The lemmatization of nouns in African languages with special reference to Sepedi and Cilubà

D.J. Prinsloo* & Gilles-Maurice de Schryver
Department of African Languages, University of Pretoria, Pretoria 0002, South Africa
prinsloo@postino.up.ac.za

October 1998

The aim of this article is to analyze traditional approaches to the lemmatization of nouns on the macrostructural level in African languages against the background of the user-perspective, the physical limitations on volume, the consideration of currently available dictionaries and the utilization of a corpus. Five basic lemmatization strategies are given due attention, namely lemmatizing nouns under stems, according to morpho-lexical fields, under both singular and plural forms, solely under singular forms, or finally on first or third letters. It is shown that any modern strategy that aims at avoiding the shortcomings and pitfalls of these five basic types whilst at the same time exploiting their virtues, will have to give full weight to target users’ desires. A means to incorporate those desires is illustrated through a small dictionary project to which the concept of ‘simultaneous feedback’ is applied. In conclusion four rules of thumb and ten general guidelines are presented.

* Author to whom correspondence should be addressed.

Introduction

There is an urgent need for lexicographical planning for the African languages in general, and for Sepedi and Cilubà in particular. The term ‘lexicographical planning’ will be used in its general sense as in Gouws (1990:53) and Wiegand (1998: 136–138) to refer to all aspects of dictionary design on the macro- and microstructural levels as well as to the planning of different types of dictionaries.

As far as the lemmatization of nouns on the macrostructural level is concerned, compilers generally fail in their efforts to lemmatize them satisfactorily within the physical limitations of a printed dictionary and to render a user-friendly product, or they err in including nouns unlikely to be looked for by the target user at the expense of essential ones. Thus although dictionaries may be available for languages like Sepedi and Cilubà, lexicographers agree that these dictionaries generally lack proper lexicographical planning. Any prospective compiler of a dictionary for these languages at this point in time, will be well advised to compile the dictionary within the framework of the general restrictions or guidelines that will be briefly outlined below, namely: the user-perspective, the physical limitations on volume, the consideration of currently available dictionaries, as well as the utilization of a corpus. These criteria are of the utmost importance and examples of failure due to the neglect of such criteria lie scattered on the dictionary market.

Previously the user-perspective was not seriously considered in the compilation of dictionaries for African languages. A typical example in this regard is the Groot Noord-Sotho woordeboek (GN). In this dictionary, although being user-unfriendly, word stems are lemmatized and a phonemic sorting principle is used instead of normal alphabetical ordering simply because the compilers regard it as ‘scientifically correct’. Modern dictionaries are, however, judged by the
success with which the user is able to retrieve the sought information and not in the first instance by 'linguistic achievement'. The lexicographer is in terms of Prinsloo & Gouws (1996:103) the mediator between linguistics and the everyday dictionary user. The modern trend in lexicography is thus to compile very practical and extremely user-friendly dictionaries. In simple terms this means that the lexicographer has to include those words which are most likely to be consulted by the target user and to lemmatize them in a user-friendly way. User-friendliness does not only relate to the central text but also to the front and back matter of the dictionary, especially the user's guide. One could say that practicality and user-friendliness virtually determine the methodology and the way in which the access structures to the dictionary are to be designed. As will be argued in this article, a means to incorporate the target users' desires, both from the very start and during the entire compilation process of a modern dictionary, is to make use of the concept of 'simultaneous feedback'.

Secondly, physical limitations on volume, mostly on the number of pages and therefore on the amount of entries that can be accommodated in a specific dictionary or sub-dictionary has a far greater impact on lemmatization in African languages than one would expect. In simple terms it boils down to the strategy of selection of words - or the lack thereof - that is which words are to be chosen and which words can be left out. This rather straightforward fact, namely the need for selection in the sense of which nouns to include or exclude from the dictionary, whether to lemmatize singular and plural forms, or singular forms only, whether to lemmatize on the noun stems, and the procedure for handling irregular forms, is the major pitfall into which compilers for African language dictionaries have so painfully stumbled.

Thirdly, compilers of dictionaries for African languages seriously err in not taking into account the nature and comprehensiveness of currently available dictionaries. Compilers of new dictionaries and lexicons act as if they are compiling the first dictionary or lexicon for the language instead of either improving on existing ones in a specific category or compiling variants for different categories for which no dictionaries or lexicons have as yet been published. In other words, they continue to compile 'general dictionaries' to be used by 'unspecified target users', often of a lower standard than existing ones, or to make things even worse, they attempt to cover three languages in a one-, two- or three-way configuration.

And fourthly, as will be shown throughout this article, today's compilers fail to seriously take into consideration the importance of a corpus as a useful tool for the description of actual language usage. As corpora have a lot to contribute towards the compilation of modern dictionaries, lexicographers cannot continue to ignore them.

Formulated differently, the major challenge regarding the lemmatization of nouns is to design a strategy according to which the user can look up nouns and find the information he/she is looking for as easily and quickly as possible.

**Noun class systems and traditional approaches to lemmatization**

The first serious problem and major decision with which the lexicographer is confronted concerns the approach towards the lemmatization of singular and plural forms of nouns, since singularity and plurality are indicated by prefixes. An oversimplified layout of the noun class systems in Sepedi and Ciluba is shown in Table 1.

Up to now lexicographers have been utilizing five basic approaches for lemmatizing nouns in African languages, namely:

1. Lemmatizing noun stems,
2. Lemmatizing nouns according to morpho-lexical fields,
3. Lemmatizing both singular and plural noun forms,
4. Lemmatizing only singular noun forms, and
5. Lemmatizing nouns on the first or the third letter.

Each type will now be given due attention.

1. **Lemmatizing noun stems**

The lemmatization of nouns under their stems is a very popular approach. Guthrie even claims that
Table 1 Layout of the noun class systems in Sepedi and Cilubà

<table>
<thead>
<tr>
<th>Class</th>
<th>Sepedi</th>
<th>Cilubà³</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>mo-</td>
<td>ma-</td>
</tr>
<tr>
<td>2</td>
<td>ba-</td>
<td>ba-</td>
</tr>
<tr>
<td>1a</td>
<td>Ø</td>
<td>Ø</td>
</tr>
<tr>
<td>2a</td>
<td>bo+</td>
<td>baa+</td>
</tr>
<tr>
<td>3</td>
<td>mo-</td>
<td>mu-</td>
</tr>
<tr>
<td>4</td>
<td>me-</td>
<td>mi-</td>
</tr>
<tr>
<td>5</td>
<td>le-</td>
<td>di-</td>
</tr>
<tr>
<td>6</td>
<td>ma-</td>
<td>ma-</td>
</tr>
<tr>
<td>7</td>
<td>se-</td>
<td>ci-</td>
</tr>
<tr>
<td>8</td>
<td>di-</td>
<td>bi-</td>
</tr>
<tr>
<td>9</td>
<td>N-/Ø</td>
<td>N-</td>
</tr>
<tr>
<td>10</td>
<td>di+</td>
<td>la-</td>
</tr>
<tr>
<td>11</td>
<td>N-</td>
<td>la-</td>
</tr>
<tr>
<td>12</td>
<td>ka-</td>
<td>bu-</td>
</tr>
<tr>
<td>13</td>
<td>tu-</td>
<td>bu-</td>
</tr>
<tr>
<td>14</td>
<td>bo-</td>
<td>ma-</td>
</tr>
<tr>
<td>15</td>
<td>go</td>
<td>go-</td>
</tr>
<tr>
<td>16</td>
<td>fa-</td>
<td>ku-</td>
</tr>
<tr>
<td>17</td>
<td>go-</td>
<td>ku-</td>
</tr>
<tr>
<td>18</td>
<td>mo-</td>
<td>mu-</td>
</tr>
</tbody>
</table>

Example  Translation Example  Translation
monna  man  muluma  man
banna  men  baamu  men
rarrangwane  paternal uncle  maamit  mother
barrangwane  paternal uncles  baamuamit  mothers
monwana  finger  muci  tree
menwana  fingers  mici  trees
lesogONA  young man  dibǒko  arm
masogona  young men  mabǒko  arms
selepe  axe  ciNtu  thing
dilepe  axes  biNtu  things
nk  sheep  nyyuyi  bird
Dinku  sheep  nyyuyi  birds
bogobe  porridge  njiti  fly
magobe  porridges  njiti  flies
mogobe  porridges  mabundu  friendship
go  to buy  mabundu  friendships
Dorka  to know  buddies  friendship
Fase  below  Buddies  friendships
Godimo  above  ku  kwisu  in the face
Morago  behind  ku  kwisu  in the face
Munda  inside

this is the only sound way for entering nouns in a dictionary:

‘... the essential word structure of Bantu languages ... requires that for indexing purposes the stem of a noun and not its concord prefix should decide its position in the alphabetic order’ (1971:358).

The Groot Noord-Soto woordeboek (GN) by Ziervogel & Mokgokong used this approach. In their dictionary these compilers state categorically that this approach is ‘the only scientific method’ (1975:87). Ten years earlier, Ziervogel suggested:

‘Entries must be arranged under their stems with cross-references where
necessary. This method is scientifically sound. A systematized survey of word-formation in the languages is given; it shows word and lexical relationship and prevents repetition' (1965:45).

The compilers of G/N deserve some credit for a brave effort towards comprehensiveness. See for instance their sincere attempt to lemmatize all the possible derivations and deverbatives for the verb reka in Figure 1.

The entry for reka is structurally analysed in terms of 18 modules in Table 2.

Ziervogel & Mokgokong managed to include most of the highly used forms, namely the 28 underlined forms in Table 2, which reflect actual usage in a corpus of 33 selected Sepedi and magazines. This corpus totals one million words and will be utilized as an instrument of evaluation in this article. However, from a user-perspective, 118 of the 146 forms given here — that’s over 80% — did not even occur once in a million words. In fact there is serious doubt among mother-tongue speakers whether many of these derivations and deverbatives are actively used.

Furthermore, the user often has to struggle through numerous columns of fine print in the dictionary to find the meaning of a word. Take for example the deverbative dithekollishano. The compilers expect him/her to know that di- is the class prefix and that it should be removed and that the word should be looked up under thekollishano. The user is then referred from thekollishano to the root, namely reka. Under reka he/she has to work through three columns of fine print to eventually find thekollishano in Module 18 — but no translation is given in Afrikaans or English. The user then moves backwards to the head of Module 18 to find at least a verbal meaning to work out a nominal meaning — to no avail, as the entire module lacks any semantic information. He/she now moves backward to the head of the previous Module 17 without the reciprocal connotation — but still no meaning is given for the entire module. The same procedure is followed for the head of Module 16 and for the head of Module 15 and for the head of Module 14. Under the head of Module 13, the user finds the meaning ‘buy back, exchange back, ask for a refund; redeem’. Now he/she has the meaning of the root + reversion transitive and has to add the causative, reciprocal and nominal connotations on his/her own anyway to eventually arrive — yet with great uncertainty — at something like ‘the result of something that individuals have mutually caused to be bought back’. And what does this really mean?

From a practical and user-friendly point of view however, this method is a disaster for Sepedi. Firstly, from examples such as given in Figure 1 and Table 2 it is clear that the compilers were obsessed to include all conceivable derived stems under verbal lemmas. It is not surprising that this ‘enter-them-all-syndrome’ resulted in an entry such as Figure 2 where the compilers concentrated so hard on completing the modular paradigm that no translation equivalents in English or Afrikaans were given (see Figure 2).

To make things worse, Van Wyk (1995:88, 91–92) has shown in a critical review of this dictionary that in following this approach the compilers did not manage to avoid repetition due to — among others — irregular forms, but rather introduced redundancy by having to resort to unnecessary cross-referencing.

‘This brings no gain in economy compared with word dictionaries. The number of entries is the same for both types, the only difference being the structure and the alphabetic classification of the entries’ (Van Wyk, 1995:88).

Secondly, the user is unnecessarily burdened with numerous problems relating to isolating the stem in many problematic instances such as ngwana (*mo-ana) ‘child’, nmswo (*mo-bašo) ‘government’, and msiti (*mo-tši) ‘smoke’ where the noun stem is no longer synchronically identifiable.

Ziervogel & Mokgokong moreover deviate from an ordinary alphabetical sorting of the entries and utilize a phonemic one, namely: A, B, BJ, D, E, F, FS, Fś, G, H, HL, I, J, K, KG, KH, L, M, N, NG, NX, NY, O, P, PH, et cetera, because this is in their opinion ‘more scientific’. To the user it is nothing more than sheer frustration to eventually find for example a word commencing on bj alphabetically after bu in the dictionary.
RÉKA (réka, rëká, rékwa, rëkúte) koop, aankoop, ruil // buy, purchase, barter; ~ polana in weede lewe // live in comfort/luxury; ~ o leledléte gofino kat in die sak koop // buy a pig in a poke; nka e rékwa masela 'n mooi geboude dame is 'n aantrekkingskrag vir jongmaas // a lady with a good figure easily attracts young men; drëkërrëkkëne (drëkërrëkkëne) verskynsellikheid gekoopte goeder // variety of things bought; lërëko, ma- (lerëko) gewoon/meiging om te koop // habit of buying, inclination to buy; morëkkë, ba- (morëkkë) pers. dev.; koper // buy, purchaser; serëkë, dr- (serëkkë) pers. dev.; lustige koper // keen buyer; serëkë, dr- (serëkkë) imper. dev.; wat gekoop word, aankoop // purchase); thëko, (n-)/dl- (thëkkë) man. dev.; koopwise, prys // manner of buying, price; RÉKANA (rékana, rëkane, rëkanwa, rëkanwe) rec.; ruil met mekaar // exchange with one another; a re rëkane, wëna o sathe hampëtë dëla, ni loo dëpp le ko go ft dëlla fe last one met mekaar ruil, jy goe my daardie hemp en t'j goe jou hierdie skoon // let us exchange, you give me that shirt, I will give you these shoes; barëkkë (borëkkë) pers. dev.; thëkkë (thëkkë) man. dev.; RÉKANTSÁH (rékantshá, rëkantshité, rëkantshwa, rëkantshítwe) caus. < RÉKANÁ; (om)ruil, wissel (goed), inruil // exchange, trade in, swap; morëkkëntë, ba- (morëkkëntë) pers. dev.; serëkëntë, dd- (serëkkëntëwa) imper. pass. dev.; thëkkëntsho, (n-)/dl- (thëkkëntshó) man. dev.; omsuiling, inruiling, wisselend // exchanging, swopping; RÉKANTA (rëkana, rëkanwa, rëkanwe, rëkanwe) imper. rec. < RÉKANA; (om)ruil, wissel (goed) // exchange, barter, swap; morëkkënyi, ba- (morëkkënyi) pers. dev.; serëkënyi, dw- (serëkkënyiwa) imper. pass. dev.; thëkanyi, (n-)/dl- (thëkkënyi) man. dev.; v. thëkkënsthò; RÉKEGA (rëkëga, rëkëglétë) nestr.; koopswa w. // b. purchaseable; REKELÉ (rékëla, rëkléše, rëkluwa, rëkluwe) appl.; koop vyr // buy for; ~ kolóe kgetting (< Afr.) kat die sak koop // buy a pig in a poke; borëkkë (borëkkë) lo. dev.; koopliep // place where things are bought; morëkkë, ba- (morëkkë) pers. dev.; morëkkëwa, ba- (morëkkëwa) pers. pass. dev.; serëkë, dr- (serëkkë) imper. dev.; iets waarin jy koop // that into which one buys; thëkkë, (n-)/dl- (thëkkë) man. dev.; maat, skaal (waarin bv. bier gekoop word) // measurement, bowl (one used for buying beer); REKELANA (rëkélana, rëkélane, rëkélane, rëkélane) appl. rec.; barëkkë (barëkkë) pers. dev.; thëkkëlo, dl- (thëkkëlo) man. dev.; RÉKILA (rékila, rëkilišë, rëkilišwa, rëkilišwe) caus. last/help koop, verkoop, van die hand sit // cause/help buy, sell; ~ ko lelémé kul, mislêd, verkeer // deceive, mislead, pervert; ~ lelémé praatlik w. skinder // gossip, b. loquacious, b. garrulous; ~ mooth a sa phela iemand kul // deceive someone; ~ mooth lebake iemand kul, 'n tervergeese belotte maak, iemand verag weens sy stekte gedrag // deceive someone, give a vain promise, despise someone because of his bad conduct; ~ segue iets aan iemand so verkoop dat hy 'n goeie slag slaan omdat hy sy vriend of familieid is, aathing ge // sell to someone at bargain price because he is your friend/relative, give discount; morëkkë, ba- (morëkkë) pers. dev.; verkoper, verkoopman, winkelier // seller, salesman, storekeeper; serëkë, dl- (serëkkë) imper. pass. dev.; thëkkëlo, (n-)/dl- (thëkkëlo) man. dev.; verkooping, uitverkooping, afset, bemarking // sale, selling, market, marketing; REKISÁNA (rëkisána, rëkisána, rëkisána, rëkisána) caus. rec.; ruil met mekaar // exchange with one another; barëkkë (barëkkë) pers. dev.; thëkkëlo, (n-)/dl- (thëkkëlo) man. dev.; REKISÉGA (rëkiséga, rëkiséglé) neur. < RÉKISÁ verkoopbaar w. // b. salable; REKÍSETSA (rëkísetsha, rëkísetshité, rëkísetshwa, rëkísetshítwe) caus. appl.; verkoop vir // sell for; borëkkëlo (borëkkëlo) io. dev.; verkoopliep // selling place, morëkkëlëshë, ba- (morëkkëlëshë) pers. dev.; tagme // f. business, agent: thëkkëlo, (n-)/dl- (thëkkëlo) man. dev.; REKÍSETSANÁ (rëkísetshána, rëkísetshána, rëkísetshána, rëkísetshána) caus. appl. rec.; sake verlig // transaction business; barëkkëlëshë (barëkkëlëshë) pers. dev.; thëkkëlo (thëkkëlo) man. dev.; thëkkëlo, (n-)/dl- (thëkkëlo) man. dev.; besigheidstransaksie // business transaction; REKOLLÁ (rëkollá, rëkollote, rëkollá, rëkolláwe) rev. tr.; terugkoop, tergoed, geld terugvry, los // buy back, exchange back, ask for a refund, redeem; morëkkëli, ba- (morëkkëli) pers. dev.; serëkëli, dd- (serëkkëliwa) imper. pass. dev.; ~ ding wat terug gekoop word // that which is bought back; thëkkëlo, (n-)/dl- (thëkkëlo) man. dev.; (Bl.) lossing // (Bl.) redemption: REKOLLÁNÁ (rëkolláná, rëkollána, rëkollána, rëkollána) rev. rec.; barëkkëi (barëkkëi) pers. dev.; thëkkëlo, (n-)/dl- (thëkkëlo) man. dev.; morëkkëli, ba- (morëkkëli) pers. dev.; thëkkëlo, (n-)/dl- (thëkkëlo) man. dev.; (Bl.) lossing // (Bl.) redemption: REKOLLÁ (rëkollá, rëkollote, rëkollá, rëkolláwe) rev. tr.; terugkoop, tergoed, geld terugvry, los // buy back, exchange back, ask for a refund, redeem; morëkkëli, ba- (morëkkëli) pers. dev.; serëkëli, dd- (serëkkëliwa) imper. pass. dev.; ~ ding wat terug gekoop word // that which is bought back; thëkkëlo, (n-)/dl- (thëkkëlo) man. dev.; thëkkëlo, (n-)/dl- (thëkkëlo) man. dev.; REKOLLÁ (rëkollá, rëkollote, rëkollá, rëkolláwe) rev. appl.; morëkkëli, ba- (morëkkëli) pers. dev.; thëkkëlo, (n-)/dl- (thëkkëlo) man. dev.; REKOLLÁNE (rëkolláñ, rëkolláñ, rëkolláñ, rëkolláñwe) rev. appl. rec.; barëkkëlëi (barëkkëlëi) pers. dev.; thëkkëlo, (n-)/dl- (thëkkëlo) man. dev.; REKOLLÁSÁ (rëkollášá, rëkollášá, rëkollášáwe) caus. rec.; barëkkëlësá (barëkkëlësá) pers. dev.; thëkkëlo, (n-)/dl- (thëkkëlo) man. dev.; thëkkëlo, (n-)/dl- (thëkkëlo) man. dev.; RÉKA v. RÉKA; rékëna, ba- v. RÉKA; rëkëntshá v. RÉKA; rëkëntshá, mo-ja- v. RÉKA; rëkëntshá, se-ja- v. RÉKA; rëkënyá v. RÉKA; rëkënyá, mo-ja- v. RÉKA; rëkënya v. RÉKA; rëkënya, mo-ja- v. RÉKA.

Figure 1 Dictionary entry for reka
<table>
<thead>
<tr>
<th>#</th>
<th>Structure</th>
<th>Derivations</th>
<th>Deverbatives</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>root + standard modifications</td>
<td><em>reka, rekile, rekwa, rekilwe</em></td>
<td>direkorekana, lercko, moreka, morekei, barekki, sorekki, direkki, sorekki, direkko, thekko, dithekko</td>
</tr>
<tr>
<td>2</td>
<td>root + reciprocal + standard modifications</td>
<td><em>rekanana, rekanane, rekananwa, rekanawe</em></td>
<td>barekani, thekano, dithekano</td>
</tr>
<tr>
<td>3</td>
<td>root + reciprocal + causative + standard modifications</td>
<td><em>rekanthinša, rekanthinšiše, rekanthinšwa, rekanthinšitśwe</em></td>
<td>morekanthinši, barekanthinši, serekatinšwa, direkanthinšwa, thekanthinšo, ditikanthinšo</td>
</tr>
<tr>
<td>4</td>
<td>root + alt. causative + standard modifications</td>
<td><em>rekaninya, rekaninše, rekaninwa, rekaninšwe</em></td>
<td>morekaninya, barekaninya, serekanywa, direkaninya, thekaninya, dithekanyo</td>
</tr>
<tr>
<td>5</td>
<td>root + neutro passive + standard modifications</td>
<td><em>rekega, rekegile</em></td>
<td>borkelo, morekedi, barekedi, moreketwa, bareketwa, sorekelo, direketo, thekela, dithekelo</td>
</tr>
<tr>
<td>6</td>
<td>root + applicative + standard modifications</td>
<td><em>rekelana, rekelane, rekelanwa, barekeloni, thekelono, dithekelonwe</em></td>
<td>rekelana, reketane, reketwa, reketšwe</td>
</tr>
<tr>
<td>7</td>
<td>root + applicative + reciprocal + standard modifications</td>
<td><em>reksiša, reksišiše, reksišwa, reksišitśwe</em></td>
<td>moreksiši, bareksiši, serekxišwa, direksišwa, theksišo, dithekxišo</td>
</tr>
<tr>
<td>8</td>
<td>root + causative + standard modifications</td>
<td><em>reksiša, reksišane, reksišanwa, reksišanwe</em></td>
<td>bareksišani, theksišano, dithekxišano</td>
</tr>
<tr>
<td>9</td>
<td>root + causative + reciprocal + standard modifications</td>
<td><em>reksišana, reksišane, reksišanwa, reksišanwe</em></td>
<td>reksišega, reksišege</td>
</tr>
<tr>
<td>10</td>
<td>root + causative + neutro passive + standard modifications</td>
<td><em>reksiša, reksišane, reksišanwa, reksišanwe</em></td>
<td>borkelo, morekedi, barekedi, moreketwa, bareketwa, sorekelo, direketo, thekela, dithekelo</td>
</tr>
<tr>
<td>11</td>
<td>root + causative + applicative + standard modifications</td>
<td><em>reksišetša, reksišeditšiše, reksišeditšwa, reksišeditšitśwe</em></td>
<td>moreksišetša, bareksišetši, serekxišetšwa, direksišetšwa, theksišeto, dithekxišeto</td>
</tr>
<tr>
<td>12</td>
<td>root + causative + applicative + reciprocal + standard modifications</td>
<td><em>reksišetšana, reksišetšane, reksišetšanwa, reksišetšanwe</em></td>
<td>bareksišetšani, theksišetšano, dithekxišetšano</td>
</tr>
<tr>
<td>13</td>
<td>root + reverse transitive + standard modifications</td>
<td><em>rekollo, rekolloše, rekollo, rekollošwe</em></td>
<td>morekolli, barekolli, serekollwa, direkollo, thekolli, dithekollo</td>
</tr>
<tr>
<td>14</td>
<td>root + reverse transitive + reciprocal + standard modifications</td>
<td><em>rekolloana, rekolloane, rekollo, rekolloane, dikollo</em></td>
<td>barekollo, thekollo, dithekollo</td>
</tr>
<tr>
<td>15</td>
<td>root + reverse transitive + applicative + standard modifications</td>
<td><em>rekollela, rekolletša, rekolletšwa, rekolletšwe</em></td>
<td>morekollela, barekollela, thekollela, dithekolela</td>
</tr>
<tr>
<td>16</td>
<td>root + reverse transitive + applicative + reciprocal + standard modifications</td>
<td><em>rekollelan, rekollelanwa, rekollelanwe</em></td>
<td>barekollelami, thekollelano, dithekolelano</td>
</tr>
<tr>
<td>17</td>
<td>root + reverse transitive + causative + standard modifications</td>
<td><em>rekollishi, rekollishiše, rekollishiwa, rekollishišwe</em></td>
<td>morekolliši, barekolliši, thekollishi, dithekollishi</td>
</tr>
<tr>
<td>18</td>
<td>root + reverse transitive + causative + reciprocal + standard modifications</td>
<td><em>rekollishana, rekollishane, rekollishanwa, rekollishane, dikollisho</em></td>
<td>barekollišani, thekollishano, dithekollisho</td>
</tr>
</tbody>
</table>
PHÉFA (-phèfa, -phèfilé, -phèfsa, -phèfiwè) (Pb.) v. PHEMA; mophéfi, ba- (mophéfi) pers. dev.; phéfo, (n-)/di- (phéfò) man. dev.; v. phemo; PHÉFANA (-phèfana, -phèfane, -phèfanwa, -phèfanwe) rec.; v. phemana; baphéfani (baphèfani) pers. dev.; phéfano, (n-)/di- (phèfanò) man. dev.; PHÉFÉGA (-phèfga, -phèfgilè) neutr.; v. phemega; PHÉFÉLA (-phèfela, -phèfèse, -phèfèwa, phèfèfse) appl.; mophéfedi, ba- (mophéfedi) pers. dev.; phèfèlo, (n-)/di- (phèfèlo) man. dev.; PHÉFÉLANA (-phèfélana, -phèfélane, -phèfélawan, -phèfélawan) appl. rec.; baphèfélani (baphèfelani) pers. dev.; phèfélano, (n-)/di- (phèfélano) man. dev.; PHÉFÉSA (-phèfèsa, -phèfèsitè, -phèfèswa, phèfèsitèswè) caus.; mophèfisì, ba- (mophèfisì) pers. dev.; phèfìso, (n-)/di- (phèfìso) man. dev.; PHÉFÉSANA (-phèfèsana, -phèfèsane, phèfèsanwa, -phèfèsanwe) caus. rec.; baphèfèsani (baphèfèsani) pers. dev.; phèfèsano, (n-)/di- (phèfèsanò) man. dev.

Figure 2 Dictionary entry for phéfa

2. Lemmatizing nouns according to morpho-lexical fields
An example of the lemmatization of nouns within the framework of morpho-lexical fields is found in the Dictionnaire Tshiluba-Français by Gabriel (TFG). Before expounding on this approach we can note that this lemmatization procedure is so obfuscating that Gabriel’s dictionary is very unpopular with its users. Kalonji summarizes the situation as follows:

‘Avec Frère Gabriel les items prétendument classés par ordre alphabétique répondent à un code qui, pour n’avoir pas été préalablement explicite, offre un bel exemple de confusion et d’amalgame’ (1993:134).

Unfortunately, the reason for this confusion and amalgamation is not the fact that the lemmatization code was not explicated – for it was, albeit in an implicit and cryptic way – but rather the fact that the very word structure of a language like Ciubè cannot successfully be approached with the concept of morpho-lexical fields. As to Gabriel’s code, it can be retrieved from the front matter of his dictionary. Three steps can be recog- nized in his reasoning. In the first step he claims that any user should be able to determine what he calls le radical of the word to be dealt with. The user should in other words know how to get rid of all kinds of affixes, after which the ‘by now abridged word’ can be looked up under the radical. In the second step he announces that he will classify the words according to their origins, that derivations and compounds follow the radical and that etymological families are brought together. In the third step Gabriel suggests, bringing steps one and two together, to present this information as a series of distinct groups. When one consults the central dictionary text, one sees that these groups are separated from each other by two blank lines. In addition, every group starts with a form that supposedly generates all the ensuing forms in the group. The result of all this is that the entire Ciubè lexicon seems to be grouped around morpho-lexical fields. Commenting upon the groups, Gabriel writes in the front matter of his dictionary: ‘Les séparations des groupes ... servent simplement à faciliter la recherche d’un mot’ (s.d. [1922]: i).

Ironically, as will be shown in the examples below, the code used in TFG anything but facilitates the quick and sound acquisition of information. This is particularly obvious in the case of nouns. Gabriel points out (s.d. [1922]: iv) that one has to make a distinction between regular nouns and derived nouns. The latter are to be found within the group headed by the radical from which they are thought to be derived, while the former are to be found at the alphabetical place of their nominal prefix (read: ‘singular nominal prefix,’ as Gabriel expects users to know that nouns should be looked up as singulars). In any case, users thus have to use two entirely different strategies when dealing with nouns. For regular nouns the user needs to determine the form of the singular noun and go to the alphabetical place of the first letter of the singular nominal prefix. However, for derived nouns the user first needs to determine the form of the stem. By studying the stem the user must then guess the form the radical that heads the group (where the user hopes to find the noun) should have. The user must then go to
the alphabetical place of the first letter of that presumed radical, and only then will he/she need to (re-)add a singular nominal prefix to try to find the noun within the group headed by that presumed radical. Since many a user is not in a position to distinguish between regular and derived nouns, a situation in which two different lemmatization strategies are used for one type of part of speech is obviously not user-friendly. As a case in point we can consider two groups, (1) and (2), taken from TFG:4

1. DIDIBA : n.s. étoffe indigène
   MUDIBA : n.s. nasse ; TSHIDIBA : n.s. sorte de liane ; MUDIBA : grande cruche
   NDIBANDIBA : n.s. paroi

2. DIMA : v.a. labourer, remuer la terre
   TSHIDIME : n.s. abatteur de besogne, (grand) fermier, cultivateur
   BIDIMENENE : n.s. mauvaises herbes
   BUDIMI : n.s. 1° champ, terrain cultivé ; 2° culture
   MUDIIMU : n.s. besogne, carrière, culture, métier
   TSHIDIIMU : n.s. saison propre à certains travaux, âge, année

   MUDIDIMA : n.s. nom d’une plante

On the one hand a user might encounter the word madiba ‘native cloths’, recognise that it is a noun from gender 5/6, and decide that it is a regular noun. In that case the user goes to the page with running heads DI to DIM in order to find the singular didiba — where it is indeed the head of a group of three entries, compare (1). On the other hand a user might encounter the word midimu ‘jobs’, recognise that it is a noun from gender 3/4, and assume that it is derived from the verb dima ‘to cultivate’. The user then goes to the pages with running heads DIM to DIN — and finds the singular mudimu listed as a deverbalizer, compare (2). Thus far everything is in line with Gabriel’s code. However, what if a user had come across mudibà ‘fish basket/large jar’ and had decided that it is a regular noun? He/she would go to the page with running heads MOT to MUN, but would not find it there. Then this user might try the page with running heads DI to DIM, and find it with some luck under didiba, compare (1). This however poses a problem. The user will not know whether didiba or mudibà heads the group. If it had been the other way round, the entire group would have had to be found under the letter M instead of the letter D. To the user there seems to be no logic as to why a specific form was chosen to head a particular group. Moreover, there is not even a semantic link between the grouped members, for example didiba ‘native cloth’ versus mudibà ‘fish basket’ versus tshidiba ‘kind of liana’ versus mudibà ‘large jar’ versus ndibandibà ‘dividing wall’. In other words, even if one knows a certain noun is not a derived one, one still is not sure where to look for it. A close examination of (1) raises even more questions. Suppose a user is searching for the noun tshidiba.5 Unfortunately, nor the letter T (regular noun type) nor the letter D (derived noun type) includes this noun as a lemma. It can only be found inside the article of mudibà which itself can only be found in the group headed by didiba. Hence, one must conclude that Gabriel also makes use of a third way to lemmatize nouns, namely within the article of other nouns. The user is unlikely to find a way through this labyrinth. Gabriel even manages to include homonyms of lemmatized nouns inside their articles, whilst different nouns separate these very homonyms, compare for example 1mudibà and 2mudibà in (1).6 And to complicate matters even further, Gabriel takes great liberty with the concept of ‘derived nouns’, as he not only derives them from verbs (deverbalizers), but also from any other part of speech, compare for example (3), or even from radicals on which he imposes a meaning, compare for example (4):

3. KADI : conj. mais, cependant, néanmoins
   DIKADI : n.s. palmier
   NKADI : n.s. amertume

   MUKAJI : n.s. 1° sorte de jeu ; 2° femme

4. NTU : quelque… ; muntu : quelqu’un ; buntu : humanité ; bumuntu : bonté ; ishuntu : quelque chose ; huntu : richesse ; kuntu, puntu : quelque part
TSHINTU(M)BINDI : esp. d’antilope

... 
MUNTUNTU : n.s. 1° le jeune mwenze avant l’apparition des ailes ; 2° saumure
TSHINTUNTU : n.s. sorte de termite

In (3) Gabriel ‘derives’ nouns from the conjunction kâdi ‘but’, and in (7) he imposes the meaning quelque ... ‘some ...’ on the radical -ntu.

Gabriel’s code in which he tries to group words around their morpho-lexical base, combined with his attempt to classify the information within those groups on etymological grounds, results in yet three other very user-unfriendly features. Firstly, some extremely important words belonging to the core vocabulary are nearly impossible to find. In (3) for instance mukâjî ‘woman’, which is one of the hundred most frequently occurring words in the Lubà corpus, is not only incorrectly derived from a conjunctive, but is not even given as the first sense! Secondly, within the groups the alphabetical arrangement is lost. This situation is obvious in (2), (3) and (4). And thirdly, in many a case TFG’s code gives way to the grouping of words that have virtually nothing in common, see (5):

5. LETA : v.n. se montrer
LETELA : v. dér. faiblir, s’affaiblir ; leto : faiblement
LEULA : v. dér. affaiblir, défaillir, exténuer ; dîleu : distraction
TSHILEWU : n.s. calebasse, gourde
LEYILA : v.a. inviter, appeler quelqu’un
KALEYA : n.s. tour
MULEYA : n.s. grande houe
TSHILEYA : n.s. 1° V. mponda ; 2° gaminerie

Unfortunately, from the third entry onwards, T seems to be freely interchangeable with U, W and Y. Likewise, in (3), where the noun mukâjî is listed under the conjunction kâdi, J takes the place of D. In other words, the user has no clue whatsoever as to why these lemmas are to be found there where one least expects to find them! To make things even worse, the verb heading the group in (5) is not listed in any other dictionary, not even in Gabriel’s own Dictionnaire Français-Tshiluba, nor do native speakers know it.

As far as the nouns in (5) are concerned, a noun such as dîleu ‘distraction’ is entered in the article of the verb -léulu ‘weaken; extinguish’, which now means that nouns are lemmatized in four different ways in TFG. Though tshilewû ‘calabash’, kâleyà ‘tower’ and muleyu ‘big hatchet’ have been provided with a translation equivalent, the user looking for the first sense of the last entry is referred to mponda. Hence, one tries the page with running heads MOT to MUN – to no avail. Then one tries the stem, thus the page with running heads POM to POT – again to no avail. The user is consequently confronted with a reference into the void, as mponda is not lemmatized in TFG. In most cases the user will have long given up by now, but if one goes on scanning groups of entries, one might come across the following group on the page with running heads PAY to PEM:

6. MPELU : n.s. rotule
PEMA : v.a. réprimander sévèrement,
admonester

... 
BIPEMBAPEMBA : V. mponda

Obviously, this situation only adds to the user’s frustration, as one cannot be sure if the last entry in (6) is a synonym of mponda, or whether mponda in (6) is a homonymous noun of mponda in (5), or whether mponda in (6) is perhaps not even a noun at all but a different part of speech. In any case, the user still has no translation equivalent(s) at hand.

Actually, both the references in (5) and (6) should have been to tshipondâ (a pejorative for
**tupondà** ‘an ear of millet’) as it is **only** in the article of this noun that the last entry from (5) and the last entry from (6) are used in connective pronoun constructions. See (7):

7. **PONDA** : v.a. écraser

... 

**LUPONDA** : n.s. épi de millet  
**TSHIPOnda** : épi de millet; mponda wa  
**Tshileya** : millet long, à épis longs, à graines volumineuses; wa mfikanyi : à tige bleue violacée; wa bipembapamba : quand les fleurs ont disparu, mais que les graines sont encore tendres; wa bikupukupu : maigre et peu productif  

... 

**MPONDOKOTO** : épine des cuisses des sauterelles

**TFG** would have gained a lot in clarity if Gabriel had, on the one hand, taken vowel quantity and tone into account, and if he had, on the other hand, restricted his groupings to morpho-semantic ones. He would have noticed that many a group he suggests in **TFG** does not consist of words that belong together. Even then, the very idea of lemmatizing words according to morpho-lexical fields, or even on morpho-semantic grounds for that matter, is not the correct starting point. The word structure of Cilubá simply does not permit a user-friendly result. Kalonji is therefore entirely in the right when he notes:

‘... retrouver un mot dans le dictionnaire de Frère Gabriel demande beaucoup de prouesse d'esprit mais aussi d'intuition, et même beaucoup de chance, car l'ordre alphabétique n'étant pas respecté, aucun autre ordre, qu'il relève du champ dérivatonnel ou sémantique, n'est systématisé jusqu'à la fin ... Autant dire que faute d'organisation et de rigueur dans la présentation de la nomenclature, chercher un mot dans le dictionnaire de Frère Gabriel relève d'une gymnastique intellectuelle et requiert une parfaite connaissance de la langue: toutes choses contraires au principe de dictionnaire et qui dispenseraient d'y recourir’ (1993: 134–136).

We are tempted to go even a little further. Only a native Lubā speaker with a solid linguistic training in his own language could hope to use **TFG** with reasonable success. Such a person however, will not take too long before seriously questioning Gabriel’s extravagant morpho-lexical groupings.

3. **Lemmatizing both singular and plural noun forms**

The lemmatization of nouns under both their singular and plural forms is the method that Kriel, the dictionary pioneer for Sepedi, used for decades in his *Pukuntšu* and *Popular* dictionaries. More recently, it was also utilized by Lombard, Barnard & Grobler in their *Sediba* trilingual dictionary. The major advantage of this lemmatization procedure is user-friendliness rendering a practical dictionary. No previous knowledge of the language is required. The user does not even have to know a single word in the language or to be able to say a single sentence as long as he/she knows the alphabet. In addition, this approach causes no problem for irregular forms.

Nevertheless, a very serious disadvantage is brought about by a **noun redundancy of almost 80%** using up vital space, which could have been utilized for the inclusion of extra entries. Even more alarming is the fact that these compilers seem to **lack a strategy of selection**, namely which nouns to include and which ones to leave out. In dictionaries such as these, compilers often randomly add words as they encounter them until the maximum dictionary volume or number of pages prescribed by the publisher has been reached. Such dictionaries therefore include lots of rarely used words that are unlikely to be looked for by the target user at the expense of crucial ones.

In *Sediba* for instance, room was found for rarely used words such as leba (you, pl.)*'unload!' (0) or lebalang! ‘forget! (you, pl.)’ (0) but not for highly used ones such as koala ‘congratulate’ (14), laalisa ‘explain’ (38), laalo ‘command’ (25), leano ‘a plan’ (60) or leba ‘go in the direction of!’ (296). Apart from the issue
essential versus non-essential, other types of imbalances occur in Sediba, namely inconsistences regarding the lemmatization of singular versus plural forms, locative versus non-locative forms, et cetera. See for example (8) for inconsistences regarding the indication of the part of speech noun:

8. maaka 'leuen(s)' lie(s) (n)
    maano 'planne' plans
    maswi 'melk (nw.)' milk (n.)

4. Lemmatizing only singular noun forms
The lemmatization of only singular noun forms is for instance followed in the latest edition of the Pukuntsu dictionary. It can be argued that it is not too much to require from the target user to know the regular productive rules in the specific African language through which singular is changed into plural and vice versa. See Table 3 for the series of Sepedi rules found in the latest Pukuntsu.

Such rules are in themselves not user-friendly at all and dictionary users are known not to consult them. But in forcing the user to look for plural forms under their singular forms, precious space is saved which can be utilized for other entries. For example, the simple rules formulated for nouns in the 1989 edition of Pukuntsu renders a saving of 30 pages in that dictionary, or more or less 10% compared to the 1983 edition. The number of pages for words commencing on ba-, di-, ma- and me- in these editions are compared in (9):

9. Pukuntsu 1983
   Number of pages
   ba- versus 7
   di- versus 18
   ma- versus 21
   me- versus 7

   Pukuntsu 1989
   Number of pages
   2
   4
   16
   2

Unfortunately it is not always that easy to apply the rule since in most cases, from a user’s point of view, there is not a straightforward one-to-one correlation. See again Table 3: di- (class 8 and class 10) has a one to two correlation, namely either se- or noun minus prefix. In simple terms it means that the user who does not know the meaning of the word dilepe ‘axes’ has to look it up under *lepe or selepe. A one to three correlation exists in the case of ma- (class 6) and a one to four correlation in the case of me- (class 4).

Say for instance that the user is confronted with the word meno ‘teeth’. The rule in Table 3 clearly states that me- should be looked up under mo-. The user finds the word mono in the dictionary as ‘finger’ and consequently concludes that meno means ‘fingers’ whilst it means ‘teeth’. The same is true for the word meetse ‘water’. Taken at face value it is a word in class 4 and according to the rule in Table 3 it should be looked up under moetse. Under moetse the user finds ‘mane, crest’ and thus concludes that meetse means ‘mane, crest’ whilst it means ‘water’. In fact meetse is in
class 6, not in class 4, and has no singular form in Sepedi. As another example, suppose the user wants to look up the word *meselane* ‘builders’. Following the rule the user tries *moselane* but finds nothing. Now the user has to accept that the word has not been included in the dictionary or try to look it up without applying the rule, which is the right thing to do for this word. In this case the rule is misleading instead of helping the user. A final illustration here is the word *mekgona* ‘rubbish’. First the user follows the rule, namely to look it up under *mo-* but does not find it. Then the user ignores the rule as had to be done in order to find the previous example, but this time the user is unsuccessful. Any lengthy search will remain in vain because this word has not been included into the dictionary due to a low frequency of usage.

The lexicographer is forced to lemmatize all irregular forms separately anyway and consequently does not succeed to avoid a certain redundancy. Moreover, the inexperienced user has to know whether the particular word that needs to be found is a noun and not something else. See Table 4 for four randomly selected words (an adjective, a verb, a demonstrative and a noun) where the rule misleads the inexperienced user in three out of four cases.

Another weakness of this approach is that extremely highly used plural forms such as *matšaši* ‘days’ (395), *basadi* ‘women’ (387) or *maoto* ‘feet’ (301) are not physically included in the dictionary so that the inexperienced user is often in doubt as to whether the right conclusions regarding meaning, circumflexes, tonal pattern, **et cetera** have been made. This is especially true in those cases where the plural forms are more frequently used than the corresponding singular forms such as *maoto* ‘feet’ (301) versus *leoto* ‘foot’ (93) or *diaparo* ‘clothes’ (141) versus *separo* ‘a piece of clothing’ (7). It is rather counterproductive to enter the less frequently used singular form just to be consistent. This is particularly evident in the case of *badimo* ‘ancestral spirits’ (296) which is lemmatised under *modimo* ‘ancestral spirit’ (9).

Finally, the diligent application of this approach also results in the formulation of rules of which the relevance is so limited that the compilers could have entered such words as separate lemmas, especially for this dictionary format. For rules such as ‘*mahj-* look up under *hj-*’, ‘*meng-* look up under *ngw-*’ and ‘*nyvit-* look up under *ngw-*,’ an ad hoc approach would probably be a better alternative here. It is this strategy that was successfully used for certain verbal suffixes in this dictionary. In simple terms it means that in sacrificing only a few pages of the dictionary, a large number of rules such as the three just enumerated could be scraped.

As far as the lemmatization of singular noun forms combined to a rule-orientated approach is concerned, it must be concluded that it succeeds in combating redundancy but unfortunately makes some unrealistic claims on the knowledge of the inexperienced user.

### 5. Lemmatizing nouns on the first or the third letter

The lemmatization of nouns on the first or the third letter is pursued by Snyman et al. in their Dikisionare ya Setswana English Afrikaans Dictionary Woordeboek (SEA) and is one in which consideration is given to the user who is not too familiar with the language. Firstly, this simple approach gives a sense of security to the inexperienced user in that there are only two possibilities. There is thus less chance of getting lost in the dictionary jungle as was the case for lemmatization approaches (1), (2) and (4). In *SEA*, irregular forms are automatically accommodated by being entered under their first letter, whilst regular

<table>
<thead>
<tr>
<th>Table 4 Four randomly selected words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example</td>
</tr>
<tr>
<td>bahedi</td>
</tr>
<tr>
<td>bota</td>
</tr>
<tr>
<td>balo</td>
</tr>
<tr>
<td>batho</td>
</tr>
</tbody>
</table>
forms are to be found under their third letter. Secondly, in contrast to lemmatization approach (3), the compilers surely succeeded in avoiding some redundancy in the case of regular nouns. See for example lokwalo ‘book’ in (10):

10. kwalø, le- ma-dev < letter // brief; los-dle, book // boek; mo-me, handwriting, orthography // handskrif, skryfwyse

With a single entry the compilers cover what might otherwise have been six separate entries.

The most important shortcoming of this approach however is the presupposition that target users have an adequate knowledge of the noun class system—which is often not quite in line with reality. Furthermore, the fact that there are only two possible lemmatization strategies for nouns instead of for instance four as in approach (2) is good news, but at the same time this also means that there are always two possible ways to look up nouns—which results in many an inexperienced user loosing a lot of time due to starting with the ‘wrong look-up strategy’. Finally, for users who are used to looking up nouns under the first letter, as for example in Kriel’s Pukunisù or Popular dictionaries, the approach used in SEA will cause lasting confusion and a longing back to the more user-friendly approach in which one always simply turns to the complete noun, thus to the prefix plus stem.

Lexicographers’ traditional approaches, linguists’ theories and target users’ desires

The above survey of lexicographers’ traditional approaches obviously brings us to the titillating question whether it is possible to formulate a lemmatization strategy for nouns which avoids the shortcomings and pitfalls mentioned for approaches (1) to (5) but which exploits the virtues of those very same approaches at the same time.

One could for instance exploit the virtues of (3) such as (a) very user-friendly, (b) no previous knowledge of the grammar required, (c) in most cases not necessary to consult the guidelines to the dictionary, and (d) solving the problems relating to irregular forms as well as instances where the plural form is more likely to be looked up than the singular form. At the same time one could avoid the major shortcoming of this approach, namely redundancy, by giving only essential information in a smaller font size at the member of the pair less likely to be consulted, with a cross reference to the other one where a full treatment is offered. See the following over-simplified example:

11. meriri n, cl 3/4 LHL hair (on the head) (plural), meriri a mesweu ke lehumo grey hair is a treasure; ~ wa titsana soft hair, motho wa ~ a reliable person

moriri n, cl 3/4 LHL (one) hair | see meriri

One could for instance capitalize on the virtues of (4) by designing user-friendly guidelines to the dictionary as well as a well-compiled front and back matter interlinked with a sound strategy of cross-referencing.

Approaches (1) to (5) could also be viewed from a different angle by regrouping them into three lexical traditions. On the one hand there is the ‘stem tradition’ in which stems of written nouns are the basis for lexical entries, compare (1). On the other hand we have the ‘word tradition’ in which complete written nouns are the basis for lexical entries, compare (3) and (4). In between there is a ‘hybrid stem or word tradition’ in which both stems of written nouns and complete written nouns can be the basis for lexical entries, compare (2) and (5). The key question is, therefore, whether one of them is perhaps more ‘scientific’ according to linguists’ theories.

Van Wyk (1995) made an in-depth study of the ‘scientific’ value of lemmatization on noun stems. He indicates clearly that the stem tradition in respect of the lemmatization of nouns is based on an incorrect assumption, namely that noun class prefixes are as productive as verb prefixes in the derivation of stems. In the case of verbs it is not disputed that lemmatizing on verbal stems is the recommended option since all verbs have series of productive affixes. This is not the case with nouns:
The basic assumption of stem dictionaries is that the morphology of the verb and the noun is identical in that prefixal elements can be attached freely to stems in both cases … This assumption is, however, wrong; the morphology of the noun differs in crucial ways from that of the verb. The noun prefix is not mobile or freely exchangeable … Any verb root can be combined with any subject marker, any modal or aspectual morpheme … None of this applies to the noun … The crucial difference with verbs is that noun class prefixes are combined largely in an ad hoc manner with stems … This results in a fundamentally different handling of verbs and nouns in stem dictionaries … This means … that separate entries must be made for each combination of a prefix plus a stem’ [italics by Van Wyk] (Van Wyk, 1995:86–88).

He finally concludes that word dictionaries are based on sounder linguistic and lexicographical principles:

“The stem principle is, in fact, less “scientific” than the word principle, since it is based on erroneous assumptions as regards noun morphology. It is not and cannot be applied consistently, as many nouns have of necessity to be lemmatized with their prefixes. Its only possible gain is that it reveals lexical relations, which is in any case questionable from the point of view of the general user. In addition, it loose in user-friendliness in that it assumes considerable linguistic sophistication on the part of the user … It has to be concluded, then, that … word dictionaries … are based on sounder linguistic and lexicographical principles’ (Van Wyk, 1995:93–95).

Whatever linguists or lexicographers might want to theorize or put into practice, every modern dictionary will have to take a third component into account, namely the target users’ desires and their competence in the language(s) under consideration. As far as the three central role players in any lexicographic activity are concerned, Lombard writes:

‘I want to depict lexicographer, dictionary user and linguist as the three angles or sides of a triangle … Under ideal conditions, the three … should form a well-integrated triangle. However, this is not always the case and more often than not, these three sides or angles are at loggerheads’ (1994:204–205).

With the notion of target users’ desires the coexistence problem between the three central role players might even be more pronounced than suggested by Lombard, as we intend those desires to be somewhat broader than the urgency to take care of target users’ needs and expectations and their competence in the language(s) under consideration.

Compilers write dictionaries to be used by target users. Target users consult the dictionaries and in this process they come to judge those very dictionaries. Dictionaries’ privileged critics are therefore the target users who were the original focus of the compilers. We have come a full circle: from compilers to target users to dictionaries and back to compilers. This full circle is not just a linguistic loop, this loop represents indispensable feedback. However, for most dictionary projects this feedback comes too late. In fact one would like to have feedback from the privileged critics (the target users) both from the very start and during the entire compilation process of the dictionary — hence, the chimera listens to ‘simultaneous feedback’ from the target users to the compilers. But is it really a chimera?

Actually, the concept of simultaneous feedback is currently applied to the Woordenboek Citibà-Nederlands (WCN). In other words, whilst WCN is being compiled it is at the same time being used by its target users — and through this loop the privileged critics guide the compilers continuously. As the whole concept of simultaneous feedback is accompanied by the introduction of a new methodology for compiling dictionaries, fully expounding on it here would go beyond the scope of the present article. We would therefore like to
refer the reader to De Schryver & Prinsloo (forthcoming) in which a thorough discussion of the basic principles will be found. Nonetheless, the gist of simultaneous feedback as applied to the compilation of WCN will be outlined here.

One can only receive helpful feedback when one knows exactly what one is trying to achieve (the target) and when one knows exactly for whom one is doing this (the target users). The target of WCN is a pocket-sized bilingual learner’s dictionary Cilubâ-Dutch based on frequency of usage. The target users are university students. At the start of the project, the ‘currently available dictionaries’ included only one with the direction Cilubâ-Dutch: the Dictionnaire Tshilubâ-Français by De Clercq & Willems (TFDW) – in which Cilubâ, French and Dutch are covered in a one-way configuration. Unfortunately, a speaker of Dutch that is not acquainted with French will not be able to use TFDW satisfactorily. Before beginning any work on WCN, the actual use of TFDW was monitored in an informal way. Later, information was gathered in a more formal way through direct questions like: ‘Ideally, what would you like to find in a Cilubâ-Dutch dictionary? And how would you like to find it?’ Feedback was thus simultaneously introduced right from the very start.

Since WCN was to derive all its data from a large Lubâ corpus, the lexicographers were faced with the impossibility to start right away with the compilation, and with the prospect of an extremely time-consuming undertaking. However, to ensure continued simultaneous feedback, the main project WCN was complemented with a series of parallel projects. A first parallel dictionary project, which had to be completed in a limited period, was set up: the Lexicon Cilubâ-Nederlands (LCN) – a small-scale unidirectional decoding learner’s dictionary. In just a few months, LCN was created from a small parallel test-corpus. From the release of LCN onwards, feedback was received from this parallel project and channelled into the time-consuming main project. From that instant, the creation of a Cilubâ-Dutch dictionary became a true ‘work in progress’ with simultaneous feedback from the target users to the compilers.

As far as the lemmatization of nouns on the macrostructural level is concerned, the analysis of LCN’s target users’ desires can be summarized into four rules of thumb:

- For intermediate and advanced learners one can safely lemmatize nouns under their singulars as long as provision is made for irregular plurals; but for elementary learners additional guidance on the noun class system will have to be provided for, if this is to be a successful approach. Alternatively, one might even have to consider lemmatizing both singulars and plurals for inexperienced users.

- Irregular plural nouns must both be mentioned within the article of their singular forms and be lemmatized under their proper alphabetical position, whilst cross-references should link both forms.

- Noun gender information is crucial and should, for advanced learners, preferably be coded using class numbers, whilst elementary and intermediate learners might wish to receive some additional guidance.

- Under no circumstances should nouns be lemmatized according to the ‘stem tradition’.

Hence, LCN’s feedback indicates that different lemmatization strategies for nouns are to be used according to the proficiency level of the target users. This conclusion was instantly implemented in a second parallel project: the Beknopt woordenboek Cilubâ-Nederlands (BCN) – basically LCN with the inclusion of a frequency of usage slot. In BCN, the feasibility of a ‘repetitive inserted text at the bottom of every page’, with the requested additional noun class system guidance for elementary learners, is tested.

In a way the parallel projects (LCN, BCN, and future ones) are thus used as experimental tools, since they can try out a plethora of strategies in order to refine the presentation of the information in the main project under construction (WCN). All this makes clear why one cannot agree with Lombard when he claims: ‘Lexicography is not a terrain in which you can experiment from one day to the next’ (1994:211).
On the contrary, complying with the target users’ desires means doing just that! And within the framework of ‘simultaneous feedback’, the combination of target users’ desires and data gathered from actual dictionary use instantly guides the compilers in their endeavour to ‘avoid the pitfalls of the traditional approaches whilst capitalizing on their virtues’.

Conclusion

Obviously, the four rules of thumb mentioned above cannot be generalized. Every dictionary project has its own objectives and target group, so that lemmatization strategies might have to vary to some degree.

In addition to those four specific items, we can extract ten useful general guidelines for the compilation of dictionaries for African languages:

- One should not compile just another dictionary or lexicon for an African language. It is cardinal to do some market research and to tackle a category for which no dictionaries or lexicons exist as yet, or if one wants to compete in a certain category, one must at least make sure to improve on existing versions.

- One should put user-friendliness and practicality on top of one’s priority list of lexicographical principles to be considered and balanced. Lemmatizing nouns in a user-friendly and practical way implies that target users must be able to look up nouns directly under their first letter.

- As space has to be used economically in printed dictionaries, one should not try to bridge more than two languages in one dictionary.

- The guidelines to the dictionary must be complete and easy to comprehend.

- As one cannot expect too much from the (inexperienced) target users, one should be careful with rule-oriented approaches.

- One must not try to create a ‘more scientific’ alphabetical sequence, but rather stick to an ordinary alphabetical ordering.

- As regards spelling (including tonal information) one has to adhere to strict orthographic rules, for the classification of the entire macrostructure depends on it.11

- One should build a corpus to embark on a study of frequency of usage and use it as a basis for inclusion or omission of words. This will ensure that highly used words are not accidentally omitted on the one hand, and will avoid that precious space is being occupied by rarely used words on the other hand. The motto throughout should be to ‘lemmatize in a user-friendly way those words most likely to be looked for by the target users’.

- One should think about including a very large number of entries with short articles rather than relatively few entries with long articles.

- In case one decides to make provision for both singular and plural nouns (to cater for elementary learners) a suggestion might be to enter the most frequently occurring form with a complete article and the less occurring counterpart (if there is any) in its appropriate alphabetical position as a separate but skeleton entry. The skeleton entry could perhaps consist of only lexical information with a cross-reference to the more frequent form. Also, one may wish to think about reducing the font size of the skeleton entry to diminish the introduced redundancy, whilst at the same time highlighting the issue essential versus non-essential by typographical means.

Notes

1. In this article we will focus our attention upon two languages from the ‘Bantu language family’: Sepedi (spoken in Southern Africa) and Cilubà (spoken in Central Africa) – respectively S32 and L31 following Malcolm Guthrie’s classification.

2. The fact that nouns carry different nominal prefixes that determine the form the concordance affixes take for grammatical agreement, is exactly one of the main characteristics of the ‘Bantu language family’.

3. The gender 1/4a for Cilubà is traditionally represented as 9/10. This numbering is correct when one only considers the nominal prefix as a classification criterion. As syntactical
concord prevails for classifying nouns in genders, it is useless to maintain classes 9/10, whose concords are the same as those of 1/4 (compare Kabuta, 1998:59–60).

4. All the Lubâ dictionary entries have been substantially shortened to highlight the details relevant to the context in which they are quoted.

5. In this dictionary the so-called ‘traditional spelling’ is utilized in which tsb is used instead of c to represent the consonant /k/. 

6. Every now and then, as for this entry, one or more part of speech slots are missing in TFG.

7. Compare previous note.

8. Frequency counts derived from the corpus have been bracketed as follows: (number of occurrences per one million words).

9. An analogous phenomenon is encountered in Cilubâ, e.g. bisaosâ ‘grass’ (200) versus cisaosâ ‘one blade of grass’ (3) or nsuki ‘hair’ (127) versus lusuki ‘one (small) hair’ (7). A short discussion of the lubâ corpus can be found in De Schryver & Kabula (1998:119).

10. The Woordenboek Cilubâ-Nederlands (Dictionary Cilubâ-Dutch) is one of the dictionary projects 'in progress' under the wings of Recall (Research Centre of African Languages and Literatures), University of Ghent, Belgium.

11. The issues concerning tonal indication in Sepedi have been discussed elsewhere (compare Prinsloo, 1994:100). In LCN and BCN, contrary to TFG and TFDW, vowel quantity and tone are used throughout the dictionary.

References


Abbreviations used in reference to dictionaries and lexicons

Beknopt woordenboek Cilubá-Nederlands (BCN)
Dictionnaire Tshiluba-Français by Gabriel (TFG)
Dictionnaire Tshilubá-Français by De Clercq & Willems (TFDW)
Dikišinare ya Setswana English Afrikaans Dictionary Woordeboek (SEA)
Groot Noord-Sotho woordeboek (GN)
Lexicon Cilubá-Nederlands (LCN)
Woordeboek Cilubá-Nederlands in progress (WCN)