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*

Gilles-Maurice de Schryver, University of Ghent

and

D.J. Prinsloo, University of Pretoria

1.0 The Problem

In 1964 Benson, summarising a century of Bantu lexicography, claimed:

> *"It is now right and proper to […] make certain suggestions which could help future compilers of dictionaries of African languages, whoever they may be, to avoid some of the more obvious pitfalls. […] there are no rules laid down for lexicographers, and whatever has been learnt by toil and sweat, by trial and error, is worth passing on. […] One cardinal principle which emerges from our study is that everything which needs to be said about a stem or root should be channelled into one single full article, complete with citations if needed" (Benson 1964: 78, 80, 82)*

However, only one year later Snoxall pointed out in a discussion of a Luganda (J15)[1] – English dictionary:

> *"even many Baganda would have little idea under what root form they should look up many of the commonest words which they use. […] The general principle of entering words in a dictionary under roots, though it was to an extent followed by some of the earlier compilers, could never be of great assistance […] It would seem therefore that, although disappointing perhaps to etymologists, a decision to enter headwords in the form in which they are used in actual speech, as words possessing meaning, […] will be welcomed by the great majority of the users of the dictionary" (Snoxall 1965: 27-8)*

Makala ya kongamano la kimataifa Kiswahili 2000. Proceedings: 216–242, 372

A sound presentation of Bantu lexica has remained a bone of contention ever since. Even two decades later, Bennett had to conclude:

> *"There has been debate as to the proper arrangement of the Bantu lexicon, and the question is far from settled. The inflection of nominals and verbals by means of prefixes, and the complex and productive derivational system, both characteristic of Bantu languages, pose difficulties [...] If items are alphabetized by prefix [...] a verb will be listed far from its nominal derivations, however transparent these may be. [...] A competing school arranges the lexicon by stem or root; this usefully groups related items, and saves on cross-referencing. Unfortunately, in such a system the user must be able to identify the stem, which given the sometimes complex morphophonemics of Bantu languages may not be easy" (Bennett 1986: 3-4)*

In addition to this debate concerning the proper presentation of a Bantu lexicon, most compilers failed to make a satisfactory selection / reduction of the numerous possible verbal and nominal derivations of a verb's *formal radical*[2], in order to stay within the physical limitations of a paper dictionary.

In this article both these problematic aspects are reviewed against the background of one of the most important principles in present-day metalexicography, viz. the user-perspective, with the specific aim to make dictionaries more *user friendly*.

2.0 User-friendliness pinpoints the needs

We can begin by looking into the required selection / reduction of the numerous possible verbal and nominal derivations of a Bantu verb's formal radical. Back in 1971 Zgusta, the father of modern lexicography, noted that:

> *"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" (Zgusta 1971: 16)*

Gove, in the preface of his revolutionary *Webster's Third*, rightly emphasised that:

> *"Selection is guided by usefulness, and usefulness is determined by the degree to which terms most likely to be looked for are included" (Gove 1961[3]: 4a)*

Therefore, if a modern Bantu dictionary is willing to be really practical and useful it must *include all the verbs and their derivations likely to be looked up*. In order to achieve this, today's lexicographers are compelled to build a tool with which the really highly used verbs and their derivations can be pinpointed.

Moving to the quest for a sound presentation of a Bantu lexicon in a printed dictionary, and still with user-friendliness as the main criteria in mind, we can once more quote Zgusta:

> *"the unsophisticated public tends to prefer compartmentalized, quickly digestible information on isolated points of immediate interest" (Zgusta 1989: 301)*

In our experiences with both Cilubà (L31) and Sepedi (S32) dictionaries, this statement, as will be shown below, has been proven very valid. Dictionaries like Gabriël's *Dictionnaire Tshiluba – Français* (*s.d*. [1922]) or Ziervogel & Mokgokong's *Groot Noord-Sotho-woordeboek* (1975) are extremely unpopular with their users, precisely because these dictionaries deviate from a straightforward alphabetical sorting in an attempt to group words on etymological grounds and/or group words under formal radicals of verbs and nouns. Therefore, if a modern Bantu dictionary is willing to be a popular source of reference, it must present verbs and their derivations *under their proper alphabetical position "in the form in which they are used in actual speech, as words possessing meaning"* (Snoxall 1965: 28). In order to achieve this however, today's lexicographers are compelled to devise a functional system of cross-references which restores the semantic and grammatical relations disrupted as a result of such an alphabetical sorting. Indeed, the scattering of semantic relations is the case in any semasiological dictionary, or thus any dictionary in which the direction is from word to explanation, rather than from concept to word.[3] For the Bantu languages, apart from this disruption of semantic relations, an alphabetical sorting has harmful consequences for grammatical relations,

since grammatically related items will be scattered all over the dictionary due to various pre- and suffixes.

3.0 Traditional approaches vs. a possible solution

So far, we looked into blocks **1.0** and **2.0** of the table shown in (1). The present section will be structured as shown in block **3.0** of (1).

(1)

1.0 Problem statement	**2.0** User-friendliness	⇒	Needs	**3.0** Traditional approaches vs. a possible solution	
No selection / reduction procedure	*include all the verbs and their derivations likely to be looked up*	⇒	build a tool with which the really highly used verbs and their derivations can be pinpointed	**3.1** Selection / reduction **3.1.1** random **3.1.2** rule-oriented **3.1.3** enter-them-all	**3.1.4** ***frequency-based***
No sound presenta-tion procedure	*present verbs and their derivations under their proper alphabetical position 'in the form in which they are used in actual speech, as words possessing meaning'*	⇒	devise a functional system of cross-references which restores the semantic and grammatical relations disrupted as a result of such an alphabetical sorting	**3.2** Presentation **3.2.1** split **3.2.2** lump **3.2.3** lump + index	**3.2.4** ***tail slots***

3.1 Selection / reduction procedures

Upon leafing through various Bantu language dictionaries, one notes that three main selection / reduction procedures have been employed until a few years ago, namely a 'random approach', a 'rule-oriented approach' and an 'enter-them-all approach'. Following a brief evaluation of these three, the 'frequency-based approach' will be advanced as a reliable method to limit the number of derivations.

3.1.1 'random approach'

In the 'random approach' the compiler seems to be unaware of (or perhaps even conveniently ignores?) the need to select / reduce the number of verbs and their derivations. In such dictionaries words are simply added whenever

they happen to cross the compiler's way, and compilation is only halted when the required number of pages is reached. Snyman, the editor of a recent Setswana (S31) – English – Afrikaans dictionary, honestly admits in the preface:

> *"The dictionary team is aware of the fact that common and even essential words may easily be omitted during the compiling of a dictionary. This can take place simply because the lexicographer had not encountered such words. We can only hope that there are not too many examples of this kind" (Snyman 1990: preface)*

The very idea that 'common and even essential words' might be excluded from that dictionary is alarming. Inevitably, the 'random approach' also leads to serious imbalances. As an example one can consider the lemmatisation of the derivations of four highly used Setswana verbs in Snyman's dictionary. In (2) **dira**, **reka**, **bona** and **rata** are listed in respect of 12 derivations, such as applicative, causative, perfect, etc.

(2)

Dikišinare ya Setswana – English – Afrikaans Dictionary / Woordeboek *Snyman 1990*				
verbal derivations	***dira*** *'do'*	***reka*** *'buy'*	***bona*** *'see'*	***rata*** *'love'*
applicative	direla	—	bonela	—
causative	dirisa	rekisa	bonesa	—
perfect	—	—	bone	ratile
neutro-passive	—	—	—	ratega
reverse transitive	—	rekolola	—	—
neutro-active	—	—	bonala	—
neutro-active + causative	—	—	bonatsa	—
causative + neutro-passive	—	rekisega	—	—
causative + perfect	—	rekisitse	—	—
neutro-active + causative	—	—	bonatshega	—
neutral + causative	diragatsa	—	—	—
causative + applicative	—	rekisetsa	—	—

It will be hard to explain why the applicative form for **bona** is given but not for **reka**, or the perfect form for **rata** and **bona** but not for **reka**. In both cases the applicative and perfect forms for **reka** (namely **rekela** and **rekile**) scored a

higher overall count than **bonela** and **ratile** in frequency studies conducted on a small-size test-corpus consisting of six Setswana books.

3.1.2 'rule-oriented approach'

In his review of a Zulu (S42) dictionary, Nkabinde argues:

> *"The inclusion of derivative nouns and verbs in a Zulu dictionary is unwieldy and cumbersome. Ideally, only those derivations that have attained an independent meaning from the primary word should be entered" (Nkabinde 1993b: 301)*

However, in doing so one runs the risk of excluding very frequently used (and thus important) derivations, and hence the risk of producing a user-unfriendly dictionary. In fact, Nkabinde's stand is only one aspect of the 'rule-oriented approach' in which one does *not physically enter all* derivations, but only tries to *cover them in theory*. To achieve this, a set of rules / guidelines presented in the dictionary's front matter must be followed whenever a word cannot be looked up directly. In the *Pukuntšu woordeboek* (Kriel, Van Wyk & Makopo 1989[4]) for instance, a Sepedi – Afrikaans dictionary, a small section of the rules is as presented in (3).

(3)

Passives		***example***	
ends with	***look up under***	***ends with***	***look up under***
-bja	-ba	tsebja	tseba
-fša	-fa	lefša	lefa
-ngwa	-ma	rongwa	roma
	-nya	fengwa	fênya
-nngwa	-nya	fenngwa	fênya
-pša	-pa	topša	tôpa
-pšha	-pha	hlopšha	hlôpha

Applicatives		***example***	
ends with	***look up under***	***ends with***	***look up under***
-êtša	-ša	tlošetša	tloša
	-tšha	tsentšhetša	tsêntšha
	-sa	lesetša	lesa
	-tswa	hlatswetša	hlatswa
	-nya	senyetša	senya
-lêtša	-tša	biletša	bitša

It goes without saying that such an approach is far from user-friendly. Firstly, users are known not to allocate much time to such prefatory matters. Secondly, while the utilisation of such rules / guidelines might not ask too much from users when they are dealing with simple extensions, from the moment sequences of three or more extensions are involved, it is very unlikely that users will be able to make the correct analysis. This is especially true for those sequences where sound changes occur, as for instance in **rekišeditšego** (< rek-iš-el-il-go). Thirdly, extremely highly used words might not be in the dictionary, *merely because they are* verbal or nominal derivations. In any case, for every single frequently used omission, inexperienced users will always be in doubt whether or not they made the right decisions while trying, first, to arrive at the formal radical (through a reversal of the grammatical rules), and second, while trying to deduct the semantic meaning (through an application of the grammatical rules).

3.1.3 'enter-them-all approach'

In the 'enter-them-all approach' the compilers are obsessed to include all conceivable nominal and verbal derivations. An example of such a brave effort towards comprehensiveness can be seen in (4), which shows the article of the verb **reka** in the *Groot Noord-Sotho-woordeboek* (Ziervogel & Mokgokong 1975), a Sepedi – Afrikaans – English dictionary.

(4) **Extract from the *Groot-Noord-Sotho-woordeboek*** (Ziervogel & Mokgokong 1975)

RÉKA (**–rêka, –rêkilê, –rêkwa, –rêkilwê**) koop, aankoop, ruil // buy, purchase, barter; ~ *polasa* in weelde lewe // live in comfort/luxury; ~ *o lebêlêtše godimo* kat in die sak koop // buy a pig in a poke; *nku e rékwa mosela* 'n mooi geboude dame is 'n aantrekkingskrag vir jongmans // a lady with a good figure easily attracts young men; **dirékárekáne (dirêkarêkane)** verskeidenheid gekoopte goedere // variety of things bought; **leréko, ma– (lerêkô)** gewoonte/neiging om te koop // habit of buying, inclination to buy; **moréki, ba– (morêki)** pers. dev.; koper // buyer, purchaser; **seréki, di– (serêki)** pers. dev.; lustige koper // keen buyer; **seréko, di– (serêkô)** impers. dev.; wat gekoop word, aankope // purchase(s); **théko, (n–)/di– (thêkô)** man. dev.; koopwyse, prys // manner of buying, price; *RÉKÁNA* **(–rêkana, –rêkane, –rêkanwa, –rêkanwe)** rec.; ruil met mekaar // exchange with one another; *a re rékanê, wéna o mphê hempê êla, nna ke go fê diêta tšê* laat ons met mekaar ruil, jy gee my daardie hemp en ek gee jou hierdie skoene // let us exchange, you give me that shirt, I will give you these shoes; **barékánr (barêkani)** pers. dev.; **thékáno, (n–)/di– (thêkanô)** man. dev.; *RÉKÁNTŠHA* **(–rêkantšha, –rêkantšhitše, –rêkantšhwa, –rêkantšhitšwê)** caus. < *RÉKÁNA*; (om)ruil, wissel (geld), inruil // exchange, barter, trade in, swop; **morékántšhi, ba– (morêkantšhi)** pers. dev.; **serékántšhwá, di– (serêkantšhwa)** impers. pass. dev.; **thékántšho, (n–)/di– (thêkantšhô)** man. dev.; omruiling, inruiling, wisseling // exchange, bartering, swopping; *RÉKÁNYA* **(–rêkanya, –rêkantše, –rêkanywa, –rêkantšwê)** caus. < *RÉKÁNA*; (om)ruil, wissel (geld) // exchange, barter, swop; **morékányi, ba– (morêkanyi)** pers. dev.; **serékanywá, di– (serêkanywa)** impers. pass. dev.; **thékányo, (n–)/di– (thêkanyô)** man. dev.; v. **thékántšho**; *RÉKÉGA* **(–rêkêga, –rêkêgilê)** neutr.; koopbaar w. // b. purchasable; *RÉKÉLA* **(–rêkêla, –rêkêtše, –rêkêlwa, –rêkêtšwe)** appl.; koop vir // buy for; ~ *kolobê kgetsing* (< Afr.) kat in die sak koop // buy a pig in a poke; **borékélo (borêkêlô)** lo. dev.; koopplek // place where things are bought; **morékédi, ba– (morêkêdi)** pers. dev.; **morékelwá, ba– (morêkêlwa)** pers. pass. dev.; **serékélo, di– (serêkêlô)** impers. dev.; iets waarin jy koop // that into which one buys; **thékélo, (n–)/di– (thêkêlô)** man. dev.; maat, skaal (waarin bv. bier gekoop word) // measurement, bowl (one used for buying beer); *RÉKÉLANA* **(–rêkêlana, –rêkêlane, –rêkêlanwa, –rêkêlanwe)** appl. rec.; **barékélani (barêkêlani)** pers. dev.; **thékélano, (n–)/di– (thêkêlanô)** man. dev.; *RÉKÍŠA* **(–rêkiša, –rêkišitše, –rêkišwa, –rêkišitšwê)** caus.; laat/help koop, verkoop, van die hand sit // cause/help buy, sell; ~ *ka leleme* kul, mislei, verdraai // deceive, mislead, pervert; ~ *leleme* praatsiek w., skinder // gossip, b. loquacious, b. garrulous; ~ *motho a sa phela* iemand kul // deceive someone; ~ *motho lebake* iemand kul, 'n tevergeefse belofte maak, iemand verag weens sy slegte gedrag // deceive someone, give a vain promise, despise someone because of his bad conduct; ~ *segaê* iets aan iemand so verkoop dat hy 'n goeie slag slaan omdat jy sy vriend of familielid is, afslag gee // sell to someone at bargain price because he is your friend/relative, give discount; **morékiši, ba– (morêkiši)** pers. dev.; verkoper, verkoopsman, winkelier // seller, salesman, storekeeper; **serékišwá, di– (serêkišwa)** impers. pass. dev.; **thékišo, (n–)/di– (thêkišô)** man. dev.; verkoping, uitverkoping, afset, bemarking // sale, selling, market, marketing; *RÉKÍŠANA* **(–rêkišana, –rêkišane, –rêkišanwa, –rêkišanwe)** caus. rec.; ruil met mekaar // exchange with one another; **barékíšani (barêkišani)** pers. dev.; **thékíšano, (n–)/di– (thêkišanô)** man. dev.; *RÉKÍŠEGA* **(–rêkišêga, –rêkišêgilê)** neutr. < *RÉKÍŠA*; verkoopbaar w. // b. sellable; *RÉKÍŠETŠA* **(–rêkišêtša, –rêkišêditšê, –rêkišêtšwa, –rêkišêditšwê)** caus. appl.; verkoop vir // sell for; **borékišetšo (borêkišêtšô)** lo. dev.; verkoopplek // selling place; **morékišetši, ba– (morêkišêtši)** pers. dev.; †agent // †(business) agent; **thékišetšo, (n–)/di– (thêkišêtšô)** man. dev.; *RÉKÍŠETŠANA* **(–rêkišêtšana, –rêkišêtšane, –rêkišêtšanwa, –rêkišêtšanwe)** caus. appl. rec.; sake verrig // transact business; **barékišetšani (barêkišêtšani)** pers. dev.; **thékišetšano, (n–)/di– (thêkišêtšanô)** man. dev.; besigheidstransaksie // business transaction; *RÉKÓLLA* **(–rêkolla, –rêkolotše, –rêkollwa, –rêkolotšwe)** rev. tr.; terugkoop, terugruil, geld terugvra, los // buy back, exchange back, ask for a refund, redeem; **morékólli, ba– (morêkolli)** pers. dev.; **serékóllwá, di– (serêkollwa)** impers. pass. dev.; 'n ding wat teruggekoop word // that which is bought back; **thékóllo, (n–)/di– (thekollô)** man. dev.; (Bl.) lossing // (Bl.) redemption; *RÉKÓLLANA* **(–rêkollana, –rêkollane, –rêkollanwa, –rêkollanwe)** rev. rec.; **barékóllani (barêkollani)** pers. dev.; **thékóllano, (n–)/di– (thêkollanô)** man. dev.; *RÉKÓLLELA* **(–rêkollêla, –rêkollêtše, –rêkollêlwa, –rêkollêtšwe)** rev. appl.; **morékólledi, ba– (morêkollêdi)** pers. dev.; **thékóllelo, (n–)/di– (thêkollêlô)** man. dev.; *RÉKÓLLELANA* **(–rêkollêlana, –rêkollêlane, –rêkollêlanwa, –rêkollêlanwe)** rev. appl. rec.; **barékóllelani (barêkollêlani)** pers. dev.; **thékóllelano, (n–)/di– (thêkollêlanô)** man. dev.; *RÉKÓLLIŠA* **(–rêkolliša, –rêkollišitše, –rêkollišwa, –rêkollišitšwê)** rev. caus.; **morékólliši, ba– (morêkolliši)** pers. dev.; **thékóllišo, (n–)/di– (thêkollišô)** man. dev.; *RÉKÓLLIŠANA* **(–rêkollišana, –rêkollišane, –rêkollišanwa, –rêkollišanwe)** rev. caus. rec.; **barékóllišani (barêkollišani)** pers. dev.; **thékóllišano, (n–)/di– (thêkollišanô)** man. dev.

rékána v. **RÉKA**
–rékáni, ba– v. **RÉKA**
rékántšha v. **RÉKA**
–rékántšhi, mo–/ba– v. **RÉKA**
–rékántšhwá, se–/di– v. **RÉKA**
rékánya v. **RÉKÁ**
–rékányi, mo–/ba– v. **RÉKA**
–rékanywá, se–/di– v. **RÉKA**

As will be illustrated in the next paragraph, it is clear that the compilers worked through a modular paradigm in order to pursue such a comprehensiveness.

3.1.4 'frequency-based approach'

From the overview of the three 'traditional approaches' above, it is clear that a sound approach would be one in which compilers do not err in allocating precious dictionary space to words unlikely to be looked up at the expense of frequently used ones. It is also clear that words should, in order not to make unrealistic claims on the knowledge of the average user, be given in full (such as **rekišeditšego** in § 3.1.2 above) so that users do not need to begin by reversing and subsequently applying often complicated grammatical rules.

Therefore, to enable a sensible selection of *all the frequently used* 'verbal and nominal derivations' of *each frequent* formal radical, one should turn to frequency counts derived from a well-designed electronic corpus of the language under study. As such a *frequency-based selection / reduction* is advanced as a reliable method to limit the number of formal radicals and their derivations one includes in a dictionary. As an illustration, we can for instance examine (5) which summarises all the derivations one finds in the article of the verb **reka** shown in (4).

(5)

#	***structure***	***verbal derivations***	***nominal derivations***
1	formal radical + standard modifications	reka, rekile, rekwa, rekilwe	direkarekane, lereko, mareko, moreki, bareki, sereki, direki, sereko, direko, theko, ditheko
2	formal radical + reciprocal + standard modifications	rekana, rekane, rekanwa, rekanwe	barekani, thekano, dithekano
3	formal radical + reciprocal + causative + standard modifications	rekantša, rekantšhitše, rekantšhwa, rekantšhitšwe	morekantšhi, barekantšhi, serekantšhwa, direkantšhwa, thekantšho, dithekantšho

4	formal radical + alternative causative + standard modifications	rekanya, rekantše, rekanywa, rekantšwe	morekanyi, barekanyi, serekanywa, direkanywa, thekanyo, dithekanyo
5	formal radical + neutro-passive + standard modifications	rekega, rekegile	
6	formal radical + applicative + standard modifications	rekela, reketše, rekelwa, reketšwe	borekelo, morekedi, barekedi, morekelwa, barekelwa, serekelo, direkelo, thekelo, dithekelo
7	formal radical + applicative + reciprocal + standard modifications	rekelana, rekelane, rekelanwa, rekelanwe	barekelani, thekelano, dithekelano
8	formal radical + causative + standard modifications	rekiša, rekišitše, rekišwa, rekišitšwe	morekiši, barekiši, serekišwa, direkišwa, thekišo, dithekišo
9	formal radical + causative + reciprocal + standard modifications	rekišana, rekišane, rekišanwa, rekišanwe	barekišani, thekišano, dithekišano
10	formal radical + causative + neutro-passive + standard modifications	rekišega, rekišegile	
11	formal radical + causative + applicative + standard modifications	rekišetša, rekišeditše, rekišetšwa, rekišeditšwe	borekišetšo, morekišetši, barekišetši, thekišetšo, dithekišetšo
12	formal radical + causative + applicative + reciprocal + standard modifications	rekišetšana, rekišetšane, rekišetšanwa, rekišetšanwe	barekišetšani, thekišetšano, dithekišetšano
13	formal radical + reverse transitive + standard modifications	rekolla, rekolotše, rekollwa, rekolotšwe	morekolli, barekolli, serekollwa, direkollwa, thekollo, dithekollo
14	formal radical + reverse transitive + reciprocal + standard modifications	rekollana, rekolla-ne, rekollanwa, rekollanwe	barekollani, thekollano, dithekollano
15	formal radical + reverse transitive + applicative + standard modifications	rekollela, rekolletše, rekollelwa, rekolletšwe	morekolledi, barekolledi, thekollelo, dithekollelo
16	formal radical + reverse transitive + applicative + reciprocal + standard modifications	rekollelana, rekollelane, rekollelanwa, rekollelanwe	barekollelani, thekollelano, dithekollelano

17	formal radical + reverse transitive + causative + standard modifications	rekolliša, rekollišitše, rekollišwa, rekollišitšwe	morekolliši, barekolliši, thekollišo, dithekollišo
18	formal radical + reverse transitive + causative + reciprocal + standard modifications	rekollišana, rekollišane, rekollišanwa, rekollišanwe	barekollišani, thekollišano, dithekollišano

Expecting from a user to know that a nominal derivation like for instance **dithekollišano** (in module 18 in (5)) should be looked up under **reka** is definitely unrealistic. Furthermore, of the 146 derivations listed under the lemma sign **reka**, only 28 (namely the underlined ones in (5)), were attested in a Sepedi corpus of one million running words. This simply means that 118 derivations – that's over 80% – did not even occur once in a million words. In fact there is also serious doubt among mother tongue speakers whether many of these derivations are actively used. We must conclude that a frequency-based approach could surely have enabled a much better use of dictionary space, for it could have cut the article **reka** down to 20% without any loss of user-friendliness, and could have used the gained 80% for the inclusion of more frequent words.

3.2 Presentation procedures

The main presentation procedures currently in use for verbs are the 'split approach', the 'lump approach' and the 'lump + index approach'. A critical overview of these three is followed by a suggestion to combine all the strong points of them through the introduction of an additional article slot, the tail slot.

3.2.1 'split approach'

In its most elementary form the 'split approach' lists the formal radicals (mostly together with their finals) and both their verbal and nominal derivations in strict alphabetical order. Such an approach is extremely user

friendly since, for disjunctively written languages users can simply turn to the first letter of the word they encountered, while for conjunctively written languages users (roughly) only need to take away the various prefixes like concords and tense markers. An example of the latter is for instance the Kiswahili (G42) word **alisema** 'he/she said; he/she spoke', where one only needs to take away the subject concord of class 1 **a-** and the past tense marker **-li-** before looking up the verb under the letter S. As such, a deverbative is lemmatised with its noun prefix, and a verb in the (second person singular of the) imperative. As far as the pitfalls and virtues of the various lemmatisation strategies for nouns are concerned, we would like to refer to a previous publication (Prinsloo & De Schryver 1999) where they are discussed at great length. As far as the lemmatisation of verbs in the imperative is concerned, Nkabinde argues that:

> *"This arrangement is flawed on two counts, firstly, […] An impression could be created to the user of the dictionary that verbs always occur in the form in which they have been entered […] Secondly, only activity and achievement verbs are amenable to use in the imperative. The entering of verbs in the imperative does not accommodate stative and process verbs" (Nkabinde 1993a: 299)*

The first problem pointed out by Nkabinde is one encountered among the conjunctively written languages. It is however standard procedure in any such language dictionary that a verbal lemma is a canonical form which represents an entire paradigm of inflected forms, because including every conceivable inflection of every verb would mean inflating the dictionary at least a hundred fold. Hence, if users would incorrectly conclude that 'verbs always occur in the form in which they have been entered', it would not be a result of bad lexicography, but a result of a lack of a dictionary culture among the users. Verily, no source of reference – no matter how user friendly it attempts to be – can be truly successful without users being trained in consulting it as well. In any case, for conjunctively written languages a hyphen can be added to the left of the imperative form to indicate that that particular lemma sign is not necessarily always encountered in that form (so for instance for the Kiswahili example used above: **-sema**).

The second problem pointed out by Nkabinde, the fact that stative and process verbs do not really lend themselves to be presented in their (so-called?) imperative form, can only be avoided through the presentation of verbs in the infinitive. Yet, this would imply entering verbs in the dictionary with their class prefix, and hence bringing all the verbs together under one particular dictionary section. Unfortunately, a strictly alphabetically ordered lemma-sign list would have no leg left to stand on at that point. For Cilubà for instance, this would mean bringing all the verbs together under the sections **ku-** (for formal radicals with an initial consonant) or **kw-** (for formal radicals with an initial vowel).[4]

Actually, the one big problem of the 'split approach' is summarised by Gouws & Prinsloo as follows:

> *"For the African languages, apart from the disruption of semantic relations, alphabetical ordering has serious detrimental consequences for grammatical relations. Many traditional compilers, although following an alphabetical ordering in principle, regard the importance of combined semantic and grammatical cohesion as too important to break" (Gouws & Prinsloo 1998: 22)*

By way of example, in the *Concise Swahili and English Dictionary* (Perrott 1965), a pocket Kiswahili – English dictionary, the verb **kusema** and its verbal and nominal derivations follow the split approach 'in principle'. Under their appropriate alphabetical positions, one finds the articles shown in (6).

(6a) **msemaji(wa),** a fluent speaker
(6b) **sema,** to say; speak; **semwa,** be said
(6c) **semeka,** to be utterable
(6d) **semezana,** to talk together
(6e) **usemaji,** fluency
(6f) **usemi,** speech

As can be seen all derivations except for the passive (here **semwa**) are entered separately without any cross-references. One can only wonder why one exception was made for the fully productive passive extension, which (throughout the dictionary) is entered within the article of the 'formal radical plus final'. Here, the 'split approach' was therefore only followed 'in principle'.

3.2.2 'lump approach'

In the 'lump approach' different derivations, sometimes a hundred or more, of a single verb are treated within the article of a formal radical (+ final) in a complex article with numerous sublemmas and sublemmatic addresses. This results in articles such as the one for **reka** presented in (4) above (and of which (5) is a structural analysis), excerpted from the *Groot Noord-Sotho-woordeboek* (Ziervogel & Mokgokong 1975), a dictionary with a very high degree of lumping.

As far as lumping is concerned, Gouws & Prinsloo state:

> *"word stems and their derivations are clustered together in one huge entry with the noun or verbal root as the lemma often containing up to eighteen levels of sublemmas. […] In this way mediostructure [that is the system of cross-referencing] is exhausted / overused for the sole purpose of maintaining structural links. Little or no realization of mediostructure as a powerful access structure is achieved" (Gouws & Prinsloo 1998: 22, 24)*

From this it is obvious that any modern approach should come up with a mediostructure which maintains the structural links, but does not clutter the dictionary with them as a result.

3.2.3 'lump + index approach'

The 'lump + index approach' is for instance encountered in the *Lexique Tembo, Tembo – Swahili du Zaïre – Japonais – Français* by Kaji (1985). In lemmatising the Tembo (M27) lexicon, Kaji strictly adhered to an arrangement of the items by formal radical. An extract from that dictionary is presented in (7).

(7) **Extract from the *Lexique Tembo*** (Kaji 1985, cited in Busane 1990: 29)

-fùm-	kúfumá		:	kupona; kuponyoka. (傷，病気が)治る；(動物が罠から)逃げる， guérir (intr.), se cicatriser (maladie, blessure); s'échapper (animal attrapé au piège).
	múfumu (1)	báfumu (2)	:	munganga, ba-. 医者， guérisseur.
	búfumu (14)	máfumu (6)	:	dawa, ma-. 薬， remède, médicament.
	kúfumyá (caus.)		:	kuponesha. (傷，病気を)治す， guérir (tr.), cicatriser.
	kúfumúkálá (intr.)		:	kupona muzuri. (人が)病気から回復する， guérir (intr.), retrouver la santé.

Here, the first column shows the formal radical, the second (and third for plurals of nouns) shows (show) both the verbal and nominal derivations, while the last column lists the different translations. All this looks very nice and straightforward. Nevertheless, Busane rightly observes:

> *"the usefulness of this presentation to the unsophisticated user is impaired by the detailed analysis itself. In many cases the phonological representation of the items differs from their morphological analysis, thus requiring the user to be aware of the morphophonological rules applicable to each case" (Busane 1990: 29)*

This has been admitted by Kaji himself, for he writes in the front matter to his dictionary:

> *"Malgré l'avantage de compiler un lexique par rangement des thèmes ou radicaux, l'inconvénient surgit quand on le consulte. Comment peut-on savoir le sens du mot múfumu rencontré dans un texte, par exemple? Il n'y aurait pas de problèmes si l'on sait que dans cette langue il y a un verbe auquel se rapporte ce nom et que son radical est -fùm. Mais cela suppose déjà une connaissance profonde de cette langue" (Kaji 1985: xii)*

Kaji solved this masterfully however, for he included an alphabetically ordered Tembo 'word'-index with the indication of each stem. Again, Busane is entirely in the right when he claims:

> *"This arrangement requires the user to find the stem in the index first before checking its meaning and that of related items in the larger context of the lexicon" (Busane 1990: 30)*

Taken at face value, this approach means that when an elementary learner wants to look up an item, the index will have to be consulted first in order to determine the formal radical, only then will this user be able to search for the relevant formal radical, and ensuing these this user will have to scrutinise the formal-radical article itself. Yet, an advanced learner will rather attempt to look up the formal-radical article right from the start.

From Kaji's own *'Mais cela suppose déjà une connaissance profonde de cette langue'* we see that Kaji's approach is not successful for elementary learners. But are the advanced learners really able to retrieve the 'hard' derivations? We are convinced that: a) deriving a new derivation from a certain formal radical is not that hard, but finding the formal radical when given a certain derivation can be truly hard – if not impossible – at times, and b) even if the meaning is inferable, the orthographic form the derivation takes is not necessarily readily inferable.

We can illustrate both for Cilubà with the phenomenon known as 'imbrication' (also called 'consonant-contraction'). In (8) the applicative of the verb **kusùùlula** is derived, and in (9) the applicative of the verb **kumòna** (cf. Kabuta 1998: 19).

(8) **-sùùlul- -il- -a** (formal radical + final: to untie; to loosen (up))

Ø
wil } *(imbrication)*

-sùùlwila (to untie for, …; to loosen (up) for, …)

(9) **-mòn- -il- -a** (formal radical + final: to see)

-èn- *(nasal assimilation)*

Ø
wèn } *(imbrication)*

-mwèna (to know as a result of, …; to recognise by, …; to see due to, …)

Example (8) illustrates that deriving a new derivation from a certain formal radical might indeed not be that hard, and that the meaning might indeed be easily inferable. However, if one is given the applicative forms from (8) or (9) one sees that our claim that finding the formal radical when given a certain derivation can be truly hard (**kusùùlwila**) – if not impossible (**kumwèna**) – at times, and that our claim that the orthographic form the derivation takes is not necessarily readily inferable (**kumwèna**), are both confirmed. In other words, in Kaji's approach, even advanced learners will need to consult the 'word'-index first for this type of items. One can wonder if going through two search-processes is indeed what users really desire to do. We are convinced that it is not.

3.2.4 'tail-slots approach'

The 'tail-slots approach' was first introduced in the *Lexicon Cilubà – Nederlands* (De Schryver & Kabuta 1997), for short LCN. This approach was an attempt to combine the strong points of the traditional presentation approaches, while avoiding their weak points. Basically, this meant devising a system in which lemmata could be lemmatised in a strictly alphabetical order while providing a transparent mediostructure.

In LCN one encounters several kinds of large networks, each one with its respective 'reference nodes'. One of those networks (when present)

is centred around the 'formal radical + final' of verbs. The idea is to use this form as a node to link all lemma signs that are connected. It goes without saying that one is only referring to lemma signs *within* the lexicon, and that nothing whatsoever is claimed about other potential possibilities in Cilubà. For this purpose the 'tail slot' was created. The tail slot is to be found at the end of an article and starts with an arrow to the right (>) and is followed by one or more lemma signs that are connected with the head of the article.

From the preceding description it should be clear that it was purposely avoided to claim that everything following the arrow to the right (>) is actually 'derived' from the head of the article. Even if the direction of derivation is actually as such in most cases, the subject of this very direction has as yet not been studied thoroughly enough to enable definite claims. Therefore, one should rather read this symbol as a 'direction' in the network, thus as 'away from node.' A good example is the verb **kudyà** presented in (10).

(10) **-dyà** [tww; cf **spw3, 5**] eten; **~ *kuukuta*** [ud] *eten en verzadigd z*
> bidyà; cidìilu; cyàkudyà; -dìika; -dìikiibwa; -dìila; -dìisha; mudì; Mudìila-mpiku

Here, one finds nine lemma signs 'away from node.' In their respective articles, all of these nine evidently contain a cross-reference back to the node, mostly through the use of an arrow pointing to the left (<), to read as 'to node'. Cross-references back can be seen in (11) where a sample of the lemma signs from (10)'s tail slot are shown.

(11a) **cidìilu** [7/8 < app **-dyà**] **1** eetzaal; **2** kribbe

(11b) **cyàkudyà** [cn sub 7/8 < **-dyà**] voedsel; ***mukàndà wà byàkudyà*** *menu(kaart)*

(11c) **-dìika** [iww, sta **-dyà**] eetbaar z; gegeten w; afgeknaagd w; verteren

(11d) **mudì** [1/2 < **-dyà**] eter; verslinder

With mainly *one simple duo*, an arrow pointing to the right (>) and an arrow pointing to the left (<), one is able to interlink the 'formal radical + final' of

verbs and all their 'derivations'. In LCN all types of learners, whether elementary, intermediate or advanced, have the possibility to look up verbs and their derivations directly under their proper alphabetical position; or if they prefer, they can start with the 'formal radical + final' and follow up the cross-references mentioned in the tail. Compared to the 'lump approach' and the 'lump + index approach' LCN certainly wins on the deal as far as quick reference is concerned. As such, the only remaining benefit of the strict grouping of derivations in one huge article in those traditional approaches is that it reveals lexical relations combined with a convergent full treatment, whereas the treatment in LCN is as full, but divergent. Compared to the 'split approach', the 'tail-slot approach' is as user friendly since it also follows an alphabetical ordering. In addition however, the one big problem of the 'split approach', namely the absence of a mediostructure, is accounted for. In a critical review of the 'tail-slot approach' introduced in LCN, Gouws & Prinsloo write:

> *"An excellent example in African language lexicography where mediostructure has been employed as a powerful access structure is the Lexicon Cilubà-Nederlands (LCN) compiled by De Schryver and Kabuta. This dictionary is highly successful in interconnecting the knowledge elements represented in different sectors of the dictionary on several levels of lexicographic description to form a network.*
>
> *In contrast to [the Groot Noord-Sotho-woordeboek], for example, the compilers of LCN are aware of the benefits of "keeping together what semantically and grammatically belong together" but also of the need (a) to avoid extremely long entries and (b) to ensure proper treatment of each derivation in terms of grammatical, tonal and lexical information. […] The compilers of LCN thus succeeded in harmonising lumping and splitting, capturing the advantages of both these approaches. It can, of course, be argued that the listing of the different derivations [in the TAIL] occupies precious space in the dictionary. However, by substantially reducing the font size, this redundancy is diminished.*
>
> *Thus the compilers not only succeeded in linking stems and derivations and treating both stems and derivations satisfactorily, they also employed a complex system of cross-referencing" (Gouws & Prinsloo 1998: 31-2)*

4.0 Frequency-based tail slots, with special reference to Kiswahili

Under §3.1 we showed that a *frequency-based selection / reduction* is a reliable method to limit the number of formal radicals and their derivations one includes in a dictionary. Under §3.2 we showed that an *alphabetical ordering with the inclusion of tail slots* can merge all the strong point of the traditional presentation approaches. The next logical step is to combine these two conclusions and to introduce the concept of 'frequency-based tail slots'. In a nutshell, a **frequency-based tail slot** is a slot at the end of an article headed by a formal radical (+ final). This tail slot lists all the frequent derivations that are linked to this head, whilst the articles of the derivations themselves are to be found under their proper alphabetical positions (with cross-references back).

In order to implement frequency-based tail slots, one will need to build an electronic corpus of the language under study. According to Atkins:

> *"Great strides have been made in dictionary making in recent years, thanks principally to the advent of computer typesetting, computer-assisted dictionary compiling, and the use by lexicographers of electronic corpora as a source of objective information about the language or languages they are describing" (Atkins 1998: 1)*

While Jeffery claims: *"It has become widely accepted that a well-designed corpus is a prerequisite for study of any language. […] and nobody nowadays would undertake a dictionary without one" (Jeffery 2000: 71)*

Compared to the other Bantu languages, Kiswahili is in a very fortunate position as reliable corpora can quickly be assembled. Indeed, Kiswahili being used daily on the World-Wide Web, one can simply download a variety of files and instantly start building an electronic corpus. In order to test the potential of such a corpus, we kick-started the *Kiswahili Internet Corpus (KIC)* in mid October 1999. In just 10 days we were able to reach one million words, and by adding daily news, this corpus stood at 1.3 million running words before the start of the New Millennium. We will now use the 1.3-million-large KIC to illustrate the possible application of frequency-based tail slots for Kiswahili.

Analysing KIC indicated that the most frequently used verb in Kiswahili (besides the verb **kuwa** 'to be') is **kusema** 'to say; to speak'. In order to make realistic comparisons between an approach based on 'frequency-based tail slots' and existing dictionaries, we choose two extremes: the *Concise Swahili and English Dictionary* (Perrott 1965), a pocket edition, and the large *The Internet Living Swahili Dictionary* with approximately 50.000 lemmata (cf. URL Kamusi). The frequency counts for the different inflections and derivations of **-sem-** in KIC are shown in (12).[5]

(12) **Frequency counts for (verbal and nominal derivations of) the formal radical -sem-**

Kiswahili Internet Corpus (KIC) *(1.271.782 tokens and 91.062 types)*							
N	**Item**	**Count**	**%**	**N**	**Item**	**Count**	**%**
25	ali**sem**a	3,722	0.293	3815	nita**sem**a	28	0.002
111	aka**sem**a	1,289	0.101	3847	wa**sem**ao	28	0.002
225	ku**sem**a	689	0.054	3860	bali**sem**a	27	0.002
259	ame**sem**a	606	0.048	4057	**sem**eni	26	0.002
267	**sem**a	586	0.046	4079	yali**sem**wa	26	0.002
341	ana**sem**a	464	0.036	4312	ume**sem**a	24	0.002
359	waka**sem**a	438	0.034	4668	iki**sem**a	21	0.002
440	wali**sem**a	342	0.027	4903	lime**sem**a	20	0.002
601	wana**sem**a	258	0.020	4970	tuna**sem**a	20	0.002
763	a**sem**a	193	0.015	5026	a**sem**e	19	0.001
843	aki**sem**a	173	0.014	5415	tuka**sem**a	18	0.001
896	hu**sem**a	163	0.013	5486	aliyo**sem**a	17	0.001
897	ili**sem**a	163	0.013	5526	haku**sem**a	17	0.001
1032	wata**sem**a	141	0.011	5552	kime**sem**a	17	0.001
1056	ina**sem**ekana	137	0.011	5626	mwa**sem**a	17	0.001
1067	waki**sem**a	136	0.011	5729	ali**sem**aka	16	0.001
1121	wame**sem**a	130	0.010	5891	m**sem**o	16	0.001
1243	ime**sem**a	114	0.009	6360	bana**sem**a	14	0.001
1342	m**sem**aji	105	0.008	6494	lili**sem**a	14	0.001
1514	zili**sem**a	91	0.007	6595	si**sem**i	14	0.001
1532	ya**sem**a	90	0.007	6629	uli**sem**a	14	0.001
1553	zina**sem**a	89	0.007	6718	alivyo**sem**a	13	0.001
1655	na**sem**a	82	0.006	6803	ika**sem**a	13	0.001
2016	ata**sem**a	64	0.005	6928	mka**sem**a	13	0.001
2037	zime**sem**a	64	0.005	6955	nika**sem**a	13	0.001
2398	mna**sem**a	52	0.004	6956	nili**sem**a	13	0.001
2474	tu**sem**e	50	0.004	7540	wa**sem**aji	12	0.001
2576	ina**sem**a	47	0.004	12447	kum**sem**ea	6	
3074	aliya**sem**a	37	0.003	17339	ku**sem**ezana	4	
3195	una**sem**a	36	0.003	17775	mi**sem**o	4	
3203	ya**sem**avyo	36	0.003	17899	m**sem**a	4	
3339	wa**sem**a	34	0.003	31386	ku**sem**ana	2	
3486	wa**sem**e	32	0.003	41753	alini**sem**esha	1	
3572	waya**sem**ayo	31	0.002	73298	**sem**eka	1	
3736	u**sem**i	29	0.002				

Bringing together the different inflections of verbs on the one hand, and singulars and (where applicable) plurals of nouns on the other, results in the first two columns of the table presented in (13).[6]

(13) **Inclusion / omission of (verbal and nominal derivations of) the formal radical -sem-**

Count in KIC	Item(s)	Frequency-Based Tail Slots	Concise Swahili-English	The Internet Living Swahili
10.862+	-sema	Y	Y	Y
137+	-semekana	Y		
117	msemaji/wasemaji	Y	Y	Y
29	usemi/semi	Y	Y	Y
26+	-semwa	Y	Y	Y
20	msemo/misemo	Y		Y
6+	-semea			Y
4+	-semezana		Y	Y
4	msema/wasema			Y
2+	-semana			Y
1+	-semesha			Y
1+	-semeka		Y	Y
0	semezano/masemezano			Y
0	msemi/wasemi			Y
0	usemaji		Y	Y

It is obvious that a user-friendly pocket dictionary should include those items that have a frequency of at least 20 in (13), while it should not allocate precious space to infrequent items at the expense of more frequent ones. Columns 4 and 5 in (13) list those verbal and nominal derivations of the formal radical -sem- that were included / omitted in the *Concise Swahili and English Dictionary* and *The Internet Living Swahili Dictionary*. This reveals the following stunning fact: *the most frequent verbal derivation of the second most frequent verb* in Kiswahili, has *not been entered* in either dictionary! Rather, extremely infrequent items, some of which did not even occur once in KIC were entered. Especially for a large dictionary like *The Internet Living Swahili Dictionary* this is unacceptable. And as far as pocket dictionaries like the *Concise Swahili and English Dictionary* are concerned, a sound use of

available dictionary space implies that the items most likely to be looked for are at least included. Unfortunately, the pocket dictionary even missed out on a frequent noun.

Therefore, if we were to compile a pocket paper dictionary Kiswahili – English using frequency-based tail slots, the article of the second most frequent verbal node and the different lemmata linked to this node would for instance be as shown in (14).

(14a) **msemaji** (*pl.* ***wasemaji;*** < **-sema**) speaker, narrator; political spokesperson

(14b) **msemo** (*pl.* ***misemo;*** < **-sema**) expression, saying, idiom, maxim, slogan

(14c) **-sema** say; speak
> msemaji; msemo; -semekana; -semwa; usemi

(14d) **-semekana** (< **-sema**) they say, what people say

(14e) **-semwa** (< **-sema**) said; spoken

(14f) **usemi** (*pl.* ***semi;*** < **-sema**) pronouncement, way of talking, manner of expression; (*grammar*) word

5.0 Conclusion

In this article frequency-based tail slots were advanced as a user-friendly tool to substantially enhance the quality of Bantu language dictionaries. Since this extra slot is frequency-based, the huge number of possible verbal and nominal derivations one includes in a small-size dictionary can be limited using sound and straightforward criteria. In addition, this method 'is highly successful in interconnecting the knowledge elements represented in different sectors of the dictionary on several levels of lexicographic description to form a network'. Since electronic corpora can quickly be assembled for Kiswahili, it was indicated that frequency-based tail slots can be implemented near-instantly in order to create better dictionaries for Kiswahili.

References

URL (Universal Resource Locator)

Kamusi http://www.yale.edu/swahili/

Atkins, B.T. Sue. (1998) Introduction. In B.T. Sue Atkins (ed.). (1998) *Using Dictionaries, Studies of dictionary use by language learners and translators*: 1-5. Tübingen: Max Niemeyer Verlag.

Bennett, Patrick R. (1986) Grammar in the Lexicon, Two Bantu cases. *Journal of African Languages and Linguistics 8/1*: 1-30.

Benson, T.G. (1964) A Century of Bantu Lexicography. *African Language Studies 5*: 64-91.

Busane, Masidake. (1990) Lexicography in Central Africa: the User Perspective, with Special Reference to Zaïre. In Reinhard R.K. Hartmann (ed.). (1990) *Lexicography in Africa, Progress reports from the Dictionary Research Centre Workshop at Exeter, 24-25 March 1989*: 19-35. Exeter: University of Exeter Press.

De Schryver, Gilles-Maurice and **Ngo S. Kabuta.** (1997) *Lexicon Cilubà-Nederlands, Een circa 2500-lemma's-tellend strikt alfabetisch geordend vertalend aanleerderslexicon met decodeer-functie ten behoeve van studenten Afrikaanse Talen & Culturen aan de Universiteit Gent*. Ghent: Recall.

De Schryver, Gilles-Maurice. (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*. (Unpublished MA dissertation, University of Ghent.)

Gabriël [Vermeersch]. (*s.d*. [1922]) *Dictionnaire Tshiluba-Français*. Brussels: Librairie Albert Dewit.

Gouws, Rufus H. and **D.J. Prinsloo.** (1998) Cross-referencing as a Lexicographic Device. *Lexikos 8*: 17-36.

Gove, Philip B. (ed.). (1961^3) *Webster's Third New International Dictionary of the English Language.* Springfield: Merriam-Webster.

Guthrie, Malcolm. (1948) *The Classification of the Bantu Languages*. London: Oxford University Press.

Hartmann, Reinhard R.K. and **Gregory James.** (1998) *Dictionary of Lexicography*. London: Routledge.

Jeffery, Chris. (2000) Projected Corpora of South Africa's Official Languages. In *Programme & Abstracts of the First International Conference on Linguistics in Southern Africa, 12-14 January 2000, University of Cape Town*: 71.

Kabuta, Ngo S. (1998) *Inleiding tot de structuur van het Cilubà*. Ghent: Recall.

Kaji, Shigeki. (1985) *Lexique Tembo, Tembo – Swahili du Zaïre – Japonais – Français.* Tokyo: Institute for the Study of Languages and Cultures of Asia and Africa.

Kriel, T.J., **Egidius B. van Wyk** and **S.A. Makopo.** (1989[4]) *Pukuntšu woordeboek, Noord-Sotho – Afrikaans, Afrikaans – Noord-Sotho*. Pretoria: J.L. van Schaik.

Nkabinde, A.C. (1993a) Review: G.R. Dent (Compiler) and C.L.S. Nyembezi (Ed.). Compact Zulu Dictionary, Eng.-Zulu – Zulu-Eng. *Lexikos 3*: 298-300.

Nkabinde, A.C. (1993b) Review: G.R. Dent and C.L.S. Nyembezi (Compilers). Scholar's Zulu Dictionary, Eng.-Zulu – Zulu-Eng. *Lexikos 3*: 301-2.

Prinsloo, D.J. (1994) Lemmatization of Verbs in Northern Sotho. *South African Journal of African Languages 14/2*: 93-102.

Prinsloo, D.J. and **Gilles-Maurice 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.

Schadeberg, Thilo C. (1992[3]) *A Sketch of Swahili Morphology*. Köln: Rüdiger Köppe Verlag.

Snoxall, R.A. (1965) Some Problems and Principles of Lexicography in Luganda. *African Language Studies 6*: 27-31.

Snyman, Jannie W. (ed.). (1990) *Dikišinare ya Setswana – English – Afrikaans Dictionary / Woordeboek*. Pretoria: Via Afrika Limited.

Zgusta, Ladislav. (1971) *Manual of Lexicography*. The Hague: Mouton.

Zgusta, Ladislav. (1989) The Influence of Scripts and Morphological Language Types on the Structure of Dictionaries. In Franz J. Hausmann, Oskar Reichmann, Herbert E. Wiegand and Ladislav Zgusta (eds.). (1989-1991) *Wörterbücher / Dictionaries / Dictionnaires, Ein internationales Handbuch zur Lexikografie / An International Encyclopedia of Lexicography / Encyclopédie internationale de lexicographie*. Berlin: Walter de Gruyter.

Ziervogel, Dirk and **Pothinus C.M. Mokgokong.** (1975) *Pukuntšu ye kgolo ya Sesotho sa Leboa, Sesotho sa Leboa – Seburu/Seisimane / Groot Noord-Sotho-woordeboek, Noord-Sotho – Afrikaans/Engels / Comprehensive Northern Sotho Dictionary, Northern Sotho – Afrikaans/English.* Pretoria: J.L. van Schaik.

[*] This article brings together elements from De Schryver's MA dissertation (De Schryver 1999), Prinsloo's study of the lemmatisation of verbs in Sepedi (Prinsloo 1994), and additional research.

[1] Codes between brackets, such as this one, refer to the (somewhat outdated) classification of the Bantu languages introduced by Guthrie (1948).

[2] A verb's 'formal radical' (Schadeberg 1992[3]: 8) is the term used in the present article to refer to what most scholars call a verb's 'root'. It should be understood as the verbal stem minus the final and minus the verbal extension(s).

[3] The counterpart of semasiological dictionaries are onomasiological ones, or thus dictionaries with a thematic order in which the direction is from concept to word, rather than from word to concept (Hartmann & James 1998: 102).

[4] This is true for all the verbs, except for the defective forms of the verb 'to be'.

[5] We decided that an item had to occur at least 12 times for inclusion in this list. This means that an item had to have a frequency of at least 0.001% in the corpus.

[6] Item number 5729 (**alisemaka**) does not show an even spreading across the different sources, as it occurs in just one interview. It appears to be a non-standard form only used in Congo Kiswahili, so it was not included in table (13).