Revisiting Lemma Lists in Swahili Dictionaries

Beata Wójtowicz, *Department of African Languages and Cultures, University of Warsaw, Warsaw, Poland (b.wojtowicz@uw.edu.pl)*

Abstract: When compiling a dictionary, a lexicographer has a set of decisions to make — starting with drawing up a lemma list to such issues as formatting a dictionary entry. Relying on corpus data while designing a lemma list and describing entries is standard in present lexicography, but there are still decisions — like the choice of a lemma or how to treat derivatives — that are often intuition-based. This article aims to investigate whether decisions put forward in Swahili dictionaries comply with users' expectations. We analyse log files from the new Swahili–Polish dictionary to investigate why looking up words goes wrong, and evaluate the choice of a lemma and the treatment of derivatives in Swahili dictionaries. Based on such data we intend to expand or modify the existing electronic dictionary to adapt to users' level of grammar and dictionary structure knowledge. During this research we identified a list of lemma lacuna that cause the majority of unsuccessful Swahili searches. The study shows that users know and understand the lemmatisation strategy of the dictionary but also reveals which word forms cause the most problems and how the lemma list of Swahili dictionaries could be expanded.

Keywords: DICTIONARY USER RESEARCH, LOG FILES ANALYSIS, SWAHILI–POLISH DICTIONARY, LEMMA LIST, DERIVATIVES

Opsomming: Die herbesoek van lemmalyste in Swahili-woordeboeke. Wanneer 'n woordeboek saamgestel word, moet 'n leksikograaf 'n reeks besluite neem — van die opstel van 'n lemmalys, tot kwessies soos die formatering van 'n woordeboekinskrywing. Om staat te maak op korpusdata wanneer 'n lemmalys opgestel en inskrywings beskryf word, is standaardpraktyk in die huidige leksikografie, maar daar is steeds dikwels besluite — soos die keuse van 'n lemma of hoe om afleidings te hanteer — wat op intuïsie gebaseer is. Hierdie artikel beoog om te ondersoek of besluite wat in Swahili-woordeboeke geneem is, voldoen aan gebruikers se verwagtings. Ons analiseer loglêers van die nuwe Swahili–Poolse woordeboek om te ondersoek waarom die opsoek van woorde skeefloop, en evalueer die lemmakeuse en die hantering van afleidings in Swahili-woordeboeke. Ons beoog om die bestaande elektroniese woordeboek op grond van hierdie data uit te brei of te wysig om aan te pas by gebruikers se vlak van kennis ten opsigte van grammatika en woordeboekstruktuur. Tydens hierdie navorsing het ons 'n lys van leemtes ten opsigte van lemmas geïdentifiseer wat die meerderheid van onsuksesvolle Swahili-soektogte veroorsaak. Die navorsing toon dat gebruikers die woordeboek se lemmatiseringstrategie ken en verstaan, maar openbaar ook watter woordvorme die meeste probleme veroorsaak en hoe die lemmalys van Swahili-woordeboeke uitgebrei kan word.

Sleutelwoorde: WOORDEBOEKGEBRUIKERSNAVORSING, LOGLÊERONTLEDING, SWAHILI–POOLSE WOORDEBOEK, LEMMALYS, AFLEIDINGS

1. Introduction

When a new dictionary is compiled a reference to a corpus is a standard procedure (cf. De Schryver et al. 2006). We use a corpus in different ways in dictionary production, either for updating an already existing dictionary or for compiling a brand new dictionary from scratch (cf. Atkins and Rundell 2008: 97). The vexing question has always been what to include in a dictionary. We may rely on a frequency list derived from a corpus during the headword selection, but already at this stage we have to make decisions as to the choice of lemma or how to treat various items, like derivatives or multiword expressions. Texts are made of word forms whereas in dictionaries we expect lemmas, the so-called dictionary headwords or citation-forms. The choice of a citation-form has direct impact on the ways the user conducts look-ups of certain items and how much knowledge of grammar is needed to successfully consult the dictionary.

The dictionary data should be presented in such a way that a user can easily access it. The success of finding the data depends on the access structure (cf. Gouws and Prinsloo 2005), that has undeniably changed in the era of electronic lexicography. In an electronic dictionary it is often the lemma itself that provides access to the article, and therefore the choice of lemma is a crucial decision the lexicographer has to make. This is especially true in case of Bantu languages where lemma is not intuitive and in some cases not identical to any word forms. The user has to learn basic grammar and know the structure of a dictionary to be able to use it. Thus far it was not known how the users manage but now, in the era of electronic lexicography, we can observe, to some extent, users' behaviour and choices. It gives us possibility to modify the shape of a lemma and a content of the lemma list to respond to users' needs.

No matter how good the theories and methods used in compiling a dictionary are, it is the user who ultimately evaluates the usefulness, efficiency and user-friendliness of a particular lexicographic work. Therefore all lexicographic decisions have to be taken with the user in mind and especially the users' skills must be taken into account (cf. Atkins and Rundell 2008, Prinsloo and De Schryver 1999). Building on this assumption, a new Swahili–Polish dictionary[1] was created and posted online as a student resource. The dictionary has been created as an electronic resource, but its printed counterpart has also been published (Wójtowicz 2013). It contains over 6 000 Swahili entries and over 7 000 entries in a searchable Polish index in the electronic version. The dictionary's lemma list is mainly based on a Helsinki Corpus of Swahili-derived frequency list (HCS 2004) of over ten thousand lemmatised entries. It is targeted at learners of Swahili, who already have some knowledge of language grammar and the structure of a Swahili dictionary. This paper aims to address the question of how well the users meet these constraints by investigating the log files of the dictionary. We intend to research whether users know how to access Swahili dictionary articles — that is their lemma choices as compared to dictionary lemma list. Various decisions concerning its macro- and microstructure were made in accordance with the Swahili lexicographic tradition. Therefore, even though we are investigating the log files of this particular dictionary, we intend to relate our findings to other dictionaries of Swahili as well.

To meet these goals, in subsequent sections we will report on the choice of citation-forms and the treatment of derivatives in various Swahili dictionaries. To make an evaluation we analyse the log files of a Swahili–Polish dictionary that was made available online four years ago. We aim to examine strings that were used during look-ups and then compare them with what the dictionaries offer as a look-up strategy. We also intend to investigate the reasons why looking up words does not always goes well.

2. Log files as a tool of dictionary user research

In his seminal work, Samuel Johnson recognises users as an integral part of the lexicographic process (Atkins and Rundell 2008: 5) and the logical expectation by now would be that research into dictionary use and user profiles is already well established (cf. Lew 2011b, Lew and De Schryver 2014, Töpel 2014). However, analysis of log files — which alongside questionnaires, protocols or observation, is one of the approaches to the study of dictionary use (Lew 2011a) — is not exploited as often as one could expect. Töpel (2014) reports on the unsatisfactory situation noting that the dictionary users and their actions are still unknown and that more research is still needed.

According to Bergenholtz and Johnsen (2005), analyses of log files may be used as a tool for improving Internet dictionaries. They consider log files to be a useful supplement to corpus-based lemma selection since they reveal lemma lacuna, frequent misspellings, frequency of searches for multiword units, etc. De Schryver and Joffe (2004: 188) opine "that an automated analysis of the log files will enable the dictionary to tailor itself to each and every particular user".

Even though such possibilities seem tempting, the limitations of this source have to be taken into consideration as well (cf. Lew 2011a). Lew (2011a) remarks that log files don't give answers on the context of dictionary use or on the user himself. With or without these limitations in mind researchers investigate log files most often to reveal which lemmas have been successfully retrieved; which lemmas have been requested but were not found; which lemmas have been looked up and how often; and which words have been used in a search field. Based on this information, one can modify the content of a dictionary to meet the users' needs.

The log files analysis was reported already in several studies that studied different aspects of dictionary use. In the study of Laufer and Hill (2000), who investigated what information is looked up and how unknown vocabulary is retained, a log file analysis was combined with a vocabulary test to check for vocabulary retention. In their study, Laufer and Hill (2000) and Nesi (2000) exploit log file analysis to compare electronic and printed dictionaries. A different approach was adopted by Lemnitzer (2001), who was interested in the reasons why looking up words goes wrong. He tested his assumptions on the log files of four bilingual electronic dictionaries. The study of De Schryver and Joffe (2004) also concerns files obtained from the normal use of an electronic dictionary rather than of a specially designed context. The authors for the first time ask a question later investigated more deeply in further studies (cf. De Schryver et al. 2006, Verlinde and Binon 2010, Koplenig et al. 2014, Trap-Jensen et al. 2014, Müller-Spitzer et al. 2015), about whether dictionary users actually look up frequent words. The research leads to an important conclusion: it seems that there is a relationship between the corpus frequency and the frequency of look-ups. Müller-Spitzer et al. (2015) further claim that "frequency does matter — even in lower frequency bands", which matters enormously in the era of corpus-lexicography. This is because dictionary compilers are provided with evidence about the value of corpus-based lexicography as the users frequently look up frequent words even beyond the first few thousands. The frequency list thus remains the main source of data especially when compiling a small dictionary lemma list. Another study of Bergenholtz and Johnsen (2005) shows how to use log file analysis method as a tool for improving electronic dictionaries. They report on the sources of unsuccessful searches and user search behaviour.

The studies show that analysing log files successfully leads to filling gaps in dictionary lemma lists and discovering other problems related to unsuccessful look-ups. However this kind of research does not address all issues concerning dictionary user studies. In order to get a full picture of the dictionary consultation process we need wider studies that combine all of the different methods. None of the methods answers all of the questions related to dictionary use.

3. Citation-forms in Swahili dictionaries

De Schryver et al. (2006: 68) notice that "not all primary speakers of Swahili can look up 'words' in their own language (as this implies being able to cut off pre- and suffixes), and even trained learners and scholars often need more than one look-up round before they find what they are looking for (as sound changes between formatives are not always predictable)".

Kosch (2013: 202) further elaborates that how well users cope with looking up words in a Bantu language dictionary and to what extent their expectations are met, is largely dependent on such factors, like consultation skills — that is, their previous exposure to dictionary pedagogy, their knowledge of the structure of a Bantu language, and the dictionary design itself. Bantu language learners often mimic their dictionary habits from non-Bantu language dictionaries and are not aware of a dictionary design owing to the agglutinative structure of the language, which calls for a specialised approach to lemmatisation.

The most important part of a consultation process concerns the choice of a citation-form. A citation-form serves different functions, but first of all, it is a form by which the user will use a dictionary in order to find other phonological, morphological, syntactic, semantic, and etymological information associated with it (cf. Kiango 2005).

In the case of languages with long lexicographical traditions, the citation-forms have long been determined and lexicographers easily avoid the problem of discrepancies between dictionaries. But the question remains of how the users manage with the lexicographers' decisions. It is they who have to lemmatise word forms found in texts. It is assumed that the user knows how to search in a dictionary. As an aid there are guidelines in the introductory part. But still, it is expected that users learn the rules before conducting a search. We can only guess how often they comply with this demand.

The methodology of lemma selection, which has been well researched for European languages and is above all based on Latin, cannot always be without proper scrutiny applied to languages for which the grammatical structures significantly differ from those of Indo-European languages (cf. Knowles and Mohd Don 2004). According to lexicographical recommendations, all forms which naturally come to mind to users when searching a dictionary should function as headwords. But due to the complex morphological structure of Bantu words, the choice of the citation form is not always obvious (cf. Kiango 2000). Prinsloo and De Schryver (1999) give a comprehensive introduction to the lemmatisation strategies in Bantu languages. There are two lexical traditions applied to the Bantu languages. These are word traditions with lemmas based on complete written words, and stem tradition with lemmas based on the stems of written words without their prefixes.

Swahili is a Bantu language and as such, it is characterised by agglutinative morphology and a noun class system, whose reflexes are manifested both lexically, on the noun, and syntactically, via agreement. This is demonstrated in the six sentences below, where the only crucial lexemic variable is the initial noun, with which all other elements have to agree (CL stands for a class marker).

(1)	a.	M-tu	huyu	m-zuri	m-moja	a-li-ye-anguka
		CL1.man	this.CL1	CL1-nice	CL1-one	CL1-PAST-CL1.REL-fall
		'this one nice man who fell down'				
	b.	Wa-tu	hawa	wa-zuri	wa-tatu	wa-li-o-anguka
		CL2.people	these.CL2	CL2-nice	CL2-three	CL2-PAST-CL2.REL-fall
		'these three nice people who fell down'				
	c.	M-fuko	huu	m-zuri	m-moja	u-li-o-anguka
		CL3.bag	this.CL3	CL3-nice	CL3-one	CL3-PAST-CL3.REL-fall
		'this one nice bag which fell down'				

d. Ma-chungwa haya ma-zuri ma-tatu ya-li-yo-anguka
CL6.oranges these.CL6 CL6-nice CL6-three CL6-PAST-CL6.REL-fall
'these three nice oranges which fell down'

e. Ki-tu hiki ki-zuri ki-moja ki-li-cho-anguka
CL7.thing this.CL7 CL7-nice CL7-one CL7-PAST-CL7.REL-fall
'this one nice thing which fell down'

f. Kalamu hizi n-zuri tatu zi-li-zo-anguka
CL10.pens these.CL10 CL10-nice CL10.three CL10-PAST-CL10.REL-fall
'these three nice pens which fell down'

Each nominal lexeme is classified as belonging to one of fifteen classes, and each of these classes has a default mapping onto the agreement pattern that holds most of the other elements of the sentence. Note that both noun-class membership and the resulting agreement are in most cases manifested by prefixal means (in the above examples only the demonstrative pronoun breaks this rule). This is also visible in (2) below, in the present-tense paradigm of the verb *kuanguka* 'to fall down'.

(2) a. ku-anguka 'to fall down'
INF-fall.down

b. ni-na-anguka 'I fall down'
1SG-PRES-fall.down

c. u-na-anguka 'you fall down'

d. a-na-anguka 'he/she/it falls down'

e. tu-na-anguka 'we fall down'

f. m-na-anguka 'you (PL) fall down'

g. wa-na-anguka 'they fall down'

All infinitives in Swahili begin with *ku-* (or *kw-*), and practically all verbal stems could be substituted for the stem *anguka* in the paradigm above. Prefixes change depending on the grammatical context, and the only element that remains unchanged is the stem.

In an established Swahili lexicographic tradition on paper dictionaries, noun headwords are introduced in singular form together with the nominal prefix, for example *mtu* 'man'. On the other hand, verbs use sequences identical to the infinitive form, but without the infinitive prefix *ku-*, for example *penda* 'love' (instead of *kupenda*). If the traditional lexicographic practice used e.g. for European languages was followed here in citation form selection, and if verbs were cited in their infinitival forms, most of the space in Swahili dictionaries (and that is also true of many other Bantu languages in which infinitives are morphologically assigned to a noun class 15) would be taken by the letter "k". In fact, all verbs would end up there, which is hardly user-friendly or sensible. A similar problem concerns adjectives: listing their fully inflected forms would result in distributing most of them under the letters of alphabet that their agreement prefixes begin with, which is predictable and regular. As in the case of inflected verbs (cf. the paradigm in example 2), using the full, inflected forms would result in massive redundancy. Therefore, adjectives, numerals, and pronouns are represented by non-prefixal forms, for example *zuri* 'good'. To inform the user that the given word necessitates the addition of a prefix in order to take on a proper form, the headwords are in some dictionaries preceded by a hyphen, for example *-zuri* 'good' (e.g. in Abdulla et al. 2002). Occasionally pronouns, especially demonstrative and possessive, are included in dictionaries in their full form, that is, with a class prefix of a noun they modify, for example *changu* 'mine, class 7' (Baba Malaika 1994). The possible ways we may treat these issues change when we switch to electronic lexicography.

Despite the Swahili lexicographic tradition, the question of the structure of citation-forms in Swahili is still vital. Kiango (2000: 25) recognizes the infinitive as the basic natural form of Bantu verbs. The only reason why we should not put them in such shape in a dictionary is the alphabetical order discussed above. As for the nouns, Kiango (2000: 31) shows that certain nominal stems are unnatural forms, and therefore a noun in a singular form with a prefix is entered as a citation-form in most dictionaries of Bantu languages. However, this method is not convenient for beginners of the language, who are not able to easily identify singular and plural forms of a noun. It could be solved by applying a method of entering both forms into a dictionary, but this violates the principle that "citation-forms should be stems from which other inflected forms could be produced" (Kiango 2000: 31-32). Both forms represent one lexeme, and therefore they cannot be entered as two separate entries. We could violate this principle when handling irregular forms, but still it is against the principle of economy. Again, it can be easily solved in an electronic version of a dictionary, where we can, for example, allow searches on plural forms, which are provided within an entry.

Indeed, the electronic form introduced novelty into accepted solutions. When the *TshwaneDJe Swahili–English Dictionary*[2] joined the market, it was the first and still remains the only corpus-driven electronic dictionary of Swahili with a new approach to the lemmatisation of headwords (De Schryver et al. 2006). The content of the dictionary is based on web-based corpus data and it includes over 16,000 entries. The most interesting innovation is that the headwords include orthographic forms in addition to stems chosen on the basis of a frequency count. The dictionary does not include morphological analysis as such, but the user may search for the frequently used word forms.

3.1 The treatment of derivatives in Swahili dictionaries

The most important issue with respect to the macrostructure in dictionaries of Bantu languages regards the handling of derivatives, sometimes referred to as the "lumping vs. splitting" debate (cf. Bański and Wójtowicz 2011). Kosch (2013) discusses the issue with reference to the demands that lexicographers place on the users. The dictionary design is then motivated by various expectations, ranging from low-level, where basic look-up skills according to the letters of the alphabet are assumed to have been mastered already; to medium-level expectations that assume the user is able to look up words in a stem-based dictionary; to high-level demands, where intuitive dictionary skills no longer suffice.

The discussion arises from the fact that Bantu derivational word-families can be extremely numerous, especially those based on verbal roots — for example, De Schryver and Prinsloo (2001: 225ff) count over 140 regular derivatives of the root *reka* 'buy, purchase' in Sepedi. The problem that we see deals with presenting the derivatives to the user, and the basic question is whether to lump them all into the entry of the root, or whether to distribute them across the dictionary giving each derivative the status of a main entry.

Radical lumping means cramming all the information into a single place in a dictionary and often denying independent status to the most commonly used words only because they happen to be derivatives; additionally, the user is required to know word-formation mechanisms (which are always less transparent than the inflectional system) in order to identify the base form of the given derivative, which means searching for e.g. *utumishi* 'civil service' in the entry for *-tuma* 'assign/give work to sb', which e.g. in Johnson's (1939/1985) dictionary takes up half a page. De Schryver and Prinsloo (2001: 224) report on a similar case in a Sepedi–Afrikaans–English dictionary, where the entry for *reka* and its derivatives takes up an entire page of dense print. There is also the possibility of partial lumping when some derivatives are kept with their roots and some are given the status of a main entry. Kiango (2000) recommends lumping only regular forms, but this does not help if it is the regular forms that happen to have high frequency and should therefore be presented to the user in separate entries.

The choice of the splitting strategy may be considered primarily practical — as Zgusta (1971: 16) notes, "[...] we must not forget that the lexicographer is doing scientific work, but that he publishes it for users whose pursuits are always more practical, at least as regarded from his own point of view." Given this, a user-friendly dictionary of Swahili should list the broadest possible range of derivatives as separate entries. It is not without reason, as Herms (1999) notes, that students of Swahili praise the "friendly" dictionary of Baba Malaika (1994), who adopts this kind of distributed approach to derivatives, while De Schryver and Prinsloo (2001) note the lack of popularity of dictionaries that group words on etymological grounds and/or under verbal or nominal roots.

The lumping approach has one undeniable virtue — it keeps word families together. In agglutinative languages, it can even show how the particular members are derived (cf. Bosch et al. 2007). Scattering derivatives across the entire dictionary means severing the lexical and semantic ties between closely related lexemes and practically hiding some of them from users looking up the root form.

Clearly, both approaches have their virtues and their disadvantages, and both of them have been suggested in the literature and used in practice. Bosch et al. (2007) argue for lumping (in South Bantu languages), and that is what

Johnson (1939/1985) did for Swahili. De Schryver and Prinsloo (2001) argue for splitting and this is what Sacleux (1939) and TUKI (2001) have implemented to varying degrees of success. Madan (1903/1992) and Abdulla et al. (2002) represent mixed approaches, whereby derivatives representing the same part of speech as the derivational base are placed inside the entry of that base form, while derivatives representing different parts of speech are listed separately.

Due to above problems it is really important that the users of a dictionary understand the lemmatisation approach used in a certain dictionary (cf. De Schryver et al. 2006).

The issue can now be, at least partially, addressed in an electronic form as showed in a Swahili–Polish dictionary where derivational families are presented to the user in the form of a searchable graph. The lemma-oriented approach is no longer necessary in electronic dictionaries as it is possible that each segment of the dictionary may be viewed in different ways by different users, thus eliminating the need for any kind of macrostructure.

4. Log files of a Swahili–Polish dictionary

The Swahili–Polish dictionary user queries have been recorded and saved since January 2013. They are noted in four log-files, namely: Swahili found entries, Swahili not-found, Polish found and Polish not-found. The users may search for Swahili and Polish headwords, but they have to choose the language before they conduct a search. The files note a string the user has typed in the search box and the number of searches for that certain string. Neither IP nor any other identification is taken into account. We assume that many of the users are university students who use the university network. This has been also attested in the data itself, as indicated below.

The data is saved in .csv files and the analysis was carried out by the author manually and with the use of regular expressions only. In the Swahili found file, additionally to the string and the number of searches, information on POS and ID of an entry that was returned to the user is provided. Additional analysis may be conducted on data from the Google Analytics that has been launched as well.

The structure of the dictionary allows a search for Swahili headwords and plural forms of nouns, which are included in the dictionary in their full forms. Derivatives like pronouns with class prefixes and irregular verbal forms are treated as separate entries and users can also search for them. The most difficult operation is the decomposition of the verbal complex. The user has to cut off all prefixal morphemes and search for the verbal root or extended root instead.

Since we are interested in revealing whether users know the search strategies of Swahili dictionaries implemented in this dictionary as well, and whether they know how to choose an appropriate lemma to find what they are looking for, we will investigate mainly the Swahili not-found log file in comparison to the Swahili found file. As the log files only record what the users have entered into the search box, it should be stressed that we can never be sure who the user was or the user's actual intentions (cf. Lew 2011a). We are also aware of the limitations of this study as we investigate a very narrow group of dictionary users — that is, mostly Polish students of Swahili. Many of them were therefore trained how to use dictionaries of Swahili; they know the grammar of the language and the structure of the dictionary under investigation. Nevertheless many unsuccessful searches were still recorded and it was interesting to reveal their sources.

Over a four-year period, up to 15 February 2017, 53,592 queries were made. This makes for an average of 36 look-ups per day, with an increase observed in searches over time — from 25 in 2013/14 to 46.5 since 2015. The number falls during holidays, especially the long summer vacations, to an average of 15 look-ups per day, which corresponds to our assumption that the dictionary is mostly used by students of the language.

The number of look-ups noted in each file is 25,466 (48%) in the Swahili found, 14,708 (27%) in the Swahili not-found, 8,052 (15%) in the Polish found, and 5,366 (10%) in the Polish not-found. Thus, the majority of the look-ups (75%) were in the Swahili–Polish direction. When we compare only the searches for Swahili entries, 63% of them are noted as found and 37% as not-found.

As for the number of unique strings searched for, 4,430 strings were noted in the Swahili found, 8,912 in the Swahili not-found, 2,366 in the Polish found, and 3,364 in the Polish not-found. The numbers are much higher in the not-found sections, but the majority of these look-ups, like 73% of the strings in the Swahili not-found, were looked up only once. In comparison, in the Swahili found file only 29% of the strings were looked up once.

4.1 The analysis of the searches

The problems identified while analysing log files may guide us towards a decision on how to improve the dictionary. By identifying missing lemmas we may improve the coverage of the dictionary, while other issues may influence our approach towards the lemmatisation strategy or the search method. For example, in their study based on log files analysis, Bergenholtz and Johnsen (2005) argue for including additional verbal forms, like imperative and passive in a Danish dictionary, or expanding it by adding some of the items identified as lemma lacuna. On the other hand, they did not find the problem of misspellings to be of an important nature.

A study of the log files from the Swahili–Polish dictionary revealed a number of specific problems encountered by the users while consulting the dictionary. They fall into several categories. To identify these categories, the top 500 strings of the Swahili not-found file were analysed one by one and annotated as a Polish word, an orthographic word, a wrongly lemmatised form, a spelling mistake, a proper name, a multiword expression and a lemma lacuna. The number of strings identified in each category is presented in figure 1.

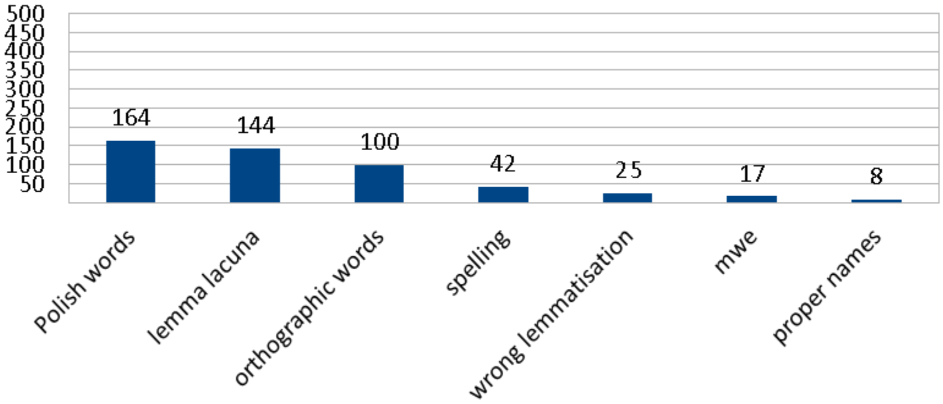

Figure 1: Number of strings per category among the top 500

The top 500 strings were searched for 3,661 times. That constitutes 26% of all searches. The remaining 74% of the searches were for the other 8,400 strings.

The first problem we notice concerns the choice of language. Since the user has to switch modes in order to look for Polish words, unfortunately as many as 11% of the searched for strings that are noted in the Swahili not-found file were identified as Polish words. They stand for 25% of the strings searched for at least twice. Among the top 100 searches in a Swahili not-found file 39 were identified as Polish words and 164 were identified among the top 500. This is also the case with the Polish not-found searches, where 25 missed words among the top 100, and 166 among the top 500, are Swahili. This shows clearly that an option of changing a language does not work well as users forget to switch the languages. Also, the search method should take into consideration all lemmas, Swahili and Polish, during every lookup process.

Out of 500, 144 (28%) strings were identified as lemma lacuna — possible new lemma candidates. Among them, 92% represent lemmas that are also present on the frequency list used to build a dictionary — mostly low-frequency words, like *burudika* 'be appeased' or *manukato* 'perfume'. Only 12 words are not on the frequency list, like *pepa* 'sway', *kauri* 'cowrie shell'. Based on this data, and an analysis of the found entries, we came to a conclusion that this dictionary should be expanded further based on the frequency-list lemmas (Wójtowicz 2017). The list was derived from a Helsinki Corpus of Swahili but only several thousand of the most frequent items were included in the dictionary so far.

100 strings (20%) were identified as orthographic words. These are the forms that the users come across in texts and were looked up in their full form, as for example *anataka* 'she/he wants', *wakicheza* 'as they play'. Among them, 27 are

inflected verbs, 22 — infinitives, like *kuja* 'to come', *kwenda* 'to go', and 51 — adjectives and pronouns with class prefixes, like *mzuri* 'good, CL.1', nouns with a locative suffix *-ni*, as *usoni* 'on the face', or nouns combined with a possessive pronoun, like *mwanangu* 'my child'. All these prefixes and suffixes should be deleted in order to perform the look-up.

There are 42 misspellings, and other unidentified words in this log. Only a few of them are spelling mistakes affected by pronunciation, as when users are spelling the word as it is pronounced, like *mojo* instead of *moyo* 'heart'; by omitting an apostrophe, like *ngombe* instead of *ng'ombe* 'cow'; or inserting a space. There are also some strings that could not be identified as Swahili or Polish words, like *nala*.

During the analysis we have identified only 4 strings out of the top 100 and 25 out of the top 500 as wrongly lemmatised Swahili words. These are the forms that the user probably knew needed to be modified somehow, and tried to lemmatise them but failed, as in *likuwa* (may be from *alikuwa*), *pokuwa* (possibly from *ijapokuwa*), *engi* (probably from *wengi*), *sanya* (part of a verb *kusanya* that looks similar to infinitive but has the morpheme *ku* as a part of a stem).

The remaining 17 strings are multiword expressions, like *baba yetu* 'our father', and 8 strings are proper nouns, like *Timon* and *Kilimanjaro*. It was also interesting to reveal in the Polish not-found file, that users looked up Polish personal names, like *Ania*, *Tomasz*, or names of cities, like *Warszawa*.

A closer examination of the top ten Swahili not-found searches reveals that seven items represent orthographic forms of entries that are present in the dictionary. These are adjectives with a class prefix *mzuri* and *nzuri* 'nice', *njema* 'good', the infinitives *kuja* 'to come', *kwenda* 'to go', *kufa* 'to die' and the inflected verbal form *ninakupenda* 'I love you'. There is also a multiword expression and the name *Mufasa* known from the movie "The Lion King". Only one item, *patia* 'to get for', may be identified as a lemma lacuna.

De Schryver et al. (2006) mention that users often seem to greet the dictionary on arrival. This was also recalled in an interview with Barak Turovsky from Google Translate (Orliński 2017). According to his data the most frequently translated expressions in all languages of the world are 'how are you' and 'I love you'. He finds 'I love you' to be among the top three searches in every language. In our data *kocham cię* or *kocham* 'I love you' in Polish, in its not lemmatised forms, is the first and the third most popular search in the Polish not-found file and the third most popular search in the Swahili not-found. Among the Polish found searches, *miłość* 'love' and *kochać* 'to love', are in the top 25 searches. *Ninakupenda* 'I love you' in Swahili is the eighth most frequently searched expression in the Swahili not-found log, and its lemmatised form *penda* is only among the top 170 in the Swahili found searches. However, since the users of the dictionary are mostly Polish, this explains why they want to express themselves in the Polish–Swahili direction. The users are also interested in greetings, and *cześć* 'hello' is the second most often looked up Polish word.

In their study on reporting look-ups of frequent words in Sepedi and English, De Schryver and Joffe (2004: 190) mention a high number of searches for the offensive and sexual sphere words: "genuine frequent words are looked up on the one hand, and then those words that only mother-tongue speakers know but, as they are taboo, *never* pronounce in public". This was also reported in a study by Bergenholtz and Johnsen (2005). In our data only eight Polish searches among the top 100 are concerned with the sexual sphere — none among Swahili words, as probably users, the learners of the language, are not familiar with these words yet.

One interesting issue to discover was that Polish users quite often, (approximately 10% of the not-found Polish searches), are looking for not lemmatised Polish words, like *lubię cię* 'I like you', *pozdrawiam* 'I greet you'. This also corresponds to the findings of De Schryver et al. (2006) who report on users treating a dictionary like a Web search engine. Our users also put longer strings in the search box, like multiword expressions or even whole sentences, sometimes in other languages. These are, for example, in Swahili *mzuri na wewe?* 'I'm fine and how are you?' with a question mark as a part of several searches, or, in Polish, *jak się masz* 'how are you', or *sto piećdziesiąt dwa* 'hundred fifty two'.

Of course the question remains how such findings should influence our decisions regarding lemma list and dictionary structure. We may easily add the missing lemmas the users are often looking up and this is what we have already done with this dictionary. We should definitely change the search method by eliminating the need for a language choice. But should we change the lemmatisation strategy based on the analysis? When we compare the number of searches in the Swahili found and Swahili not-found files we see that the majority of not-found searches were carried only once, so they may not be significant.

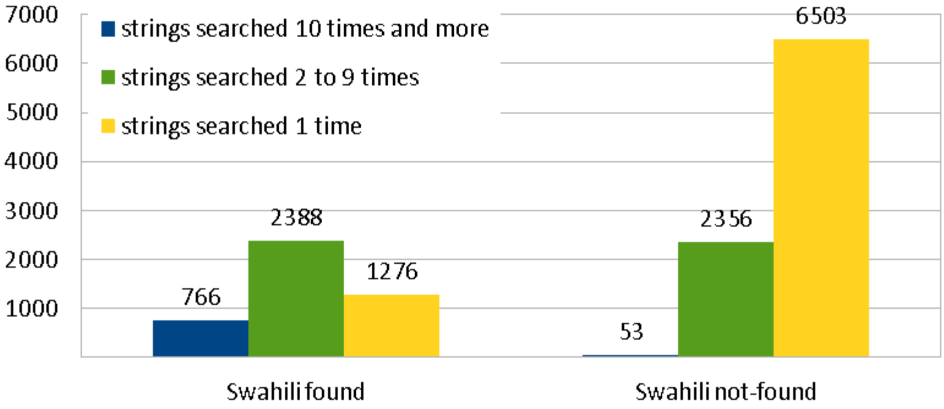

Figure 2: Frequency of searches per string

The top 500 strings in the Swahili found file were carried 11,300 times, while in the Swahili not-found only 3,600 times. The overall number of searches is probably too small to make reliable assumptions. When we analyse the first fifty missed strings that were looked up at least 10 times in the Swahili not-found file, we see that Polish words provide the biggest group, and we find orthographic words next in line in volume. Again, as in the bigger sample, the majority of forms are those with class prefixes attached, the next group are infinitives and then finally we see inflected verbs in only two instances. So based on this data, the next step of expanding the dictionary could be adding adjectives with class prefixes and then infinitival forms of verbs.

When we aim at evaluating the strategy of giving the derivatives a status of main entries, we must stress that the great majority of identified verbal lemma lacuna are verbs with an extension suffix. On the other hand among the top 100 Swahili found searches we have 50 verbs, but only 9 with an extension. When we enlarge the number of analysed strings to 500, among 168 verbs, 49 (30%) have extensions. Therefore, verbal derivatives are not among the most frequently looked up items, but they are searched for quite often, and including them on a headword list proves to be a good, user friendly strategy. Besides frequency, there is also an argument that such a strategy does not leave users in doubt as to the conclusions regarding meaning of derivatives (cf. Gouws and Prinsloo 2005).

As for the nouns, out of over 200 nouns identified among the top 500 most frequently searched items, only 13 are for plural forms. A dozen or so searches are for full forms of pronouns, like *yangu* 'mine', *vile* 'those', *pale* 'there'. But there are only two such searches among the top 100: these are *hivyo* 'that' and *nyingi* 'many'. None of the top searches are for the stem of a pronoun. When we expand the analysis to all of the strings, it appears that there were only a few searches for chosen stems, like *ako* 'your', and *le* 'that'.

Overall, evaluation should take into consideration the number of searches in each file. The majority of the look-ups were successfully retrieved Swahili headwords. The users seem to understand and successfully apply the lemmatisation procedure to find translations of the word forms found in texts. However, they look up full forms of pronouns rather than their stems — these are the forms that are the most difficult to lemmatise, especially the demonstrative pronouns, which are often used and lexicalised. Based on the not-found searches we shall consider expanding the dictionary with full forms of adjectives, which seem to be the most searched for from among orthographic words. The other group represents infinitival forms of verbs. It is possible that users use this form, as this is the lemma for Polish verbs and they mimic their habits from Polish dictionaries. As the last addition we should consider full forms of verbs.

5. Conclusion

Log file analysis helps us to reveal how Internet dictionaries are used. While monitoring user queries, we may keep track of which lemmas are looked up how often, and investigate the reasons why looking up words goes wrong. In our study of log files from the Swahili–Polish dictionary we revealed some problems caused by the searching method that requires users to choose the language of their search and to lemmatise word forms. As dictionaries of Swahili demand some knowledge of the grammar in order to conduct a successful search, we were interested to uncover how well the users manage their queries.

We have identified a list of lemma lacuna that cause the majority of unsuccessful Swahili searches. The study shows that the users seem to understand the lemmatisation strategy and successfully apply it to the word forms they want to find translations for. They often search for extended verbs and full forms of pronouns, so these need to be treated as headwords. If we were to expand the dictionary based on this data, we should consider adding full forms of adjectives, and infinitival forms of verbs as these two groups, besides lemma lacuna, compose the most missed searches. This came as a surprise to us, as we assumed that the most difficult operation is the decomposition of the verbal complex, when the user has to cut off all prefixal morphemes and search for the verbal root or extended root instead. Searches for full forms of verbs also caused some unsuccessful searches, but they were not as many as we expected.

Endnotes

1. http://kamusi.pl/ [accessed 20.04.2017]
2. http://africanlanguages.com/kiswahili/ [accessed 20.04.2017]

References

A. Dictionaries and corpora

Abdulla, A., R. Halme, L. Harjula and M. Pesari-Pajunen (Eds.). 2002. *Swahili–Suomi–Swahili-sanakirja*. Helsinki: Suomalaisen Kirjallisuuden Seura.

Baba Malaika. 1994. *Modern Swahili Modern English.* Arusha: Danish Training Center for Development Co-operation.

(HCS) Helsinki Corpus of Swahili. 2004. Compilers: Institute for Asian and African Studies (University of Helsinki) and CSC — Scientific Computing Ltd.

Johnson, F. (Ed.). 1985 [1939]. *A Standard Swahili–English Dictionary (founded on Madan's Swahili–English Dictionary).* Oxford: Oxford University Press.

Madan, A.C. 1992 [1903]. *Swahili–English Dictionary*. New Delhi: Asian Educational Services.

Sacleux, Ch. (Ed.). 1939. *Dictionnaire Swahili–Français*. Paris: Institut d'Ethnologie.

(TUKI) Taasisi ya Uchunguzi wa Kiswahili. 2001. *Kamusi ya Kiswahili–Kiingereza. Swahili–English Dictionary.* Dar es Salaam: Chuo Kikuu cha Dar es Salaam.

Wójtowicz, B. 2013. *Słownik suahili–polski*. [Swahili–Polish Dictionary]. Warszawa: Elipsa.

B. Other literature

Atkins, B.T.S. and M. Rundell. 2008. *The Oxford Guide to Practical Lexicography*. Oxford/New York: Oxford University Press.

Bański, P. and B. Wójtowicz. 2011. New XML-encoded Swahili–Polish Dictionary: Micro- and Macrostructure. Goźdź-Roszkowski, S. (Ed.). 2011. *Explorations across Languages and Corpora. PALC 2009:* 497-514. Frankfurt a. Main: Peter Lang.

Bergenholtz, H. and M. Johnsen. 2005. Log Files as a Tool for Improving Internet Dictionaries. *Hermes, Journal of Linguistics* 34: 117-141.

Bosch, S.E., L. Pretorius and J. Jones. 2007. Towards Machine-Readable Lexicons for South African Bantu Languages. *Nordic Journal of African Studies* 16(2): 131-145.

De Schryver, G.-M. and D.J. Prinsloo. 2001. Towards a Sound Lemmatisation Strategy for the Bantu Verb through the Use of *Frequency-based Tail Slots* — With Special Reference to Cilubà, Sepedi and Kiswahili. Mdee, J.S. and H.J.M. Mwansoko (Eds.). 2001. *Makala ya kongamano la kimataifa Kiswahili 2000: Proceedings*: 216-242, 372. Dar es Salaam: TUKI, Chuo Kikuu cha Dar es Salaam.

De Schryver, G.-M. and D. Joffe. 2004. On How Electronic Dictionaries are Really Used. Williams, G. and S. Vessier (Eds.). 2004. *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004, Lorient, France, July 6–10, 2004*: 187-196. Lorient: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud.

De Schryver, G.-M., D. Joffe, P. Joffe and S. Hillewaert. 2006. Do Dictionary Users Really Look Up Frequent Words? — On the Overestimation of the Value of Corpus-based Lexicography. *Lexikos* 16: 67-83.

Gouws, R.H. and D.J. Prinsloo. 2005. *Principles and Practice of South African Lexicography.* Stellenbosch: SUN PReSS.

Herms, I. 1999. Swahili Lexicography: The Swahili–German Dictionary. Wolff, H.E. (Ed.). 1999. *Contributions to Bantu Lexicography. Languages and Literatures* 10: 1-8. Leipzig: University of Leipzig Papers on Africa.

Kiango, J.G. 2000. *Bantu Lexicography: A Critical Survey of the Principles and Process of Constructing Dictionary Entries*. Tokyo: Institute for the Study of Languages and Cultures of Asia and Africa, Tokyo University of Foreign Studies.

Kiango, J.G. 2005. Problems of Citation Forms in Dictionaries of Bantu Languages. *Nordic Journal of African Studies* 14(3): 255-273.

Knowles, G. and Z. Mohd Don. 2004. The Notion of a "Lemma". Headwords, Roots and Lexical Sets. *International Journal of Corpus Linguistics* 9(1): 69-81.

Koplenig, A., P. Meyer and C. Müller-Spitzer. 2014. Dictionary Users Do Look up Frequent Words. A Log File Analysis. Müller-Spitzer, C. (Ed.). 2014. *Using Online Dictionaries:* 229-249. Berlin: Walter de Gruyter. (Lexicographica Series Maior 145.)

Kosch, I. 2013. Expectation Levels in Dictionary Consultation and Compilation. *Lexikos* 23: 201-208.

Laufer, B. and M. Hill. 2000. What Lexical Information Do L2 Learners Select in a Call Dictionary and How Does It Affect Word Retention? *Language Learning & Technology* 3(2): 58-76.

Lemnitzer, L. 2001. Das Internet als Medium für die Wörterbuchbenutzungsforschung. Lemberg, I., B. Schröder and A. Storrer (Eds.). 2001. *Chancen und Perspektiven computergestützer Lexikographie. Hypertext, Internet und SGML/XML für die Produktion und Publikation digitaler Wörterbücher:* 247-254. Tübingen: Max Niemeyer Verlag.

Lew, R. 2011a. User Studies: Opportunities and Limitations. Akasu, K. and S. Uchida. (Eds). 2011. *ASIALEX2011 Proceedings. Lexicography: Theoretical and Practical Perspectives. Papers Submitted to the Seventh ASIALEX Biennial International Conference, Kyota, Japan, August 22–24, 2011:* 7-16. Kyoto: Asian Association for Lexicography.

Lew, R. 2011b. Studies in Dictionary Use: Recent Developments. *International Journal of Lexicography* 24(1): 1-4.

Lew, R. and G.-M. de Schryver. 2014. Dictionary Users in the Digital Revolution. *International Journal of Lexicography* 27(4): 341-359.

Müller-Spitzer, C., S. Wolfer and A. Koplenig. 2015. Observing Online Dictionary Users: Studies Using Wiktionary Log Files. *International Journal of Lexicography* 28(1): 1-26.

Nesi, H. 2000. On Screen or in Print? Students' Use of a Learner's Dictionary on CD-ROM and in Book Form. Howarth, P. and R. Herington (Eds.). 2000. *EAP Learning Technologies:* 106-114. Leeds: Leeds University Press.

Orliński, W. 2017, March 17. Sztuczna Inteligencja jest Głupsza od Trzylatka. *Gazeta Wyborcza.* http://wyborcza.pl/magazyn/7,124059,21512640,sztuczna-inteligencja-jest-glupsza-od-trzylatka-orlinski.html?disableRedirects=true. [Accessed 27.04.2017.]

Prinsloo, D.J. and G.-M. de Schryver. 1999. The Lemmatization of Nouns in African Languages with Special Reference to Sepedi and Cilubà. *South African Journal of African Languages* 19(4): 258-275.

Töpel, A. 2014. Review of Research into the Use of Electronic Dictionaries. Müller-Spitzer, C. (Ed.). 2014. *Using Online Dictionaries:* 13-54. Berlin/New York: De Gruyter.

Trap-Jensen, L., H. Lorentzen and N.H. Sørensen. 2014. An Odd Couple — Corpus Frequency and Look-up Frequency: What Relationship? *Slovenščina* 2.0, 2(2): 94-113. https://dsl.dk/medarbejdere/medarbejdere-publikationer-m-m/ltj/an-odd-couple. [Accessed 05.04.2017.]

Verlinde, S. and J. Binon. 2010. Monitoring Dictionary Use in the Electronic Age. Dykstra, A. and T. Schoonheim (Eds). 2010. *Proceedings of the XIV Euralex International Congress, Leeuwarden, 6–10 July 2010:* 1144-1151. Ljouwert: Fryske Akademy — Afûk.

Wójtowicz, B. 2017. Evaluating a 12-Million-Word Corpus as a Source of Dictionary Data. *International Journal of Lexicography* 2017: 1-15. doi: 10.1093/ijl/ecx011.

Zgusta, L. 1971. *Manual of Lexicography.* Prague: Academia.