Learner Features in a New Corpus-based Swahili Dictionary

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

Abstract: As far as traditionally published Swahili language dictionaries are concerned, throughout the long history of Swahili lexicography, most new dictionaries were based on their predecessors. Thus far the only innovative traditionally printed corpus-based dictionary has been published by Finnish scholars (Abdulla et al. 2002). This dictionary takes advantage of the Helsinki Corpus of Swahili (HCS 2004) and was an inspiration for the new Swahili–Polish dictionary project described in this paper. The Swahili–Polish language pair, which without doubt can be called *less resourced*, does not have many lexicographical antecedents on which a new dictionary could be based. Therefore, the new dictionary relies on data from the Swahili corpus.

In this paper, we present a new, corpus-based Swahili–Polish dictionary that has been published online and printed. The paper introduces the resources used to build the dictionary, its learner-oriented features and grammatical assumptions, with a focus on the idea of the visualisation of derivational hierarchies.

Keywords: SWAHILI, POLISH, BILINGUAL LEXICOGRAPHY, DICTIONARY PROJECT, CORPUS-BASED LEXICOGRAPHY

Opsomming: Aanleerderskenmerke in 'n Nuwe Korpusgebaseerde Swahiliwoordeboek. Dwarsdeur die lang geskiedenis van Swahilileksikografie is die meeste nuwe woordeboeke, veral tradisioneel gepubliseerde Swahili taalwoordeboeke, op hulle voorgangers gebaseer. Tot dusver is die enigste innoverende tradisioneel gedrukte korpusgebaseerde woordeboek gepubliseer deur Finse vakkundiges (Abdulla et al. 2002). Hierdie woordeboek trek voordeel uit die Helsinki Korpus van Swahili" (HCS 2004) en was 'n inspirasie vir die nuwe Swahili–Poolse woordeboekprojek wat in hierdie artikel beskryf word. Die Swahili–Poolse taalpaar, wat sonder twyfel beskryf kan word as *gebrekkig aan hulpbronne*, het nie baie leksikografiese voorgangers waarop 'n nuwe woordeboek gebaseer kan word nie. Daarom maak die nuwe woordeboek staat op data uit die Swahilikorpus.

In hierdie artikel lê ons 'n nuwe, korpusgebaseerde Swahili–Poolse woordeboek voor wat aanlyn gepubliseer sowel as gedruk is. Die artikel stel die bronne bekend wat gebruik is om die woordeboek op te bou, sy aanleerdersgerigte kenmerke en taalkundige veronderstellings, met 'n fokus op die begrip van visualisering van afleidingshiërargieë.

Sleutelwoorde: SWAHILI, POOLS, TWEETALIGE LEKSIKOGRAFIE, WOORDEBOEKPROJEK, KORPUSGEBASEERDE LEKSIKOGRAFIE

1. Introduction

Swahili is one of Africa's major languages. It is the most widely-used language of Sub-Saharan Africa with tens of millions of people using it as a *lingua franca* throughout the East African region. It is the national and official language of Kenya and Tanzania (alongside English). Due to its status, the number of speakers and the interest of scholars, numerous Swahili dictionaries have been compiled all over the world, with electronic (online) dictionaries joining the market in the 1990s.

Both at European and American universities, African languages are recognized as less-commonly taught languages. This often goes hand in hand with a scarcity of lexicographic and teaching materials. This is also true for Swahili in non-English speaking parts of the world. Although Swahili has been taught alongside Hausa at the University of Warsaw in Poland since the 1960s, only a few teaching resources have been compiled during this time.

Even though numerous bilingual Swahili dictionaries exist, the only publication of this kind targeting Polish was the small Swahili–Polish and Polish–Swahili dictionary by Stopa and Garlicki (1966), which has been out of stock for a number of years. While students can manage with various English-language textbooks available for studying Swahili, they find it difficult without access to a bilingual dictionary in their native language, especially during the first years of their studies. Until recently, Polish students were forced to use non-Polish dictionaries that are not easily available for purchase. Therefore, they have often taken advantage of various resources that are available on the Internet. Two electronic Swahili–English dictionaries were especially popular and widely-used: for years this included the *Internet Living Swahili Dictionary*[1] — the largest such dictionary and on-line community-based initiative that has now evolved into the *Global Online Living Dictionary*, as well as the *TshwaneDJe Swahili–English Dictionary*[2] (Hillewaert et al. 2012) — the first, and so far the only, corpus-driven electronic Swahili online dictionary. Although the *Internet Living Swahili Dictionary* contained over 50 thousand entries and has been the leading dictionary for years, it suffered as a result of the community-based mode of its creation. It tended to list many minor meanings of the entry in an unsorted manner among the more important (frequent) ones, which beginner learners often found misleading. The *TshwaneDJe Swahili–English Dictionary* is recognised by Polish students as the best dictionary of Swahili at present. It is the first corpus-driven dictionary of Swahili with a new approach to the lemmatisation of headwords (cf. De Schryver et al. 2006). The content is based on web-based corpus data (as the authors themselves describe it, *a balanced and representative Swahili corpus of around fifteen million running words* (De Schryver et al. 2006: 70).

As far as traditionally published dictionaries are concerned, throughout the long history of Swahili lexicography, most new dictionaries were based on their predecessors; thus far the only innovative corpus-based dictionary has been published by Finnish scholars (Abdulla et al. 2002). It was based on data from the Helsinki Corpus of Swahili, the only annotated and publicly available, with free access for researchers, electronic corpus of the language. This dictionary was a source of inspiration for the project described in this paper.

2. The new Swahili–Polish dictionary project

Over the last several years, we have observed a rising interest in African studies at the University of Warsaw, especially in the field of Swahili language studies.

Given the growing interest in learning Swahili, the fact that Swahili dictionaries are not easily available for purchase in Poland, and that the only Swahili–Polish dictionary is now simply out-dated, the compilation of a new Swahili–Polish (and Polish–Swahili) dictionary became a necessity. Due to the shortage of educational materials, a new project involving the creation of a dictionary was proposed by the Department of African Languages and Cultures, and the idea was then approved by the Polish Ministry of Science.

The project aimed to create a new dictionary that would be developed in accordance with recent lexicographic practices. Therefore, the macrostructure as well as the description of entries are based on corpus data from the Helsinki Corpus of Swahili (HCS 2004). The electronic version of the dictionary[3], which is discussed in this article, is primary to the printed version that has also already been published (Wójtowicz 2013). The focus within this article is on the Swahili–Polish direction, as the reverse-language part in the electronic version of the dictionary is delivered in the form of a structured index — a standard in this type of resource.

The dictionary is aimed primarily at Polish students of Swahili. The other target groups include tourists visiting East Africa and anyone interested in the Swahili language and culture. The needs of these three groups partially overlap since the culture-oriented programme of African studies ensures that first-year students acquire vocabulary needed in day-to-day interaction with native speakers of Swahili.

While there are no other Polish resources the students can use, we aimed at creating a dictionary that could also serve as a language learning aid and supplement other foreign teaching materials used in language classes, as well as help students with homework. Furthermore, we also regard the dictionary as a source of cultural knowledge.

The project website, apart from being a dictionary interface, provides sections on *how to use the dictionary*, *the language* — a general introduction to the language and its grammar, a *Swahili phrasebook* — a list of popular and useful phrases, and *Swahili proverbs* that present over 60 different Swahili sayings and proverbs translated into Polish.

The following sections briefly outline the main concepts of the dictionary, as well as the data and software that were used to build it.

3. The Helsinki Corpus of Swahili — the source of dictionary data

The skeleton of the dictionary was derived from the Helsinki Corpus of Swahili (HCS). The HCS is the only large and annotated corpus of Standard Swahili available to the linguistic community. Although it is neither a representative nor balanced corpus, it seems to contain appropriate data considering our dictionary target users. It consists of over 12 million words taken from numerous literary books and current news sources, i.e. texts that an average student of Swahili comes across in his day-to-day learning process. Literary works are read in class, while news items written in Swahili are the most popular and easily accessible texts on the Internet. The size of the corpus may be estimated as small and insufficient as compared to the big corpora available for many European languages. Nevertheless, the HCS is one of the biggest corpora for African or other lesser-resourced languages, and the findings of Prinsloo (2015) prove that a corpus of even a small size should supply enough data to compile a good dictionary for our target group.

The annotation layer of the corpus was provided by SALAMA (Swahili Language Manager, cf. Hurskeinen 2008), an environment for the computational processing of the Swahili language. SALAMA includes a comprehensive language analyser of Swahili text, including morphological analysis, morphological and semantic disambiguation, syntactic analysis, and a bilingual translation module from Swahili to English, making it possible to compile a corpus-based dictionary. The SALAMA Dictionary Compiler (SALAMA-DC), a by-product of SALAMA, is a comprehensive system for producing dictionary entries from any word-form in Swahili. It produces entries with appropriate linguistic information, single-word headwords, multiword headwords, various types of cross-references, and a selection of usage examples in context. Furthermore, the example texts are translated into English and attached to each entry. The whole process from raw text to dictionary entries takes place without manual editing in between, which speeds up the compilation of the dictionary greatly. The output is provided in text format, all as a single file. The dictionary compiled in such an automated process needs manual editing and correction; however, the work needed during this step is only a fraction of the work carried out in manual dictionary compilation.

The dictionary data has three main types of information:

— headword including relevant information (within square brackets)
[kiasi-N-7/8] {quantity, amount, measure} AR 3384

— cross-references (with double square bracket), *taz*. stands for 'cf.'
[kiasi-N-7/8]] {amount} taz. [kadri-N-9/10] [kima-N-7/8]
[kiasi-N-7/8]] {quantity} taz. [idadi-N-9/10]

— examples in context (with triple square bracket)
[kiasi-N-7/8]]] Alifungua bomba, akakinga kiasi [kiasi-N-7/8] na kisha akanywa hali makofi (Opened the pipe, protected the quantity and then drank it is not the slaps)

Since the primary data included in the dictionary has been derived from a tagged corpus, all entries were accompanied by basic linguistic information, depending on the word category, right from the beginning. In this primary data, each entry contains POS information, and furthermore, nouns are described by their class number and animacy categorization, while verbs are described by the type of derivation and references to their base/root in the case of derivatives, or to derivatives in the case of roots. Additionally, all entries are accompanied by their English equivalents. The step of translation into Polish has been further facilitated by the automatic concatenation[4] of the Swahili–English dictionary with the English–Polish dictionary by Piotrowski and Saloni (1992).

Apart from the grammatical information and English glosses, the raw data contains additional information on verb features, frequency counts, the etymology of borrowings, and references to synonyms. What follows are three typical dictionary entries produced by SALAMA-DC. An automatic concatenation with the English–Polish dictionary has been carried out but no further editing was done. In this dictionary compilation process, all verbal extensions have been given the status of separate entries, as in (2).

(1) Noun:
[ardhi-N-9/10] {land, soil (gleba, ziemia)} AR 2247
[ardhi-N-9/10]] {soil} taz. [chafua-V] [chafulia-V] [chafuliwa-V] [dongo-N-5/6] [udongo-N-11]

(2) Verb:
[apa-V] [apa] {swear (przysięgać, przeklinać, kląć), take_an_oath} SV 131
[apa-V]] {swear} taz. [apia-V] [apisha-V] [apishwa-V]
[apia-V] [apa] {swear (przysięgać, przeklinać, kląć), take_an_oath} SVO APPL 9
[apia-V]] {swear} taz. [apa-V] [apisha-V] [apishwa-V]

(3) Adjective:
[bandia-ADJ] A-UNINFL {artificial (sztuczny), spurious (pozorny, fałszywy), counterfeit} AR 153

The status of separate entries was also given to idioms, or more generally, multi-word expressions.

(4)
[fanya_kazi] V SVO IDIOM-V {work} 2584
[unga_mkono] V SVO IDIOM-V {support, be of the same opinion} 951
[makao_kuu] N 6SG MWE {headquarters} 893
[ona_raha-V] SVO IDIOM-V {rejoice (radować się)} 15
[uchaguzi_kuu] N 11/10 MWE {general elections} 791

A great advantage of the SALAMA-DC system is the ability to include into the dictionary multiword examples, isolate them, and produce appropriate translations. Most entries are accompanied by several examples, and each of them is translated into English, as in example (5). Example sentences were cut on both sides of the keyword, preferably in clause boundaries, but if such boundaries were too far, the sentence was cut after a certain number of words. If the headword has more than one interpretation (different part of speech or noun class), each interpretation has its own examples separately, as in (6). Translations are only given to provide some clue of the meaning, and in cases where translations need to be retained, they should undergo extensive editing. Since long examples were cut, the translation sometimes suffers as a result of incomplete sentences, but it undoubtedly sheds light on what the sentence is about. The English translation is a word-to-word gloss of Swahili text.

(5)

[piga_chafya-V] SVO IDIOM-V {sneeze} 31

[piga_chafya-V]]] Hata hivyo mwenzake hakuweza kupiga [piga_chafya-V] chafya wala kujitingisha. (However the countryman did not be able to sneeze nor to encircle self.)

[piga_chafya-V]]] Pia kutema ovyo na kupiga [piga_chafya-V] chafya karibu na mtoto ni hatari (Also to cut carelessly and to sneeze near the child is the danger)

(6)

[mpaka-N-3/4] {border, boundary, frontier} 1390

[mpaka-N-3/4]] {border} taz. [pakana-V]

[mpaka-N-3/4]]] Ila bila msaada unaovuka mipaka [mpaka-N-3/4] ya mwanga ungeweza kupotea njia (Except without the assistance whom it crosses the borders of the light it would be able to be lost the way)

[mpaka-PREP] {until, till} 2140

[mpaka-PREP]] {till} taz. [hadi-PREP]

[mpaka-PREP]] {until} taz. [hadi-PREP] [hata-ADV] [lama-ADV] [mpaka-CONJ]

[mpaka-PREP]]] Aliamua kuwa asitelemke Makambako mpaka [mpaka-PREP] baadaye hivyo alipitiliza moja kwamoja (He decided be should not descend Makambako until afterwards in this way surpassed one kwamoja)

[mpaka-PREP]]] Aliendelea kuzunguka huko jikoni, [mpaka-PREP] na mara akaiangalia saa: (He continued to go around there kitchen in, and immediately looked at it the hour:)

The lemma-sign list that was used to build a dictionary consisted of over 10 thousand entries. Most of them, app. 50% were nouns, over 20% were verbs, 10% adjectives, and 6% adverbs plus other parts of speech and multiword expressions. This list was compared with vocabulary from students' books like *Colloquial Swahili* (McGrath and Marten 2003) and *Tusome Kiswahili* (Muaka and Muaka 2006). All vocabulary from the first book was present on the list and around 50 entries were identified as missing from the other book. These were mainly multiword expressions, like *kitinda mimba* 'last born child', *baba wa kambo* 'step father', *chama cha siasa* 'political party', *mwandishi wa habari* 'journalist' and were added to the dictionary. Furthermore closed sets like days of the week, months, and pronouns were verified and names of countries and continents added. Additional vocabulary was also verified and supplemented by students who worked on chosen sets they found useful in their studies, like animals, musical instruments, or means of transport.

The first phase of dictionary editing was cleaning up messy entries with the help of regular expressions. Several entries were merged together with no space in between, several had wrong examples attached, as the last example in (7) — for *toka* instead of *tohara*. Some entries were wrongly identified, had no POS information or senses attached. But the main work was devoted to identifying wrong translations from English to Polish. In example (8) *akili* is well translated into English as 'intelligence, intellect' but further into Polish as *inteligencja* 'intelligence' and *wywiad* that means in Polish both 'intelligence service' and 'interview', but none of them is an appropriate translation of the Swahili word *akili* and only the first translation into Polish, *inteligencja*, should make it into the dictionary. In (9) English 'interest' is translated into Polish as 'interests, hobby' and the right interpretation of 'bank interest' is completely missing.

(7)

[tohara-N-9/10] {circumcision (obrzezanie), purity (czystość), cleanliness (schludność)} AR 76

[tohara-N-9/10]] {cleanliness} taz. [unadhifu-N-11] [usafi-N-11]

[tohara-N-9/10]]] <ALA> akasema kampeni dhidi ya tohara [tohara-N-9/10] kwa wanawake mbali na kuhitaji (Said give him/her against the circumcision with the women apart from to need)

[tohara-N-9/10]]] <ALA> wazazi kuharakisha kuwapeleka kwenye tohara

[toka-MWE]]] <ART> jumla ya sh milioni 10,826,740 [toka-MWE] katika wilaya zake na kwamba (The total of sh million in his/her/its districts and that)

(8)

[akili-N-9/10] {intelligence (inteligencja, wywiad), intellect (intelekt, inteligencja), nous} AR 1614

(9)

[riba-N-9/10] {interest (zainteresowanie, hobby, interes), usury (lichwa)} AR 146

Already identified senses were further researched in the corpus and completed when needed. For example, based on more corpus examples, *ghorofa* 'floor' got a second sense of 'seat, office'. Also noun class membership was verified and supplemented when needed. *Dawa* 'medicine' was initially described as belonging to class 5/6 but based on corpus examples class 9/10 membership was also added.

4. FieldWorks Language Explorer

Originally, one of the aims of the project was the encoding of the dictionary in XML conformant with TEI P5 Guidelines (TEI Consortium 2015). The virtues of such a choice lie in its extensibility, interchangeability, manageability, ease of maintenance, and the possibility of creating virtually any kind of output.

To enable easy editing, the data has been converted to the LIFT format for the time being and imported into the FieldWorks Language Explorer (FLEx), free software available from SIL[5]. FieldWorks Language Explorer consists of tools that help manage linguistic and cultural data. It is a powerful tool aimed at helping field linguists to perform many common language documentation and analysis tasks, among others, the dictionary development process. To mention but a few of its features, the lexical data can be formatted within FLEx for previewing before publication, in the print-oriented or electronic dictionary view, the order and formatting of fields can be changed, and it supports root-based or stem-based arrangements of entries. New custom fields can be added, existing ones omitted, reordered, formatted, and example sentences selected, so different publications can be derived from one database. With the possibility to deselect each entry, sense, and example sentence we can for instance create anything from a pocket edition to a full dictionary. It supports an extended system of cross-references, illustrations can be attached and data exported among others to XML. It also offers semantic domain-based, built-in descriptions of the lexicon, which could be the possible next step for the development of the dictionary under presentation. The main drawback of the system is the lack of a website publication module. It relies on the plug-in *Pathway*[6] to create dictionary documents in various formats, and *Webonary*[7] to publish the dictionary online. The data needs to be exported in XHTML format and then imported by *Webonary* that supports the creation of the dictionary website.

5. The concept of the new Swahili–Polish dictionary

Given the linguistic orientation of African studies, this general student-oriented dictionary contains extended grammatical information, above and beyond the scope of what is minimally necessary. It is assumed that even though the primary function of this Swahili–Polish dictionary is to provide help in the process of text reception, it may also provide grammatical information that might be helpful in text production.

The dictionary is not strictly translational but rather of a descriptive-translational type. As the cultural differences are substantial and the contacts between the two cultures slight, we aimed at providing culturally-bound entries with broad explanations and possibly pictures in the future. For this reason the list of entries also includes geographical names.

The dictionary is regarded as a work in progress and by closely monitoring user queries, new entries are constantly being added while the existing ones are updated. The dictionary may be searched in both directions with nearly 6100 Swahili entries and over 7000 entries in the searchable Polish index. The searches are carried out on headwords and plural forms of headwords for Swahili nouns. If a user looks up *wapishi* 'cooks', he will be directed to the entry for *mpishi* 'cook'.

5.1 The macrostructure and the treatment of derivatives

The selection of headwords for the Swahili–Polish dictionary (ultimately 10,000 entries, published incrementally) has been made primarily on the basis of a frequency list derived from the corpus, and further expanded with words from various Swahili textbooks and other material used in language classes. The dictionary headword list also includes geographical names and grammatical morphemes. Since it is not possible to translate the morphemes into Polish, their function in Swahili is explained.

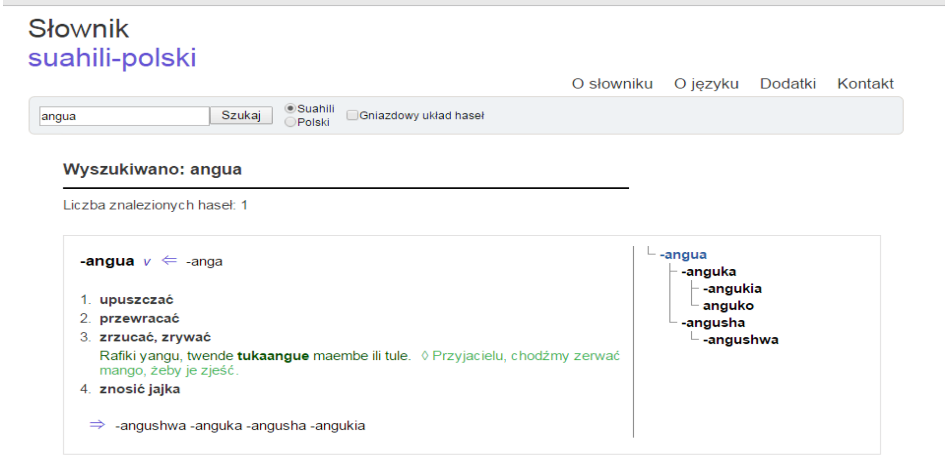
While working with a Bantu language we had to address problems not experienced by lexicographers working with European languages. These problems are connected primarily to two issues: the form of headwords and the presentation of the numerous derivatives of a single root.

As a Bantu language, Swahili is an agglutinative language, which means that morphemes are juxtaposed to form words. Within the Swahili lexicographic tradition, the accepted lemmatisation strategy is to list nouns in their full forms with class prefixes, whereas the prefixes of verbs, numerals, pronouns and inflected adjectives are ignored and the stems alone are listed (see Kiango 2000 for a thorough discussion of this issue, and De Schryver et al. 2006 for a novel approach of lemmatising full orthographic words in addition to stems). Given the dictionary culture of our students, we have decided to follow chosen solutions of the lexicographic tradition, used within nearly all Swahili dictionaries, in regard to the process of lemmatization. Therefore, we ignore prefixes and list the stems alone for verbs, numerals, and inflected adjectives. On the other hand, in response to beginner learners' needs, we reject this tradition in some other cases and list pronouns in their full forms (with the stems also included as separate entries).

Another important issue with respect to macrostructure is related to the handling of derivatives, sometimes referred to as the "lumping vs. splitting" debate (cf. Bański and Wójtowicz 2011). Derivation in Swahili is very robust and typically creates dozens of complex lexemes from a single root since Bantu derivational word-families can be extremely numerous, especially those based on verbal roots — for example, De Schryver and Prinsloo (2001: 225ff) estimate that there are over 140 regular derivatives of the root *reka* 'buy, purchase' in Sepedi.

To meet the expectations of the users who find it difficult to cope with root-based dictionaries of Bantu languages, we follow the so-called splitting approach as the default method: rather than lumping all related lexemes in a single entry headed by the root form, we place derivatives of verbs in separate entries, thereby breaking the semantic and lexical connections between the individual derivatives and their respective bases. In order to maintain a system whereby derivatives have the status of headwords, while simultaneously the derivational and semantic relationships between forms are preserved, the dictionary uses an extended mechanism of cross-entry references. Also in the process of implementing the structure in the form of XML, we realised that what we should do is treat the lumping vs. splitting debate not as a deep issue concerning the structuring of the data, but rather as a surface issue on how to present the data to the user. Therefore, emphasis is placed on the presentation and visualisation of derivational families, as this is regarded as an educational feature useful for the development of our students' linguistic skills.

Traditional cross-entry references, especially among word-families, offer a one-sided view of derivational relationships (derivative → root). Introducing run-on entries offers a view from the opposite side (root → derivatives). Typically, however, word-families feature more than two generations of words, and quite often the link between the ends of the chain (root ↔ complex derivative) is either unclear to the average speaker or at least not as important as the relationship between the immediately related lexemes. Therefore, as an addition to traditional cross-entry references, we introduce a visualisation of the derivational families showing explicitly how the lexemes relate to each other. In the derivational tree, the derivational bases point to the next level of the derivational hierarchy only, and derivatives point to their derivational bases, which crucially need not be the same as their roots. The user is presented with information on the root or its verbal derivatives within the entry and the whole derivational family tree in the form of an information graph accompanying the entry.



angua

● Suahili
○ Polski

□ Gniazdowy układ haseł

Wyszukiwano: angua

Liczba znalezionych haseł: 1

-angua *v* ⇐ -anga

1. **upuszczać**
2. **przewracać**
3. **zrzucać, zrywać**
   Rafiki yangu, twende **tukaangue** maembe ili tule. ◊ Przyjacielu, chodźmy zerwać mango, żeby je zjeść.
4. **znosić jajka**

⇒ -angushwa -anguka -angusha -angukia

```
└ -angua
   ├ -anguka
   │  ├ -angukia
   │  └ anguko
   └ -angusha
      └ -angushwa
```

Making users aware of the structure of the hierarchy in one case reinforces their knowledge of the possible derivational patterns that can be productively applied in other cases, i.e. to the creation of new forms or to the analysis of newly encountered words, which need not be present in the dictionary due to their low text frequency.

To overcome the lump or split debate and exploit the possibilities offered by the new media, it is also possible to lump or split entries of verbal derivatives. The user may choose if he wants to see nested entries (all derivatives presented under their roots) or all derivatives as separate entries (no other derivatives are then attached but the information about the root is still provided).

As for homonyms, they are determined on the basis of their morphological features, i.e. as different parts of speech.

5.2 The microstructure

The structure of an entry includes the following elements: the headword, variant/variants of the headword; grammatical information dependent on the POS and the properties of the individual lexical item; equivalents, definitions[8]; examples of use; idioms that the headword is part of; collocations; synonyms; etymological information.

Each dictionary entry is described with the basic grammatical information given in most of the dictionaries, such as the part of speech, noun class number, type of verbal derivation and the root — if applicable.

-lisha *v causative* [< -la] feed

Additionally, the dictionary also marks different grammatical features that the beginner student may find helpful, like the animal/human distinction for nouns from non-human classes that aims at drawing the user's attention to the agreement structure. To facilitate interpreting, subsequent parts of dictionary entries are presented using different colours, e.g. grammatical information is provided in blue and examples in green.

The main additional learner-oriented features of the dictionary are the following:

— explicitly marked class prefixes on nouns and pronouns and full plural forms of nouns:

m | pishi (*pl* ***wapishi***) *n 1/2* cook

— animacy of nouns from non-human classes (the distinction animate/human is introduced):

n | dege (*pl ndege*) *n 9/10* ***animate*** bird

— irregular verbal forms (imperative or retaining of the augment *-ku-*):

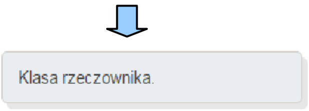
-ja *v* ***ku*** **(imp. njoo)** come

— highlighting of the most frequently used words (the entry is red and labelled with a red dot):

-pata● *v* get

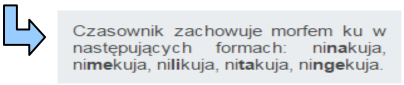
In order to present the information more explicitly, pop-up windows and links are exploited. Moving a cursor over numbers in noun entries evokes a pop-up window with information that these represent the class membership of a given noun. Additionally, the same numbers are also hyperlinked to the agreement chart, where agreement patterns for animate/human subjects from non-human classes have been added. Pop-up windows also inform the users about the relation of the augment *-ku-* with different tenses or why some entries are printed in red.

m | pishi (*pl wapishi*) *n 1/2* cook[9]



Klasa rzeczownika.

-ja *v ku* (imp. njoo) come[10]



Czasownik zachowuje morfem ku w następujących formach: ni**na**kuja, ni**me**kuja, ni**li**kuja, ni**ta**kuja, ni**nge**kuja.

A cross-reference system is also exploited to link synonymous entries, like *ndege* ↔ *eropleni* 'airplane'. Subsequent senses are listed based on their frequency and specialised senses are described with field labels. Real examples extracted from the corpus accompany many entries, and where needed useful phrases and multiword expressions have also been added.

6. Conclusion

The new Swahili–Polish dictionary presented in this article aims at solving the issue of a shortage of Swahili-language learning materials in the Polish market. Created with students of Swahili as the main target group, it provides a variety of learner-oriented features that will possibly simplify the task of acquiring the language and understanding Bantu morphology. The introduction of hierarchical derivational trees reinforces their knowledge of derivational patterns and provides an additional access route to the lexicon.

To keep the dictionary up-to-date, its macrostructure was based on the corpus-driven data from the Helsinki Corpus of Swahili. Most of the features of this data have been preserved in the final version of the dictionary. Within this paper, the primary dictionary data has been presented, while also discussing the main dictionary aims and assumptions already addressed.

Endnotes

1. The dictionary used to be available at http://kamusi.org.
2. http://africanlanguages.com/swahili. Accessed on 06/05/2016.

3. http://kamusi.pl. Accessed on 06/05/2016.
4. With the invaluable assistance of Prof. Arvi Hurskainen, to whom I am very thankful.
5. http://fieldworks.sil.org/. Accessed on 06/05/2016.
6. http://pathway.sil.org/. Accessed on 06/05/2016.
7. http://www.webonary.org/. Accessed on 06/05/2016.
8. When additional explanation is needed, for example of family relations not present in Polish as older brother, or different names for sisters/brothers of our parents.
9. Text in the box: 'noun class'.
10. Text in the box: 'the verb retains the morpheme *ku* in the following forms *ninakuja*, *nimekuja*, *nilikuja*, *nitakuja*, *ningekuja*'.

References

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

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.

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., 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.

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

Hillewaert, S., P. Joffe and G.-M. de Schryver. 2012. *Kamusi ya Kiswahili–Kiingereza Katika Mtandao/ Online Swahili–English Dictionary.* Available from: http://africanlanguages.com/swahili [01/09/ 2016].

Hurskainen, A. 2008. SALAMA Dictionary Compiler - A Method for Corpus-Based Dictionary Compilation. *Technical Reports in Language Technology,* Report No 2, 2008. Available from: http://www.njas.helsinki.fi/salama/salama-dictionary-compiler.pdf [01/09/2016].

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.

Muaka, L. and A. Muaka. 2006. *Tusome Kiswahili. Let's Read Swahili: Intermediate Level.* Madison Wisconsin: NALRC Press.

McGrath, D. and L. Marten. 2003. *Colloquial Swahili: The Complete Course for Beginners.* London: Routledge.

Piotrowski, T. and Z. Saloni. 1992. *Nowy słownik angielsko–polski polsko–angielski* [New English–Polish, Polish–English Dictionary]. Warszawa: Editions Spotkania. Electronic version available from: http://clip.ipipan.waw.pl/Nowy_slownik_angielsko-polski [01.09.2016].

Prinsloo, D.J. 2015. Corpus-based Lexicography for Lesser-resourced Languages — Maximizing the Limited Corpus. *Lexikos* 25: 285–300.

Stopa, R. and B. Garlicki. 1966. *Mały słownik suahilijsko–polski i polsko–suahilijski* [A small Swahili–Polish and Polish–Swahili Dictionary]. Warszawa: Wiedza Powszechna.

TEI Consortium (Eds.). 2015. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. [2.9.1] Available from: http://www.tei-c.org/Guidelines/P5/ [10/11/2015].

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