we do not need very detailed information.

6. DESCRIPTIVE SKETCH OF A LANGUAGE

By Prof. Ruth Beshu, Kiswahili Department, UDSM

6.1 Introduction

In the context of the "Languages of Tanzania" project, a description of a language has a specific focus. It must help the analyst to make a decision about the "status" of the language compared to others, which are assumed to be similar. In the end the project should be able to answer the question: how many languages are there in Tanzania, and which are they. It must be able to cluster the present assumed languages into clusters of relationship as well.

The sketch should, therefore, bring out, not only the major features which the analyst thinks are similar to other languages in the cluster, but must also deal at length with those features which are unique to the language under description. In dealing with Bantu languages, in particular, there is a lot which is taken to be common, so there is a real danger of stressing the more common features which are taken to be "universals" in these languages. While no grammar, no matter how sketchy, can ignore these basic features, on their own, they will not be very interesting, in the sense of being the basis of developing theories. It means they should also give some real insight into those languages.

This position is very traditional, as it has its basis in the concept of "linguistic relativity". While the world is one, and human beings have many shared views, these human beings organize their world differently in their own space and time relations, and these differences are embodied in the system of their language. So it must be the case that each group of human beings, collectively, make certain distinctions through their language, which are not necessarily replicated in other languages. This will mean treating each language as an individual entity, in the real structural sense, and providing as much information as possible. Only when we have useful descriptions will we be able to make useful statements about the languages of Tanzania. In the following sections, these aspects will be further elaborated.

6.2 The major Aspects in a Descriptive Grammar

6.2.1 The Sounds and Sound system of a language

The phonemes of a language have to be established, as well as the phonemic variations, and the allophonic alternations. For example, even if it is established that a language has five vowel phonemes, the grammarian should still search for the different contexts in which those vowels occur, note any peculiarities, and make statements on any predictable occurrences permitted and blocked sequences, and position in a word. The same has to be done for the consonants as well.

Tone is very important in many Bantu languages, and it is known that many of these languages have both lexical and grammatical tone. This feature has to be given a lot of weight in any description.

Phonological processes/changes must be described, and these are among the most interesting features, which could bring out the uniqueness of

each language system. The morphophonological processes which occur are central to the language structure and will form a good base for comparison. These should be carefully dealt with.

6.2.2 The Structure of the Words

Augmentative and diminutive word forms are problematic, and it would be interesting to note the basis for their classification in particular classes.

The noun: the classes have to be established. But the basis for this classification has to be clearly shown, without any theoretical pre-judgements. For example, the semantic basis of the nominal classes seems to be "pervasive" in many languages, and should be explored; the kinship terms are another area.

The verb: the morphology of verbs, and the verbal extensions in particular, has received a lot of attention. While the forms are important and must be described thoroughly, the description should note the possible restrictions in the occurrence of these forms with particular verbs, as well as the differences in meaning that result.

Other word classes/categories: these include the adjectival words (adjectives, demonstratives, possessives, etc), the adverbials, locatives, and others. These should be subjected to similar scrutiny like in the above categories.

Other grammatical categories: here I have in mind the categories of tense, aspect, and mood. This is an area which has not been the focus of attention for many of the languages, Kiswahili being one of the few exceptions.

6.2.3 The Grammatical constructions

The structure of the sentence: here attention should be drawn to the basic word order and permitted variations, cohesive markers in complex sentences, subordinate and coordinate structures, and others. Aspects needing particular attention include relative and conditional structures as these show a wealth of distinctions.

The verbal construction: this is a rich area which needs to be described; one could deal with the interesting features of animacy if these arise in the particular language, for example, in the area of object markers.

In the sample of a sketchy outline grammar of Kishambala, some of the points noted above are illustrated. Although it is not as comprehensive as I would have wished, I hope it can form a basis for discussion

6.3 Discussion

6.3.1 General Discussion

It was pointed out that we must get decided on how to analyse our data and how we want to treat each language aspect that we will be dealing with. Pre-nasals and nasalization were given as examples, and a question was asked as to whether we should look at these as biphonemic or

monophonematic. The answer partly depends on the expected output as well as the target audience. Purely descriptive grammars that raise intricate theoretical issues (just for linguists) may not be the desired outputs. Products should also be accessible to school teachers and students and other researchers who may not be linguists. While the sketches should not be very complicated, they should also not be too simplistic. It was agreed that the outputs expected are not pedagogical grammars.

It is doubtful that we could work out a grammatical sketch that would be acceptable and reproduceable with different languages. It would be helpful to have as many sketches as possible; these would provide some models that other linguists/researchers could consult and use in their various situations.

It was observed that a project like this one has got various stakeholders, each group with their own interests and positions. This should not be seen as a problem; it is fine and it will continue to be the case. What is important is for the Coordinators to decide on how to reconcile the different positions by setting priorities.

6.3.2 Discussion of specific questions

Question 1: *What common framework/model is appropriate for describing the different languages?*

It was pointed out that what we are aiming at is not a grammar book, and as such there is no grading of the materials. We should make it as simple as possible. As for models/frameworks, it was pointed out that, because the languages involved are different, it is most likely that there will be different formats. Descriptions should be written in the context of the relevant languages. It was noted for example, that these descriptions must take into account aspects like: is the language relevant, is it living, or is it dying. This, it was observed, will give the descriptions authenticity and relevance. As far as phonology is concerned, we need to have a phonological system based on phonetic transcription. This implies that people must know the phonology of the language. It was stressed that this does not imply data analysis; it merely refers to transcription.

Regarding orthography, which is a system used to communicate in writing, it is obvious that for most of the languages it is still lacking. Perhaps here we should adopt what is already there, modify what is there or work out new systems. It was suggested that a combination of informed linguists and people trained to be linguists must help in solving the problems related to orthography. This should be accompanied by the presentation of tonology.

With morphology, there is need to work out a new definition of a 'word', and this calls for stating the criteria for such a definition. Perhaps we need a small section in our description on the sentence structure before describing morphology. It should also be born in mind that word classes/categories are very fluid, specifically due to the changing nature of language. In examining the word structure, its various components will have to be made clear; for example, some parts are inflectional, while some others

are derivational. One has to be able to decide which are the former and which are the latter.

Syntactic considerations will have to be made and various levels will have to be considered. This will be done at sentence structure level, clause level, phrase structure level, and so on. It will also have to be made clear as to whether one is dealing with phrase structures (NPs, VPs, PPs, APs, etc) or constituent structures (VP = V+NP+; or NP = (Det) + N+ A, etc.). Statements will also have to be made regarding the functions of the elements (e.g. Subject, object, complement, etc). Clauses will have to be presented along the same lines: whether they are equivalent to sentences, their function, nature, grammatical positions, and the like (subordinate, complex, relativization, etc).

Other elements that will have to be included are tense, aspect, and mood; meaning and semantics (which is going beyond syntax), taking us all the way through to information packaging, thematization, cohesion and coherence. In all this examples are needed and they must be glossed accordingly. We will have to use the same abbreviations, cross-references, index, tables of contents, and nice constructions for our structures. All this should be done with the connection/relationship between morphology, phonology, syntax, and semantics kept in mind. A concern was raised that morphophonology has been traditionally given peripheral treatment. It is almost always hidden in a chapter on morphology. It was pointed out that perhaps this is the right time to correct this anomaly.

7. THE LANGUAGE ATLAS

By Prof. K. Legere, Gothenburg University, Sweden

7.1 Introduction

The first point of departure for this project should be the contribution of linguistics to the description of languages. Very little has been done in the area of systematic language description. While we admire activists working hard to save endangered animals, efforts to save endangered languages leave a lot to be desired.

The second point of departure is the Harare Plan of Action for the Promotion of National Languages in Africa of 1997. The Plan has ten priority activities. Activity number 4 says that the objective should be a typology and inventory of all African languages. In this regard, our starting point should be to establish an inventory and put it on the map. UNESCO and OAU have also urged the production of language maps and our project is therefore pointing in the right direction.

It should be cautioned, however, that it is very important that we decide right from the beginning as to what we actually want from the project. There are seven possible outputs that we may want to come up with, namely:

- A. A language/linguistic map.
- B. A particular language map.
- C. A one language (e.g. Kiswahili) map, and this could cover more than one country, e.g. Kiswahili in Kenya, Tanzania, Uganda, Zaire, Burundi, etc.
- D. Various language maps.
- E. Historical and comparative language maps.
- F. Kiswahili dialects on the map.
- G. A map showing the dynamics of language shift, death, etc.

7.2 Questions to be addressed

There are basic questions that the project should address. The main question is: what kind of a map do we want for Tanzania? We need an inventory as a first step, but this also leads to other questions such as do we have an inventory on all languages? The answer to this question is definitely not. The next question to this would be: do we have any sources where we can obtain language data to determine this inventory? We should remember that the last reliable source was the 1967 census; it is doubtful whether we can still rely on this information. We couldn't rely on ethnologies because sometimes people claim to be what they are not. It is unfortunate that in the past people never took ethnic languages as 'languages'. This is the reason why people would concern themselves with being literate in English, Kiswahili and any other language which is not an ethnic community language (ECL). We could take the example of Mozambique, where in the past everybody was proud of being a Portuguese speaker, but now the situation has changed in favour of the ECLs.

Regarding reliable sources for information, we could even form a pressure group to lobby for the inclusion of language questions in the coming national census. It is important that the workshop participants discuss this issue and pursue it accordingly.

It is important to reiterate the earlier discussion about the inclusion or otherwise of Kiswahili on the map. Any map of the languages of Tanzania will be incomplete without Kiswahili. Perhaps what is lacking at the moment is appropriate terminology for ECLs; for instance, of late we have heard people using all sorts of words to refer to minority groups in the population, such as Ethnic Albanians in Europe. As far as data is concerned, we should concentrate more on the data that is necessary for a meaningful description than on language internal matters that may be seen as trivial.

Finally, let us note that currently the French have been asked to work on the language atlas project for Africa on behalf of the OAU and UNESCO.

7.3 Discussion

7.3.1 General Discussion

It was observed that conferences make excellent statements (e.g. the Harare Initiative) but, due to lack of resources, nothing is done to implement them. Still such statements may serve a purpose as reference points for committing resources to a particular cause. A contrast was drawn between countries like Botswana, which is moving from monolingualism to multilingualism, and Tanzania, where a monolingual preference for Kiswahili is dominant. In Botswana, each language is to be assigned a role, but in Tanzania, the project aim is preservation since no roles are being assigned to the ethnic community languages in the society. Preservation, it was noted, is not synonymous with prevention of language death; what linguists can do is simply to record the status of the languages and thereby provide data for future work on these languages. Only the speech communities can keep their languages alive or allow them to die.

The possibility of redrawing existing maps and renaming/regrouping existing speech communities so as to encourage/promote larger language groups was discussed. For instance, it was suggested that, in line with proposals by scholars at Makerere University, a *Runyakitara* language could be recognised. It would put together such 'dialects' as Runyankore, Ruciga, Runyoro, and Rutooro (in Uganda), and Runyambo, Ruhaya, and Rusubi (in Tanzania). However, it was felt that the project should desist from such radical departures and stick more or less to existing groupings. It was observed that combining languages might amount to denying people their right to claim that they belong to some ethnic grouping. We as researchers should respect the people's wish, and whenever possible we should show the mutual intelligibility among the language in the descriptions. A further observation was made that some of the languages are actually groupings based on activities, clans, etc. The question is whether we should get into the complexity of all this? It was noted that we should only engage in splitting

languages if there is enough linguistic proof of the differences at language level.

7.3.2 Discussion of specific questions

Question 1: *What kind of Atlas do we want?*

Several levels of maps were identified which would contribute to the language atlas of Tanzania:

- A country-wide map of all the languages in the country.
- Regional maps showing major language groups (i.e. Bantu, Nilotic, Cushitic, Khoisan) as well as their sub-divisions
- Regional/district maps showing specific languages and their approximate locations
- Maps showing languages and their respective dialects

It was generally agreed that the degree of detail in making these maps will depend on time and resources available. Thus, priority should be on the national/regional maps showing language groupings and specific languages. Language dynamics (e.g. the influence of big languages like Kiswahili, Kingoni, Kisukuma, etc on smaller languages) should be captured on the maps. There should also be a way of showing areas of overlapping languages/dialects on the maps

Question 2: *What data do we already have?*

It was emphasized both in the group discussion and the plenary session that it is important to get data, which is already available. Such data would include an inventory of all the languages of Tanzania, work done by other researchers in this area (e.g. maps drawn by the Summer Institute of Linguistics etc.) Consolidation work could then be undertaken to add to what is already available.

Question 3: *How do we proceed getting new data?*

Language descriptions and other data obtained during the data collection phase of the project will be an important input into the language atlas. Further, data could be obtained through a questionnaire distributed to as many schools/colleges in Tanzania as possible, by using University of Dar-es-Salaam students during Teaching Practice sessions. It was however cautioned that unsupervised students may “cook” data. If this method has to be used, it should be thought out carefully lest we get unreliable data.

Data obtained from the field will have to be given to cartographers as input into the drawing of the maps. It was suggested that local cartographers should be contacted in the first instance (e.g. IRA and Geography Department of the University). Only when local resources are not available or are inadequate should we seek outside help. It was also suggested that outside institutions working in Tanzania (e.g. SIL) could be contacted for possible collaboration/assistance.

Question 4: *Do we need a language question in the forthcoming national census?*

It was generally agreed that a language question in the forthcoming census was important for the project. The Project Management Committee should therefore find ways of lobbying for such question(s) to be included in the census. The justification for inclusion of a language question into the census would be two-fold:

- a. The *Cultural Policy* which was adopted by the Government in 1997 explicitly says the ethnic community languages of Tanzania are important as our cultural heritage. To preserve this heritage, we need to document these languages.
- b. The ethnic community languages have a *pedagogical importance*: we need to know the linguistic resources that children come to school with in order to help them better.

It was suggested that the Project Coordinators together with one or two "lobbyists" should work out the appropriate question(s) and try to sell it/them to the people in charge of the census.

Whether or not there is a language question in the census, demographic data collected by the census could still be used to estimate the number of speakers for each language.

Question 5: *How do we treat multilingual urban centres?*

It was agreed that it is impossible to represent language distribution by location in urban centres because people do not live according to ethnic groups! Only the number of speakers of different languages can be documented.

Question 6: *How should Kiswahili be treated?*

Individual maps can be made showing areas where Kiswahili is spoken as a mother tongue, where it is spoken as a second language, as well as areas with a strong or weak presence of Kiswahili.

8. WORKSHOP OVERVIEW

By Prof. H.M. Batibo, University of Botswana

8.1 The expected outputs fall under three categories

- a. The maps: these will include the main map, the regional/district maps, and other language specific maps.
- b. Descriptive Grammars: it was agreed that these should be comprehensive enough to capture the major features of the languages.
- c. Classified Vocabularies: there was a consensus among participants that there is need to expand the vocabulary lists to go beyond the 3,000 minimum required.

8.2 Workshop Evaluation

This was a very resourceful workshop. Its strengths were derived from a well organized format, well chosen and researched presentations, and plenty of discussion time. Two issues need further consideration: The first concerns capacity building, in terms of bringing together the existing experts and giving them a short training. This way, we can be sure of having reliable people in a short time. It should be cautioned that we should not concentrate or rely only on the M.A. training component because this is not capacity building in the sense that we need it for this project. The second issue concerns the need to revise the project schedule. The possibility of including more languages to be covered in the first phase of the project should be seriously considered. It may be possible to spread out the study to more parts of the country instead of focusing on the North West.
