TAKING DICTIONARIES FOR BANTU LANGUAGES INTO THE NEW MILLENNIUM – WITH SPECIAL REFERENCE TO KISWAHILI, SEPEDI AND ISIZULU

D.J. Prinsloo, University of Pretoria

and

Gilles-Maurice de Schryver, University of Ghent

1.0 Introduction

The modern lexicographer is constantly looking for ways in which the dictionary can be improved to increase the success of information retrieval by the target users, giving especially the encoding user maximum guidance within the physical limitations of a paper dictionary. For the compiler of dictionaries for the Bantu languages the challenge is even greater since he/she is the mediator between a complicated grammatical system on the one hand and the often inexperienced dictionary user on the other hand.

For ages lexicographers battled to increase the quality of dictionaries. The corpus, however, has suddenly opened up new horizons for dictionary makers just as the word processor superseded the typewriter in word processing. In using a corpus the lexicographer can substantially enhance the quality of dictionary text in many ways, of which some will be briefly outlined below.

One could say that the basic aim of the lexicographer is to guide the user in respect of the properties/features/characteristics/use/meaning of the lemma, i.e. to *know* the word. Laufer formulates this basic aim as follows:

> *"knowing a word would ideally imply familiarity with all its properties […] When a person 'knows' a word, he/she knows the word's pronunciation, its spelling, its morphological components, if any, the words that are*

Makala ya kongamano la kimataifa Kiswahili 2000. Proceedings: 188-215

> *morphologically related to it, the word's syntactic behaviour in a sentence, the full range of the word's meaning, the appropriate situations for using the word, its collocational restrictions, its distribution and the relation between the word and other words within a lexical set" (Laufer 1992: 71)*

A large corpus which is as well-balanced and as representative as possible is the first requirement for corpus-based dictionaries. However, a corpus without advanced corpus query tools, is of no use. Corpus tools must be able to provide at least two basic outputs namely word-frequency counts and concordance lines, as well as the capability of analysing problematic contexts.

2.0 Corpora and the compilation of the lemma-sign list

The first major problem with which a lexicographer is confronted on the macrostructural level is well echoed in the literature:

> *"One of the basic problems of lexicography is to decide what to put in the dictionary and what to exclude" (Tomaszczyk 1983: 51)*

> *"The decision what to include in the dictionary still has to be made by the lexicographer himself, however, and this depends in turn upon the nature and size of the dictionary and its intended users. In this respect lemmatised frequency-lists can be a further help, […] we have reached a stage where co-operation between man and machine is useful and perhaps indispensable in making better dictionaries" (Martin et al. 1983: 81-2, 87)*

Formulated differently, in order to decide what to put in and what to exclude from a *useful* dictionary, lemmatised frequency lists are advanced as a guidance. Thus, on the macrostructural level the first useful output of a corpus are word-frequency counts. Compare, for example (1), which lists the hundred most frequently used words in Kiswahili (G42)[1] in a corpus consisting of 507,370 running words taken from newspaper and magazine texts. The latter is but a sub-corpus of the 1.3-million-large *Kiswahili Internet Corpus* (KIC) currently under construction, cf. also De Schryver & Prinsloo (2001).

(1) Kiswahili (Counts in the Newspaper & Magazine Sub-corpus of KIC)

N	Item	Count	%
1	na	24,309	4.79
2	ya	20,614	4.06
3	wa	15,612	3.08
4	kwa	10,379	2.05
5	kuwa	6,897	1.36
6	katika	5,350	1.05
7	ni	5,118	1.01
8	za	4,364	0.86
9	la	3,749	0.74
10	hiyo	3,748	0.74
11	alisema	3,345	0.66
12	huyo	3,280	0.65
13	bw	3,008	0.59
14	cha	2,996	0.59
15	kwamba	2,389	0.47
16	kama	2,329	0.46
17	yake	2,286	0.45
18	baada	1,899	0.37
19	hilo	1,860	0.37
20	huo	1,786	0.35
21	hao	1,682	0.33
22	hata	1,626	0.32
23	watu	1,615	0.32
24	wake	1,612	0.32
25	hivyo	1,585	0.31
26	mwaka	1,582	0.31
27	lakini	1,568	0.31
28	wakati	1,519	0.30
29	ambaye	1,333	0.26
30	serikali	1,318	0.26
31	kwenye	1,293	0.25
32	ili	1,260	0.25
33	vya	1,189	0.23
34	siku	1,186	0.23
35	hayo	1,169	0.23
36	sasa	1,134	0.22
37	mmoja	1,118	0.22
38	alikuwa	1,091	0.22
39	habari	1,089	0.21
40	pia	1,086	0.21
41	hicho	1,074	0.21
42	hizo	1,065	0.21
43	polisi	1,037	0.20
44	jana	1,033	0.20
45	pamoja	1,026	0.20
46	mkuu	1,012	0.20
47	moja	990	0.20
48	kutoka	961	0.19
49	yao	923	0.18
50	huu	878	0.17
51	mtu	873	0.17
52	hadi	866	0.17
53	huko	865	0.17
54	chama	859	0.17
55	kesi	842	0.17
56	ambao	834	0.16
57	kazi	834	0.16
58	mahakama	820	0.16
59	nchini	812	0.16
60	tu	810	0.16
61	mara	799	0.16
62	mama	795	0.16
63	kutokana	786	0.15
64	hali	778	0.15
65	kila	773	0.15
66	mtoto	767	0.15
67	au	763	0.15
68	ambayo	730	0.14
69	bila	725	0.14
70	watoto	723	0.14
71	wao	722	0.14
72	baadhi	712	0.14
73	sana	707	0.14
74	zaidi	683	0.13
75	nchi	678	0.13
76	rais	669	0.13
77	taifa	666	0.13
78	hapa	660	0.13
79	dar	642	0.13
80	jijini	641	0.13
81	hivi	637	0.13
82	wengine	628	0.12
83	novemba	619	0.12
84	ambapo	618	0.12
85	tena	616	0.12
86	huku	610	0.12
87	tanzania	610	0.12
88	kufanya	609	0.12
89	sababu	609	0.12
90	wananchi	585	0.12
91	ndani	584	0.12
92	hakimu	581	0.11
93	yeye	576	0.11
94	sh	574	0.11
95	taarifa	573	0.11
96	akasema	570	0.11
97	hapo	567	0.11
98	fedha	566	0.11
99	muda	561	0.11
100	mambo	558	0.11

It is important to obtain a total count of a word, that is its overall occurrence in all the sources (or sub-corpora) taken together. Apart from seriously considering the *total count*, the lexicographer should also take the *spreading* of a specific word across the different sources into consideration. Knowles states that a word must occur evenly across a broad spectrum of miscellaneous data corpora:

> *"a word must occur evenly in a large number of the stratified sub-samples rather than excessively often in a small number of them, given that these two very different cases could show identical 'total-corpus' frequencies" (Knowles 1983: 188)*

Hence, the dictionary compiler should look out for words that may have a high total count but occur only in a single or in a limited number of sources. In such cases the lexicographer has to decide on inclusion or omission depending on the target user group of the dictionary. It is also important to look at low frequency counts or even zero occurrences, which have to be considered for omission since they take up space in the dictionary which can be better utilised for more frequent lemmata. Compare the following example from Setswana (S31) in (2).

(2) Setswana (Total count vs. spreading across different sources)

Word	Total	Source 1	Source 2	Source 3	Source 4	Source 5
letlapa	**31**		16	2	7	6
letsatsi	**168**	53	47	29		39
letshogo	**16**	3	3	2	4	4
lona	**158**	29	15	20	28	66
maabane	**28**	5	1	1	3	18
mabedi	**23**	2	2	3	2	14
mabogo	**19**	2	5	6	4	2
madi	**247**	7	8	111	65	54
mafoko	**125**	20	5	2	24	74
mafura	**25**		15	3	3	4
maikutlo	**20**	3	3	2	7	5
maina	**238**	13	1	3	9	212

mainakgopolo	**10**					10
maineng	**9**					9
malatsi	**31**	1	12	9	2	7
mane	**15**	1	1	4	2	7
mang	**90**	17	11	16	4	42
mangwe	**69**	7	6	10	3	43
marapo	**15**	6	3	2	1	3
mathata	**12**	1	3	1	1	6
matlho	**92**	17	8	31	29	7
matlhong	**11**	1	1	3	5	1

From (2) it is clear that a word such as **matlho** 'eyes' not only has a high total count, namely 92, but also that these 92 occurrences have a very good spreading across the different sources, namely 17, 8, 31, 29 and 7 respectively. Conversely, the word **mainakgopolo** 'abstract nouns', although having a relatively high total count, occurs only in one of the sources. In this case the lexicographer, depending on the target user group of his/her dictionary, has to decide whether it should be included in or omitted from the dictionary.

3.0 Corpora and the battle against inconsistencies

One very unfortunate tradition in the compilation of dictionaries for many a Bantu language is to enter words as they cross the compiler's way. This approach results in serious inconsistencies. Two types of inconsistencies will be dealt with, namely (a) inconsistencies regarding *inclusion/omission of lemmata* and (b) inconsistencies regarding the *lemmatisation of derivations*, especially in the case of reflexives.

Firstly, in (3) a random section of the lemma-sign list of the English side of a Setswana – English – Setswana dictionary is compared to the respective English sections of another Setswana – English – Setswana dictionary, a Sepedi (S32) – English – Sepedi dictionary, and an Afrikaans – English – Afrikaans dictionary. The first dictionary is a small desktop one, the other three are pocket editions.

(3) English / Setswana / Sepedi / Afrikaans (Inclusion/omission of lemmata)

English – Setswana *Snyman 1990*	**English – Setswana** *Brown 1925*	**English – Sepedi** *Kriel 1988*[3]	**English – Afrikaans** *Kromhout 1997*[xiii]
dab	—	dab	dab
—	Dabble	dabble	dabble
—	Dad	dad	dad/daddy
—	—	—	daffodil
—	Daft	—	daft
dagga	Dagga	dagga	—
dagga-pipe	Dagga-pipe	—	—
—	Dagger	dagger	dagger
—	—	dahlia	dahlia
—	Daily	daily	daily
—	Dainty	dainty	dainty
—	Dainties	—	—
—	Dairy	dairy	dairy
—	Dais	—	dais
—	Daisy	daisy	daisy
—	Dale	dale	dale
—	Dally	—	—
dam	Dam	dam	dam

The editor of the dictionary in column one honestly admits in the preface:

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

The absence of commonly used words in column one such as **dad**, **daily** and **dairy** surely proves his point.[2] If his dictionary team had utilised frequency counts – even if these had been based on a relatively small-size corpus – this would not have happened.

As a second example of inconsistencies, one can compare the tables shown in (4) and (5). In (4) the ten *most frequently* used reflexives in Sepedi are listed, while (5) shows *all* the reflexives that were lemmatised in the *Klein Noord-Sotho woordeboek* (Ziervogel & Mokgokong 1988[4]).

(4) Sepedi (Lemmatisation of reflexives – part 1)

Sepedi Corpus, Phase 2 (SC2)		**Count in SC2**
Reflexive	*Translation*	*Total*
ikemišeditše(go)	'intended'	42
ikhwetša(go)	'find oneself'	25
ikwa(go)	'feel/hear oneself'	33
ipha(go)	'give to oneself'	32
iphile	'gave to oneself'	24
ipona(go)	'see oneself'	41
ipotšiša	'ask oneself'	34
ithuta(go)	'teach oneself'	69
itokišetša	'prepare oneself for'	41
itshola	'blame oneself'	33

(5) Sepedi (Lemmatisation of reflexives – part 2)

Klein Noord-Sotho woordeboek *Ziervogel & Mokgokong 1988*[4]		**Count in SC2**
Reflexive	*Translation*	*Total*
ikgata	'tread on oneself'	2
ikola	(no clear translation)	0
ikwela	'fall (for) oneself'	1
ipea	'place oneself'	11
ithuta	'teach oneself'	69
itiša	'take care of oneself'	2
itshelela	'seek food for oneself'	1
itshwara	'behave oneself'	19
itsomarela	(no translation)	0
itswalanya	'associate oneself with'	4

The frequency counts in (4) and (5) are derived from the *Sepedi Corpus, Phase 2* (SC2). SC2 was built from fifteen randomly selected Sepedi literary works and magazines, totalling circa 220,000 words. From (5) one sees that the *Klein Noord-Sotho woordeboek* lemmatised *only ten* reflexives. Such an ad hoc decision is totally acceptable if it is done to reflect extremely high usage, say for example the ten reflexives listed in (4). However, with the exception of **ithuta** and **itshwara** the likeliness of these words to be looked up by the target user is *highly questionable*. Compare their occurrences or even total absence in SC2, as shown in (5). One cannot but deplore the fact that precious space has been allocated to reflexives which are unlikely to be looked up by the target users whilst highly used reflexives were omitted.

These two examples amply support the view expressed by Gouws now a decade ago:

> *"lexicographical activities on the various indigenous African languages […] has resulted in a wide range of dictionaries. Unfortunately, the majority of these dictionaries are the products of limited efforts not reflecting a high standard of lexicographical achievement" (Gouws 1990: 55)*

All these inconsistencies can be *avoided* if lexicographers base the lemma-sign lists of their dictionaries on frequency counts derived from corpora.

4.0 Corpora as an aid for conjunctively written languages

Through the use of a corpus and advanced query tools the lexicographer can *bring together all inflections and derivations of a verb which are otherwise scattered all over the dictionary*. Sensible decisions can then much more easily be made in respect of different options for lemmatisation. The isiZulu (S42) verb **ukuhamba** 'to walk; to go' for instance, is used with a single affix or combinations of affixes as shown in (6).

(6) isiZulu (Conjunctive orthography – part 1)

i**hamb**e, uku**hamb**a, kayi**hamb**i, aye**hamb**a, sebe**hamb**a, ngangili**hamb**a, nginga**hamb**a, **hamb**ani, eku**hamb**eni, ube**hamb**ele, ngizi**hamb**ela, owaye**hamb**ele, wam**hamb**isa, ayengasa**hamb**eli, zi**hamb**ayo, ngi**hamb**ile, kaba**hamb**anga

With an aid such as (7) the lexicographer can bring all the inflections and derivations together, study them in context, and decide on a lemmatisation strategy.

(7) isiZulu (Conjunctive orthography – part 2)

njengoba sengishilo ukuze indaba	i	**-hamb-**	e	igijime kesizwe nabo bekhuluma
yikuba ngiyilandelise kahle konke	uku	**-hamb-**	a	kukashaka impela kusuka ekuzalweni
ethunywa ngumbengi wenguga	kayi	**-hamb-**	i	yodwana belu ihamba nomfana
inyama adle nomfana lowo wakhe	aye	**-hamb-**	a	naye ngenxa yomusa lowo wakhe
kwayihlaba indoda ithe nalapho	sebe	**-hamb-**	a	nomfana beqonde enkosini
ubona umusa ongakaya mfana kade	ngangili	**-hamb-**	a	nje izwe kangibonange ngiwubone
kakhulu yasimnye nyezela ithi	nginga	**-hamb-**	a	nawe ngokukhombisa nkosazana
bakubikele usenza ngakhona lokho athi		**-hamb-**	ani	niyozibiza bazibize abafana zize
umzukulu kandaba uzalwa ngunonkwelo	eku	**-hamb-**	eni	kwabo lapha kwamthethwa
ubehlasele ubaba bekungaliwa ndaba	ube	**-hamb-**	ele	khona wafika wabu lawa athule
kababa izikhalela indoda yeka mina	ngizi	**-hamb-**	ela	ngiziqhubela izimbuzi zami kanti
nabathwa ayefike nomlungu lowo	owaye	**-hamb-**	ele	kushaka kwathi ngelinye ilanga
wasezibisini wamxhoxha amehlo	wam	**-hamb-**	isa	ngokhalo olukhulu lukankume

sempi yakhe ngoba lawa maxhegu	ayengasa	**-hamb-**	eli	phezulu njenga mabutho akhe
kazi bonwa muntu zifihliwe nalapho	zi	**-hamb-**	ayo	ziya emfuleni noma ziphume
nje imikhuba yabo bacabanga ukuthi	ngi	**-hamb-**	ile	sizobafica kahle bonke sibabambe
uma wayeke wezwa nje ushaka ukuthi	kaba	**-hamb-**	anga	nempi babeyokufa kabi futhi

5.0 Corpora as the key to writing better dictionary articles

On the microstructural level concordance lines which are derived from the corpus by means of concordancing tools such as those provided by *WordSmith Tools* (cf. URL WordSmith), form the basis for information retrieval for the lexicographer. These lines are indispensable as an aid to sense distinction for the writing of better definitions (monolingual dictionaries) or for the selection of suitable translation equivalents (bilingual dictionaries).

Concordance lines also reflect typical collocations, clusters, idioms, proverbs and examples of usage. Such concordances on data sources from the living language supplement and support the lexicographer's (mother tongue) intuition. It takes him/her to the *heart of actual usage* of words through the display of the word in context, seeing up to 30 contexts at a glance. Compare (8) as an example for a Kiswahili noun, **mwanamume** 'man', and (9) for a verb, **kupika** 'to cook' (including some of its verbal derivations).

(8) Kiswahili (Concordance lines for a noun)

wakiwa wanabishana. Alisema ghafla yule	mwanamume	alikimbia na hapo Bw. Hatibu alitoka nje
dereva aliyehusika. Katika ajali nyingine,	mwanamume	ambaye hajafahamika alikufa papo hapo baada
gari ambalo lilitoweka baada ya tukio. Aidha	mwanamume	ambaye hakuweza kufahamika anayekadiriwa
mahakama ya mwanzo Magomeni ilimwamuru	mwanamume	anayehusika kutoa sh. 15,000 kila mwezi kwa
huyo pia alidai anataka alipwe sh. 7000 ambazo	mwanamume	huyo alichukua siku moja wakati walipokuwa
Aidha inasemekana kuwa baada ya tukio hilo	mwanamume	huyo alikimbilia sehemu ambayo haijulikani,
ya mwanzo Magomeni baada ya kuona	mwanamume	huyo amekaa kimya.
Alidai kuwa mahakama iliwahi kuamuru	mwanamume	huyo atoe Sh. 15,000 kila mwezi, ili kusaidia
10 nyumbani kwa mzazi mwenzake, kwa vile	mwanamume	huyo hatoi pesa za matumizi. Amandu (24)
habari na kwamba kitendo hicho kimemfanya	mwanamume	huyo kuhama nyumbani kwake na kwenda
na mimba hakuwa akipata msaada kutoka kwa	mwanamume	huyo na kwamba mtoto baada ya kuzaliwa
kupitia dirishani na kuwaona marehemu na	mwanamume	mmoja wakiwa uchi wakiwa wanabishana.
na wahudumu wawili, mwanamke na	mwanamume	, ambao hawawezi kutoa huduma yo yote ya

(9) Kiswahili (Concordance lines for a verb + some of its important verbal derivations)

...aliamua kususa kuingia jikoni	kupika	eti kwa sababu nimezidi...
...aliamua kususa kuingia jikoni	kupika	eti kwa sababu nimezidi...
...i wa vyakula. Imeelezwa kuwa	kupika	chakula ndio shughuli p...
...asa linaibuka hilo lingine la	kupika	nyama ya mtu kama kito...
...huku na mtoto akiwa mgongoni.	kupika	na kila kitu. Wakati h...
...a, kufua chupi yake mwenyewe,	kupika	uji wa mtoto hadi kutu...
...kufanya ni kuingia jikoni na	kupika	tu basi. Hata kama mt...
...amoja na kuwaogesha watoto na	kupika	chakula harakaharaka na...
...oyote, kwa hiyo kazi yangu ni	kupika	gongo na hii ndiyo ina...
...ishi peke yangu? Najua siwezi	kupika	lakini nitafanyaje? Ka...
...ma bali ni kazi ya vyuo vikuu	kupika	wasomi ili wawe, watend...
...jana Juma kuwa hatua hiyo, ya	kupika	solo kwa maji hayo, in...
...kitaifa. Nilianza shughuli za	kupika	pale chakula saa 11.30...
...ma kawaida, watu wakipika vya	kupika	, wakiimba nyimbo na ku...
...aa watoto au kufua nguo, wala	kupika	, isipokuwa kumfurahisha...
...arusi hiyo, hasa kwa kusadiai	kupika	. Nikaenda harusini hapo...
...kuni, kwenda mashine kusaga,	kupika	.Yeye hagusi kitu. Kibay...
...hayo yanatumika kwa kunywa na	kupikia	alianza kunipiga ngumi...
...ala hicho ni kituko kama vile	kupikia	chakula chooni. Kwa h...
...ameomba apelekewe vyombo vya	kupikia	ili aweze kujipikia mw...
...ja ya kuwaeleza wale wapendao	kupikia	na nazi jinsi bei ya...
...ya chai, biskuti na mafuta ya	kupikia	kutoka nchi jirani za...
...isha tani 30,000 za mafuta ya	kupikia	kwa mwaka. Baadhi ya...
...ila siku kwa ajili ya kufulia,	kupikia	, kunywa na kuoga. "K...
...afara huo walikuta vyombo vya	kupikia	, sufuria sita, dumu l...
...ampuni ya kusindika mafuta ya	kupikia	, Murzah Oil Mills Lim...
...a majumbani ikiwemo kunywa na	kupikia	. Akasema kwa kutumia...
...i hiyo ya kusindika mafuta ya	kupikia	. Akizungumza wakati w...
...be kama vile gongo inaendelea	kupikwa	na kuuzwa huku ikisaba...
...a soda, au ndizi mbivu ama za	kupikwa	. Amalizapo kula hurudi...

6.0 Corpora as an aid to sense distinction for the writing of better definitions (monolingual dictionaries) or for the selection of suitable translation equivalents (bilingual dictionaries)

The lexicographer is always in doubt whether he/she has covered all the relative senses of a lemma in the definition or in selecting a translation equivalent paradigm. A corpus helps him/her considerably to ascertain whether all relevant senses of a particular lemma have been covered. A simple word such as **run** contains 82 different senses and 350 sub-senses in *The Oxford English Dictionary* (1992^2). See (10) which are but tiny extracts from the entry for **run**.

(10) English (Tiny extracts from the entry for **run** in *The Oxford English Dictionary* (1992^2))

> **I.** Intransitive senses.
>
> The conjugation of the perfect and pluperfect tenses with *be* instead of *have* (as *is run*, *was run*, etc.) is occasionally found in literary use down to the end of the 18th century.
>
> ***** *Of persons and animals, in literal or fig. senses.*
>
> **1. a.** To move the legs quickly (the one foot being lifted before the other is set down) so as to go at a faster pace than walking; to cover the ground, make one's way, rapidly in this manner.
>
> *Run* may be construed with a large number of preps. and advs., as *about*, *after*, *against*, *at*, etc. Some idiomatic uses arising from such phrases are treated under III and IV, and others will be found under some other distinctive word in the phrase (as random *n.* 3).
>
> **b.** In various fig. contexts.
>
> **c.** *Sc.* Contrasted with *ride*. (Cf. go *v.* 1.)
>
> **d.** Used to denote (hurried) travelling or going about, esp. to distant places.
>
> **e.** In proverbs and proverbial phrases.
>
> ***that he who runs may read*** is an alteration of *Habakkuk* ii. 2, 'That he may run that readeth it'.

f. Used allusively, with reference to the legs (in contrast to the wings) of game or poultry.

g. ***to run counter (to)***: see counter *adv.* 1 and 3.

h. *Cricket*. To act as a runner (runner 1 f) *for* (a disabled batsman).

i. *colloq.* To suffer pressingly from diarrhoea. Cf. run *n.*1 14 f.

2. a. To go about freely, without being restrained or checked in any way. Freq. with *about*; also const. *with*, and with adjs. as *wild*.

b. Of animals. Also with *in*.

c. ***to run (a)round***: to associate or consort *with* (someone, esp. of the opposite sex); to court, have an affair with; similarly with *together*. Also in general sense, to go about hurriedly with no fixed goal; to go from one place or person to another. Also *transf.* ***to run (a)round in circles***: see circle *n.* 1 c.

• • •

51. a. To cause (a conveyance, vehicle, vessel, etc.) to ply from place to place, or between two places, or to move in a particular direction, or to a specified destination.

b. To keep (a mechanical contrivance, etc.) moving or working; *spec.* to keep, use, and maintain (a road vehicle).

c. To direct, conduct, carry on (a business, etc.). orig. *U.S.* Also in various extended uses. In *transf.* use *esp.* to look after, manage, or control (someone, *spec.* a spy). Also *refl.* (said of a business or other organization): to function smoothly, to require little administrative interference. ***to run the show***: see show *n.*1 16.

transf.

d. To introduce or push (a person) in society.

e. *U.S.* To support or provide for (a person or family).

f. orig. *U.S.* To publish or print in a newspaper or magazine; *spec.* to publish repeatedly or successively (an advertisement, article, etc., or a series of such items). Also *transf.* of broadcast items.

g. To be suffering from (a fever or high temperature).

h. ***to run a book***

i. To show (a film or television recording); to set (a film camera) in action. Also with *through*.

j. To perform (a test, analysis, experiment, or the like); to subject (something) to, or measure (a property) by means of, an experimental procedure.

k. *Computers*. To perform (a computation), execute (a program or other task), investigate (a problem), etc., on a computer.

52. a. ***run one's face for***, to get (an article) on credit. *U.S.* See also face *n.* 7 b.

b. To put or set up as a candidate. orig. *U.S.*

c. *U.S.* and *Austral.* To tease, nag, or vex.

Characterized by Webster (1879) as 'Colloq. or low'.

d. To prosecute (a person); to bring (one) in *for* damages.

e. *slang.* To report or hand over (someone) to the police, etc.; *spec.* in *Mil.* use, to bring a charge against (someone).

f. To manipulate or falsify, esp. in phr. ***to run the odds***.

g. ***to run one's mouth***, to talk profusely or excessively, to chatter; to complain. Cf. *to shoot* (*off*) *one's mouth* s.v. shoot *v.* 23 g. *U.S.* and *Black slang.*

h. ***to run a game***: to obtain money by deceit or trickery; freq. const. *on*. *U.S. Blacks*.

• • •

81. run up. (See also 11 a.)

* *intr.*

a. (a) To shoot up; to grow rapidly.

(b) To grow up *to*, arrive at, manhood.

(c) To increase, mount up.

† **b.** To land; to arrive on shore. *Obs.*

• • •

(b) To accumulate (a bill, debt, etc.) against oneself or another.

(c) To bid against (a person) at an auction in order to compel him to pay more.

(d) To cause (prices) to rise; to force (a thing) up to a higher price.

h. To trace or follow up in some way.

i. (a) To cause to ascend or rise, to lead, bring, or force up, *to* some point.

(b) To build, erect, set up (a wall, etc.).

(c) To bring (a gun) up to the firing position.

(d) *Austral.* To fetch or bring (a horse) from pasture, etc.

(e) To raise (a flag) to the top of a mast, etc. Also *fig.* (see quot. 1962).

(f) To run (an aircraft engine) quickly while it is out of gear in order to warm it up. Also *intr.*

j. (a) To build or construct rapidly or hurriedly (and unsubstantially).

transf.

(b) To add up (a column of figures, etc.) rapidly.

transf.

(c) To sew quickly (and loosely). Now usu. to make (a garment, etc.) by sewing quickly or simply.

k. To cut up (a tree) as sound wood.

l. *Printing*. (See quot.)

V. 82. In various collocations used attributively or as ns., as

run and fell *Needlework* (see quot. 1968); also *attrib.*;

run-and-read, given to hasty reading (see 1 e);

run-flat *a.*, applied to a kind of tyre on which a vehicle may run after a puncture has occurred;

run-over, due to being run over by a vehicle;

run-sheep(y)-run *N. Amer.* and *Sc.*, a children's hiding game (see quot. 1909);

run-the-hedge, a vagabond;

runther(e)out (only in Sc. form *rin*-), a vagabond, roving person; also *attrib.*;

run-through, applied to a particular stroke in billiards.

The chances of a dictionary compiler gathering all of these senses and sub-senses on intuition is zero. However, by studying corpus lines as in the oversimplified examples (11) for **crawl** and (12) for **sepela**, the various senses and sub-senses can easily be determined.

(11) English (Corpus lines for **crawl***, cited in Atkins et al. (1997: Slide 6abc2))

You have to	**crawl**	along these tunnels.
Exhausted fugitives	**crawl**	from the lake.
Too tired even to read, he	**crawled**	into bed.
A two-mile tail-back	**crawled**	towards the Auditorium.
...as if a gigantic spider had just	**crawled**	across the table.
You've got little brown insects	**crawling**	about all over you.
The whole kitchen was	**crawling**	with ants.
East Germany is	**crawling**	with spies and traitors.
Angela Morgan's car was being	**crawled**	over inch by inch by a forensic team.
Let's stop trying to get women to support us by	**crawling**	to them.
Dark heavy clouds were	**crawling**	across the sky.
There was a little sheep trail	**crawling**	up the hillside.
She was having little chats as she	**crawled**	down the list.
The days before then seemed to	**crawl**	past.

(12) Sepedi (Corpus lines for **sepela**)

...loiwa goba a ba kotsing: "Mma re	**sepela**	bjalo ka wena. Re go...
...bona gore na ditaba di ile tša	**sepela**	**bjang** mabakeng ao a fet...
...tša mohuta wa tsela ye a tlogo e	**sepela**	bophelong bja gagwe ka m...
...O be a šetše a ngenegile, ebile go	**sepela**	ga gagwe e le go goga maoto...
...bilego nke o a kgamega." "Go	**sepela**	gona 0 sepetše bjang?"...
...e, Mašilo, a Iwala, a thoma go	**sepela**	ka dikoloi tša batho ba ban...
...o ka se kgone go tseba gore na e	**sepela**	ka dillo goba lethabo goba...
...go ba tlile ba etšwa Gathe, ba	**sepela**	**ka dinao** ba eta kgoši...
...a a mo rata, ba a mo hlompha ba tla	**sepela**	**ka ditaelo** tša gagwe. Ga...
..., bona ba re ke "setafo". Bao ba go	**sepela**	ka ditimela tša bogego ba n...
...a go goweletša ka mokgwa wo? O	**sepela**	ka ditsebe mošaa?" "Aowa ta...
...etšwa Polokwane ga boMaria. O be a	**sepela**	ka klase ya bobedi, yo- na...
...a ka iri. Ke be ke sa kgahlwe ke go	**sepela**	**ka lebelo** la mma mmati ka...
...enaneo ke be ke bone gore tšohle di	**sepela**	ka lenaneo. Semaka ke...

...a yo mongwe yo a bego a tlwaetše go	**sepela**	**ka maoto** ge a eya ka toro...
...ponela ge re etšwa mo, Gauteng re	**sepela**	**ka molao**. Ge le sa dire bja...
...a mohlang woo ka re: Ke tla lesa go	**sepela**	**ka paesekele** bošego goba...
...ba He sepetlele, ba ile ba kwana go	**sepela**	ka pese. Ke nnete, batho ba...
...a Bohlabela go ya Bodikela. Ge a ka	**sepela**	**ka sefofagodimo** gape, a tl...
...e matona a rena a Lebowa a be a ka	**sepela**	ka setimela nako le nako,...
...ka kokopaneng. Mmalo mošmo o be o	**sepela**	**ka tshwanelo**. Bagologolo ba...
...elo tša mantšiboa ao. Banna! Go	**sepela**	ke go bona, sogana le sa et...

Some prominent senses of **crawl** such as 'moving on hands and feet', 'slow-moving traffic', 'time passing slowly', etc. immediately come to the fore to be considered by the lexicographer for inclusion or omission. The same holds true for **sepela** where the lexicographer's attention is drawn to the fact that 'walk', 'go', 'ride', 'obey', 'follow', etc. are to be considered as possible senses.

7.0 **Corpora as an aid to finding typical collocations, idioms and proverbs**

A glance at concordance lines for the Sepedi word **ipona** 'to see oneself' reveals the typical *collocations* shown in (13). Collocations are words that occur more often than not in the neighbourhood of a specific word.

(13) Sepedi (Collocations of **ipona** 'to see oneself' – part 1)

ipona	botlaela
ipona	botlaela
……	……
ipona	botlaela

'to see oneself being foolish'

ipona	molato
ipona	molato
……	……
ipona	molato

'to see oneself being guilty'

ipona	phošo
ipona	phošo
……	……
ipona	phošo

'to see oneself being wrong'

Advanced corpus query tools such as the one based upon Microsoft Access, which was developed at the University of Pretoria, can even statistically analyse such collocations, in giving total counts for each pattern as in (14) below, in addition to listing the patterns as in (13) above.

(14) Sepedi (Collocations of **ipona** 'to see oneself' – part 2)

Word	Expr1	Expr2
ipona	botlaela	6
ipona	molato	19
ipona	phošo	13

Deciding on the most typical *idioms* and *proverbs* is also no longer a guessing exercise for the lexicographer. A quick glance at the corpus immediately reveals the most commonly used ones, as shown in (15), which make them candidates for inclusion.

(15) Sepedi (Typical idioms and proverbs)

ga gabo e be e le gore o na le moko le maatla a go	**swara**	**tau ka mariri** Aowa ka nnete Pelompe monna wa
ge a ba laodišetša ka tša Madibamaso tšona tša go	**swara**	**tau ka mariri** Mosebjadi mekgolokwane a hlaba
nwele meetse ka kgolwa ka gore o kgonne le go	**swara**	**tau ka mariri** wa iphetla molala Ngwanake bjale
gona e tloga e laeditše gore mmagongwana o	**swara**	**thipa ka bogaleng** Ga go motswadi yo a ka se
Sepedi le sona se re mmagongwana o	**swara**	**thipa ka bogaleng** ge o bona le wena o gopola
moka Aowa ke nnete magongwana o	**swara**	**thipa ka bogaleng** Morwedi wa Ketladireng o tlo
le motho Kganthe lehono mosadi ga a sa	**swara**	**thipa ka bogaleng** naa MOTŠHELO WA

These concordance lines clearly indicate the use of the verb **swara** 'to grab' in **go swara tau ka mariri** 'to tackle the bull by the horns' or in **go swara thipa ka bogaleng** 'to get into trouble'.

8.0 Corpora as an aid in pinpointing clusters and choosing better examples of usage

Most dictionaries offer examples of usage which were made up by the author(s). Often, made-up examples can be described as functional but dull as in (16), taken from the *Groot Noord-Sotho-woordeboek* (Ziervogel and Mokgokong 1975).

(16) Sepedi (Made-up example of usage)

swanêtše … *o swanêtše go šoma ka maatla* 'you ought/must work hard'

O swanetše go šoma ka maatla is not a bad choice since the cluster **swanetše go šoma** occurred eight times in the corpus, as shown in (17). Clusters are groups of words which frequently follow each other, hence groups of words which seem to *cluster* together.

(17) Sepedi (Corpus lines with the cluster **swanetše go šoma**)

...sa setšhaba se a tsomega. Re	**swanetše go šoma**	mmogo go...
...kolong bana ba lapile gomme ba	**swanetše go šoma**	mešomo ya...
...mo Lebowa motho mang le mang o	**swanetše go šoma**	ka maatla...
...e laetša gore barutiši ba	**swanetše go šoma**	ka maatla....
...Moswana o re mong le mong o	**swanetše go šoma**	goba...
...di sa lemoge gore di	**swanetše go šoma**	gammogo ka...
...ye e tlilego le nna ya gore o	**swanetše go šoma**	...
...Ka ntle go be go	**swanetše go šoma**	badiredi ba...

However, an in-depth study of the use of **swanetše** reveals a much more interesting/problematic situation. In cases such as these, where it is quite tough to find typical uses at a first glance, the lexicographer can zoom in on the word in order to detect vital co-occurrences with words which are not immediately preceding or following the word in question. Indeed, any mother tongue speaker of Sepedi can tell immediately that **swanetše** is always followed by **go**, see (18). If pressed for further information they

might guess words such as **šoma**, **sepela**, **bolela**, etc. which are not bad guesses. However, zooming in 2 levels further with the help of sophisticated query tools, reveals that **ba** is the *most typical* word following **swanetše go**.

(18) Sepedi (Counts for clusters with **swanetše**)

Word	**Expr1**	**Expr2**	**Expr3**
swanetše	go	**ba**	237
swanetše	go	no	45
swanetše	go	mo	37
swanetše	go	ya	31
swanetše	go	tseba	29

Still, this **ba** can be an object concord or a copulative verb stem. Going down yet another level reveals that the use of the object concord **ba** of class 2 as in (19) is relatively infrequent.

(19) Sepedi (Disambiguating clusters with **swanetše** – part 1)

…gomme a bona a swanetše go **ba** begela ditaba…
'…and he realised that he had to report to them the issues…'

Rather, it is the *copulative verb stem* **ba** that is frequently used with **swanetše go**. Table (20) reveals yet one more important aspect, namely that the copulative verb stem **ba** is used in most cases with the conjunctive particle **le**.

(20) Sepedi (Disambiguating clusters with **swanetše** – part 2)

Word	**Expr1**	**Expr2**	**Expr3**	**Expr4**
swanetše	go	ba	banna	2
swanetše	go	ba	gona	14
swanetše	go	**ba**	**le**	34
swanetše	go	ba	sejo	2
swanetše	go	ba	yena	3

With all this information at his/her disposal, the lexicographer is in a position to treat the entry **swanetše** in such a way that the information most likely to be looked for by his/her target user is presented in the dictionary. Target users, especially learners of the languages will surely be confronted with **swanetše go ba** or **swanetše go ba le** on page two of their new prescribed Sepedi book, if not on page one! Thus examples such as (21) and (22) would immediately help the user to understand the most frequent use(s) of **swanetše**.

(21) Sepedi (Corpus-based example of usage for **swanetše** – part 1)

Moahlodi a ka thušwa bjang ka gobane e swanetše go ba yena a tsebago melao.
'How can the judge be assisted because he is the one who must know the law.'

This example is very natural and it clearly illustrates the use of "must be".

(22) Sepedi (Corpus-based example of usage for **swanetše** – part 2)

Bona bao hlogo ya sekolo e swanetše go ba eletša ka mo e ka kgonago.
'Those in particular, the headmaster must warn them as well as he can.'

This example is also very natural and clearly illustrates the use of "must (do something to) them".

9.0 Integration of different corpus query tools in order to enhance the quality of data presentation in the dictionary

With good query tools at his/her disposal, the lexicographer can combine different tools such as word-frequency counts and concordance lines. For instance, by means of word-frequency counts the lexicographer can determine that the second most frequently used verb in Kiswahili is **kusema** (cf. also De Schryver & Prinsloo 2001). From (1) above it is clear that the most frequent *inflection* for this verb is **alisema** 'he/she said; he/she spoke'. Now this highly used inflected form of the verb can firstly be studied in terms of the typical *clusters* in which it occurs in the 1.3-million-large *Kiswahili Internet Corpus* (KIC). The results are shown in (23).

(23) Kiswahili (Clusters with **alisema** in KIC)

N	Clusters in KIC	Count
1	huyo **alisema** kuwa	61
2	**alisema** baada ya	48
3	**alisema** pamoja na	34
4	rais mkapa **alisema**	30
5	gewe **alisema** kuwa	24
6	**alisema** hata hivyo	23
7	**alisema** jana kuwa	23
8	**alisema** kutokana na	23
9	kamanda gewe **alisema**	23
10	waziri mkuu **alisema**	21
11	**alisema** hatua hiyo	20
12	**alisema** kuwa kwa	20

From (23) we see that the most frequent cluster in KIC is **huyo alisema kuwa** 'he/she said that'. The second most frequent cluster is **alisema baada ya** 'he/she said after', followed by **alisema pamoja na** 'he/she said in addition to', etc.

Secondly, the ability of the query tool to analyse an item in terms of its typical *collocations* can be employed for **alisema**. In (24) *WordSmith Tools* was asked to search for collocations starting five places to the left of **alisema** up to five places to the right. In order to achieve this, the 'collocate horizons' were put to L5-R5. The twentieth most frequent collocate of **alisema** for instance, **serikali** 'government', collocates 188 times within this range. 65 of those occur to the left, and 123 occur to the right. Furthermore, by way of example, (24) also shows that **serikali** appears 23 times 4 places to the right of **alisema**.

(24) Kiswahili (Collocations of **alisema** in KIC)

N	Item	Total	Left	Right	L5	L4	L3	L2	L1	*	R1	R2	R3	R4	R5
1	**alisema**	4146	154	273	18	17	65	21	33	3719	26	64	76	57	50
2	na	1239	645	594	213	157	132	140	3	0	17	100	148	150	179
3	ya	1225	618	607	170	191	106	151	0	0	0	183	151	129	144
4	wa	1015	585	430	170	160	106	149	0	0	0	81	115	94	140
5	kuwa	965	65	900	19	18	16	8	4	0	670	54	52	63	61
6	bw	887	641	246	19	37	148	395	42	0	161	29	30	12	14
7	kwa	715	292	423	87	86	72	47	0	0	105	74	67	83	94
8	huyo	412	235	177	18	10	25	7	175	0	0	68	54	34	21
9	hiyo	389	166	223	28	36	43	11	48	0	7	67	48	52	49
10	katika	365	162	203	84	41	31	6	0	0	52	31	34	44	42
11	ni	350	111	239	40	27	31	13	0	0	34	40	60	50	55
12	za	262	148	114	35	35	29	49	0	0	0	16	24	30	44
13	la	248	147	101	39	38	33	35	2	0	0	27	21	21	32
14	huo	228	87	141	6	24	28	10	19	0	2	41	39	36	23
15	hivyo	222	128	94	6	8	49	7	58	0	11	26	22	25	10
16	kwamba	215	23	192	10	3	6	2	2	0	100	17	33	15	27
17	hata	202	111	91	9	49	10	43	0	0	29	23	24	3	12
18	hayo	200	59	141	10	14	17	8	10	0	50	24	25	25	17
19	hilo	193	102	91	11	11	30	11	39	0	3	21	26	22	19
20	**serikali**	188	65	123	16	9	11	9	20	0	47	19	20	23	14
21	yake	171	82	89	9	16	25	8	24	0	0	33	26	14	16
22	hao	161	34	127	6	4	8	7	9	0	0	36	30	37	24
23	cha	154	72	82	22	27	9	14	0	0	0	22	21	20	19
24	jana	140	68	72	5	4	26	14	19	0	35	12	10	7	8
25	kama	133	46	87	13	15	5	13	0	0	30	16	11	15	15
26	**rais**	129	83	46	8	6	9	33	27	0	16	7	9	9	5
27	baada	128	10	118	5	5	0	0	0	0	46	30	8	18	16
28	sasa	128	31	97	9	8	2	7	5	0	2	46	20	21	8
29	pia	127	52	75	6	2	5	3	36	0	31	10	7	21	6
30	**mkuu**	123	85	38	20	16	9	9	31	0	0	7	8	12	11
31	dk	120	77	43	2	5	9	50	11	0	28	4	6	2	3
32	hizo	119	47	72	8	6	14	7	12	0	0	24	19	16	13
33	**waziri**	119	87	32	5	2	5	61	14	0	11	3	4	10	4
34	**watu**	110	38	72	13	14	4	5	2	0	17	12	19	14	10
35	wake	106	66	40	8	12	19	9	18	0	0	8	13	8	11

From (24) one must conclude that it is not the word **watu** 'people' which is the noun collocating most frequently with **alisema**, as one might have expected, but rather nouns depicting 'high-ranking civil servants or bodies' like **serikali** 'government', **rais** 'president', **mkuu** 'district commissioner; leader' or **waziri** 'minister'.

Finally, with all this available corpus data it is now very easy for the lexicographer to select a *typical example of usage* for inclusion into the

dictionary by simply glancing at the output of one or more concordance-line screens. Going through the concordance lines for **alisema**, one can for instance choose (25) as a very typical example of usage to illustrate the verb **kusema**.

(25) Kiswahili (Typical example of the usage of the verb **kusema** extracted from KIC)

Balozi huyo alisema kuwa jitihada za serikali za kutokomeza rushwa hazina budi kuungwa mkono.
'The ambassador said that government efforts to combat corruption must be supported.'[3]

This example sentence to illustrate the usage of the verb **kusema** has the huge advantage that: (a) it uses the most frequent inflection (**alisema**) of the verb to be illustrated, (b) it reflects the most frequent cluster (**huyo alisema kuwa**) with this most frequent inflection, and (c) it contains the most frequent noun (**serikali**) collocating with this most frequent cluster.

10.0 Conclusion

In this article it has been argued that *corpora* and powerful *corpus query tools* are indispensable for the compilation of modern Bantu dictionaries. On the macrostructural level corpus data enables the lexicographer to solve the most challenging question, namely what to include in and what to omit from the lemma-sign list of a dictionary. Furthermore, it enables him/her to combat different types of lemmatisation inconsistencies which commonly occur in dictionaries which are compiled without a corpus. On the microstructural level corpora are indispensable as an aid to sense distinction for the writing of better definitions (monolingual dictionaries) or for the selection of suitable translation equivalents (bilingual dictionaries). Concordance lines also reflect typical collocations, clusters, idioms, proverbs and examples of usage. Finally, it was also shown that different corpus query tools can even be *combined* to maximise information retrieval by the lexicographer.

References

URL (Universal Resource Locator)

WordSmith http://www.liv.ac.uk/~ms2928/wordsmith/screenshots

Atkins, B.T. Sue, **Michael Rundell** and **Edmund Weiner.** (1997) Salex'97, A training course in the compilation of monolingual dictionaries. (Unpublished course material of a tutorial held at the Dictionary Unit for South African English, Rhodes University, Grahamstown, 15-26 September 1997.)

Brown, J. Tom. (1925) *Secwana Dictionary, Secwana – English and English – Secwana*. Lobatsi: South Africa District Committee of the London Missionary Society.

De Schryver, Gilles-Maurice and **D.J. Prinsloo.** (2001) Towards a Sound Lemmatisation Strategy for the Bantu Verb through the Use of *Frequency-based Tail Slots* – with special reference to Cilubà, Sepedi and Kiswahili. In J.S. Mdee & H.J.M. Mwansoko (eds.). (2001) *Makala ya kongamano la kimataifa Kiswahili 2000. Proceedings*: 216–242, 372. Dar es Salaam: TUKI, Chuo Kikuu cha Dar es Salaam.

Gouws, Rufus H. (1990) Information Categories in Dictionaries, with Special Reference to Southern Africa. In Reinhard R.K. Hartmann (ed.). (1990) *Lexicography in Africa, Progress reports from the Dictionary Research Centre Workshop at Exeter, 24-25 March 1989*: 52-65. Exeter: University of Exeter Press.

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

Hartmann, Reinhard R.K. (ed.). (1983) *Lexicography: Principles and Practice*. London: Academic Press.

Knowles, Frank. (1983) Towards the Machine Dictionary, 'Mechanical' dictionaries. In Reinhard R.K. Hartmann (ed.). (1983): 181-93.

Kriel, T.J. (1988[3]) *Popular Northern Sotho Dictionary, N. Sotho – English, English – N. Sotho*. Pretoria: J.L. van Schaik.

Kromhout, Jan. (1997[xiii]) *Klein woordeboek / Little Dictionary, Afrikaans – Engels, English – Afrikaans*. Cape Town: Pharos.

Laufer, Batia. (1992) Corpus-based versus Lexicographer Examples in Comprehension and Production of New Words. In Tommola, H. et al. (eds.). *Proceedings of the Fifth Euralex International Congress, 4-9 August 1992*: 71-6. Tampere: University of Tampere.

Martin, Willy J.R., **Bernard P.F. Al** and **Piet J.G. van Sterkenburg**. (1983) On the Processing of a Text Corpus, From textual data to lexicographical information. In Reinhard R.K. Hartmann (ed.). (1983): 77-87.

Sinclair, John M. (ed.). (1995[2]) *Collins COBUILD English Dictionary*. London: HarperCollins Publishers.

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

The Oxford English Dictionary, Second Edition on Compact Disk. (1992[2]) Oxford: Oxford University Press.

Tomaszczyk, J. (1983). On Bilingual Dictionaries, The case for bilingual dictionaries for foreign language learners. In Reinhard R.K. Hartmann (ed.). (1983): 41-51.

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

Ziervogel, Dirk and **Pothinus C.M. Mokgokong.** (1988[4]) *Klein Noord-Sotho woordeboek, N.-Sotho – Afrikaans – English, Afrikaans – N.-Sotho, English – N.-Sotho*. Pretoria: J.L. van Schaik.

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

[2] In COBUILD2 (Sinclair 1995[2]) information on frequency is indicated in the Extra Column. Five 'frequency bands', shown by black diamonds, are used – where the most frequent words have five diamonds, the next most frequent four, etc. Less frequent words do not have any black diamonds. For **dad** three diamonds are given, for **daily** four, and for **dairy** two.

[3] The source of this corpus example is *Nipashe*, the most widely read Kiswahili daily tabloid, covering a wide range of well-balanced local and foreign news and analysis. The original read: **Balozi huyo alisema kuwa jihada za serikali za kutokomeza rushwa hazina budi kuungwa mkono ...**. However, the standard spelling for the word 'effort' is **jitihada** and not **jihada**.