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

Opsomming: Die aanvulling van subinskrywings in woordeboeke met meerwoordige eenhede uit konkordansiereëls. Leksikografie is hoofsaaklik gemoeid met die weergawe van woorde en hul betekenisse in woordeboeke. Met *woorde* verwys die meeste woordeboekgebruikers en leksikograwe na 'n kombinasie van lettertekens afgegrens deur spasies aan beide kante. Hierdie artikel bespreek die swakheid van hierdie benadering by die keuse van woordeboekinskrywings. Deur 'n ondersoek van konkordansiereëls gegenereer uit 'n multimiljoen-Setswanakorpus, word daar geredeneer en verduidelik hoe meerwoordige eenhede (MWE's), ook bekend as meerwoordige uitdrukkings (MWU's), uit konkordansiereëls onttrek kan word om woordeboekinskrywings aan te vul. Daar word aangetoon hoe sowel eentalige as meertalige Setswanawoordeboeke uitgebrei kan word deur die toevoeging van MWU's as subinskrywings.

Sleutelwoorde: SETSWANA, LEKSIKOGRAFIE, MEERWOORDIGE EENHEID, KORPUS, KONKORDANSIE, MEERWOORDIGE UITDRUKKING, KOLLOKASIE, WOORD, SUBINSKRYWINGS, WOORDEBOEK

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/dinaô* (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.

houseboat	*house-boat*	*house boat*
tradeoff	*trade-off*	*trade off*
khuduthamaga	*khudu-thamaga*	*khudu thamaga*
pelotshetlha	*pelo-tshetlha*	*pelo tshetlha*
rampatshetlha	*rampa-tshetlha*	*rampa tshetlha*
motshwaradiphala	*motshwara-diphala*	*motshwara diphala*
kgakalakgakala	*kgakala-kgakala*	*kgakala kgakala*

The examples *houseboat* and *kgakalakgakala* 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 (Otlogetswe 2007: 56). We illustrate this below with the example of the word *pelo* (heart).

Figure 1: Concordance results for the word *pelo*

```
o ka kgopolo ya gore Morwadi o tlaa wela pelo. A mo gaupanya. ka legofl fa ga re
 ngwatiaka, O se tshoge bono wa ka, Wela pelo ga o seitaodi re Use rotlhe, O sek
gang, o tla e rola morago o sena go wela pelo. '/'r~ a emelela, o b~ua a le esi)
a ka seatla. "0 sale sentle, moratiwa wa pelo ya me. Ga ke itse gore ke~ tla go
se o lela jalo? Ke a go rata moratiwa wa pelo ya me." Fa a sa ntse a e phimola,
ne a ithuta ona. "Gomotsega, moratiwa wa pelo, ya me." Mosele a didimala, mme go
 rebe la Mokwena, a buledisa moratiwa wa pelo ya gagwe. O ne a tsamaya ka bonya,
a. : Ke go reile ka re o seka wa utlwisa pelo botlhoko tlhe rra! 0 a itse gore b
g mo matlhong a gago. O se ka wa utlwisa pelo ya gago botlhoko ka nna, ke swetse
 gatlhisa thata., Ba,utlwil ba mo tswela pelo tota.. ,I, Mmaago Molebi a mo roma
be, a di phailela kwa, a re ba mo tswela pelo. Le ene Pule tota tsala ya gagwe y
hegelwa ke moratiwa, e seng go mo tswela pelo kgotsa go mo tlhoafalela. Seno se
sadi yo montle, mme phokojwe a mo tswela pelo. Phokojwe a leka maano a le mantis
wana wa kgosi, noga ya bona peba ya tswa pelo ka ene e ntse lobaka lo lo leele e
wana wa kgosi, noga ya bona peba ya tswa pelo ka ene e ntse lobaka lo lo leele e
ro ya gagwe. NtsVwa Mosetsanyana ya tswa pelo, ya metsa mathe a keletso. Saitsan
swe ke marago a tanka e nngwe. A tshwara pelo monnamogolo wa batho mme a ntse a
gotla-tshekelo. Mmaagwe Sereri a tshwara pelo ya gagwe, a bua ka tidimalo le bad
e motlha mongwe lokwalo lo tla tshikinya pelo ya ga Mmatheebe. A kwalela kgarebe
ologeletsweng morwadi wa we la tshikinya pelo ya gagwe gore a bale ka mabogo a a
 nala kwa Naledi a tla a ratile go tlola pelo. Mogoma e re ntlhomane a feta fa M
o ngwega Uncle Boot 0 ne a rata go tlola pelo. Mmaagwe ene, a ipega fa a ne a le
r ." Bikibiki a re,"Ngwana yo o tlhomola pelo. Lefatshe le mo itaya ka ntlha ya
 re ruri rre Rapitso wa batho o tlhomola pelo ka tshenyo e e leng ka mo supamake
 le bobotlana. Motho wa tsona o tlhomola pelo. Keikepetse o ntse jalo mo teramen
ka go tena Ontefile. Ketshedile a tlhapa pelo ka o ne a sa ntse a senka leano la
a. Bofelo a mo tsepela leitlho. A tlhapa pelo gonne fa go rata Bofelo, tsotlhe d
ng. jaanong Morwadi a nametsega a tlhapa pelo. A tsaya tlhogo a e latsa mo sehub
otlhapelo a ngwana wa mpa. "Nnake, tiisa pelo. O sa ntse o na le mogomotsi e bon
 bosigo. Tlogela go nna legatlapa. Tiisa pelo. Ga o sa tlhola o le mosimane. Ka
a gagamatsa thamo ya gagwe, a thatafatsa pelo ya gagwe, mo a bileng a se ka a bo
2Mme le ka sebaka seo Farao a thatafatsa pelo ya gagwe gape, a se ka a naya mora
wa gee." "Ke mang?" A botsa a swegaswega pelo. Ngwananyana a bolela fa ene a bon
a ke matlhagatlhaga, e bile a swegaswega pelo. O ne a batla go balela kwa pele.
```

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
betwa ke pelo
bofa pelo
bolawa ke pelo
bolwetse 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 sena letsapa le fisang pelo
gonolwa ke pelo
isa pelo mafisa
itaya pelo
itse pelo
kgwaralatsa pelo
lala ka pelo e rotha madi
mabetwa-e-pelo
masetla pelo
matlhomola pelo
matlhotlha-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 botlhoko
pelo e e letlapa
pelo e ja serati
pelo e khibidu
pelo e rotha madi
pelo e rutha
pelo e setlhogo
pelo e thata
pelo khutshwane
pelo namagadi
pelo ntsho
pelo pedi
pelo pholwana e a golegwa
pelo potsane e a golegwa
pelo tshweu
pelo yotlhe
pelo-e-thata
pelo-kgale
pelo-telele
pelo-tlhomogi
pelo-tshetlha
phatlola pelo
ritibatsa pelo
sephiri sa pelo
sera pelo
sethunya sa pelo
sisa pelo
sulafatsa pelo
swa pelo
swegaswega pelo
thiba maroba a pelo
thuba pelo
tlala pelo
tlalelana pelo
tlhomola pelo
tlola pelo
tshwara ka pelo
tshwara pelo
tswa pelo
tswela pelo
uba pelo
wa pelo
wela pelo

In Table 1 below, only 10 of these are explained.

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

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

In Setswana, the phenomenon of idiomaticity when considering a word and its collocates is not unique to *pelo*. Words like *molomo* (mouth), *mpa* (stomach), *nkô* (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), *lonaô* (foot) and *matlhô* (eyes) respectively which have been extracted through studying concordance lines.

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

Collocates	Literal translation	Meaning
bolwetsi jwa tlhako le molomo	disease of hoof and mouth	foot and mouth disease
itoma molomo wa tlase	bite the lower mouth	be determined
itshwara molomo	hold/touch the mouth	be shocked
ntsha ka molomo	release with the mouth	speak
pula molomo	that which opens the mouth	money paid before someone speaks in lobola negotiations
pipa-molomo	that which covers the mouth	a bribe
rwala molomo	carry the mouth on your head	be angry and tight-lipped
roka molomo	sew the mouth	remain quiet
tswa molomo	grow a mouth	speak
tlhoka molomo	lack a mouth	have nothing to say

Table 3: Corpus-derived possible sub-entries for the entry *mpa*

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

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

Collocates	Literal translation	Meaning
apaya ka lonaô	cook with a foot	avoid cooking and eat at other people's homes instead
goga dinaô	drag the feet	move slowly
fodisa dinaô	cool the feet	have a rest
motsamaya ka dinaô	one who walks with the feet	a pedestrian
ngotla dinaô	reduce the feet	walk slower
tlhatlosa dinaô	raise the feet	walk faster
baya lonaô	put a foot	be in a place
tsholetsa dinaô	lift the feet	walk faster
kgwele ya dinaô	ball of the feet	football
tsosa dinaô	wake up the feet	walk faster; hurry up
tiisa dinaô	strengthen the feet	walk faster

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

Collocates	Literal translation	Meaning
bula matlhô	open the eyes	educate; make aware; open the eyes
diga matlhô	drop the eyes	look down
digalase tsa matlhô	glasses of the eyes	spectacles; sunglasses

latlhela matlhô	throw the eyes	look briefly
matlhô a phage a lebane	the eyes of a wild cat face to face	face to face
kala matlhô	measure the eyes	confuse
tlodisa matlhô	make the eyes jump	overlook someone or something
kgarakgaratsha matlhô	make the eyes move from one place to another	look from one place to another
tlhatlosa matlhô	raise the eyes	look up
tlhaetsa matlhô	shorten the eyes from	despise someone

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 molomo*, the first month of the Sechuana year; the month of eating first-fruits. *Go cwa molomo*, to open the mouth, in speaking.

Kgasa (1976: 71) does not list any sub-entry for *molomo*. It may be that Kgasa'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 'ganong 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):

> **mo•lomo** TTT *ln./3.* me-. phatlha e e tswalwang ke dipounama tse pedi e go tsenngwang dijô ka yônê 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, 9Ø-, 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 lefatshe/naga:** in the middle of nowhere ◘ **mpa ya sebete:** 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):

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

Matumo's entry of *lonaô* with a single sub-entry may be improved with the addition of eleven sub-entries as follows:

lonaô *n.* a foot ◘ **apaya ka lonaô:** 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 ◘ **tlhatlosa dinaô:** walk faster ◘ **baya lonaô:** set foot in a place ◘ **tsholetsa dinaô:** walk faster ◘ **kgwele ya dinaô:** football ◘ **tsosa dinaô:** walk faster ◘ **tiisa dinaô:** walk faster

Matumo (1993: 232):

matlhô N. CL. 6 *ma-*, PL OF CL. *leitlhô*; *maitlhô* 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 ◘ **kgarakgaratsha matlhô:** look from one place to another ◘ **tlhatlosa 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 tsenya mo diatleng tsa gago 2. inêêla ka molaô ga monna go tshela le mosadi; nyala ♠ *go tsaya seditse = go dumêlwa ke ba bangwe 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 nngwe **4.** nyala **5.** nna le sengwe; tshola ◘ **tsaya botshelo:** 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**

motlhofo: nyatsa ◘ **tsaya puso:** simolola go etelela mmuso ◘ **tsaya phekelo e sele:** go senyegela pele ◘ **tsaya tshwetso:** dira mogopolo ◘ **tsaya nako:** iketle ◘ **tsaya motlhala:** kopa sengwe se se siameng ◘ **tsaya mosadi:** nyala ◘ **tsaya mogote:** tlhola selekanyo sa mogote mo mongweng ◘ **tsaya matsapa:** dira sengwe mo nakong e telele ◘ **tsaya tsia:** tlotla; tlhokomela ◘ **tsaya malatsi:** ikhutsa; nna o ye tirong ◘ **tsaya malebela:** kopisa sengwe se se ntle ◘ **tsaya maikarabelo:** nna wena o tshwereng sengwe ◘ **tsaya loeto:** eta ◘ **tsaya maemo:** simolola maemo ◘ **tsaya setshwantsho:** dirisa khamera go tshwantsha ◘ **tsaya sekgele:** fenya ◘ **tsaya sebaka:** sengwe sa nako e telele

5. 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

Aitchison, J. 1992. *Teach Yourself Linguistics.* London: Hodder & Stoughton.

Bannard, C. 2007. A Measure of Syntactic Flexibility for Automatically Identifying Multi-word Expressions in Corpora. *Proceedings of the ACL Workshop on a Broader Perspective on Multi-word Expressions, Prague, Czech Republic, June 2007*: 1-8.

Biber, D., S. Conrad and R. Reppen. 1998. *Corpus Linguistics: Investigating Language Structure and Usage*. Cambridge: Cambridge University Press.

Brown, T.J. 1925. *English Dictionary*. Johannesburg: Pula Press.

Brunner, A. and K. Steyner. 2008. *Corpus-Driven Study of Multi-Word Expressions Based on Collocations from a Very Large Corpus*. Paper presented at the Fourth Inter-Varietal Applied Corpus Studies (IVACS) Conference, University of Limerick, Ireland, 13–14 June 2008.

Dash, N.S. and B.B. Chaudhuri. 2000. The Process of Designing a Multidisciplinary Monolingual Sample Corpus. *International Journal of Corpus Linguistics* 5(2): 179-197.

Fazly, A. and S. Stevenson. 2007. Distinguishing Subtypes of Multiword Expressions Using Linguistically-motivated Statistical Measures. *Proceedings of the ACL Workshop on a Broader Perspective on Multi-word Expressions, Prague, Czech Republic, June 2007*: 9-16.

Finch, G. 2000. *Linguistic Terms and Concepts*. Basingstoke: Macmillan Press.

Jackendoff, R. 1997. *The Architecture of the Language Faculty*. Cambridge, MA: MIT Press.

Kgasa, M.L.A. 1976. *Thanodi ya Setswana ya Dikole*. Cape Town: Longman.

Kgasa, M.L.A. and J. Tsonope. 1998. *Thanodi ya Setswana*. Gaborone: Longman.

Leech, G., M. Deuchar and R. Hoogenraad. 1982. *English Grammar for Today: A New Introduction*. Basingstoke: Macmillan Press.

Matumo, Z.I. 1993. *Setswana–English–Setswana Dictionary*. Gaborone: Macmillan.

McArthur, T. 1998. *Living Words: Language, Lexicography and the Knowledge Revolution*. Exeter: University of Exeter.

Moon, R. 1998. *Fixed Expressions and Idioms in English: A Corpus-based Approach*. Oxford: Oxford University Press.

Oflazer, K. and Ö. Çetinoğlu. 2004. Integrating Morphology with Multi-word Expression Processing in Turkish. *Second ACL Workshop on Multi-word Expressions: Integrating Processing, Barcelona, Spain, July 2004*: 64-71.

Otlogetswe, T.J. 2007. *Corpus Design for Setswana Lexicography*. Unpublished Ph.D. Thesis. Pretoria: University of Pretoria.

Pearsall, J. 1998. *The New Oxford Dictionary of English*. Oxford: Oxford University Press.

Renouf, A. 1987. Corpus Development. Sinclair, J.M. (Ed.). 1987. *Looking Up. An Account of the COBUILD Project in Lexical Computing and the Development of the Collins COBUILD English Language Dictionary*. London/Glasgow: COBUILD.

Sag, I.A., T. Baldwin, F. Bond, A. Copestake and D. Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002), Mexico City, Mexico, March 2002*: 1–15.

Schone, P. and D. Jurafsky. 2001. Is Knowledge-free Induction of Multi-word Unit Dictionary Headwords a Solved Problem? *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing, Pittsburgh, PA*: 100-108.

Scott, M. 2004–2006. *Oxford WordSmith Tools Version 4*. Oxford: Oxford University Press.

Sharoff, S. 2004. What is at Stake: A Case Study of Russian Expressions Starting with a Preposition. *Second ACL Workshop on Multi-word Expressions: Integrating Processing, Barcelona, Spain, July 2004*: 17-23.

Snyman, J.W., J.S. Shole and J.C. le Roux. 1990. *Dikisinare ya Setswana–English–Afrikaans Dictionary/Woordeboek.* Pretoria: Via Afrika.

Villavicencio, A., A. Copestake, B. Waldron and F. Lambeau. 2004. Lexical Encoding of MWEs. *Second ACL Workshop on Multi-word Expressions: Integrating Processing, Barcelona, Spain, July 2004*: 80-87.