Technical Reports on Language Technology
Report No 72, 2021
http://www.njas.helsinki.fi/salama

Linguistic distance of Swahili speech varieties[1]

Arvi Hurskainen
Department of Languages
FIN-00014 University of Helsinki, Finland
arvi.hurskainen@helsinki.fi

Abstract

In this report I study the speech varieties of Swahili in the coastal area of Tanzania and on the nearby islands of the Indian Ocean. The study material consists of the tape-recorded speech of various people. The material was collected as part of the so-called DAHE project in 1989 - 1991. The aim of the project was to initiate the Computer Archives of Swahili Language and Folklore. The counterparts of the project were the University of Helsinki and the University of Dar-es-Salaam. The project produced two types of material, (a) recorded speech on various topics, and (b) wordlists from various speech areas, 620 words in each list. In this report, the first type of material is discussed.

Key Words: *linguistic distance.*

1 Introduction

In the DAHE project we wanted to study the variation of speech in the coastal Swahili-speaking area of Tanzania. The speech variety of Zanzibar town is generally considered the standard form of Swahili. When we move away from the city, we find speech forms that deviate from the standard form. There has been tendency to classify the deviant speech forms as dialects. This concept is, however, problematic, because often the speech changes gradually, when we move further from the centre. It is often not possible to demarcate the boundary between two speech varieties. This is particularly the case when the speech varieties belong to the same language group, such as Bantu languages. The case is very different, if the speech varieties belong to different language groups. For example, the difference between Swahili and Maasai is very clear. They have hardly anything in common, and they can be classified as distinct languages.

The situation with speech varieties among the Bantu-speaking people on the coast is different. There are several speech varieties, and it is difficult to decide whether a particular speech variety is a dialect or not. Traditionally, such speech varieties of Swahili as Kimakunduchi (or Kikae), Kipemba and Kitumbatu have been identified as dialects of Swahili. However, no systematic study has been done for supporting the classification.

[1] The report is issued under licence CC BY-NC

In this report, I will make the comparison of the speech varieties using the Swahili morphological analyser for finding out the degree of difference between the Standard Swahili and the particular speech variety. The analyser was first constructed in 1985 and it has been constantly improved and updated. Therefore, it can be considered as a reliable criterion for concluding whether a word is a standard Swahili word or not.

In the DAHE database, each speech episode of the speaker forms a single line. In front of the line, there are three kinds of code, (a) one for identifying the language (or speech variety) of the speaker, (b) one for identifying the speaker, and (c) one for identifying the sex of the speaker.

In the case of interviewees, the identity of the speaker is encrypted, and only the code can be seen. The encryption was done for ensuring the privacy of each speaker. On the part of the interviewers, encryption was not done, and I speak about them using their real names.

In all, there is speech of 253 interviewees in the corpus. There are seven main interviewers, and also some ad hoc interviewers.

2. Speech varieties

The study of speech varieties on the basis of the corpus material is problematic, because there are two ways of dealing with speech varieties. In one method, the speaker is asked to use the particular speech variety when answering questions, or stelling histories, stories and so on. This material is ideal for studying speech varieties, and such sections were encoded using the relevant language code.

In another method, the interviewer and the respondent discuss about speech varieties using Standard Swahili. In these episodes, standard language and local speech variety are mixed. The latter types of speech were encoded as Standard Swahili, although they contained also non-standard words.

In all, the corpus contains 11 labelled speech varieties. As we can see in Table 1, the list also contains speech varieties that are not varieties of Swahili. Kikae, Kipemba, Kitumbatu and Kinungwi can be considered as Swahili dialects, due to their closeness to Zanzibar town. The other speech varieties are from southern coastal area, and people do not count them as varieties of Swahili. On the basis of this assumption, it is expected that Swahili dialects are closer to Standard Swahili than the other Bantu speech varieties.

However, the results in Table 1 do not seem to support this assumption. We cannot make direct conclusions on the basis of Table 1, because the source material has several sources of bias.

First, the speech sections marked as a certain speech variety are not necessarily examples of that speech variety only, because they often contain words and even speech sections in Standard Swahili. The problem in recording the material was that none of the interviewers spoke any of the speech varieties except Standard Swahili. This affected the discussions, and often the interviewee slipped to using Standard Swahili, and the interviewer reminded that the interviewee should speak the local speech variety, although the interviewer used Standard Swahili. The speech situation was unnatural, and the interviewee, being fully fluent in Standard Swahili, slipped to using it.

There are also songs in local speech, and these are good material for representing the speech variety. However, such material is too scarce for proper analysis, and there are songs only in part of the speech varieties.

Second, there were problems in marking the speech variety of each speech section. It was assumed that if the interviewer asked the respondent to use the local speech variety, the subsequent speech sections of the respondent were marked with the code of that speech variety. It may have happened that the respondent slipped to using Standard Swahili until the interviewer reminded about it. Also the inability of the interviewer to understand the speech variety forced the respondent to use Standard Swahili for clarifying the message of the speech.

Third, the ability of the person to speak the given speech variety differs. It often turned out that the local people were well aware of those, who mastered well the local speech, while others mastered it only partly, and many expressions were just Standard Swahili.

Table 1. Distance of speech varieties from Standard Swahili. Columns 2-4 contain the number of words in the corpus.

Language	Different	Similar	Total	% different
Kikae (Ka)	14554	30223	44777	32.50
Kipemba (Pe)	492	2255	2747	17.91
Kitumbatu (Tu)	2899	6740	9639	30.07
Kinungwi (Nu)	1108	2073	3181	34.83
Kimwera (Mw)	239	1000	1239	19.29
Kimalaba (La)	3242	4877	8119	39.93
Kimakonde (Ko)	428	2920	3348	12.78
Kimtwara (Mt)	300	512	812	36.95
Kimasoko (So)	2223	10418	12641	17.59
Kingindo (Ng)	65	69	134	48.51
Kiswahili (Ki)	25658	91266	116924	21.94

Table 1 above displays the size of each speech variety in DAHE corpus. In the leftmost column is the name of the speech variety and its code in the corpus. In the next column is the number of deviant words. In the following column is the number of such words that are identical with Standard Swahili. The following column shows the total number of words. The last column shows, how many percent points each speech variety has non-standard words.

The data were modified so that all such utterances in speech that are not real words were removed. Also codes, punctuation marks, and diacritics were removed, so that only real words were left. The analysis was made on the basis of real words.

When we look at Table 1, we must take the above limitations into consideration. The speech varieties Kikae, Kipemba, Kitumbatu and Kinungwi should be closest to Standard Swahili. Three of them have more than 30 percent of their vocabulary non-standard Swahili. Kipemba is an exception, and this is due to the data, which is not ideal for comparison. Kikae has more data than any other speech variety, and its results are quite reliable. That is, a third of vocabulary in running speech deviates from Standard Swahili.

When we look at other speech varieties, we make a surprising observation. They do not differ much from the Swahili dialects. The data on the part of these languages are heavily biased. When we look at the speech sections marked as representing a given speech variety, only occasionally the speech is that language that it should be. The speech concerns the language, but the language used in discussion is often Standard Swahili.

We see that even the sections marked as Standard Swahili (Ki) have more than 20 percent non-standard words. This group includes the speech of all interviewers as well as much of the speech of the respondents. Also interviewers used often non-standard words, when they worked with the word lists and confirmed aloud what the respondent had said. Such words became part of the speech of the interviewer. We will discuss about this more in conjunction with Table 2.

Kingindo has half of its vocabulary non-standard Swahili. It does not mean that it would be more distant from Standard Swahili than other speech varieties. The text in Kingindo just happened to be songs in that language, and it was not distorted by intervening Swahili.

3. The language of interviewers

The second aim of this report is to study the language of the interviewers. It is assumed that each of them uses Standard Swahili. Therefore, the share of non-standard forms should be minimal. However, the results in Table 2 show big variation. Differences are often not due to insufficient language skills. I will discuss each interviewer individually.

Table 2. Don-standard words in the speech sections of the interviewers. Columns 2-4 contain the number of words in the corpus.

Interviewer	Different	Similar	Total	% different
Madumulla	631	6725	7356	8.58
Sengo	737	14928	15665	4.70
Massamba	516	4592	5108	10.10
Hurskainen (a)	151	3492	3643	4.14
Hurskainen (b)	99	6189	6288	1.57
Stude (Ki)	411	2199	2610	15.75
Mlacha	68	589	657	10.35
Yambi	6	476	482	1.24

Madumulla: He was working together with Stude, a Finnish researcher, who still was improving her command of Swahili. Madumulla used sometimes English for making the message understandable for Stude. This largely explains the rather high percentage 8.58.

Sengo: Sengo is the most productive interviewer, and the percentage 4.70 of non-standard words is reasonably low, but not low enough for passing without comments. Sengo uses such Arabic expressions as *ewallah*, *alhamdullillah*, *taib*, etc. in discussions. Also defective words that appear in speech contribute to the result.

Massamba: Massamba has 10.10 percentage points non-standard words. This is largely due to his tendency to go through wordlists with respondents. He often repeats the word that the respondent said, to make sure that he had heard correctly. This adds to the number of non-standard words in his speech.

Hurskainen: Hyrskainen has a rather low percentage, but it could be still lower. When we inspect the data, we see that there is a section, where a wordlist is handled, and Hurskainen uses non-standard words when confirming that he has heard correctly. If we remove this section, the result will be 1.57 percent non-standard words.

Stude: Stude has a high percentage of non-standard words. There are such sections in the corpus, where Stude uses English. Such sections were marked with the code *En*, and they are not included into this study. Nevertheless, the percentage is over 15. This is mostly due to the fact that also in the Swahili sections there are English words.

Mlacha: The speech sections of Mlacha contain non-standard words (10.35%), when he discusses about Pemba speech varieties. This explains the high percentage.

Yambi: The speech sections of Yambi are without discussions on speech varieties. As a result, the percentage of non-standard words is very low (1.24%).

4. Discussion and conclusion

In this report, I made a test on how the morphological analyser could be used for extracting information on speech varieties in settings, where people use different varieties of speech. Many of the speech situations were such, that the respondents were specifically asked to use local speech varieties in discussions. The speech material so compiled was a mixture of speech varieties, where the speech in Standard Swahili is mixed with local speech varieties.

The result of the test is that it is very difficult to extract reliable quantitative results from this material. It was difficult to mark the speech sections uniquely, so that the code would fully correspond to the content of the speech section. Fully clean speech sections were too rare for making any reliable conclusions on the distance of each speech variety from Standard Swahili.

It was assumed that each of the interviewers would speak Standard Swahili. However, this was true only in one case, while the interviewer did not handle local speech varieties in any way. All other interviewers had non-standard words, mostly due to repeating the word that the respondent said.

There is another possibility to study the distance of these speech varieties from Standard Swahili. The lexical lists of 620 words each would offer a more suitable material for this kind of study. I will do this study in another technical report.