Revising Matumo's *Setswana–English–Setswana Dictionary*

D.J. Prinsloo, *Department of African Languages, University of Pretoria, Pretoria, Republic of South Africa (danie.prinsloo@up.ac.za)*

Abstract: The aim of this article is to design a revision strategy for the Setswana to English side of the *Setswana–English–Setswana Dictionary* compiled by Z.I. Matumo in 1993. An existing general organic Setswana corpus as well as a dedicated corpus compiled for the purposes of the revision will be used as a basis for macro- and microstructural aspects of the proposed revision. Lemma candidate lists for inclusion in and omission from the existing dictionary will be generated from these corpora, existing articles will be critically analysed and models for revised/updated articles will be presented. Key components of the revision strategy include the design and use of a multi-dimensional Ruler and Block System for the measurement and balancing of alphabetical stretches for the revised dictionary in terms of time, average length of articles and number of pages per alphabetical category. It is not possible to present all aspects of the revision within the scope of a journal article but the most prominent ones as well as a selection of typical issues will be dealt with.

Keywords: LEXICOGRAPHY, LEMMATISATION, REVISION, INFORMATION RETRIEVAL, MACROSTRUCTURE, MICROSTRUCTURE, RULER, BLOCK SYSTEM, DICTIONARY, AFRICAN LANGUAGES, SETSWANA (TSWANA)

Opsomming: Hersiening van Matumo se *Setswana–English–Setswana Dictionary*. Die doel van hierdie artikel is om 'n hersieningstrategie te ontwerp vir die Setswana na Engelse kant van die *Setswana–English–Setswana Dictionary* wat in 1993 deur Z.I. Matumo saamgestel is. 'n Bestaande algemene Setswanakorpus asook 'n spesifieke korpus wat saamgestel is vir die doel van die hersiening sal as basis vir mikro- en makrostrukturele aspekte van die voorgestelde hersiening gebruik word. Lemmakandidaatlyste vir insluiting in en weglating uit die bestaande woordeboek sal vanuit hierdie korpusse gegenereer word, bestaande artikels sal krities ontleed word en modelle vir die hersiene bygewerkte artikels sal aangebied word. Sleutelkomponente van die hersieningstrategie sluit die ontwerp en gebruik van 'n multi-dimensionele Liniaal en Bloksisteem in vir die meting en balansering van alfabetiese reekse vir die hersiene woordeboek in terme van tyd, gemiddelde lengte van artikels en aantal bladsye per alfabetiese kategorie. Dit is nie moontlik om alle aspekte van die hersiening binne die bestek van 'n tydskrifartikel aan te bied nie maar die vernaamstes, asook 'n aantal tipiese kwessies, sal behandel word.

Sleutelwoorde: LEKSIKOGRAFIE, LEMMATISERING, HERSIENING, INLIGTINGSONTSLUITING, MAKROSTRUKTUUR, MIKROSTRUKTUUR, LINIAAL, BLOKSISTEEM, WOORDEBOEK, AFRIKATALE, SETSWANA (TSWANA)

Introduction

Substantial revision and updating of a dictionary require detailed and meticulous planning on microstructural and macrostructural levels and is not less laborious than the planning and design of a new dictionary. Lexicographers often err in tackling such revisions in a haphazard way; eager to simply add new words to the dictionary rather than to take an holistic approach towards delivering a well-balanced and improved product.

> Many people think that the bulk of the work done by lexicographers, or dictionary makers, is that of collecting new words and defining them. Inclusion of the latest words is indeed a major part of our work, but no less important is the revising and updating of the entries for words that are already in our dictionaries. . ... During revision every aspect of a dictionary entry is examined and if necessary changed. (Stevenson 2004)

> The most obvious way the dictionary will develop is by the addition of more words. We already have a small list of words for inclusion in the next edition, and we look forward to obtaining more from our readers as well as from our own researchers. (Matumo 1993: ix)

Landau (2001) distinguishes between *updating* and *revision* of a dictionary. He regards updating as an exercise which should ideally be performed annually or biennially while substantial revision or in his terms *a complete re-examination of the previous edition* should be performed about every ten years.

> Dictionaries may be updated by the substitution of some new entries for old entries, and for the first few years after publication, such a procedure may work very well. But when a dictionary passes the ten- or fifteen-year-old mark, updating takes on a desperate character. (Landau 2001: 397)

The envisaged revision of Matumo's *Setswana–English–Setswana Dictionary* (henceforth referred to as MSD), published by Macmillan in 1993, thus qualifies in terms of Landau for such substantial revision.

On macrostructural level, the most prominent issue in the revision of a dictionary remains the decisions on lemmas to be included or excluded as echoed by Busane (1990: 30):

> One of the basic problems of lexicography is to decide what to put in the dictionary and what to exclude.

On microstructural level, the proposed revision of MSD will focus on a critical analysis of the data types and microstructural architecture with a view to creating a more user-friendly design with enhanced quality based on corpus data.

Background and original dictionary

MSD (1993) is the fourth edition of what is titled since 1993 the *Setswana–English–Setswana Dictionary*. The first edition dates back to approximately 1875, the second to 1895, and the third to 1925, entitled *Secwana–English Dictionary*. The latter was compiled by J. Tom Brown and formed the basis for MSD.

The features of the 'new' (1993) edition are summarized as follows:

- Completely reset in the most up-to-date orthography.
- Greatly increased number of headwords.
- Grammatical details in contemporary dictionary style.
- Tables of noun classes, concords and prefixes.
- References to many Setswana traditions.
- Proverbs quoted to illustrate delicate shades of meaning.
- Descriptive, not prescriptive, particularly with regard to borrowed or coined words. (Matumo 1993: Back cover)

In the Introduction Matumo says:

> I am as conscious as anyone else that there are shortcomings in this dictionary. Language is a fluid and developing organism, and a dictionary freezes it momentarily so that its vocabulary can be studied. This means that in an important sense a dictionary is already out of date on its day of publication. (Matumo 1993: ix)

Electronic Setswana corpora

The proposed revision of MSD is based on two Setswana electronic corpora. Firstly, the general Setswana Pretoria Corpus, compiled at the University of Pretoria, consisting of a variety of printed matter totalling 4.5 million running words (tokens) and 131 000 different words (types). Secondly, a dedicated Setswana corpus consisting of publications most likely to be studied by the target users of the revised dictionary, of approximately 1 million running words and 50 000 types.

Macrostructural revision strategies

As far as the choice of lemmata is concerned, the challenge to the lexicographer is the question as to whether, on the one hand, lemmas most likely to be looked for by the target users are included, and, on the other hand, whether all lemmas currently included in MSD can be justified in terms of such a likelihood. If frequency of use is an important criterion as is the case in the revision of MSD, the question is whether frequently used words were not accidentally left out or whether all the lemmas included in MSD deserve a place in the dictionary. Further the question could be raised if the space they occupy should rather be more fruitfully used for other words that either have a high frequency in the general corpus or a high frequency in the dedicated one. (See De Schryver and Prinsloo (2003) for a detailed discussion of the issue of balancing out general corpora and dedicated corpora in an effort to compile a lemma list for a restricted dictionary.) Even if the lexicographer ignores frequency counts and decides on the basis of his/her intuition that current entries should be retained, the question is whether they should be lemmas in their own right or treated in the articles of other lemmas.

Consider the following examples of words that occur more than a thousand times in the general corpus, frequently in the dedicated corpus and which were entered as translation equivalents in the English–Setswana side of MSD but that were not lemmatised in the Setswana–English side.

Table 1: Frequently used words not included as lemmas in MSD

Lemma	Freq. Gen. Corpus	Freq. Ded. Corpus	Meaning
le	12 5616	8 851	and
fa	56 463	2 987	here
bona	16 697	1 431	they; see
batho	8 424	899	people
botlhe	1 662	167	all
bosigo	1 478	95	night
sekolo	1 218	123	school
bonala	1 105	22	visible
tseo	1 064	58	those
otlhe	1 055	93	all

The occurrence of such instances underline the view of De Schryver and Prinsloo (2000) that utilization of a corpus is indispensable in assuring that words most likely to be looked for by target users are not omitted simply because they did not cross the compilers' way.

Different types of omissions/inconsistencies are apparent in Table 1. Firstly, a common failure is to complete a typical paradigm of which only a limited number of elements exist, e.g. quantitatives (cf. Gouws and Prinsloo (1997: 47) for a perspective on limited versus unlimited elements). The forms for classes 8 or 10 *tsotlhe* (2 336), class 15 *gotlhe* (397), 1st pers. plural *rotlhe* (217) and class 14 *jotlhe* (183) are given, but not classes 2 *botlhe* (1 662), class 6 *otlhe* (1 055), class 5 *lotlhe* (409), class 7 *sotlhe* (67), etc.

A second example in this regard is the demonstrative second position, class 7: *seo* 'that one' (1 301) is given, but not classes 8 or 10 *tseo* 'those' (1 064), class 5, *leo* 'that one' (949), etc. All demonstratives given in the guidelines to the dictionary should be treated in the central text. (See Prinsloo (1996) for a discussion on dead references pertaining to words given in the guidelines to a dictionary.)

In order to combat what Gouws and Prinsloo (1998: 21) call the decontextualisation of lexical items, brought about by the alphabetical sorting of lemmas in a dictionary, tables such as those given for the quantitatives and demonstratives in the front matter fulfil a valuable function in restoring such lexical and grammatical relations. It is however imperative that the members of such a paradigm be lemmatised in the central text and that appropriate and correct reference be made from each individual lemma to the tables as reference addresses. Compare also in this regard the inclusion of numerous colour plates of different trees and cattle in the back matter of Kgasa and Tsonope (1995) without cross-referencing from the articles of these trees and cattle in the central text.

When candidates for deletion from the lemma list of MSD must be decided on, consider the following extract from a list of multiword lemmas in MSD.

Table 2: A selection of multiword lemmas in MSD

ka go dira	ke gone	ka thelelo	ka mmanene	tolwana ya leitlho
mokgatha-thete	ke a	kgaphasetse	ka kgaga	thini ya tsebe
ke ne ke	kaololwa	kgeleisitse	ka ke	tladi mothwana
ke mong	kago e e godileng	ka moso	ka jeno	tladi ya tlapana
ke mang	kabayanya	ka mmanete	ka gope	nkgiwa

Singled out for attention here are the numerous clusters presented as multiword lemmas. The lemmatisation of multiword items such as *ke ne ke* 'I was', *ka go dira* 'by acting', *kago e e godileng* 'a building that is high or tall', etc. cannot be critisized in principle. Gouws (1991) and Zgusta (1971) emphasize that there are numerous multiwords that should be regarded as single lexical items and therefore be presented as multiword lemmas in the central text of the dictionary. However, in MSD multiword lexical units are often confused with frequently used free combinations.

The potential for the successful retrieval of information by target users is also low for most lemmas in Table 2. Of the 330 occurrences of *kago* 'the process of building, a building' in the corpus, *kago e e godileng* occurred only once and clusters such as *kago ya phemelo* 'protection building/structure' and *kago ya bokgoni* 'successful structure' occur more frequently with counts of 17 and 10 respectively but were not lemmatised as multiword lemmas. Since *kago* was lemmatised, no real harm is done in lemmatising *kago e e godileng* as well because alphabetically it directly follows the article of the lemma *kago* and may therefore catch the eye of the user. In the case of *ka go dira*, and many other similar ones, the value of the entry is however questionable since it is unlikely that the user will know how to look it up in the alphabetical stretch for K especially since no cross-referencing is provided from the article for *dira* to *ka go dira*.

Even if users do consult lemmas starting with or consisting of *ka*, they are confronted with another problematic aspect of lemmatisation in MSD, i.e. extensive stacking of a large number of lemmas, in this case 38, consisting of 12 lemmas for *ka* and 26 lemmas for *ka* plus a noun, verb, etc. Even a cross-reference to these 38 possible lemmas that are not marked as homonyms, e.g. by superscript homonym markers, would be user-unfriendly. A much better solution would be to treat frequent clusters such as *ka go dira* (167) *ke go dira* (87) and *kgona go dira* (51) in the article of *dira*.

Building and applying a multi-dimensional Ruler

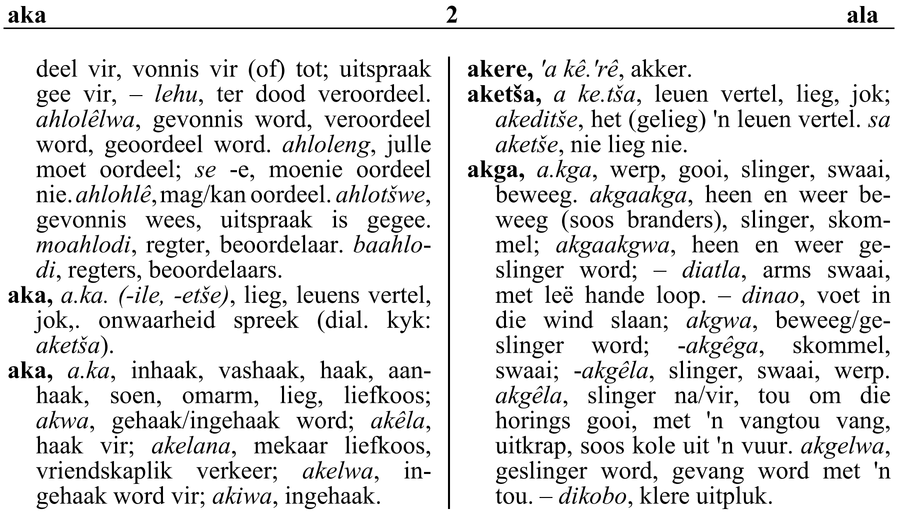
Apart from the macrostructural aspect relating to inclusion versus omission of individual lemmata, such control should be exercised in terms of balancing out entire alphabetical categories in the dictionary as a whole.

> Nothing is more difficult to predict or control than a dictionary begun from scratch. (Landau 2001: 398)

This remark is equally applicable to dictionaries that were compiled without the availability of a corpus. (See De Schryver and Prinsloo (2000) and Prinsloo and De Schryver (2003) for numerous examples of inconsistencies regarding over- and undertreatment in terms of alphabetical categories.)

Consider the following example where substantial inconsistency between the length of articles in the first few alphabetical categories compared to the last few in Kriel (1983) is apparent even to the naked eye, without any help from measuring instruments.

(1)



aka **2** **ala**

deel vir, vonnis vir (of) tot; uitspraak gee vir, – *lehu*, ter dood veroordeel. *ahlolêlwa*, gevonnis word, veroordeel word, geoordeel word. *ahloleng*, julle moet oordeel; *se* -e, moenie oordeel nie. *ahlohlê*, mag/kan oordeel. *ahlotšwe*, gevonnis wees, uitspraak is gegee. *moahlodi*, regter, beoordelaar. *baahlodi*, regters, beoordelaars.

aka, *a.ka. (-ile, -etše)*, lieg, leuens vertel, jok,. onwaarheid spreek (dial. kyk: *aketša*).

aka, *a.ka*, inhaak, vashaak, haak, aanhaak, soen, omarm, lieg, liefkoos; *akwa*, gehaak/ingehaak word; *akêla*, haak vir; *akelana*, mekaar liefkoos, vriendskaplik verkeer; *akelwa*, ingehaak word vir; *akiwa*, ingehaak.

akere, *'a kê.'rê*, akker.

aketša, *a ke.tša*, leuen vertel, lieg, jok; *akeditše*, het (gelieg) 'n leuen vertel. *sa aketše*, nie lieg nie.

akga, *a.kga*, werp, gooi, slinger, swaai, beweeg. *akgaakga*, heen en weer beweeg (soos branders), slinger, skommel; *akgaakgwa*, heen en weer geslinger word; – *diatla*, arms swaai, met leë hande loop. – *dinao*, voet in die wind slaan; *akgwa*, beweeg/geslinger word; -*akgêga*, skommel, swaai; -*akgêla*, slinger, swaai, werp. *akgêla*, slinger na/vir, tou om die horings gooi, met 'n vangtou vang, uitkrap, soos kole uit 'n vuur. *akgelwa*, geslinger word, gevang word met 'n tou. – *dikobo*, klere uitpluk.

(2)

ribega, *ri bê.ga*, onderstebo keer, toemaak; bedek.
ribegetša, *'ri be ge.tša*, onderstebo draai.
ribesela, *ri bê sê.la*, omslaan, omkeer.
ribete, *'re bê.tê*, klinknael.
ribetela, *ri bê tê.la*, klink, vasnael, vasklink.
riboga, *'ri bo ga*, ontvang, swanger word.
ribogolla, *'ri bo go l.la*, regop draai, *ribolla*, openbaar, oopmaak, blootlê.
ribolla, *ri bo l.la*, omgekeerde oprig/regstel.
rifa, *ri.fa*, bedek, toemaak.
rifi, *'ri.fi*, rif, rotsbank; – *ya gauta*, goudrif.
robetše, *ro be.tše*, slaap, het ontslaap, *yo a robetšego*, ontslapene.
robja, *rô.bja*, gebreek word.
roboka, *rô'bô.ka*, smok, aanrand.
robong, *ro'bo.ng*, nege.
roborobo, *rô bô rô bô*, spoorwegbus
roboto, *ro bô tô*, robot, verkeerslig.
roga, *'ro.ga*, vloek, skel, skeltaal gebruik; swets, beledig.
rogaka, *ro ga.ka*, vloek, skel; *-wa*, gevloek, vervloek word; *-ile*, het vervloek; – iša, laat vloek.
rogana, *ro ga.na*, vloek, mekaar uitskel.
roganela, *ro ga nê.la*, vloek terwille van.
rogo, *'rô.gô*, rog, *mo-*, bredie.

In order to address such inconsistencies on the macrostructural level, Prinsloo and De Schryver (2002, 2003) and De Schryver (2003), studied the balance between alphabetical categories for English, Afrikaans and a number of African languages.

The question was whether a specific distribution, preferably one that could accurately be measured, exists between the different categories in a given language. They found that this is indeed possible. A remarkable consistency in respect of the balance between alphabetical stretches has been detected by comparing dictionaries and corpora. This consistency is observed with regard to, on the one hand, the number of lemmas treated for or the number of pages dedicated to each alphabetical category, and, on the other hand, the lemmatised as well as unlemmatised alphabetical word lists culled from corpora. For purposes of the revision of MSD, Rulers were compiled from the general corpus as well as from the dedicated corpus.

The concept *Ruler* is defined as a practical instrument of measurement for the relative length of alphabetical stretches in alphabetically ordered dictionaries. They are designed according to the generally accepted principle that alphabetical categories in any given language do not contain an equal number of words. For example, a single glance at a few popular English dictionaries reveals that the alphabetical categories or alphabetical stretches for A, B, D, M, R and especially C and S, contain large numbers of lemmas, occupying almost 50% of the dictionary, while categories such as J, K, Q, U, V, X, Y and Z are relatively small, and consequently fill only a few pages. For a dictionary such as the *Macmillan English Dictionary* (Rundell 2002), where the alphabetical categories are marked with coloured thumb tags, one does not even have to open the dictionary in order to appreciate this breakdown which can also literally be measured by putting an ordinary ruler against the dictionary to roughly measure the 'thickness' of each alphabetical stretch in millimetres. Likewise, an alphabetical list of types generated from the Sesotho sa Leboa corpus shows that roughly 17% of all words in this language fall under the single category M while categories such as C, J, Q, U, V, W, X, Y and Z are virtually empty.

Consider the Ruler for Setswana in Figure 1, based on the average of the percentage breakdown of types in (a) the general Setswana corpus and (b) the dedicated Setswana corpus.

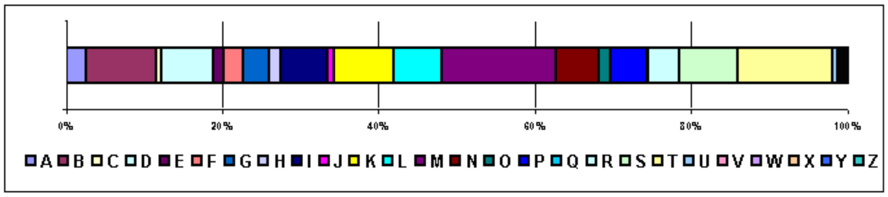


Figure 1: A Ruler for Setswana

For the revision of MSD, the focus is shifted from an alphabetical breakdown in the sense of the balance between the 26 letters of the alphabet (A to Z) by reorganising the data given in Figure 1 into a *percentage* breakdown in the form referred to as a Block System in Table 3.

Table 3: A Block System for Setswana

1	**ALAF**		21	**FELE**		41	**KOUS**		61	**MOTL**		81	**SELE**
2	**AROG**		22	**FOLO**		42	**LAEL**		62	**MPHE**		82	**SERA**
3	**BADI**		23	**GAGW**		43	**LEBO**		63	**NATE**		83	**SETO**
4	**BANN**		24	**GATS**		44	**LEKI**		64	**NGWA**		84	**SIMO**
5	**BATW**		25	**GOLO**		45	**LERI**		65	**NKUK**		85	**SUAS**
6	**BIRO**		26	**GWET**		46	**LETS**		66	**NTEM**		86	**TALE**
7	**BOGA**		27	**HUBE**		47	**LOKO**		67	**NTSH**		87	**THAA**
8	**BOLA**		28	**IJES**		48	**MAAD**		68	**NYOR**		88	**THIB**
9	**BONK**		29	**IKGO**		49	**MAHA**		69	**OOMA**		89	**THWE**
10	**BORU**		30	**INOL**		50	**MALE**		70	**PANT**		90	**TLAM**
11	**BOUT**		31	**IPUS**		51	**MARA**		71	**PHAK**		91	**TLHA**
12	**DAAM**		32	**ITIS**		52	**MATL**		72	**PHIM**		92	**TLHO**
13	**DIFA**		33	**ITSH**		53	**MEFA**		73	**PITL**		93	**TLWA**
14	**DIKG**		34	**JOKO**		54	**MESU**		74	**PUDU**		94	**TSAP**
15	**DINK**		35	**KANY**		55	**MMAL**		75	**RAMO**		95	**TSHE**
16	**DIRA**		36	**KERO**		56	**MMOL**		76	**RENG**		96	**TSHW**
17	**DITH**		37	**KGAR**		57	**MOFI**		77	**ROKG**		97	**TSUN**
18	**DITU**		38	**KGOM**		58	**MOKG**		78	**RURU**		98	**UBAU**
19	**EGEP**		39	**KHAN**		59	**MONG**		79	**SEBA**		99	**WABO**
20	**ETLH**		40	**KODU**		60	**MORW**		80	**SEHI**		100	**ZIMB**

While based on the same statistics, the Block System opens the door to a number of very practical applications and a multi-dimensional utilization in the revision process of MSD. For lexicographers and editors it gives clear guidance in terms of page allocation, average length of articles, progress in terms of time and even remuneration intervals for part-time compilers.

With the prescribed number of pages set at roughly 300 for each side of the dictionary, it means that 3 pages should roughly correlate with each block/ percentage point; the average article length should be 3 lines, and the average compilation time per article 10 minutes. Even remuneration scheduled at the markers 25% **GOLO**, 50% **MALE**, 75% **RAMO**, and 100% **ZIMB**, is being negotiated.

An actual compilation test was performed by treating a selection of 100 typical lemmas logging the average length and time used for the compilation of each article, with and without consultation of the corpora.

It is important that a sound perspective be maintained on the value of the multidimensional Ruler and Block System as dictionary compilation tools. They should not be regarded as absolute or precision instruments of measurement. The real value of the Ruler lies in the fact that it focuses the attention of the compiler on potential ill-balanced areas. This will now be illustrated for MSD.

Table 4: MSD lemma and page breakdown versus the Setswana Ruler

	MSD: Lemmas %	**MSD: Pages %**	**Setswana Ruler**	**MSD lemmas vs the Ruler**
A	1.2	1.3	2.6	-1.4
B	**4.7**	**4.6**	**9.0**	**-4.3**
C	0.0	0.0	0.6	-0.6
D	6.0	6.4	6.6	-0.6
E	1.2	1.3	1.4	-0.2
F	3.7	3.3	2.4	1.3
G	5.2	5.3	3.4	1.8
H	0.9	0.9	1.5	-0.6
I	5.3	4.9	5.9	-0.6
J	0.7	0.7	0.8	-0.1
K	**12.2**	**11.9**	**7.7**	**4.5**
L	6.7	6.8	6.1	0.6
M	12.5	13.7	14.6	-2.1
N	4.0	4.0	5.5	-1.5
O	1.3	1.3	1.6	-0.3
P	5.9	6.0	4.6	1.3
Q	0.0	0.2	0.2	-0.2
R	3.9	3.5	3.9	0.0
S	8.5	8.6	7.5	1.0
T	**15.4**	**14.1**	**12.2**	**3.2**
U	0.5	0.4	0.6	-0.1
V	0.0	0.0	0.3	-0.3

W	0.1	0.2	0.4	-0.3
X	0.0	0.2	0.1	-0.1
Y	0.1	0.0	0.3	-0.2
Z	0.0	0.0	0.2	-0.2
		99.8	100	

In the revision of MSD, the Ruler suggested under-treatment of the alphabetical stretch B and over-treatment of the stretches K and T in terms of the number of lemmas treated and the number of pages allocated to these categories. It is now the lexicographer's task to analyse these categories in order to ascertain why these alphabetical categories deviate from the Ruler and if corrective action is required. The corpora supply further assistance in the form of candidate lists for inclusion and for omission discussed above.

In the case of the presumed under-treatment of B in MSD, the lexicographer should particularly study the list of candidates for *inclusion* to see if frequently used words were not left out. In the case of K and T the focus should primarily be on the candidate lists for omission to determine whether inclusion of words that do not occur even once in the corpora are justified or not. A detailed analysis of these stretches cannot be given here but a brief analysis will be attempted. By analysing B on suspected under-treatment, gross inconsistencies and omissions were indeed and immediately detected.

Table 5: Frequently used words in the alphabetical stretch B not included as lemmas in MSD

Lemma	**Translation**	**Tot. freq. in both corpora**	**In MSD: Yes/No**	**In Brown: Yes/No**	**Transl. Eq. in MSD Eng.▸Sets.: Yes/No**
banna	men	2 341	No	Yes	Yes
batho	people	9 323	No	Yes	Yes
bona	they; see	18 128	No	Yes	Yes
botlhe	all	1 829	No	Yes	Yes
batsadi	parents	1 516	Yes	Yes	No

The policy of MSD is to include plural forms as lemmas, e.g. *batsadi*. However, lemmas such as *banna*, *batho*, *bona*, *botlhe* and *bosigo* were excluded even though they

(a) occur more than a thousand times in the corpora,

(b) were included in the 1925 *Secwana–English Dictionary* of Brown of which MSD is a revision, and

(c) are given as translation equivalents in the reverse side of MSD.

For the alphabetical stretches K and T, the lexicographer should critically evaluate the huge number of hapaxes (words occurring once only in a corpus) and zero frequencies given in the candidate lists for deletion in MSD, i.e. 1 664 lemmas (56.7% of all lemmas) for K and 1 812 (49.2%) for T.

The use of Rulers and Block Systems in the compilation or revision of dictionaries, does not mean, however, that the status of *hapax* or *zero-occurrence* in corpora is per definition a directive for omission. In the compilation of a lemma list for a restricted dictionary for very specific target users, De Schryver and Prinsloo (2003: 42-44) justified an extreme case of lemma selection/omission by including words that have a zero frequency in the dedicated corpus as lemmas but excluding words occurring up to nine times in the dedicated corpus.

Microstructural revision strategies

On the microstructural level, comment on semantics is the most important component or data type that, for a bilingual dictionary, should be presented mainly in the form of translation equivalent paradigms. Gouws (1989: 113) states that it is the information type most generally consulted by target users, most substantial and considered as the central component of the article.

> Vir die deursneewoordeboekgebruiker is betekenis die inligtingstipe wat die algemeenste in woordeboeke nageslaan word. As 'n mens na die struktuur van 'n woordeboekartikel kyk, is dit ook duidellik dat betekenisbeskrywing nie net die omvangrykste komponent van die artikel is nie maar dat dit ook as die sentrale deel van 'n woordeboekartikel beskou moet word.

In MSD, this is clearly not the case. Translation equivalents are to a large extent overshadowed by morphological and grammatical information, by the piling up of source language synonyms, etc. Compare the first few articles taken from a single, randomly selected page in MSD.

(3)

matlhagatlhaga ABS. N. CL. 6 *ma-*, NO SING., industriousness; activity.
matlhajana N. CL. 6 *ma-*, PL. OF *letlhajana*, shelves.
matlhaje N. CL. 6 *ma-*, PL. OF *letlhaje*, same as *matlhajwa*, a species of berry-yielding bush; Diospyros lycioides.
matlhakang N. CL. 6 *ma-*, NO SING., DER. F. *tlhakana*, a mixed, or motley lot.
matlhakola N. CL. 6 *ma-*, COLL. PL. OF *letlhakola*, Euclea spp.
matlhakola N. CL. 6 *ma-*, NO SING., DER. F. *tlhakodisa*, a remnant. ID. EXPR., *matlhakola a a dipêpa*, a bare remnant.
matlhaku N. CL. 6 ma-, PL. OF *letlhaku*, cut branches.

It is clear from (3) that comment on semantics takes a secondary place to detailed comment on form made even more prominent by the use of capital letters and to the piling up of source language synonyms sometimes even resulting in the total omission of any comment on semantics:

(4)

todi N. CL. 9N-, SING. OF *ditodi*, same as *lelodi* and *kgobati*.

Another aspect that should be corrected in the revision of MSD is inconsistent labelling and grammatical descriptions:

(5)

gotlhe ENUM. QUAL. CL. 15 < go-, see tab. p. xviii, all; altogether; entirely.
jotlhe CL. 14 QUANT. S., all; the whole, see tab. p. xix.
rotlhe QUAT. USED WITH SUBST. AND IN PARTICULAR THE ABST. PROV., rona, all of us.
tsotlhe QUANT. QUAL., USED WITH CONSTRUCTIONS OF CLS. 8 AND 10, all.
yotlhe QUANT. QUAL. USED TO QUAL. NOUNS AND PRON. CL. 9, all; nama yotlhe, all the meat.

In (5), a variety of grammatical labels, abbreviations and treatment styles are used to refer to quantitatives including punctuation errors and incorrect cross-references. As for punctuation, errors that need to be corrected include double commas, double full stops, grammar labels not followed by a full stop, etc.

For articles such as (6) that contain a translation equivalent paradigm of unrelated meanings, a homonymic approach should be considered as in (7).

(6)

ntlha N. CL. 9N-, SING. OF *dintlha*, a point; an item; a side; the first. INTERJ. EXPR., surprise; wonder.

It could be argued that translation equivalents such as 'a point', 'a side', 'the first' and 'idea' are not merely different senses but unrelated meanings that should accordingly be treated as homonyms:

(7)*

ntlha^1 *num.* **1** first: ***ke motho wa ntlha go nwa tee***, *he is the first person to drink tea;* **2** beginning: ***lwa ntlha o ne a itumetse***, *in the beginning he was happy*
ntlha^2 *conj.* but, by the way: ***ntlha e ka re re a latlhega***, *but it seems we are getting lost*
ntlha^3 *n.* side: ***o ntse ntlha ya lokotswana***, *he is sitting on the side of the wall*
ntlha^4 *n.* end, point: ***o bone ntlha ya teng e bogale***, *beware of its end, it is sharp*
ntlha^5 *n.* point, idea: ***o tsile ka ntlha e e botlhokwa***, *he came with a good point*

For the lemmatisation of verbs, the treatment of a randomly selected verb, *gagaba* in MSD as well as in a few other Setswana dictionaries can be considered.

(8)

(a) *Dikišinare ya Setswana–English–Afrikaans* (Snyman 1990)
gágábā, slither (eg a snake) // seil (eg 'n slang)

(b) *Thanodi ya Setswana* (Kgasa and Tsonope 1995)
gagaba GGG tpt. –ile. Tsamaya ka diatla le mangole

(c) *Secwana–English Dictionary* (Brown 1925)
Gagaba, v.i., pft. gagabile, creep or crawl, on hands and knees; crawl, as a cat hunting.

(d) MSD
gagaba V. S. SIMP., same as gogoba, creep or crawl, on hands and knees; crawl, as a cat hunting.

In comparison, consider the following extract from the concordance lines generated for *gagaba* from the corpora:

Table 6: Concordance lines for *gagaba*

..aetsega a tsena mo lobaleng **a**	**gagaba**	**ka mangole le diatla** a reedi
ba a sale a tshwana le **noga, a**	**gagaba**	**ka mpa** mo loroleng
aana mmoki a boka a ba a sala a	**gagaba**	**ka dimpa** fa fatshe. Moji a ts..
a tlhoka: **Maru a bo a tlhaga a**	**gagaba**	go tswa borwa. Botsho ba matl
go ikatametsa fa go tsona ka go	**gagaba**	**ka matsogo le mangole.** O ne a
se bonela mo lefifing se tla se	**gagaba**	**jaaka katse e ratela legotlo,**
tsaya motlhala wa mo mosong. Ba	**gagaba**	**ka iketlo, dikoloi di tletse**

A single glance at these concordance lines reveals that *creep* or *crawl* are indeed core senses of *gagaba* in relation to humans, animals and reptiles but also senses such as *slow movement* of e.g. clouds or traffic.

In (8)(a) the translation equivalent *slither* with reference to 'snake' is given but not in any of (8)(b)–(8)(d). In (8)(b) the definition is limited to 'move with hands and knees' which defines one of the core senses of *gagaba* but excludes this kind of movement for all animals and reptiles. In (8)(c) and (8)(d) movement of humans and animals are well captured but not that of reptiles nor the sense of slow movement. In an attempt to improve on MSD's article for *gagaba*, and in fact on all of (8)(a)–(8)(d), the following treatment is suggested for the lemma *gagaba*.

(9)

gagaba *v.* **1** crawl, creep: ***~ ka diatla le mangole***, crawl on hands and knees; ***~ jaaka katse e ratela legotlo*** crawl like a cat stalking a mouse **2** slither: ***noga e ~ ka mpa mo loroleng*** *the snake slithers on its belly in the dust*; **3** move slowly; ***maru a ~ go tswa borwa*** *clouds move in from the South*

Articles (7) and (9) represent an attempt to improve on typical articles for nouns and verbs in MSD such as (6) and (8) by putting much more emphasis on the comment on semantic, less on the comment on form, and to maximally use corpus data for sense distinction, frequent collocations, authentic examples, etc. in the treatment of such lemmas.

Conclusion

In this article an attempt has been made to formulate a typical revision strategy for substantial revision of a Setswana dictionary representing a case where in Landau's terms, revising should take on *a desperate character*. In all the official African languages of South Africa, many dictionaries exist that are outdated and in need of such a fundamental revision. Since electronic corpora exist for these languages, the strategies presented here could be considered for such revisions. Much emphasis has been placed on revision on the macrostructural level because it is believed that the dilemma of what to include in or exclude from the lemma list of especially a single-volume paper dictionary in terms of Busane (1990), is likely to remain 'forever'. It is therefore imperative for the lexicographer to be able to motivate inclusion/omission of lemmas in terms of sound lexicographic and statistical principles and only then to proceed to maximally utilise concordance lines to enhance microstructural treatment of these lemmas.

Endnote

* The original draft of this article for the lemma *ntlha* is credited to Mr Thapelo Otlogetswe.

References

Brown, J.T. 1925. *Secwana–English Dictionary*. Lobatsi: London Missionary Society.

Busane, M. 1990. Lexicography in Central Africa: The User Perspective with Special Reference to Zaïre. Hartmann, R.R.K. (Ed.). 1990. *Lexicography in Africa*: 19-35. Exeter Linguistic Studies 15. Exeter: Exeter University Press.

De Schryver, G.-M. 2003. Drawing up the Macrostructure of a Nguni Dictionary, with Special Reference to isiNdebele. *South African Journal of African Languages* 23.

De Schryver, G.-M. and D.J. Prinsloo. 2000. Electronic Corpora as a Basis for the Compilation of African-language Dictionaries, Part 1: The *Macrostructure*. *South African Journal of African Languages* 20(4): 291-309.

De Schryver, G.-M. and D.J. Prinsloo. 2003 Compiling a Lemma-sign List for a Specific Target User Group: The Junior Dictionary as a Case in Point. *Dictionaries* 24: 28-58.

Gouws, R.H. 1989. *Leksikografie*. Cape Town: Academica.

Gouws, R.H. 1991. Toward a Lexicon-based Lexicography. *Dictionaries* 13: 75-90.

Gouws, R.H. and D.J. Prinsloo. 1997. Lemmatisation of Adjectives in Sepedi. *Lexikos* 7: 45-57.

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

Kgasa, M.L.A. and J. Tsonope. 1995. *Thanodi ya Setswana.* Botswana: Longman.

Kriel, T.J. 1983. *Pukuntšu Dictionary*. Pretoria: J.L. van Schaik.

Landau, S.I. 2001. *Dictionaries: The Art and Craft of Lexicography.* Cambridge: Cambridge University Press.

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

Prinsloo, D.J. 1996. Review: Robert Botne and Andrew Tilimbe Kulemeka: *A Learner's Chichewa and English Dictionary* (Afrikawissenschaftliche Lehrbücher 9). *Journal of African Languages and Linguistics* 17(2): 199-202.

Prinsloo, D.J. and G.-M. de Schryver. 2002. Designing a Measurement Instrument for the Relative Length of Alphabetical Stretches in Dictionaries, with Special Reference to Afrikaans and English. Braasch, A. and A. and C. Povlsen (Eds.). 2002. *Proceedings of the Tenth EURALEX International Congress, EURALEX 2002, Copenhagen, Denmark, August 13–17, 2002*: 483-494. Copenhagen: Center for Sprogteknologi, University of Copenhagen.

Prinsloo, D.J. and G.-M. de Schryver. 2003. Effektiewe vordering met die *Woordeboek van die Afrikaanse Taal* soos gemeet in terme van 'n multidimensionele Liniaal [Effective Progress with the *Woordeboek van die Afrikaanse Taal* as Measured in Terms of a Multidimensional Ruler]. Botha, W. (Ed.). 2003. *'n Man wat beur. Huldigingsbundel vir Dirk van Schalkwyk*: 106-126. Stellenbosch: Buro van die WAT.

Rundell, M. (Ed.). 2002. *Macmillan English Dictionary for Advanced Learners*. Oxford: Macmillan.

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

Stevenson A. 2004. *Revising a Dictionary* [online]. Available at <http://www.askoxford.com/worldofwords/worddetectives/revising/?view=uk>. [Accessed 28 May 2004.]

Zgusta, L. 1971. *Manual of Lexicography*. The Hague: Mouton.