Multi-level Reference Hierarchies in a Dictionary of Swahili

Piotr Bański
Beata Wójtowicz
University of Warsaw

This paper can be classified into at least two categories: Computational lexicographty and Reports on lexicographical projects, bordering on yet another, the dictionary-making process. The context is a lexicographic project that creates an electronic, TEI XML-encoded Swahili-Polish learner dictionary-with a goal of 10,000 entries in the first stage. Here, we focus on one of the innovative features that we want to introduce in the dictionary, at a relatively small cost-due to the way the dictionary will be compiled out of a Swahili corpus: explicit visualization of derivational hierarchies-essentially a learner-oriented feature, but also serves as a basis for further lexicographic/lexicological applications. We primarily discuss our motivation for this idea and its XML implementation. Nevertheless, by the Conference date, we should also be able to present an actual visualization of it, going beyond a mere set of colourful hyperlinks, which is the way it is presented in our test dictionary-composed of 300 hundred selected illustrative entries, currently being expanded to 1,500, for database testing.

1. Introduction

The present paper describes the linguistic and practical basis for introducing an innovative learner-oriented feature, allowing the user to trace and traverse the derivational history of complex lexemes. The idea is to visualize the structure of derivational families of Swahili words, thus making it easier for users both to perceive the morpholexical regularities and to browse the dictionary along the branches of derivational trees. The proposal relies on the high degree of regularity of Swahili derivational morphology and is illustrated by examples drawn from a small Swahili-English electronic dictionary that serves as testing ground for developing the architecture of a new Swahili-Polish dictionary, which we describe in Bański and Wójtowicz (in print). The dictionary is encoded in XML conformant with the TEI P5 Guidelines (Sperberg-McQueen and Burnard 2007).[1]

We begin by discussing the basic features of Swahili morphology and showing what lexicographic decisions they prompt (section 2). Next, we focus on the way to encode the relationships between derived lexemes and their roots (section 3), and finally, in section 4, we present our own proposal for extending this system of relationships as well as the advantages that such an extension may bring. Section 5 concludes the discussion and sketches the directions for future research.

2. Swahili inflection and derivation as guidelines for architectural decisions

As a Bantu language, Swahili is characterized by agglutinative morphology and concord classes; inflection is primarily prefixal, while derivation—primarily suffixal, with a small degree of allomorphy. Due to their complex derivational and inflectional systems, Bantu languages pose problems not experienced by lexicographers working with European languages. These problems concern primarily two issues: the form of headwords and the presentation of the numerous derivatives of a single root.

2.1. *Inflection: trimming the headwords*

With respect to the former issue, our dictionary follows the traditional lexicographic solutions introduced in most of Swahili dictionaries published so far (see Kiango 2000 for thorough discussion) and rather than listing fully

[1] There has been a drastic set of changes in the dictionary-related part of TEI specifications introduced around September 2007, on the way to the final release of TEI P5 1.0. While we are not extremely comfortable with all of them, we are going to eventually convert the dictionary to this new(er) format, although for the time being we feel more at ease working with a customized schema that is intermediate between versions 0.6 and 1.0 of the Guidelines.

inflected adjectives, numerals, some pronouns and verbs, it lists their stems, with grammatical prefixes removed. The necessity of this move is illustrated by the paradigm below.

(1) a. ku-leta "to bring, to fetch"
INF-bring

b. ni-na-leta "I bring" e. tu-na-leta "we bring"
1SG-PRES-bring
c. u-na-leta "you bring" f. m-na-leta "you (PL) bring"
d. a-na-leta "he/she/it brings" g. wa-na-leta "they bring"

If the infinitive were chosen as the citation form, all verbs would end up under the letter "k". Listing all inflectional forms would lead to massive redundancy, which we are going to reduce by delegating the task of analysing such complex forms to a separate component of the dictionary (part of the user interface) that will query the dictionary itself for bare stems listed in the form exemplified below (where the leading hyphen indicates that the headword is a stem).

(2) a. **-leta** *v* (imp. *lete*) bring; fetch

Having dealt with the form of headwords, we move to the discussion of their distribution and interrelatedness, which forms the basis on which our proposal rests.

2.2. *Derivation: splitting word-families*

Swahili is an agglutinative language, which means that in the majority of cases a single morpheme signals a single lexical operation or a single grammatical feature. Derivation in Swahili is very robust and typically creates tens of complex lexemes from a single root. Derivational operations can be multiplied, as illustrated in the example of the verb *-timilizishiana* "cause to carry out for each other's benefit" that derives from *-timu* "be complete", from Schadeberg (1992:10).

(3) -tim-il-iz-ish-i-an-a < -timu "be complete"
-tim-appl-caus-caus-appl-rec-a
"cause to carry out for each other's benefit"

This naturally brings up the question concerning the way in which the derived forms should be presented—whether e.g. *-timilizishiana*, along with the other derivatives of *-timu*, should be listed under the entry describing the root, or whether it should head a separate entry. In other words, is it better to lump, or to split.[2]

We follow the so-called splitting approach as the default: 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, and distributing word families across the entire dictionary, partly in the form of stems.[3] The alternative, i.e. keeping all derivatives together in a single entry, is not a feasible approach, given the robustness of Bantu derivational morphology (see De Schryver and Prinsloo (2001) for arguments against lumping in Bantu). Eliminating the cost of splitting word-families is the topic of the next section.

[2] The modern reflex of the lumping vs. splitting debate for highly derivational agglutinative languages in the context of electronic dictionaries can be traced to Weber's (2002) reaction to the postulates put forward by Bell and Bird (2000). Weber argues, contra Bell and Bird, for lumping Quechua derivatives inside the entries for their roots. Bosh et al. (2007) follow Weber's example for South Bantu, while e.g. De Schryver and Prinsloo (2001) take the opposite approach. We follow De Schryver and Prinsloo to some extent, noting however, that the lumping vs. splitting opposition can be treated as an issue of data presentation rather than data architecture (see Bański and Wójtowicz (in print). Our default approach is to split, but we leave the final choice to the users, who can decide on the format of entries displayed to them.

[3] Not all derivatives are verbs that are mostly derived by suffixation and that therefore would usually cluster around the entry of the root. Some of them are nouns with various class prefixes, as shall be demonstrated below. Nouns are listed together with their prefixes (see Kiango 2000 for arguments) and therefore must be located in different parts of the dictionary.

3. Inter-lexeme links: reuniting word-families

Example (4) below presents two related entries (simplified after transformation from the original XML): an entry for the stem *-leta* “bring” in (4a), and its irregular imperative form in (4b). Each of them contains a textual reference to the other.

(4) a. -**leta** *v* (imp. *lete*) bring; fetch
b. **lete** *v* imp. of *-leta*

Cross-entry references of this type have been present in dictionaries since the dawn of lexicography. The development of online dictionary systems has, in a natural way, resulted in turning textual cross-references into hyperlinks, which are becoming a standard solution in most online/electronic dictionaries. In the examples of intra-lexeme relations above, the references go both ways: from the stem to the inflected form, and the other way round. Swahili lexicographic tradition has done the same with respect to inter-lexeme relations: for example, Johnson (1939/1985) and Abdulla et al. (2002) use references pointing from derivatives to their roots, while Sacleux (1939) and TUKI (2001) use so-called run-on entries that point from roots to the most important derivatives.[4] Online Swahili dictionaries, Kamusi[5] and Swahili-English Dictionary[6], use (often unsystematically) pointers in both directions.[7] Such references are exemplified below in the entries for the verb *-sema* ‘speak’ and its derivatives; the beginning of the run-on list (referred to as the “tail-slot” by De Schryver and Prinsloo 2001, who supplied this example) is signalled by the “>” sign.[8]

(5) -**sema** *v* speak, say > msemaji; msemo; -semekana; -semwa; usemi

(6) a. **msemaji** *n* [< -sema] spokesperson
b. **msemo** *n* [< -sema] saying, slogan
c. -**semekana** *v* [< -sema] be said, be claimed
d. **usemi** *n* [< -sema] style of speaking, (*gram.*) word

Given the number of possible derivatives for a typical Bantu verb (example 6 contains only selected high-frequency derivatives of *-sema*), such run-on entries can significantly increase the volume, and therefore the cost, of traditional print dictionaries. In electronic dictionaries, they come cheap. Figure 1 below shows two trimmed XML representations: of the root verb with the derivatives listed as content of `<ref>` elements at the end, and of the noun *msemaji* in (6a).

```
<entry n="sema" type="root">
  (...)
  <xr type="run-on">
    <ref target="msemaji">msemaji</ref>
    <ref target="msemo">msemo</ref>
    <ref target="semekana">-semekana</ref>
    <ref target="usemi">usemi</ref>
  </xr>
</entry>

<entry type="derived" n="msemaji">
```

[4] Abdulla et al. (2002) use in fact a hybrid approach: the entries for verbal roots list verbal derivatives.

[5] *http://www.kamusiproject.org/*.

[6] *http://africanlanguages.com/swahili/*.

[7] As has been pointed to us by an anonymous reviewer, some modern print dictionaries (e.g. De Schryver and Mogodi 2007) also make it possible for the user to construct derivational hierarchies. We happily take this as confirmation of the general methodology advocated here and note that visual presentation of these hierarchies should be even more attractive to the user.

[8] De Schryver and Prinsloo (2001), following De Schryver (1999), argue for “frequency-based” tail-slots, in order to restrict their number—Bantu verbal roots can easily have over a 100 regular derivatives. In the dictionary discussed here, the frequency factor is taken into consideration in the process of compiling the dictionary from the Helsinki Corpus of Swahili: tail-slots list only the derivatives that have made it into the dictionary.

```
  (...)
  <xr type="derived_from">
    <ref target="#sema">-sema</ref>
  </xr>
  <sense><trans><tr>spokesperson</tr></trans></sense>
</entry>
```

Fig. 1. Fragment of the entry for the root *-sema* “speak”, illustrating forward references from the root to all of the important derivatives, followed by the entry for the noun *msemaji* “spokesperson”, derived from *-sema*.

4. Extending the system: multi-level word-family hierarchies

Traditional cross-entry references, especially among word-families, offer 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.

As an example, consider the example of the word *establishmentarian*: while the information on the root, *establish*, may in some way be helpful to dictionary users, it is more critical for them to know that the word has a lot to do with the intermediate form, *establishment*, as *establishmentarian* derives its meaning only very indirectly from the root and crucially depends on the semantic drift that resulted in one of the lexicalised meanings of *establishment*. The semantic connection between *establish* and *antidisestablishmentarianism* is even thinner. Allowing the interested user to look directly at the hierarchy of intermediate forms would certainly be of value.

The two sets of examples that follow illustrate similar phenomena in Swahili.

(7)	a.	enda	– endesha	– mwendeshaji
		“go”	“drive”	“driver; administrator”
	b.	zaa	– zalisha	– mzalishaji
		“give birth”	“assist at childirth; produce”	“producer”
	c.	tengenea	– tengeneza	– mtengenezaji
		“be arranged”	“manufacture, prepare”	“manufacturer, producer”

The examples in (7) present a regular sequence of root verb—causative verb—agent noun derivations, with a tendency for a semantic drift in the middle form. The agent noun often builds on the extra enrichment of meaning that the causative undergoes, so presenting the intermediate form to the user explicitly, apart from the root, may ensure better understanding of the lexical and semantic regularities.

A slightly different set of problems is manifested below, for a sequence of root—conversive—stative verbs:

(8)	a.	anga	–	angua	– anguka
		“fly”		“drop”	“be down, fall down”
	b.	funga	–	fungua	– funguka
		“close”		“open”	“be openable, loose”
	c.	ziba	–	zibua	– zibuka
		“block”		“unblock”	“be opened up, open”
	d.	panga	–	pangua	– panguka
		“arrange”		“scatter”	“be scattered”

Stating that *-anguka* “fall down” in (8a) derives from *-anga* “fly” can confuse learners rather than help them: *-anguka* is a regular stative derivation from *-angua* (whereby a [k] is added to the stem; the final [a] is a Bantu verbal ending), while *-angua* is a regular conversive derivation from *-anga*.[9]

“Regularity” is the keyword here: each step of the derivation presents a pattern that the learner should internalise, because it is highly regular and it may reapply at a later stage of the derivation. Making users aware of the structure of the hierarchy in one case reinforces their knowledge of the possible derivational patterns that

[9] The Swahili-English Dictionary at africanlanguages.com goes around this difficulty by presenting *-angua* as the root of *-anguka* (and listing *-anga* separately). In contrast, the relationship between *-fungua* and *-funga* is explicitly noted there. We feel that more can be done about this for the benefit of the user.

can be productively applied in other cases: 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.

Given the facts presented above, we propose for run-on entries of derivational bases to point to the next level of the derivational hierarchy only, and for derivatives to point to their derivational bases, which, crucially, need not be the same as their roots. Example (9) below shows a fairly simple hierarchy where the verb *-la* "eat" is the root. Some of the lexemes are derived from the causative verb *-lisha* "feed", which in turn derives from *-la*:

(9)	**-la** *v* eat	
	-liwa *v pass* be eaten	(<la)
	-lika *v stat* be edible	(<la)
	-lisha *v caus* feed	(<la)
	-lishisha *v dcaus* feed with, cause to eat	(<lisha)
	mlisha *n 1/2* waiter/waitress	(<lisha)
	mlishi *n 1/2* one who feeds	(<lisha)
	mlisho *n 3/4* feeding	(<lisha)
	mla *n 1/2* eater	(<la)
	mlaji *n 1/2* consumer	(<la)
	mlo *n 3/4* meal	(<la)
	ulaji *n 11* eating	(<la)

A fragmentary tree visualizing these dependencies is shown below in Figure 2. It is important to stress that the resulting system can be folded into a flat dependency tree created by the flat-hierarchy approach illustrated in (5)-(6). Thus, our proposal does not cause any information of the flat-hierarchy approach to be lost, while offering a more fine-graded view of the inter-lexeme relationships within a word-family.

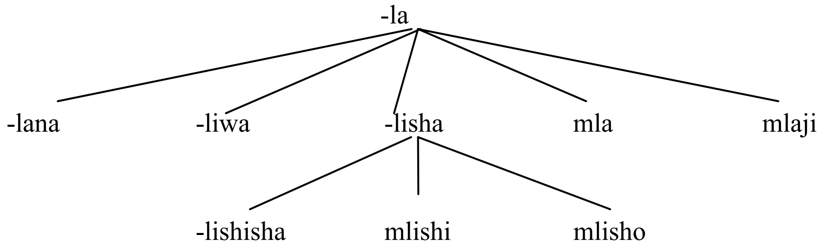

Fig. 2. Fragmentary tree of word-family relationships for the root *-la*

What is planned for the web version of the dictionary is to provide optional mouse-over visualization of the place of the given word in the entire word-family, thus allowing the user to easily navigate among the particular lexemes.[10] This is the ultimate step in restoring the cohesion of split word-families without lumping them all into single dictionary entries. At the same time, the robust derivational patterns of Swahili are made more accessible to the learner.

5. Conclusion and future research

The arguments for implementing visualisation of the derivational patterns of Swahili are of two kinds: systemic (re-establishing the broken links between closely related lexemes, in order to aggregate the necessary information or to navigate the dictionary) and didactic (making the student aware of the regularities in the maze of Bantu morphology).

Providing explicit links of the form advocated here makes it possible for us to effectively ignore the splitting vs. lumping debate and to treat it as a question of delivery (i.e. of data display) rather than a fundamental question of macrostructural decisions: the default that we work with when constructing the dictionary (and to which we are effectively forced by the electronic resource that the entire dictionary is based on—the Helsinki Corpus of

[10] We view this as a question of customizing the existing packages for visualizing semantic hierarchies, such as Visuwords (*http://www.visuwords.com/*) or Jambalaya (*http://www.thechiselgroup.org/jambalaya*). We are currently experimenting with both.

Swahili (HCS 2004)) is a splitting approach. But nothing prevents the user from folding the scattered derivatives into a single, structured, lumped entry, at a single click. The visualization component can in practice be built on either approach, as long as they feature explicit inter-lexeme links.

Naturally, it is not always possible to establish the linguistically correct hierarchy, due to many diverse factors. Sometimes, the intermediate forms may be missing in the dictionary because they do not exist or are below the frequency threshold that would qualify them for inclusion. In such cases, it is possible to fall back to the safe system of flat relationships for the more difficult parts of the tree. The question of distinguishing between fully productive and unproductive processes (in a *learner* dictionary) is also something that deserves attention.

It is such problematic cases that we intend to focus on in our subsequent research, to make sure that the endeavour is worth introducing (and recommending) in dictionaries of Swahili and morphologically similar languages. Another direction for future research involves looking at languages with regular semantic relationships between derivatives but less transparent morphological signals of the particular derivations, with a view towards adding more meaning-oriented visualisations of derivational hierarchies in *active* dictionaries whose overall organisation remains semasiological.

References

Abdulla, A. et al. (2002). *Swahili-Suomi-Swahili sa-nakirja*. Helsinki: Suomalaisen Kirjallisuuden Seura.

Bański, P.; Wójtowicz, B. (in print). "New XML-encoded Swahili-Polish dictionary: macro- and microstructure". In Lewandowska-Tomaszczyk, B. (ed.). *Practical Applications in Language and Computers: PALC 2007*. Frankfurt am Main: Peter Lang.

Bell, J.; Bird, S. (2000). "A Preliminary Study of the Structure of Lexicon Entries" [online]. Paper presented at the workshop on Web-Based Language Documentation and Description. 12-15 December 2000, Philadelphia. *http://www.ldc.upenn.edu/exploration/expl2000/papers/bell/bell.html* [Access date: 30 March 2008].

Bosch, S. E.; Pretorius, L.; Jones, J. (2007). "Towards Machine-Readable Lexicons for South African Bantu languages" [online]. *Nordic Journal of African Studies* 16 (2). *http://www.njas.helsinki.fi* [Access date: 30 March 2008].

De Schryver, G. M. (1999). Bantu Lexicography and the Concept of Simultaneous Feedback, Some preliminary observations on the introduction of a new methodology for the compilation of dictionaries with special reference to a bilingual learner's dictionary Cilubà-Dutch. Ghent University (Belgium) and University of Pretoria (South Africa). Unpublished M.A. thesis.

De Schryver, G. M.; Mogodi, M. (2007). *Oxford Bilingual School Dictionary: Northern Sotho and English*. Cape Town: OUP Southern Africa.

De Schryver, G. M.; Prinsloo, D. J. (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". In Mdee, J. S.; Mwansoko, H. J. M. (eds.). *Makala ya kongamano la kimataifa Kiswahili 2000*. Dar es Salaam: TUKI, Chuo Kikuu cha Dar es Salaam. 188–215.

[HCS]. *Helsinki Corpus of Swahili* (2004). Compilers: Institute for Asian and African Studies (University of Helsinki) and CSC – Scientific Computing Ltd. See *http://www.aakkl.helsinki.fi/cameel/corpus/intro.htm.*

Johnson, F. (1939/1985). A Standard Swahili-English Dictionary (founded on Madan's Swahili-English Dictionary). Oxford: Oxford University Press.

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.

Sacleux, Ch. (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.

Schadeberg, T. C. (1992). *A sketch of Swahili Morphology*. Köln: Rüdiger Köppe Verlag.

Sperberg-McQueen, C. M.; Burnard, L. (eds.) (2007). TEI P5: Guidelines for Electronic Text Encoding and Interchange [online]. *http://www.tei-c.org/Guidelines/P5/index.xml* [Access date: 30 March 2008].

Weber, D. J. (2002). Reflections on the Huallaga Quechua dictionary: derived forms as subentries [online]. *http://emeld.org/workshop/2002/presentations/weber/emeld.pdf* [Access date: 30 March 2008].