Populating Sub-entries in Dictionaries with Multi-word Units from Concordance Lines

Thapelo J. Otlogetswe, Department of English, University of Botswana, Gaborone, Botswana (otlogets@mopipi.ub.bw)

Abstract: Lexicography is primarily concerned with the representation of words and their senses in dictionaries. By words most dictionary users and lexicographers refer to a combination of characters delineated by spaces on both sides. This article discusses the weakness of this approach in the selection of dictionary entries. Through an inspection of concordance lines generated from a multi-million Setswana corpus, it is argued and demonstrated how multi-word units (MWUs), also known as multi-word expressions (MWEs), may be extracted from concordance lines to supplement dictionary entries. It is illustrated how both monolingual and bilingual Setswana dictionaries may be enhanced by the addition of MWEs as sub-entries.

Keywords: SETSWANA, LEXICOGRAPHY, MULTI-WORD UNIT, CORPUS, CONCORDANCE, MULTI-WORD EXPRESSION, COLLOCATION, WORD, SUB-ENTRIES, DICTIONARY

1. Introduction

At the centre of lexicography lies the problem of what constitutes a word. The problem is not only a lexicographic one. It is also a linguistic one. McArthur (1998: 45-47) identifies eight types of words: orthographic, phonological, morphological, lexical, grammatical, onomastic, lexicographical and statistical words.
What constitutes words is critical in corpus linguistics, since it translates into the problem of what gets counted by the computer. Lexicographically, those ones considered as words are listed in the dictionary.

2. The word problem

In frequency analysis, there is therefore a need to clarify what constitutes a word in a language and how words get counted. In linguistic literature, the term word is defined in a variety of ways. Some of these definitions, while useful for theoretical linguistics, are useless for computational word counts. Finch (2000: 132) defines a word as "a unit of expression which native speakers intuitively recognize in both spoken and written language" and adds that "there is a certain indeterminacy about the definition of a word". Finch's definition is unhelpful in that "a unit of expression" could be anything from a word, a phrase, a clause or a sentence. His definition also leaves the determination of what a word is to a speaker's intuition which may vary from one speaker to another. Aitchison (1992: 49) points out that "the best-known definition of a word is the one proposed by the American linguist Bloomfield who defined it as a minimum free form, that is, the smallest form that can occur by itself". She further argues that distinctions must be made between lexical items, syntactic words and phonological words. If we consider lexical items, a form such as fly represents at least two words:

fly [noun]: an insect with two wings.
fly [verb]: to move through the air in a controlled manner.

The two lexical items have different syntactic forms associated with them. The noun could either be singular (fly) or plural (flies). The verb on the other hand could occur as fly, flying, flies, flew and flown. This therefore raises problems for the Bloomfieldian approach.

Leech et al. (1982: 27) consider a word as "delimited, for most purposes by a space (or punctuation mark other than a hyphen or apostrophe) on each side". This is known in linguistic literature as an orthographic word. However they also acknowledge that "the boundaries of words ... are not always clear; e.g. we can write the sequence piggy + bank in three ways: piggy bank, piggy-bank, or piggybank".

In most computational processes, a word is treated as a "minimal free form, the smallest unit that can exist on its own" (Dash and Chaudhuri 2000: 189) and "delimited by a space ... on each side" (Leech et al. 1982: 27). This approach is helpful if one is studying forms delineated by spaces. However, in this article, larger units which have spaces within them are studied. Moon (1998) calls these fixed expressions and idioms. In other literature they are called multi-word units or MWUs (Schone and Jurafsky 2001) or multi-word expressions or MWEs (Sharroff 2004; Oflazer and Çetinoğlu 2004; Villavicencio
et al. 2004; Fazly and Stevenson 2007). Bannard (2007: 1) gives the following definition:

A multi-word unit is usually taken to be any word combination (adjacent or otherwise) that has some feature (syntactic, semantic or purely statistical) that cannot be predicted on the basis of its component words and/or the combinatorial processes of the language. Such units need to be included in any language description that hopes to account for actual usage.

Sag et al. (2002: 2) characterize MWEs as "idiosyncratic interpretations that cross word boundaries (or spaces)". And if Jackendoff's (1997: 156) estimate that the number of MWEs in a speaker's lexicon is of the same order of magnitude as the number of single words, then MWEs deserve focus and will significantly enhance dictionary entries.

MWEs therefore include idioms, phrasal verbs, proverbs, compound words, etc. English examples are by and large, kick the bucket, in step, take up, take off, shake up, telephone booth, pull strings, fresh air, fish and chips, salt and pepper, etc. Setswana examples are solegela molemo (benefit), kukega maikutlo (be upset), iphaga dikoro (involve oneself in other people's business), tsholetsa maoto/dinao (walk faster), opisa tlhogo (cause trouble), tsaya karolo (participate), tsaya tsia (pay attention), nna le seabe (take part), ja monate (enjoy), etc. The immediate problem arises with their identification, since they can be written in diverse and inconsistent ways. Take for example the following different spellings which are acceptable in both English and Setswana as written in Botswana.

<table>
<thead>
<tr>
<th>English</th>
<th>Botswana</th>
<th>Setswana</th>
</tr>
</thead>
<tbody>
<tr>
<td>houseboat</td>
<td>house-boat</td>
<td>house boat</td>
</tr>
<tr>
<td>tradeoff</td>
<td>trade-off</td>
<td>trade off</td>
</tr>
<tr>
<td>khuduthamaga</td>
<td>khudu-thamaga</td>
<td>khudu thamaga</td>
</tr>
<tr>
<td>pelo tsheitha</td>
<td>pelo-tsheitha</td>
<td>pelo tshetlha</td>
</tr>
<tr>
<td>rampa tsheitha</td>
<td>rampa-tsheitha</td>
<td>rampa tshetlha</td>
</tr>
<tr>
<td>motshwaradiphala</td>
<td>motshwara-diphala</td>
<td>motshwara diphala</td>
</tr>
<tr>
<td>kgakala kgakala</td>
<td>kgakala-kgakala</td>
<td>kgakala kgakala</td>
</tr>
</tbody>
</table>

The examples houseboat and kgakala kgakala will each constitute a single token, while house boat and kgakala kgakala will form two tokens each. Words joined by a hyphen can either be recognized as single words or as two separate words depending on the tokenizing program. The difference is not trivial in statistical linguistics, since the number of tokens will vary significantly depending on what is counted.

3. Methodology and experiments

For our experiments, we follow Brunner and Steyner (2008) and use corpus data. By a corpus is meant, according to Renouf (1987: 1), "a collection of texts,
of written or spoken words, which is stored and processed on computer for the purpose of linguistic research. The Setswana corpus used for the experiments is just over 13 million tokens. The software employed is Oxford Wordsmith Tools Version 4 (Scott 2004–2006). It is applied to study a specific word in context in some detail in terms of co-texts to its left and to its right. This is achieved by generating a key word in context (KWIC), often referred to as concordance lines. Dash and Chaudhuri (2000: 190) give the following definition:

A concordance is an index of the surface word forms in a text. It is a collection of the occurrences of a word form, each in its own textual environment.

A concordance reveals the context of a word, its collocates, and thereby reveals meanings and usages which are hard to recover through mental recall (Otlo-getswe 2007: 56). We illustrate this below with the example of the word *pelo* (heart).

**Figure 1: Concordance results for the word *pelo*”

In the above concordance lines, *pelo* together with its collocates, is rarely used to convey the meaning of the physical heart, “a hollow muscular organ that pumps the blood through the circulatory system by rhythmic contraction and
dilation” (Pearsall 1998: 847). In the first lines, *wela pelo*, which literally translates as “have your heart fall down”, means “be at peace or be settled”. In the next lines, *moratiwa wa pelo* (the loved one of the heart) is equivalent to “sweetheart” or “beloved”. Further on, *tshwara pelo* (handle or hold the heart) means “be in control of your emotions”.

It is by inspecting collocates that we can uncover different MWEs such as proverbs, compounds, idioms, sayings, phrasal verbs, etc. Such structures can then be entered into dictionaries as sub-entries. Through the use of computer programs or concordance software, it is relatively easy to obtain a list of all the co-occurrences of a particular word in context and see all the meanings associated with the word (Biber et al. 1998: 27). The concordance lines above reveal the different subtle meanings associated with the word *pelo*. From such a study of concordance lines, a possible 84 sub-entries of the headword *pelo* have been extracted:

- *ama pelo*
- *balabala ka pelo*
- *baya pelo*
- *beta pelo*
- *betwes ke pelo*
- *bofa pelo*
- *bolawo ke pelo*
- *boliwetse jwa pelo*
- *bona pelo*
- *bongwefela jwa pelo*
- *bonosi jwa pelo*
- *boteng jwa pelo*
- *bua ka pelo*
- *bula pelo*
- *busa pelo*
- *fela pelo*
- *feretlha pelo*
- *fetola pelo*
- *gapa pelo*
- *garoga pelo*
- *kgaoga pelo*
- *go sera letsapa le fisang pelo*
- *gomolwa ke pelo*
- *isa pelo mafisa*
- *itaya pelo*
- *itse pelo*
- *kgvaradatsa pelo*
- *lala ka pelo e rotha madi*

- *mabetswe-e-pelo*
- *masetla pelo*
- *matlhomola pelo*
- *matlholiha-pelo*
- *nametsa pelo*
- *ngomola pelo*
- *ngona pelo*
- *nna pelo*
- *nona pelo ka mathe*
- *ntsha pelo*
- *ntsha pelo pelaelo*
- *pateletsa pelo*
- *pelo e boela mannong*
- *pelo e e boltlokho*
- *pelo e e letlapa*
- *pelo e ja serati*
- *pelo e khibidu*
- *pelo e rotha madi*
- *pelo e rotha*
- *pelo e setlhogo*
- *pelo e nhata*
- *pelo khotshuane*
- *pelo namagadi*
- *pelo ntshi*
- *pelo pedi*
- *pelo pholiwana e a golegwa*
- *pelo potsane e a golegwa*
- *pelo tshwe*
- *pelo yetlhe*
- *pelo-e-thata*
- *pelo-kgale*
- *pelo-tele*
- *pelo-tlhomogi*
- *phatloa pelo*
- *ritibatsa pelo*
- *sethunja sa pelo*
- *sera pelo*
- *sethunja sa pelo*
- *sisa pelo*
- *sulafatsa pelo*
- *sua pelo*
- *svegaswega pelo*
- *thiba maroba a pelo*
- *thuba pelo*
- *tlaala pelo*
- *tlaeletana pelo*
- *tlhomola pelo*
- *tlola pelo*
- *tsava pelo*
- *tswela pelo*
- *uba pelo*
- *wa pelo*
- *wele pelo*
In Table 1 below, only 10 of these are explained.

Table 1: Corpus-derived possible sub-entries for the entry *pelo*

<table>
<thead>
<tr>
<th>Collocates</th>
<th>Literal translation</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>ama pelo</td>
<td>touch the heart</td>
<td>hurt someone</td>
</tr>
<tr>
<td>balabala ka pelo</td>
<td>speak too much by the heart</td>
<td>talk aloud to yourself; be absent-minded</td>
</tr>
<tr>
<td>baya pelo</td>
<td>put the heart</td>
<td>relax; lay back</td>
</tr>
<tr>
<td>beta pelo</td>
<td>suffocate the heart</td>
<td>persevere</td>
</tr>
<tr>
<td>betwa ke pelo</td>
<td>be choked by the heart</td>
<td>be very angry</td>
</tr>
<tr>
<td>bofa pelo</td>
<td>tie the heart</td>
<td>restrain yourself</td>
</tr>
<tr>
<td>bolawa ke pelo</td>
<td>be killed by the heart</td>
<td>desire something but be unable to acquire it</td>
</tr>
<tr>
<td>bolwetsi jwa pelo</td>
<td>the disease of the heart</td>
<td>heart attack</td>
</tr>
<tr>
<td>bona pelo</td>
<td>see the heart</td>
<td>see somebody’s intentions or thoughts</td>
</tr>
<tr>
<td>bua ka pelo</td>
<td>speak with the heart</td>
<td>be troubled to the extent that you talk to yourself</td>
</tr>
</tbody>
</table>

In Setswana, the phenomenon of idiomaticity when considering a word and its collocates is not unique to *pelo*. Words like *molomo* (mouth), *mpa* (stomach), *nkọ* (nose), *monwana* (finger), *kgomo* (cow), and many others display similar characteristics. Such idiomatic expressions can enrich dictionary entries as sub-entries. Tables 2–5 present the idiomatic expressions for *molomo* (mouth), *mpa* (stomach), *lonaọ* (foot) and *matlhọ* (eyes) respectively which have been extracted through studying concordance lines.

Table 2: Corpus-derived possible sub-entries of the entry *molomo*

<table>
<thead>
<tr>
<th>Collocates</th>
<th>Literal translation</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>bolwetsi jwa tlhako le molomo</td>
<td>disease of hoof and mouth</td>
<td>foot and mouth disease</td>
</tr>
<tr>
<td>itoma molomo wa tase</td>
<td>bite the lower mouth</td>
<td>be determined</td>
</tr>
<tr>
<td>itshwara molomo</td>
<td>hold/touch the mouth</td>
<td>be shocked</td>
</tr>
<tr>
<td>ntsha ka molomo</td>
<td>release with the mouth</td>
<td>speak</td>
</tr>
<tr>
<td>pula molomo</td>
<td>that which opens the mouth</td>
<td>money paid before someone speaks in lobola negotiations</td>
</tr>
<tr>
<td>pipsa-molomo</td>
<td>that which covers the mouth</td>
<td>a bribe</td>
</tr>
<tr>
<td>ruwala molomo</td>
<td>carry the mouth on your head</td>
<td>be angry and tight-lipped</td>
</tr>
<tr>
<td>roka molomo</td>
<td>sew the mouth</td>
<td>remain quiet</td>
</tr>
<tr>
<td>tsaga molomo</td>
<td>grow a mouth</td>
<td>speak</td>
</tr>
<tr>
<td>tlhoka molomo</td>
<td>lack a mouth</td>
<td>have nothing to say</td>
</tr>
</tbody>
</table>
Table 3: Corpus-derived possible sub-entries for the entry *mpa*

<table>
<thead>
<tr>
<th>Collocates</th>
<th>Literal translation</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>bana ba mpa</td>
<td>children of a stomach</td>
<td>relatives</td>
</tr>
<tr>
<td>bipa mpa ka mabele</td>
<td>cover the stomach with the</td>
<td>withhold bad information to protect a relative or friend</td>
</tr>
<tr>
<td></td>
<td>breasts</td>
<td></td>
</tr>
<tr>
<td>gare ga mpa ya bosigo</td>
<td>in the centre of the belly of</td>
<td>in the middle of the night</td>
</tr>
<tr>
<td></td>
<td>the night</td>
<td></td>
</tr>
<tr>
<td>gare ga mpa ya lefatshe</td>
<td>in the centre of the</td>
<td>in the middle of nowhere</td>
</tr>
<tr>
<td></td>
<td>stomach of the world</td>
<td></td>
</tr>
<tr>
<td>gare ga mpa ya naga</td>
<td>in the centre of the</td>
<td>in the middle of nowhere</td>
</tr>
<tr>
<td></td>
<td>belly of the wilderness</td>
<td></td>
</tr>
<tr>
<td>mpa ya sebete</td>
<td>the belly of the liver</td>
<td>flat on the stomach</td>
</tr>
<tr>
<td>mpa e tuka molelo</td>
<td>a belly burning fire</td>
<td>filled stomach</td>
</tr>
<tr>
<td>go ja ka mpa tsoopedi</td>
<td>eat with two stomachs</td>
<td>eat until the stomach is full</td>
</tr>
<tr>
<td>ntsha mpa</td>
<td>take out a stomach</td>
<td>commit abortion</td>
</tr>
<tr>
<td>imelwa ke mpa</td>
<td>be overladen with a belly</td>
<td>have a full stomach</td>
</tr>
</tbody>
</table>

Table 4: Corpus-derived possible sub-entries of the entry *lonaô/dinaô*

<table>
<thead>
<tr>
<th>Collocates</th>
<th>Literal translation</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>apaya ka lonaô</td>
<td>cook with a foot</td>
<td>avoid cooking and eat at other people's homes instead</td>
</tr>
<tr>
<td>goga dinaô</td>
<td>drag the feet</td>
<td>move slowly</td>
</tr>
<tr>
<td>kodisa dinaô</td>
<td>cool the feet</td>
<td>have a rest</td>
</tr>
<tr>
<td>motsamaqua ka dinaô</td>
<td>one who walks with the feet</td>
<td>a pedestrian</td>
</tr>
<tr>
<td>ngotla dinaô</td>
<td>reduce the feet</td>
<td>walk slower</td>
</tr>
<tr>
<td>tlhlatlosa dinaô</td>
<td>raise the feet</td>
<td>walk faster</td>
</tr>
<tr>
<td>baqa lonaô</td>
<td>put a foot</td>
<td>be in a place</td>
</tr>
<tr>
<td>kgoelo elo dinaô</td>
<td>lift the feet</td>
<td>walk faster</td>
</tr>
<tr>
<td>kgwele ya dinaô</td>
<td>ball of the feet</td>
<td>football</td>
</tr>
<tr>
<td>tsisa dinaô</td>
<td>wake up the feet</td>
<td>walk faster; hurry up</td>
</tr>
<tr>
<td>tiisa dinaô</td>
<td>strengthen the feet</td>
<td>walk faster</td>
</tr>
</tbody>
</table>

Table 5: Corpus-derived possible sub-entries of the entry *matlhô*

<table>
<thead>
<tr>
<th>Collocates</th>
<th>Literal translation</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>bula matlhô</td>
<td>open the eyes</td>
<td>educate; make aware; open the eyes</td>
</tr>
<tr>
<td>diga matlhô</td>
<td>drop the eyes</td>
<td>look down</td>
</tr>
<tr>
<td>digalase tsa matlhô</td>
<td>glasses of the eyes</td>
<td>spectacles; sunglasses</td>
</tr>
</tbody>
</table>
4. Treatment of multi-word units in Setswana dictionaries

When idiomatic collocates are treated as sub-entries in dictionaries, it is important that the type of dictionary should be kept in mind. Normally general dictionaries, which have a more inclusive nature can accommodate more sub-entries than standard or school dictionaries, which, because of their smaller nature, have to exclude many sub-entries. In the case of very economical, restrictive and selective dictionaries, all sub-entries will have to be omitted. When, in the following discussion, we therefore indicate how the sub-entries in some Setswana dictionaries may be increased, it does not necessarily mean that all these sub-entries should be included. It merely shows what are available. When a choice has to be made, which sub-entries have to be included in accordance with a specific type of dictionary, corpus evidence will be helpful to indicate which idiomatic collocates are the most commonly and generally used.

Setswana dictionaries have attempted to include sub-entries based on the idiomaticity of collocates. However, some of these have been few because of a lack of sufficient corpus evidence. Above we have shown that 84 sub-entries for pelo could be extracted from a corpus. When the entry pelo in Matumo (1993: 306-3007) is referred to, we can see that he lists only 20 sub-entries. Presented below are examples of how the entry molomo has been treated in Setswana dictionaries to illustrate the nature and extent of this.

Brown (1925: 210) identifies only two sub-entries kgwedi ya molomo and go cwa molomo:

- **Molomo**, n., pl. melomo, A mouth (outside); a beak of a bird; a foreskin. *Kgwedi ea melomo*, the first month of the Sechuana year; the month of eating first-fruits. *Go cwa molomo*, to open the mouth, in speaking.

Kgasá (1976: 71) does not list any sub-entry for *molomo*. It may be that Kgasá’s dictionary, which was aimed at primary schools was simplified for this reason; he might have seen no need to complicate entries with sub-entries:

- **molomo(me)** kgôrô e dijô di yang mo ’ganoŋ ka yônê.
Kgasa and Tsonope (1998: 171) list only a single sub-entry: *molomo o tlola noka e tletse* (a claim is easy to make):

*molomo* TTT In.,/β. me- phatlha e e tswalwang ke dipounama tse pedi e go tsengwngang dijô ka yôné go ya ko mpeng le go bua. • *molomo o tlola noka e tletse* = motho o kgôna go bua dilô tse di ntsi tse a ka di dirang mme ntswa a se ka ke a kgôna

While Snyman et al. (1990) do not enter *molomo* in their dictionary at all, Matumo (1993: 260) lists only two sub-entries, *go tswa molomo* and *sejô sennye ga se fete molomo*:

*molomo*, N. CL, 3 mo-. SING. OF *melomo*, a mouth; lip; a beak of a bird; an opening, as a tube, piping or tunnel; a foreskin. ID. EXPR., *go tswa molomo*, to open the mouth in speaking. PROV., *sejô sennye ga se fete molomo.*

All the above dictionary treatments of the entry *molomo* are deficient and will benefit greatly from the use of corpus evidence. For instance, the definition from Matumo (1993) may be revised in the following way, • being used to mark a sub-entry. This shows how the study of collocations can enrich dictionary entries.

*molomo*, n. 1. mouth 2. a lip 3. a beak 4. an object opening, as that of a bottle • bolwetsi jwa tlhako le molomo: foot and mouth disease • itoma molomo wa tlase: be determined • itshwara molomo: be shocked • ntsha ka molomo: speak; express an opinion; express a view • pula molomo: money paid before someone speaks during lobola negotiations • pipa molomo: a bribe • rwala molomo: be angry and tight-lipped • roka molomo: remain quiet • tswa molomo: speak; say something; contribute; express an opinion • tlhoka molomo: have nothing to say; be dumbstruck; be rendered speechless • molomo o tlola noka e tletse: it is easy for someone to claim that they can achieve what they cannot do

We conclude this section by illustrating how dictionary entries for *mpa*, *lonaô* and *matlhô* could be enriched by means of information in Tables 2–5 derived from a corpus. The proposed entries in each case are compared with entries from Matumo (1993).

Matumo (1993: 276):

*mpa* N. CL, 9θ-, SING. OF *dimpa*, a belly; a stomach. ID. EXPR. *mpa ya lentswê*, the middle of a hill; *mpa ya lonao*, the sole of a foot. PROV., *seboba re bata sa mokwatla sa mpa re a mpampetsa.*

Matumo’s entry of *mpa* with only three sub-entries may be improved in the following manner with the addition of nine sub-entries:

*mpa* n. a belly; a stomach • *bana ba mpa*: relatives • *bipa mpa ka mabele*: withhold bad information to protect a relative or friend • *gare ga mpa ya bosigo*:
in the middle of the night • gare ga mpa ya lefatshenate: flat on the stomach • mpa e tuka molelo: with a full stomach • go ja ka mpa tsoopedi: eat until the stomach is full • ntsha (senya) mpa: commit abortion • imelwa ke mpa: have a full stomach

Matumo (1993: 212):

**lonäô** N. CL. 11 lo-, SING OF dinaô, a foot. ID EXPR, go baba lonaô.

Matumo’s entry of lonäô with a single sub-entry may be improved with the addition of eleven sub-entries as follows:

lonäô n. a foot • apaya ka lonäô: avoid cooking and eat at other people’s homes instead • goga dinaô: move slowly • fodisa dinaô: take a rest • motsamaya ka dinaô: a pedestrian • ngotla dinaô: walk slower • tihatlosa dinaô: walk faster • baya lonäô: set foot in a place • tsholetsa dinaô: walk faster • ya dinaô: football • tsosa dinaô: walk faster

Matumo (1993: 232):

**matlhô** N. CL. 6 ma-, PL OF CL. leitlhô; matlhô is still used in a few areas, eyes.

Matumo’s entry of matlhô which lacks any sub-entry, may be improved by the addition of nine sub-entries:

matlhô n. eyes • bula matlhô: educate, make aware, enlighten • diga matlhô: look down • digalase tsa matlhô: spectacles, sunglasses • latlhêla matlhô: look briefly • matlhô a phagê a lebane: face to face • kala matlhô: confuse • tlodisa matlhô: overlook someone or something • kgaragaratsha matlhô: look from one place to another • tihatlosa matlhô: look up

The updating does not only apply to the bilingual dictionaries. Monolingual Setswana dictionaries could be enhanced in a similar manner, as the following example of tsaya (take) from Kgasa and Tsonope (1998: 303):

tsaya GT|tseile tpt. -ile. 1. tlosa sengwe fa se ntseng se le teng ka go se tsenywa mo diatleng tsa ga go se le teng 2. inêêla ka molaô ga monna go tshela le mosadi; nyala • go tsaya seditse = go dumêlwa ke la bongwe mo go se o se buileng

Kgasa and Tsonope’s treatment of tsaya with a single sub-entry may be revised in the following comprehensive manner with the aid of concordance lines to add 28 sub-entries:

tsaya ld. 1. amogela mo diatleng 2. sutisa sengwe fa se neng se le teng 3. tsamaya ka tselana; ya ntlheng ngwwe 4. nyala 5. nna le sengwe; tshola • tsaya botshe-lo: bolaya • tsaya dinopolo: utswa diphiri • tsaya ditaelo: sala morago melawana • tsaya ka motlhala: sala morago • tsaya dipilisi: metsa dipilisi • tsaya karolo: nna le seabe • tsaya ka letsogo la molema: sotla; nyatsa; kgetholola • tsaya kgakololo: amogela kgakololo • tsaya kgato: dira sengwe • tsaya lobaka: go diragala mo nakong e telele • tsaya mongwe/sengwe
Conclusion

In this article, we have attempted to illustrate what could be achieved by a simple study of concordance lines to extract MWEs for the significant improvement of dictionary entries. Considering only single words as candidates for dictionary entry impoverishes a dictionary and betrays a rudimentary understanding of what constitutes a word in language. If Jackendoff’s estimate that the number of MWEs in a speaker’s lexicon is of the same order of magnitude as the number of single words is accurate, then MWEs in African languages deserve intensive study, which they have hitherto not received. To generate concordance lines is inexpensive, and free concordance programs are available online to aid researchers explore the complexity of texts. Dictionaries of African languages would therefore benefit greatly from populating sub-entries with MWEs harvested from concordance lines.

References