Technical Reports on Language Technology
Report No 73, 2021
http://www.njas.helsinki.fi/salama

Lexical and phonological differences of Swahili speech varieties[1]

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

Abstract

The topic of the Technical Report No. 71 was the linguistic distance between Swahili speech varieties. It was based on the transcriptions of tape-recorded speech, which was collected as part of the DAHE project in 1889 - 1991. The current report is based on the word lists, compiled in conjunction with the same project. All the 16 collected lists have the same format. Each word, written in Standard Swahili, has a unique identification number, ranging between 001 and 620. Some of the lists are defective and do not have all entries. Each entry has the corresponding form in local speech variety, and it has also four different codes. The first code is the identification number of the line. The second code marks the speech variety. The third code marks the informant. The fourth code contains information on what ways the word differs from Standard Swahili.

With the help of these lists and various types of codes in them, I will carry out various computational processes for finding out interesting differences and similarities between speech varieties.

Key Words: *lexical difference, phonological difference.*

1 Introduction

The material for the DAHE project was collected in conjunction with field trips in 1989 - 1991. The research team consisted of researchers from the universities of Dar-es-Salaam and Helsinki. Also local people assisted in finding suitable informants.

We did not manage to find full filled-in lists from all language varieties we wanted. The result reflects the many difficulties that field workers usually encounter. The result is not in balance. For example, there are 10 lists from Kikae, but from some speech varieties we have only one full list or in some cases only a partial list. These limitations make the reliable comparison difficult.

The filled-in lists were manually edited and encoded. Particularly important is the encoding of how the local gloss differs from the Standard Swahili word. Although there are many kinds of differences, I grouped the differences into four classes. If the gloss was

the same as in Standard Swahili, the code is *do* (abbreviation from Latin *ditto*). If the gloss has the same root in both speech varieties and the difference is phonological, the code is *ph* (= phonological). If the gloss has a different root than Standard Swahili, the code is *le* (= lexical).

There are also such cases which are difficult to classify. Particularly problematic are such nouns, which have the same root, but which have a different noun class, and which, in addition, have phonological differences. These cases fulfil the criteria for being classified as phonologically different, but they have also other differences. These cases were given the code *leph* (le + ph) indicating that they have elements of both categories.[2]

Some Standard Swahili words had more than one gloss in local speech variety. Each variety was given its own line with the same word number. Therefore, if information is retrieved using the word number as a search key, all varieties will be retrieved, and they are on separate lines. This solution was made, because often the separate glosses were given different distinction codes.

One further border case were such verbs, where the Standard Swahili verb had the same base as the local variety but a different extension. These cases were given the tag *le* (= lexical).

The current study of speech varieties is defective in many ways. One shortcoming is that some lists are only partially complete, and many entries have no answer. It was also hard to judge whether an empty answer meant that the gloss is the same as in Standard Swahili or whether no answer was given. Some lists were filled in under supervision of a researcher, while others were filled in without the help of the researcher. This also affects the homogeneity of the results.[3]

2 What can we learn from the comparative word lists of Swahili speech varieties?

Below I try to elucidate the points, where the statistical calculations of various types of differences might lead to wrong conclusions. They show reliably the main trends, but one should avoid making conclusions on the basis of fine-grained differences.

There are two types of tables below. In one type (Tables 1 - 3), if there were more than one list filled in, the joint breakdown of all the lists was calculated. In another type of tables (Tables 4 - 21), the breakdown of each individual list was calculated. For example, there are ten lists of Kikae, but only one list of Kitumbatu.

The two different table types differ also in how the numbers were calculated. From the joint lists, where all lists of the same speech variety were joined together, each numbered word with the same code was listed only once. For example, the word number 026, which in Standard Swahili is *kivi*, has four different glosses with lexical difference in Kikae (1). Yet in the joint table of Kikae, all these were considered as one single type, because all of them have the same difference code *le*. All of them are lexically different from Standard Swahili, although in different ways.

---

[2] The list of abbreviation is in the attached appendix.
[3] The entire corpus of word lists is in the address:
http://www.njas.helsinki.fi/salama/DAHE-list.pdf

(1)
026 kae at *kivi | kigudi | le
026 kae ay *kivi | kigudi | le
026 kae hc *kivi | kigudi | le
026 kae mh *kivi | futi | le
026 kae mi *kivi | kigodi | le
026 kae nn *kivi | kisugudi | le
026 kae uh *kivi | kigudi | le

Correspondingly, if the numbered word has more than one difference code, such as in (2), each code was counted once. Therefore, the entries in (2) yield *do* (two), *ph* (two), and *le* (one).

(2)
019 kae hc *jino | jino | do
019 kae mh *jino | gino | ph
019 kae mh *jino | jino | do
019 kae uh *jino | gino, magino | ph
019 kae mi *jino | gego, magego (back teeth) | le

When the calculation was done as described above, we got the statistics of all difference codes found in all lists of each type of speech variety. When we look at the joint table of Kikae (Table 1), we see that 519 words out of 620 were considered as equal with Standard Swahili by at least one informant. There were 240 cases, where at least one informant considered the word lexically different. In 93 cases, at least one informant considered the word phonologically different. And in 24 cases the word was considered lexically and phonologically different.

When we look at the joint tables of Kikae, Kipemba, and Kimalaba, we see that in all of them the sum of all four code types far exceeds 620, which is the number or Standard Swahili words in each list. The reason is that a numbered word may have varying encodings, and each type is calculated as a separate entry.

Finally, each code type in all lists was counted, and the sum is displayed in Table 22.

Table 1. Joint list of Kikae

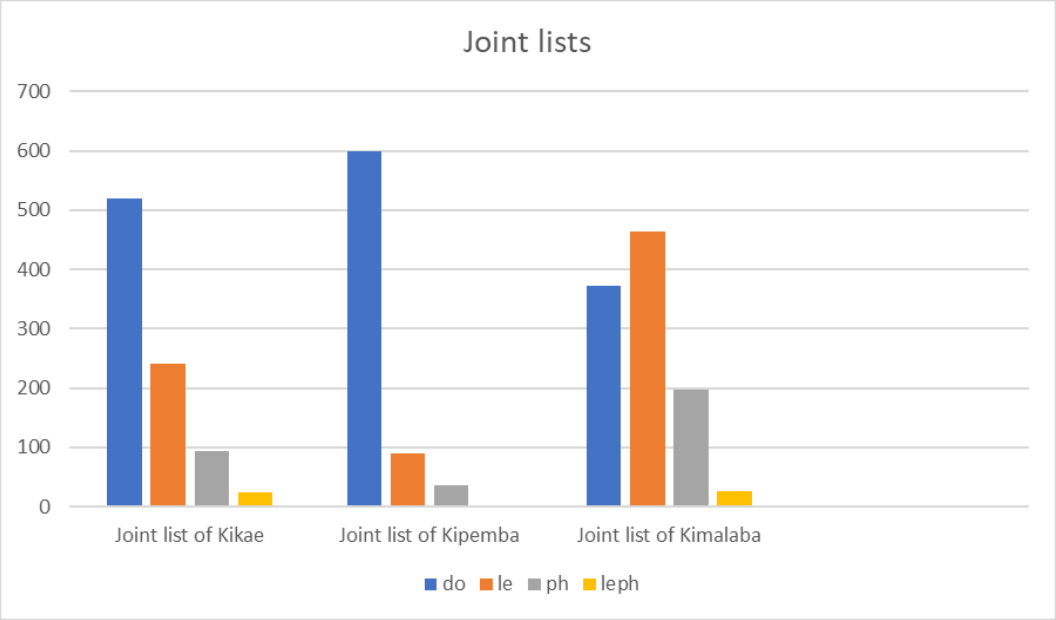
Code	**Number**
do	519
le	240
ph	93
leph	24

Table 2. Joint list of Kipemba

Code	Number
do	599
le	90
ph	35
leph	2

Table 3. Joint list of Kimalaba

Code	Number
do	372
le	464
ph	198
leph	25



The tables based on individual lists were constructed so that the sum of each code type was simply calculated. In theory, the total sum of the four code types should be 620. However, none of the tables reaches the target.

There are two reasons for this. First, there are lists, where only part of the words was given glosses in local speech variety. In these cases, the sum of the codes is less than 620. Second, there are words, for which more than one gloss was given. The alternative glosses were calculated as separate entries. In these cases, the total sum may exceed 620, provided that all words were given at least one gloss.

Table 4. Kikae, List of hc

Code	Number
do	499
le	54
ph	31
leph	1

Table 5. Kikae, list of ay

Code	Number
do	7
le	53
ph	18
leph	1

Table 6. Kikae, list of at

Code	Number
do	10
le	46
ph	26
leph	1

Table 7. Kikae, list of hs

Code	Number
do	11
le	51
ph	30
leph	5

Table 8. Kikae, list of sm

Code	Number
do	18
le	102
ph	26
leph	15

Table 9. Kikae, list of mh

Code	Number
do	15
le	58
ph	26
leph	6

Table 10. Kikae, list of uh

Code	Number
do	14
le	54
ph	20
leph	8

Table 11. Kikae, list of ma

Code	Number
do	11
le	85
ph	26
leph	9

Table 12. Kikae, list of mi

Code	Number
do	16
le	44
ph	12
leph	1

Table 13. Kikae, list of nn

Code	Number
do	4
le	22
ph	10
leph	4

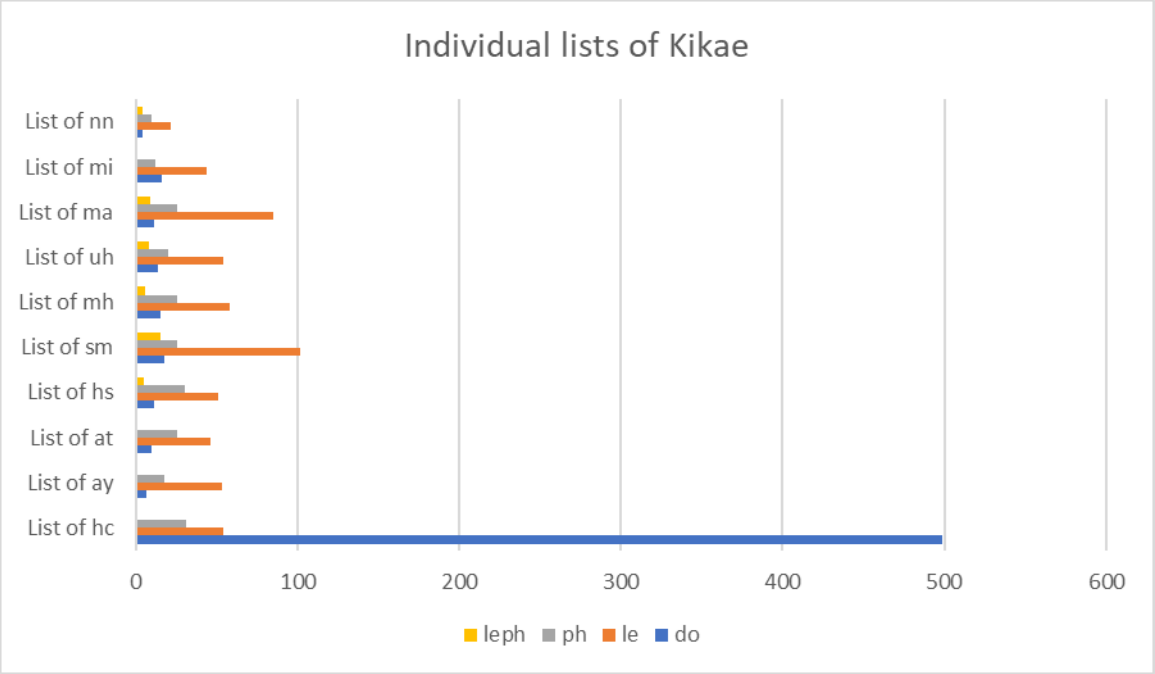


Table 14. Kitumbatu, list of nt

Code	Number
do	444
le	100
ph	69
leph	5

Table 15. Kipemba, list of ss

Code	Number
do	561
le	45
ph	17
leph	1

Table 16. Kipemba, list of hh

Code	Number
do	539
le	67
ph	24
leph	1

Table 17. Kimalaba, list of nn

Code	Number
do	95
le	287
ph	55
leph	17

Table 18. Kimalaba, list of mm

Code	Number
do	281
le	235
ph	112
leph	6

Table 19. Kimalaba, list of st

Code	Number
do	191
le	309
ph	125
leph	2

Table 20. Kinyagatwa, list of ad

Code	Number
do	249
le	199
ph	185
leph	3

Table 21. Kimakonde, list of mc/dn

Code	Number
do	88
le	389
ph	127
leph	8

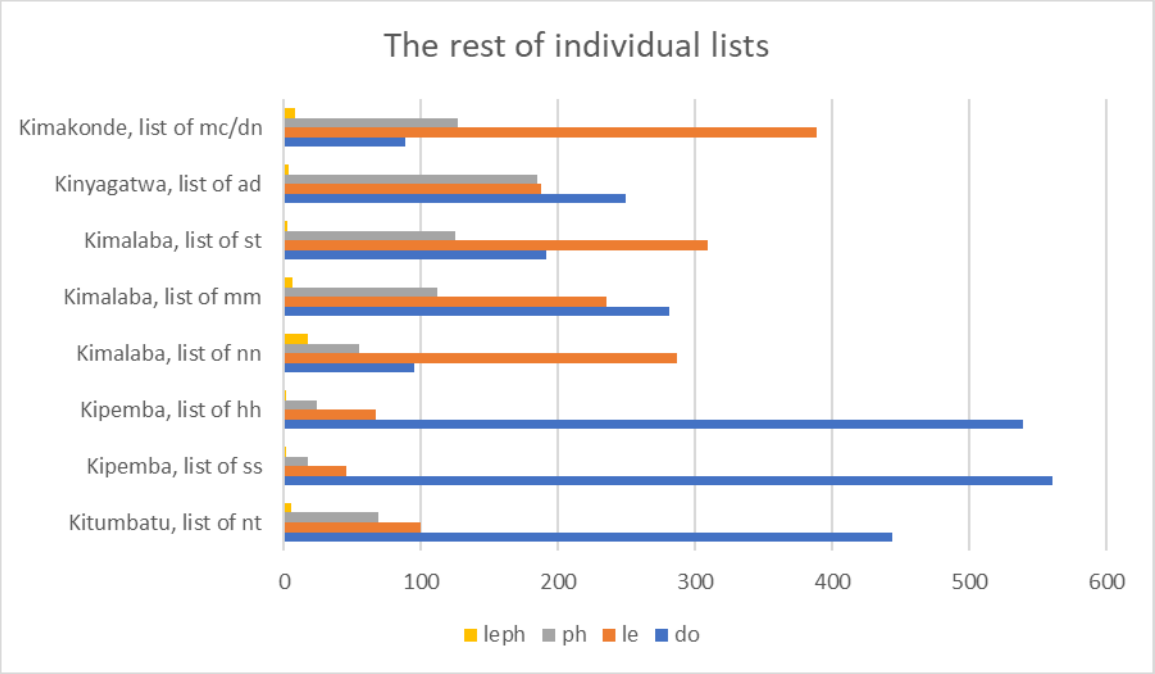
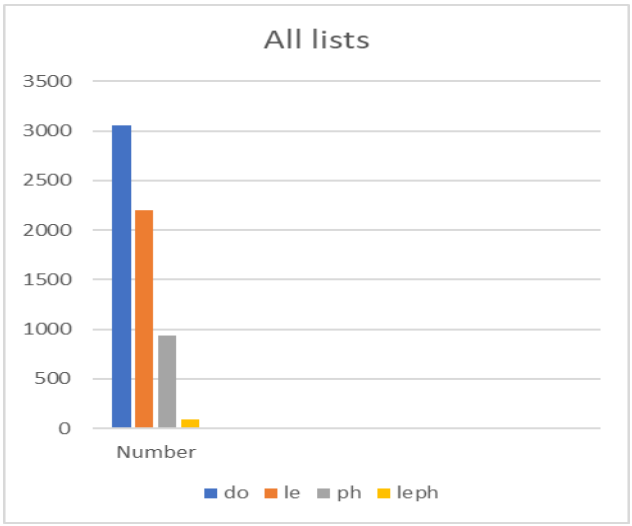


Table 22. All lists

Code	Number
do	3052
le	2203
ph	939
leph	94



3 Discussion

It would be expected that the results of a specific speech variety would be similar in all lists. In the case of Kikae the differences are, however, very big. Two main reasons contribute to the results. The first reason is that respondents had various degrees of competence in the speech variety. Many Kikae speakers used normally Standard Swahili, and only with difficulty they were able to find a true Kikae expression. The origin of Haji Chum was in Kae area and he was a Swahili scholar, with expertise in Kikae. Yet his list (Table 4) has 499 similar words with Standard Swahili, and 54 words with lexical differences. The list of Amina Yusufu (Table 5) has only 5 similar words, but 53 words with lexical differences. Her list was filled in only selectively.

The former list was fully filled in and it had glosses for all words. The latter list was selective, and only part of the list was covered. Amina had a reputation of knowing Kikae exceptionally well, and she wanted to excel with rare Kikae words. Most of the Kikae lists have glosses for only part of the words. It was also difficult to judge whether the entry left without answer should be interpreted as similar, or that the respondent did not know.

The speech varieties such as Kikae, Kitumbatu and Kipemba are considered as Swahili dialects. When we look at the statistics of these speech varieties (Tables 1, 2 and 14), we see that in all of them the share of similar glosses is high. This indicates that they are closely related to Standard Swahili.

When we look at the statistics of speech varieties on the southern coastal area, such as Kimalaba (Table 3), Kinyagatwa (Table 20), and Kimakonde (Table 21), we see that in them the share of lexical differences is very high, higher than of phonological differences. Also the share of similar glosses is small. This gives reason to the conclusion, that they are clearly more distant from Standard Swahili than the speech varieties spoken in Zanzibar and its vicinity. Linguistically most distant seems to be Kimakonde.

When we calculate the total sum of the codes in all lists (Table 23), we see that most common are the cases where the gloss is the same as in Standard Swahili (3052). About the difference types, the lexical difference is most common (2203). Then follows the phonological difference (939). The least common is the type that includes lexical and phonological differences (94).

4 Conclusion

The report discusses the possibilities for extracting information from the comparative lists of various speech varieties. The point of reference is Standard Swahili, and each item in the lists is given one of the four values, which describe the type of difference. Because the lists were not homogenous, conclusions could be made only taking into consideration the identified limitations.

The speech varieties spoken in Zanzibar and its nearby islands were close to Standard Swahili in respect to all four variables. The other speech varieties in the south were more

distant in respect to the variables. These varieties are not considered as dialects of Swahili. They are considered as separate Bantu languages.

APPENDIX

Below is a key to abbreviations that occur in text and in the word lists.

kae = the speech variety of Kikae
pem = the speech variety of Kipemba
tum = the speech variety of Kitumbatu
mal = the speech variety of Kimalaba
nya = the speech variety of Kinyagatwa
mak = the speech variety of Kimakonde
do = ditto, the gloss is identical with Standard Swahili
le = the difference is lexical
ph = the difference is phonological
leph = the difference is lexical and phonological

In addition, there are several encodings of informants. For privacy reasons the names are encrypted.