doi:10.1093/ijl/ecp009

CURRENT LEXICOGRAPHY PRACTICE IN BANTU WITH SPECIFIC REFERENCE TO THE *OXFORD NORTHERN SOTHO SCHOOL DICTIONARY*

D.J. Prinsloo: *Department of African Languages, University of Pretoria, (Pretoria 0002, South Africa, danie.prinsloo@up.ac.za)*

Abstract

The aim of this article is to provide a perspective on lexicographic traditions, lemmatisation strategies and lemmatisation approaches in Bantu language dictionaries from a South African point of view. It will be argued that Bantu language lexicography reflects a complex interplay of lexicographic traditions and lemmatisation approaches. The focus will be on Sepedi[1] – English dictionaries and on the analysis of the *Oxford Northern Sotho School Dictionary*, henceforth (ONSD). The ONSD will be evaluated in terms of the presumed best practices in terms of lemmatisation and against the background of the user-perspective.

1. Introduction

Lexicography of the Bantu languages is in a developmental phase. Gouws's (1990) statement that Bantu languages generally lack lexicographic quality is to a large extent still applicable after almost two decades.

> 'Lexicographical activities on the various indigenous African languages [. . . have] 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 lexicographic achievement.' (Gouws: 1990: 55)

Gouws (2007: 314), however, says that a shift has taken place from externally motivated compilation of dictionaries, for example by missionaries, to an internal drive by mother-tongue speakers of the languages to take responsibility for the compilation of dictionaries. Target users of dictionaries

International Journal of Lexicography, Vol. 22 No. 2. Advance access publication 5 April 2009

for the Bantu languages are also increasingly realizing the value of dictionaries and the South African government actively promotes the compilation of dictionaries for all eleven official languages in South Africa by means of government-funded National Lexicography Units (NLUs). Publishing houses also make a major contribution by publishing dictionaries for these languages compiled by individuals and the NLUs.

Since Gouws's 1990 observation re the status of Bantu language lexicography, lexicographic knowledge has benefited from a number of workshops, numerous publications on problematic aspects of Bantu language lexicography, the establishment of the just mentioned National Lexicography Units and the dawn of the corpus era for Bantu languages. Central to Bantu language lexicography is lexicographic debate and decisions in respect of

(a) lemmatisation approaches
(b) orthography of the language
(c) lexicographic traditions and
(d) lemmatisation strategies

that are unique to the Bantu languages. The Bantu language lexicographer not only has to deal with all of these aspects, but he or she also has to consider the complex interplay within (a) to (d) for each dictionary to be compiled in order to fulfil the needs of the respective target users. The aim of this article is thus to contextualise lemmatisation approaches, lexicographic traditions and lemmatisation strategies in terms of the relevant issues in each case. In addition, the article suggests how those approaches, traditions and strategies could be harmonised, especially in terms of the lemmatisation of nouns and verbs in Bantu languages which represent by far the most lemmas in Bantu languages dictionaries. The article also attempts to position and evaluate the ONSD in terms of these aspects.

2. Lexicographic traditions, lemmatisation approaches and lemmatisation strategies

Given the strictures of length, these issues will only be briefly outlined in order to enable categorization of select Sepedi—English dictionaries and the ONSD in particular. Table 1 reflects the most relevant relations categorically in terms of columns A–E and rows 1–5.

2.1 *Lemmatisation approaches*

What is referred to, for lack of a better term, as the *traditional* approach is a situation where a dictionary compiler adds words to the dictionary as he or she

Table 1: Lemmatisation approaches, lexicographic traditions and lemmatisation strategies in Bantu languages

	A Lemmatisation approaches	B Orthography of the language	C Lexicographic traditions	D Lemmatisation strategies : verbs	E Lemmatisation strategies : nouns
1	Traditional	Disjunctive	Stem tradition	Strict stem	Strict stem
2	Rule-orientated	Conjunctive	Word tradition	Left-expanded stem	Left-expanded stem
3	Paradigm				Singular only
4	Frequency				Singular and plural
5					First and 3rd letter

Table 2: Guidelines for looking up derived forms of verbs in the PUKU 2 (Preface)

Suffix:	Perfect form:	Look up under present tense form:
-dile:	badile	bala
-ditše:	biditše	bitša
-etše:	rapetše	rapela
	robetše	robala
-itše:	bešitše	beša
	bontšhitše	bontšha
	lesitše	lesa
	hlatswitše	hlatswa

encounters them. De Schryver and Prinsloo (2000a) provide examples of the consequent inconsistency in the treatment and obvious omissions in the lemma lists of dictionaries compiled without a corpus. *Rule-orientated* dictionaries, by contrast, deliberately limit lemmatisation, especially the treatment of derivations, by such strategies as lemmatising stem forms and giving sets of derivation rules which, if applied correctly, should at least guide the user to the stem form from where he or she can start the information retrieval process. Table 2 cites a subset of rules given in the Preface of *Pukuntšu* (Kriel and Van Wyk 1989, henceforth PUKU 2) that are required for looking up derived forms of verbs. In this case, perfect suffixes need to be stripped (with the help of

the guidelines provided) in order to isolate the stem which can then be looked up.

The *paradigm* approach could be described as an urge to physically include all derivations either as lemmas or as sub-lemmas as in Ziervogel and Mokgokong's *Comprehensive Northern Sotho Dictionary*, 1975 (CNSD) as in (1).

(1) CNSD

BALA (**–bala, –badilê, –balwa, –badilwê**), cf. **PÁLA, THOMA, ŠIMOLLA,** begin, aanvang, uittart // begin, commence, provoke; *ga se ba ~ le go lema* hulle het nog nie eers begin ploeg nie // they have not even started ploughing; *re lwelê ka gobane o ilê a mpala* ons het baklei omdat hy my uitgetart het // we fought because he provoked me; **mmádi** pl. **babádi** pers. dev.; beginner, uittarter // beginner, provoker; **pálo, (n–)/di– (palô)** man. dev.; begin, aanvang, uittarting // beginning, commencement, provocation; *BÁDÍŠA* **(–badiša, –badišitšê, –badišwa, –badišitšwê)** caus.; **mmádíši** pl. **babádíši** pers. dev.; **pádíšo, (n–)/di– (padišô)** man. dev.; *BÁDÍŠANA* **(–badišana, –badišane, –badišanwa, –badišanwe)** caus. rec.; **babádíšani** pers. dev.; **pádíšano, (n–)/di– (padišanô)** man. dev.; *BÁLÁNA* **(–balana, –balane, –balanwa, –balanwe)** rec.; mekaar pla, mekaar uittart // provoke one another, tease one another; **babáláni** pers. dev.; **páláno, (n–)/di– (palanô)** man. dev.; *BÁLÉLA* **(–balêla, –balêtše, –balêlwa, –balêtšwe)** appl.; **mmálédi (mmalêdi)** pl. **babálédi** pers. dev.; **pálélo, (n–)/di– (palêlô)** man. dev.; *BÁLÉLANA*

In (1) the lexicographer attempts to give all derived forms of *bala*, for example, *badiša*, *balana*, *balela* and *balelana* as well as their respective perfect, passive and passive plus perfect forms. It is not surprising that semantic information tends to get lost in the process. There are, for instance, no translation equivalents for *badiša, badišana* or *balela.*

Lexicographers following a *frequency* approach shown in Table 1 select lemmas, and especially derived forms, on their frequency in the corpus, cf. detailed discussion in terms of the ONSD below.

2.2 *Orthography of the language*

A conjunctive orthography versus a disjunctive way of writing has major implications for lemmatisation. For disjunctively written languages, such as Sepedi, Setswana, Sesotho, Tshivenda and Xitsonga, lemmatisation is non-problematic and the ratio of token versus lemma is almost 1-1. In Table 3 the

Table 3: Conjunctivism versus disjunctivism

Sepedi	ba a mo thuša	*ba*	*a*	*mo*	*thuša*	
	'They help him/her'	they	[pres.]	him/her	help	
	go be go le motho	go	be	go	le	motho
	'There was a person'	there	was	there	is	a person
isiZulu	bayamsiza	*ba-*	*-ya-*	*-m-*	*-siza*	
	'They help him/her'	they	[pres.]	him/her	help	
	kwakungumuntu	kwa	(be)	ku	ng(u)	umuntu
	'There was a person'	there	was	there	is	a person

four orthographic words/tokens *ba a mo thuša* in the disjunctively written Sepedi orthography have a single orthographic word *bayamsiza* as equivalent in the conjunctive isiZulu orthography. These four Sepedi tokens also correspond to four separate lemmas in Sepedi dictionaries namely *ba, a, mo*, and *thuša*. In the case of *bayamsiza*, one orthographic word corresponds to the four lemmas *ba-*, *ya- m-*, and *-siza*. The same applies to *go be go le motho* versus *kwakungumuntu*.

For the conjunctively written languages, for example, isiZulu, isiNdebele, isiXhosa and Siswati, complex lemmatisation processes to isolate stems, affixes and concords are required. In most cases orthography has a direct bearing on lexicographic traditions in Bantu lexicography.

2.3 *Lexicographic traditions*

The *word* tradition is followed for most dictionaries of the disjunctively written languages and a *stem* tradition for the conjunctively written ones. A perception that stem lemmatisation is somewhat superior to word lemmatisation has resulted in a number of dictionaries of disjunctively written languages also being compiled on a stem principle. Van Wyk (1995) strongly condemns this perception and is supported by Prinsloo and De Schryver (1999) and Gouws and Prinsloo (2005a), who point out that the stem approach is not only user-unfriendly but also unnecessarily introduces difficulties regarding stem identification in disjunctively written languages.

2.4 *Lemmatisation of verbs*

There is no tension between the stem and word traditions in respect of the lemmatisation of verbs. Lexicographers of conjunctively as well as disjunctively

Table 4: Infinitive versus imperative stem forms in isiZulu and Sepedi

Stem	Infinitive	Imperative
IsiZulu: -hamba 'go, walk'	ukuhamba 'to walk'	Hamba! 'Go!'
Sepedi: sepela 'go, walk'	Go sepela 'to walk'	Sepela! 'Go!'

written languages agree that stem lemmatisation is the best option. Lemmatising stem forms of verbs in particular makes sense for the conjunctively written languages, because a huge number of prefixes combine *freely* and *productively* with verbs in a conjunctive orthography, such as subject concords, object concords, negative morphemes, the progressive, the potential, future, etc. It would be totally redundant to attempt lemmatising each verb stem plus prefixes separately. So, for example, the forms *ngiyafunda* 'I am studying', *bayafunda* 'they are studying' *asifundi* 'we are not studying', *uzofunda* 'he will study', etc., in isiZulu are all lemmatised under the stem *-funda* 'learn'. Likewise for *bayamsiza* as shown in Table 3 the lemma would be *-siza*. The traditional view is that the infinitive forms of verbs should be lemmatised. This approach is debatable because the imperative forms may also be chosen for this purpose since these resemble the basic stem form more closely as shown in Table 4.

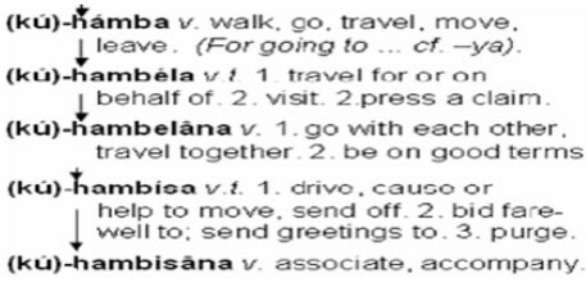
Alternatively, a total abstraction option could be utilised, that is, *hamba* and *sepela* not linked to any modal category.

In the case of verbal suffixes however, verb stems plus suffixes should be lemmatised separately to avoid very long articles where treatment of the numerous derivations is attempted under a single stem form, for example, as in (2) in the *Popular Northern Sotho Dictionary* (POP) in contrast to (1) above.

(2) POP
badiša cause to read/count . . .
bala read; count, reckon; include
balêga be counted
balêgê, go se ~ innumerable
balêla read/count for . . .
balola recount . . .
balwa be read, counted, ~ **le** including

Left-expanded stem lemmatisation for verbs as described by Gouws and Prinsloo (2005) is the lemmatisation of the verb stem with the infinitive prefix, for example, *kuhamba* 'to walk' in Siswati. The alphabetical ordering runs on the first letter of the stem with the infinitive prefix left expanded as for *hamba* and its derivations in Rycroft's *Concise SiSwati Dictionary* (CSD) in (3).

(3) CSD



The ONSD treats verb stem forms as well as derivations as separate lemmas. This is also the best approach in a school dictionary. From user surveys it became clear that learners generally lack sufficient knowledge of the morphology of verbs to isolate the verb stem, cf. Gouws and Prinsloo (2005a: 40) for a detailed discussion. In addition, the compilers also did not hesitate to include verbs with the relative suffix *-go* on an ad hoc basis justified by very high frequency of use in Sepedi.[2] The ONSD also utilises the so-called *ga*/*sa*/*se* convention designed by Prinsloo and Gouws (1996) and introduced in the fourth revision of the Popular dictionary (POP) as well as in the *New Sepedi* (NSE) and *Nuwe Sepedi* (NSA) articles. This convention covers, in a user-friendly way, the eleven possible meanings that could be conveyed by Sepedi verb stems ending in *-e*, for example, *thuše* in (4) and (5).

(4)

a. ... not helping; b. if/while ... not helping; c. who are not helping; d. so that ... must help; e. so that ... must not help; f. not to help; g. ... usually help; h. ... usually do not help; i. and then ... did not help; j. help him!; k. do not help him! (Prinsloo and Gouws 1996: 102)

(5) NSE

thuše, thušê must help; ..ga/sa/se.. not help

The *ga*/*sa*/*se* convention is utilised in POP and ONSD for the lemmatisation of highly used inflected forms of verbs. For example, the verb *thuša* 'help' as well as its frequently used inflected form *thuše* will be lemmatised in dictionaries where the target users are presumed not to be familiar with the modal system, negation and inflection strategies of the language.

2.5 *Lemmatisation of nouns*

Tension exists between the word and the stem traditions in respect of the lemmatisation of nouns. Unlike verbs, prefixes do not combine *freely* and *productively* with nouns, but the possible combinations are limited to but a few in each case. Van Wyk (1995) pays detailed attention to this misconception and

possible other reasons why lexicographers assume that verbs and nouns have to be treated in the same way, namely, to lemmatise nouns in conjunctively written and even disjunctively written languages on their stem form. He says that it is important to note the difference between nouns and verbs when it comes to affixes (prefixes and suffixes). First, only a very limited number of prefixes can combine with noun stems and, secondly, it is not wise to remove nominal prefixes in the disjunctively written languages in the process of lemmatisation.

> 'The basic assumption of stem dictionaries is that the morphology of the verb and the noun is identical in that prefixal elements can be attached freely to stems in both cases [...] This assumption is, however, wrong; *the morphology of the noun differs in crucial ways from that of the verb. The noun prefix is not mobile or freely exchangeable* [...] Any verb root can be combined with any subject marker, any modal or aspectual morpheme [...] None of this applies to the noun [...] The crucial difference with verbs is that *noun class prefixes are combined largely in an ad hoc manner with stems* [...] This results in *a fundamentally different handling of verbs and nouns in stem dictionaries* [...] This means [...] that separate entries must be made for each combination of a prefix plus a stem.' (Van Wyk 1995: 86–88, original emphases)

Lemmatising noun stems is not user-friendly especially for inexperienced users and learners of the language and it introduces unnecessary problems in respect of stem identification. More importantly, Van Wyk (1995: 88, 91–92) has shown in a critical review of CNSD that in following this approach the compilers did *not* manage to avoid repetition due to – among others – irregular forms, but rather introduced redundancy by having to resort to unnecessary cross-referencing.

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

Prinsloo and De Schryver (1999: 261) point out that the user is unnecessarily burdened with numerous problems relating to isolating the stem in many problematic instances such as *ngwana* (**mo-ana*) 'child', *mmušo* (**mo-bušo*) 'government', *muši* (**mo-uši*) 'smoke', where the noun stem is no longer synchronically identifiable. In some cases, (such as stems containing the nasal prefix of class nine or aspirated and non-aspirated noun stems), it is simply not possible for either the user or the lexicographer to determine unambiguously what the form of the isolated stem is.

Lexicographers for the disjunctively written languages need not follow the stem lemmatisation tradition for the sake of tradition, nor should they assume that stem lemmatisation is more 'scientific' than word lemmatisation. Van Wyk (1995: 85 and 95) rejects the validity of such an assumption with detailed explication.

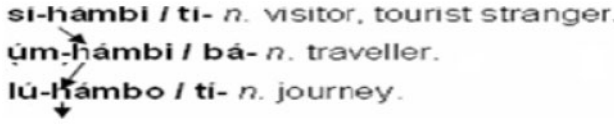
Strict stem lemmatisation entails the lemmatising of nominal stems and, generally, the addition of the singular and plural prefixes as in (6) from the *Scholar's Zulu Dictionary* (SZD).

(6) SZD

-bhashu (isi- izi-) (n) burnt patch.
-bhasi (i- ama-) (n) bus.
-bhasikidi (u- o-) (n) basket.

Left-expanded stem lemmatisation of nouns entails lemmatisation of the full noun but with the alphabetical ordering running on the stem with nominal prefixes left expanded as for *sihambi, umhambi* and *luhambo* in Siswati in (7).

(7) CSD



Gouws and Prinsloo (2005: 44) state that left-expanded article structures offer a solution to cases where stem identification is difficult or impossible.

Lemmatising only *singular forms* of nouns substantially combats redundancy but is heavily dependent on the application of sets of rules as in PUKU 2 given in Table 5 to enable successful information retrieval especially by inexperienced learners (cf. De Schryver and Prinsloo 2000a for a detailed discussion.) At face value, rules guiding the user from the plural to the singular do not appear to be complicated. However, in the case of the Class six plurals, *ma-* in Table 5, corresponding singular forms could be lemmatised in three different alphabetical stretches namely *le-*, *bo-* and *bj-*, and the situation is complicated by substitution of plural prefix with the singular prefix versus mere omission of the plural prefix.

Lemmatising *singular and plural* forms is user-friendly, especially for the inexperienced learners. However, redundancy becomes a factor, especially in dictionaries that offer treatment of both the singular and plural form as in (8).

(8) NEN

ba'sadi, n. pl., of **mosadi**, women.
mo'sadi, n. a woman, a wife . . .

For the lemmatisation of nouns the compilers of the ONSD opted for the most user-friendly option as (9a), that is, lemmatising both singular and plural forms of nouns as suggested by Prinsloo and De Schryver (1999) and Gouws and Prinsloo (2005a: 84–85). Compare (9b).

Table 5: Rules for looking up nouns in the PUKU 2

Rule		Example	
word starts with	*look word up under*	*word starts with*	*look word up under*
ba-	mo-	basadi	mosadi
bab-	mm-	babetli	mmetli
bo-	(the stem)	bomalome	malome
di-	se-	dilepe	selepe
(the stem)	dikgomo	kgomo	
ma-	le-	maleme	leleme
bo-	maleke	boleke	
mabj-	bj-	mabjang	bjang
mabo-	bo-	mabothata	bothata
me-	mo-	mello	mollo
meb-	mm-	mebutla	mmutla
mef-	mph-	mefoma	mphoma
mengw-	ngw-	mengwaga	ngwaga
nyw-	ngw-	nywako	ngwako

(9)

a. ONSD	b.
moriri *noun* ***3/4*** (*pl.* **meriri**) ■ hair • Mokgaetši o kama **moriri** ka sekamo. *Mokgaetši combs her hair with a comb.* **meriri** *pl. noun 3/4* See sg. MORIRI	**meriri** n. cl. **3/4** LHL hair (on the head) (plural), *meriri e mešweu ke lehumo* grey hair is a treasure; **~ wa titsana** soft hair, **motho wa ~** a reliable person **moriri** n. cl. **3/4** LHL (one) hair | see **meriri**

Note that in (9b) it is suggested that the treatment be given for the most frequent member of the singular/plural pair and even that the less frequent member be given in a *smaller font* with skeleton treatment of the lemma.

Singular forms of nouns are treated in the ONSD. However, if the plural form is overwhelmingly more frequent, treatment is given at the plural form. This approach is in line with the more radical approach suggested by Gouws and Prinsloo (2005a), giving the treatment at the more frequently used member of the pair. For example, for *meriri* versus *moriri*/*meriri* in 9b treatment is given at the plural form which is more frequent in the Pretoria Sepedi Corpus (PSC)[3] than the singular form.

Lemmatising both singular and plural forms is especially recommended for learners dictionaries. This, however, comes at a huge price in terms of redundancy of space taken up by lemmatising the other member of the pair, usually the plural forms. Once again the compilers of the ONSD took the best option, that is, lemmatising the plural forms, and instead of treating them, they supplied a cross-reference to the singular form as in (9a).

Lemmatising plural forms with cross-referencing to the singular forms results in overuse of the mediostructure as lexicographic device, rendering sections that consist entirely of cross-references as in (10).

(10) ONSD

a.	b.
diphedi * *pl. noun 7/**8*** See sg. SEPHEDI **diphefo** *pl. noun 9/**10*** See sg. PHEFO **diphego** *pl. noun 9/**10*** See sg. PHEGO[1] **dipheko** *pl. noun 9/**10*** See sg. PHEKO **dipheta** *pl. noun 9/**10*** See sg. PHETA **diphetho** *pl. noun 7/**8*** See sg. SEPHETHO **diphetogo** *pl. noun 9/**10*** See sg. PHETOGO	**maphodisa** ** *pl. noun 5/**6*** See sg. LEPHODISA **maphoto** *pl. noun 5/**6*** See sg. LEPHOTO **mapogo** *pl. noun 5/**6*** See sg. LEPOGO **marago** *pl. noun 5/**6*** See sg. LERAGO

This, however, is defensible. First, very little space is used; often not exceeding a single column-line. Spelling and frequency guidance are given together with other morphological information, showing how each form is linked with the correct singular form. The relation among different forms of a word is a problem in dictionaries, such as the PUKU 2 where users are misled by the rules given as to how to look up plural nouns under their singular forms: for instance, *meno* 'teeth' > *mono* 'finger' and *meetse* 'water' > *moetse* 'mane'. Here, the inexperienced user is misguided from *teeth* to *finger* and from *water* to *mane* as a result of irregular singular/plural forms of the nouns. (See Prinsloo 1990 for a detailed discussion.)

Lemmatising on the first and third letter is an experiment by Snyman (1990) in *Dikišinare ya Setswana English Afrikaans* (DS). It has certain advantages for the inexperienced learner of Setswana, but can be frustrating to the user, because there are always two options to choose from when looking up nouns.

(11) DS

a. Lemmatised under third letter:
kwálô, le- ma- *dev* < ***kwala***, letter//brief; **lo- di-**, book//boek; **mo- me-**, handwriting, orthography//handskrif, skryfwyse

b. Lemmatised under first letter:
mmútla *pl* **mebútla**, hare//haas

In terms of Table 1, given earlier in this article, the ONSD can be classified as A4:B1:C2:D1:E4, that is, Frequency : disjunctive : word tradition : strict stem (verbs) : singular and plural (nouns). Consider also the classifications of selected dictionaries by Kriel, Van Wyk, Ziervogel and Mokgokong, Mabille and Dieterlin and Rycroft in terms of these criteria:

- **Kriel: *Pukuntšu* (PUKU 1) and *Popular* (POP) dictionaries:** A1:B1:C2:D1:E4, that is, Traditional : disjunctive : word tradition : strict stem (verbs) : singular and plural (nouns)

- **Ziervogel and Mokgokong:** ***Comprehensive Northern Sotho Dictionary*** **(CNSD)**: A3:B1:C1:D1:E1, that is, Paradigm : disjunctive : stem tradition : strict stem (verbs) : strict stem (nouns)
- **Rycroft:** ***Concise SiSwati Dictionary*** **(CSD):** A1:B2:C1:D2:E2, that is, Traditional : conjunctive : stem tradition : left-expanded stem (verbs) : left-expanded stem (nouns)
- **Mabille and Dieterlen:** ***Sesotho Dictionary*** **(SSED):** A1:B1:C1:D1:E2, that is, Traditional : disjunctive : stem tradition : strict stem (verbs) : left-expanded stem (nouns)
- **Kriel and Van Wyk:** ***Pukuntšu Dictionary*** **(PUKU 2):** A2:B1:C2:D1:E3, that is, Rule-orientated : disjunctive : word tradition : strict stem (verbs) : singular only (nouns)

3. A brief synopsis of available Sepedi — English dictionaries

The *Oxford Bilingual School Dictionary: Northern Sotho and English* (ONSD) is the latest addition to the bidirectional English — Sepedi bilingual dictionary market. A comprehensive list of Sepedi dictionaries is given in Prinsloo and De Schryver (2007). The dictionaries of the pioneer T.J. Kriel, especially the *New Northern Sotho Dictionary* and the numerous editions of the *Popular* dictionary dominated the scene for many years. These dictionaries were supplemented by a small dictionary, the *New Sepedi English dictionary* (NSE) by Prinsloo and Sathekge in 1997. The latest addition prior to the ONSD is the *Sesotho sa Leboa/English Pukuntšu dictionary* of the Sesotho sa Leboa National Lexicography Unit. By far the most comprehensive Sepedi dictionary to be compiled is the *Comprehensive Northern Sotho Dictionary* (CNSD) by Ziervogel and Mokgokong (1975), a monodirectional Sepedi — English/Afrikaans dictionary.

4. Affordability as limiting factor for Bantu language dictionaries

Bidirectional dictionaries bridging English with a Bantu language in South Africa are currently caught up in a triangulation of number of lemmas versus exhaustiveness of treatment versus price. This simply means that 500–600 pages are the default limit within which the compiler can operate as prescribed by the publishers. In principle, these limitations leave the compiler with two basic options: the inclusion of a large number (e.g., 20,000–30,000) of lemmas with limited (e.g., 1–2 lines double column) treatment, or a limited number (e.g., 10,000) of lemmas with more exhaustive (e.g., 5–7 line) treatment. The market price is normally limited to R100 per dictionary. The *Popular* dictionaries, for example, include an impressive 28,000 lemmas (14,000 for each section of the dictionary), but the treatment is limited to one or more

translation equivalent. Thus it is only suitable for basic decoding (text reception) purposes. The ONSD provides extended/exhaustive treatment but consequently lemmas are limited to approximately 5,000 in the Sepedi to English section and 5,000 lemmas in the English to Sepedi section.

Consider the randomly selected section starting with *ntlo* 'house, hut' and its treatment in the POP, NEN, CNSD and *Sesotho sa Leboa/English Pukuntšu Dictionary* (SLEPD) versus the ONSD in (12).

(12)

a. POP

ntlo house, hut
ntlogêla go away from me, leave me alone
ntlogêle, ntlogêlê(ng) must go away from me, leave me alone; **..ga/sa/se..**~ not leave me alone
ntlong in/at the hut/house
ntlwana little house/hut
ntlwanêng in/at the little hut/house
ntô a single object
ntôbiša cause me to suffer a loss
ntoma bite me; ~ **tsebe** tell me a secret
ntoo one
ntoto penis
ntotoma pile
ntotompane five-stones (a game); puffed up
ntša eat (rob) me

b. NEN

n'tlo, n.., pl., **di** — or **ma** —, or **matlo,** house, hut, family.
ntloditše, v., has anointed me, has smeared something on me.
n'tlokgolo, n., big house, palace.
n'tlo'maleke'leke, n., sky-scraper.
n'tlo'phahlo, n., store, shed, store-room.
n'tlose'edi, light-house.
n'tlose'etša, light-house.
n'tlwana, n., small house, (building).
n'tobiša, v., cause me to suffer a loss.
n'to'o, adj., one.
n'to'tolo, nto to. lô, n., ash-heap, scrap heap.
n'toto'mana, adj., big, large, puffed up n., heap, mound.
n'toto'mpane, n., five-stones (a game played by girls).

c. SLEPD

ntlo *n* house; home
ntlogele *v* leave me
ntlogeleng *v* leave me
ntlokgethwa *n* church
ntloleswiswi *n* prison; gaol
ntlong *adv* in the house
ntlongkgethwa *adv* in the church
ntlwana[1] *n* small house
ntlwana[2] *n* toilet
ntlwaneng *adv* at the toilet

d. ONSD

ntlo *** *noun* **9**/10 (*pl.* **dintlo**) ■ house
• Khunedi o agile **ntlo** ye kgolo kgauswi le sekolo. *Khunedi has built a big house next to the school.*
▸**ntlong** /ntlông/ (*pl.* **dintlong**) ■ to/in the house
ntlogela /ntlogêla/ *object concord 1p sg + verb + intrans.-reversive (og) + applicative (el)*
c TLA[1] ■ leave me behind; leave me alone
• Le se ke la **ntlogela,** ke tloga le lena. *Don't leave me behind; I'm going with you.*
ntlogele /ntlogêle, ntlogêlê/ *object concord 1p sg + verb + intrans.-reversive (og) + applicative (el)* c TLA[1] ■ (must) leave me behind; (must) leave me alone
◇**ga/sa/se** (...) **ntlogele** ■ not leave me behind; not leave me alone • Le se **ntlogele** hle, ke sepela le lena nkemeleng. *Please do*

e. CNSD

-TLO, n–/**di**–, Pb. **ntlo,** Sek. **ntlô,** cf. **NGWAKO,** hut, huis // hut, house; ~ *ga e je yê nngwê* familiebesittings behoort net aan hulle alleen, 'n mens moet vir homself sorg; die hemp is nader as die rok // family property cannot be shared by others; one must fend for oneself; charity begins at home; ~ *ya lerole ga e tswale kgôši* eendrag maak mag // unity is strength; †~ *–bojêlô* eetsaal // dining hall; †~ *–kgêthwa* tabernakel, kerk // tabernacle, church

Commercially the ONSD, selling at a very reasonable price of approximately R100 (€ 8), is in competition with the POP, NSE and the SLEPD in particular. The POP is cheaper than the ONSD and offers three times as many lemmas but is limited to offering only minimum receptive information. The NSE offers fewer lemmas than the ONSD and only minimum receptive information but is half the price of the ONSD. Finally, the SLEPD contains fractionally more lemmas than the ONSD but also provides minimum receptive information, and the open line between articles wastes valuable dictionary space.

What is thus seriously missing in Sepedi – English bilingual lexicography are dictionaries, or at least one dictionary covering the top 15,000–20,000 words on each side with a fairly rich microstructure suitable for text production purposes. Until such a dictionary is compiled and is affordable to the target users, all other dictionaries in the lower categories will be expected and exploited to fill this publishing gap, and may be unfairly judged for what they cannot be for the user. One hardly needs a survey amongst dictionary users to know that dictionaries that are so limited in scope will not satisfy the needs of learners of Sepedi – no learner of English, French or German, for example, will be satisfied if the most comprehensive dictionary available for their text production needs contains a maximum of 5,000 lemmas, which can hardly cover the highest frequencies marked with diamonds and stars in *Macmillan English Dictionary for Advanced Learners* (MED) and *Collins COBUILD English Dictionary* (COBUILD).

Against this background, the ONSD is a dictionary of limited coverage in terms of the number of lemmas for both the Sepedi and English components, but it is none the less a work of exceptional achievement in the category of 'school dictionary' for which it was designed. It will furthermore be argued below that this dictionary is of high quality in terms of implementing sound strategies for lemmatisation as well as of practically implementing the latest insights into lexicographic principles and practice for Sepedi.[4]

An evaluation of the ONSD in terms of the feature set Frequency : disjunctive : word tradition : singular and plural (nouns) : strict stem (verbs) given at the outset in Table 1 follows.

5. Frequency considerations

The significance of frequency as an important criterion is contestable but the following statistics for English and Sepedi, for example, underline the significance of frequency in the selection of lemmata. De Schryver and Prinsloo (2000, 2000a and 2000b), De Schryver and Joffe (2004), all emphasize the importance of frequency of use for the compilation of dictionaries. In COBUILD, the most frequent 14,700 lemmas are marked by means of filled diamonds on a scale of five filled diamonds to one filled diamond in descending order.

Table 6: Summary of frequency band values in COBUILD

Number of filled diamonds	Lemmas per category	Totals	% of all written and spoken English
5	700		
4	1200		
(Total 5 + 4)		1900	75
3	1500		
2	3200		
1	8100		
(Total 3 + 2 + 1)		12800	20
(Total 5 + 4 + 3 + 2 + 1)		14700	95

Table 7: Types versus tokens in Sepedi

Types (Number of different words)	Total frequencies (Sum of all counts)	Tokens (Total number of words in the corpus)	% of tokens
Top 1,000	4,615,053	5,957,553	77. 5
Top 5,000	**5,250,768**	**5,957,553**	**88.1**
Top 10,000	5,462,500	5,957,553	91.7

From Table 6, it is clear that the top 1,900 lemmas represent 75% of English (tokens) and the top 14,700 an astonishing 95%. For Sepedi, the top 1,000 types represent 77.5% of the tokens and the top 10,000 types 91.7% in the PSC as in Table 7. In terms of the PSC, the ONSD with its 5,000 Sepedi lemmas has the potential to cover almost 90% of the corpus or, if generalised, 90% of Sepedi in a given context[5] and roughly the same for English coverage in terms of Table 7.

For the compilation of lemmalists for new dictionaries or for the revision of existing dictionaries, frequency lists can play a vital role in ascertaining that, on the one hand, frequently used words are not accidentally omitted and, on the other hand, that dictionary space is not consumed by articles of lemmas unlikely to be looked for by the majority of target users.

The analysis of log files reflecting the actual lookups by dictionary users (De Schryver and Joffe 2004) strongly supports the assumption that frequently used words are, in principle, the ones most likely to be looked up.

> 'If one compares the top 100 Sesotho sa Leboa searches with the ranks of the corresponding items in a frequency list derived from a 6.1-million-word

Table 8: Frequently used verbal derivations in the PSC

root →	***bolela***	***dira***	***hwetša***	***rata***	***reka***	***tseba***
↓	**(5,735)**	**(5,475)**	**(3,371)**	**(2,786)**	**(551)**	**(5,851)**
derivation						
+ *applicative*	*bolelela*	*direla*	—	**RATELA**	*rekela*	**TSEBELA**
	(76)	(508)	(0)	**(11)**	(88)	**(47)**
+ *passive*	**BOLELWA**	*dirwa*	*hwetšwa*	*ratiwa* (5),	*rekwa*	*tsebja*
	(408)	(636)	(260)	*ratwa* (126)	(122)	(441)
+ *applicative*	**BOLELELWA**	**DIRELWA**	—	—	*rekelwa*	—
& *passive*	**(6)**	**(40)**	(0)	(0)	(19)	(0)
+ *perfectum*	*boletše*	*dirile*	*hweditše*	*ratile*	**REKILE**	*tsebile*
	(767)	(910)	(671)	(151)	**(90)**	(234)
+ *perfectum*	*boletšwe*	*dirilwe*	**HWEDITŠWE**	**RATILWE**	*rekilwe*	**TSEBILWE**
& *passive*	(44)	(137)	**(57)**	**(13)**	(17)	**(10)**
+ *causative*	**BOLEDIŠA**	*diriša*	—	—	*rekiša*	*tsebiša*
	(72)	(200)	(0)	(0)	(223)	(376)
+ *causative*	**BOLEDIŠWA**	**DIRIŠWA**	—	—	*rekišwa*	*tsebišwa*
& *passive*	**(45)**	**(72)**	(0)	(0)	(27)	(63)

(De Schryver and Prinsloo 2000a: 296)

> Sesotho sa Leboa corpus, then one notices that 30 of the top 100 searches can also be found in the corpus top 100, while as many as 63 can be found in the corpus top 1 000. Clearly, users indeed look up the frequent words of the language . . ..'
>
> 'An analogous study of the top 100 English searches reveals a similar pattern . . ..' (De Schryver and Joffe 2004: 190)

Frequency of use considerations are also useful in the selection of verbal derivations in Bantu languages given the fact that several hundreds of derivations can occur for each verb stem and that many frequently used forms were omitted from Bantu language dictionaries simply because they were accidentally overlooked (cf. DS, and De Schryver and Prinsloo 2000). Table 8, for instance, reflects inconsistent lemmatisation of derived forms of the verbs *bolela* 'speak', *dira* 'do', *hwetša*, 'find', *rata* 'love', *reka* 'buy' and *tseba* 'know' where frequently used derivations given in boldface and in capital letters were omitted from the lemma list of a Sepedi dictionary.

It is clear that frequency of use also forms the basis for all lexicographic activities in the ONSD – compilation of the lemma lists, selection of examples, cross-referencing and frequency indications all point to frequency considerations as the main criterion. A comparison of three randomly selected

Table 9: Comparison of the ONSD's categories A, K, L on the Sepedi – English side with frequency counts in the PSC

Alphabetical stretch	No. of lemmas in ONSD	Lemmas >=50 in PSC	%
A	95	108	88
K	310	376	82
L (la-leletša)	221	246	90

alphabetical stretches A, K and a section of L indicates that between 82% and 90% of the ONSD's Sepedi lemmas occur 50 times or more in the PSC, cf. Table 9.

With regard to the English – Sepedi side, a comparison of the ONSD and the MED's star-rated lemmas for the alphabetical stretch G indicates that 60% of English lemmas are star-rated in the MED. There are 7,500 star-rated words in the MED: the 2,500 most common and basic English words are marked with three stars. Three-starred words in the MED not lemmatised in the ONSD in the alphabetical stretch G are *gap, gently, growing* and *growth*. Two-starred words in this same stretch not in the ONSD are *gardener, gay, genetic, giant, good-looking, governor, grace, graphics, greatly* and *guidelines*. By contrast, lemmas in the ONSD in the alphabetical stretch G not lemmatised in the MED are *Gauteng, gave up, gender equity, genet, get out/up, give up, go back/down/into/on/out/round/towards/up/with, God, good fortune/person, grain basket, grazing ground, great-grandchild, greenness, greetings, grinding stone* and *guideline*.

Ideally, the corpus lexicographer should be able to justify the inclusion or omission of each and every lemma in the dictionary. Such justification becomes quite relevant, especially when lemma lists have to be compiled for very specific or narrowly defined target-user groups, when the number of lemmas are severely restricted. Say, for example, a lemma list restricted to 3,000 lemmas has to be compiled for a dictionary for primary school children to be used mainly for reception and production purposes in respect of their prescribed text books. The lexicographer has to find a sound balance in terms of the selection of lemmata between words likely to be looked up by the target users from their prescribed work and those from general usage.

What proved to be a sound strategy was to compile a so-called domain-specific corpus for the prescribed material and then to compare frequency counts from this domain-specific corpus with frequency counts from the general corpus of the language in order to select a lemma list. De Schryver and Prinsloo (2003) in preparation of a suggested lemma list for the compilation of the *Nuwe woordeboek sonder grense* (NWSG) selected all words occurring

Table 10: Positive keys in a comparison, domain-specific versus general corpus, calculated with WordSmith Tools

WORD	FREQUENCY Domain-specific corpus	FREQUENCY General corpus	KEYNESS
LEARNERS	10,722	4	46,363.0
ACTIVITY	6,461	375	25,150.3
LEARNER	5,289	6	22,797.2
ASSESSMENT	2,580	30	10,841.3
ANSWERS	2,721	295	9,912.1
WRITE	3,190	1,455	8,381.4
HOW	7,123	12,403	8,230.9
GROUP	2,736	810	8,223.4
SCIENCES	2,064	147	7,883.6
QUESTIONS	2,468	1,002	6,750.3
ASSESS	1,504	8	6,407.9
DISCUSS	1,602	154	5,923.6
OUTCOMES	1,340	0	5,796.2

nine times or more in the domain-specific corpus and those occurring 100 times or more in the general corpus. In effect, this means that even words with zero occurrence in the general corpus were considered for inclusion in the lemma list on the basis of relatively frequent occurrence in the domain-specific corpus. This strategy has since been applied for a few other dictionary projects with similar target-user groups. Compare, for example, a domain-specific corpus of prescribed textbooks in English for junior learners against a general English corpus. All the words in Table 10 especially *learner(s)*, *assess(ment)*, and *outcomes* occur much more frequently than expected in the domain-specific corpus compared to the general corpus, and should be exhaustively treated. All the words in Table 10 are lemmatised and/or satisfactorily treated in the ONSD.

The ONSD is generally effective in terms of the treatment of homonyms and disambiguation of concords with multi-grammatical functions, such as *-a-*, *-o-*, *-le-*, etc. For *-a-*, the most ambiguous orthographic word in Sepedi, no fewer than eight lemmas are included and exhaustively treated, that is, a^1 subject concord, a^2 object concord, a^3 possessive concord, a^4 demonstrative, a^5 present tense morpheme, a^6 question particle, a^7 hortative particle and a^8 past tense morpheme — all most likely to be consulted by the target users, especially for productive use.

Guidance from incorrect to correct in the case of typical errors related, for example, to word division and spelling is given, such as *kamoka* → *ka moka* 'all', *kgaufsi* → *kgauswi* 'near', *kwišiša* to *kwešiša* 'understand', etc.

Greater sensitivity to words and meanings frequently used in oral communication could have been shown, for example, treatment of *dumela*(*ng*)! as a greeting term could be improved by including translations, such as 'be greeted!', 'good morning/afternoon/evening'. Guidance in terms of *good morning*, *good afternoon*, *good evening* should also be given, since no separate greeting terms are used in these instances. The lemma *hello* is given with translation equivalents *dumela* and *dumelang*, but the reversibility principle is not followed in this case, that is, giving 'hello' also as a translation equivalent for *dumela* in the Sepedi to English section. Translation of the example *ba mo phorole* under the lemma *mošwang* should, more accurately, be 'her' and not 'the woman'.

Isolated instances of questionable inclusion/omission of lemmata can be found in cases such as *websaeteng* 'on the Web site' but not *websaete* 'Web site'; inclusion of *meanness* (not in the MED) and the absence of *mad* (three out of five stars in the COBUILD, two out of three stars in the MED).

6. Balance in alphabetical stretches

Prinsloo and De Schryver (2002, 2005, 2007) and Prinsloo (2004) have designed so-called lexicographic rulers for regulation and measurement of alphabetical stretches. They define a ruler as a practical instrument of measurement for the relative length of alphabetical stretches in alphabetically ordered dictionaries. Rulers are designed according to the generally accepted fact that alphabetical categories in any given language do not contain an equal number of words. For example, a cursory glance at a few popular English dictionaries reveals that the alphabetical categories or alphabetical stretches for A, B, D, M, R, and C and S in particular, 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 take up only a few pages. Likewise, an alphabetical list of types generated from the PSC 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. The Sepedi Ruler is shown in Figure 1

With the apparent exception of the alphabetical stretches D, M and L, the ONSD compares well to the ruler with less than 1% deviation from the Sepedi Ruler, as shown in Table 11.

For the alphabetical stretches D and M, which are under-represented, and L, over-represented in terms of the ruler, the deviation can be explained in terms of the lemmatisation strategy for nouns. The categories D and M contain the plural class prefixes *di-*, *me-* and *ma-* and these plural forms are

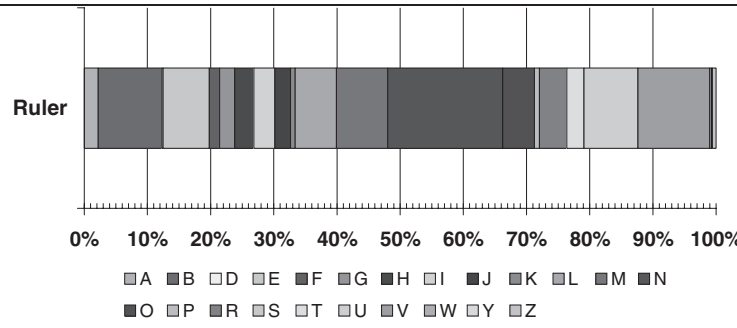


Figure 1: Sepedi Ruler based upon tokens occurring 50 times or more in the PSC.

Table 11: Alphabethical stretches in the ONSD compared to Ruler in the Sepedi – English side

	Pages ONSD	% ONSD	Ruler	ONSD vs Ruler
A	6.5	2.6	2.2	0.4
B	25.4	10.3	10.0	0.3
C	0	0	0.2	−0.2
D	9.5	3.9	7.3	**−3.4**
E	4	1.6	1.6	0.0
F	7.5	3	2.4	0.6
G	8.3	3.4	3.0	0.4
H	8.9	3.6	3.3	0.3
I	5.9	2.4	2.5	−0.1
J	0.7	0.3	0.7	−0.4
K	18.8	7.6	6.5	1.1
L	25.5	10.3	8.1	**2.2**
M	36.5	14.8	18.1	**−3.3**
N	12.4	5	5.0	0.0
O	1.8	0.7	0.8	−0.1
P	11.4	4.6	4.3	0.3
Q	0	0	0.0	0.0
R	6	2.4	2.7	−0.3
S	22.3	9	8.5	0.5
T	30.8	12.5	11.3	1.2
U	0.8	0.3	0.3	0.0
V	0	0	0.1	−0.1
W	1.7	0.7	0.6	0.1
X	0	0	0.0	0.0
Y	2	0.8	0.5	0.3
Z	0	0	0.0	0.0

cross-referenced to their singular forms (cf. 13), where elaborate treatment, also in respect of the plural forms, is given.

a.	b.
diphedi * *pl. noun 7/8* See sg. SEPHEDI **diphefo** *pl. noun 9/10* See sg. PHEFO **diphego** *pl. noun 9/10* See sg. PHEGO[1] **dipheko** *pl. noun 9/10* See sg. PHEKO **dipheta** *pl. noun 9/10* See sg. PHETA **diphetho** *pl. noun 7/8* See sg. SEPHETHO **diphetogo** *pl. noun 9/10* See sg. PHETOGO	**maphodisa** ** *pl. noun 5/6* See sg. LEPHODISA **maphoto** *pl. noun 5/6* See sg. LEPHOTO **mapogo** *pl. noun 5/6* See sg. LEPOGO **marago** *pl. noun 5/6* See sg. LERAGO

In the case of L, many cross-references from the plural class M have to be accommodated and often receive additional treatment, for example, *matswele* in both *letswele*[1] and *letswele*[2] are treated in the alphabetical stretch L instead of M according to the editorial policy of treating singular forms as given in (14).

(14)

letswele[1] *noun 5/6* (*pl.* **matswele**) ■ fist ◆ O ba bethile ka **matswele** le ka dithupa. *He hit them with his fists and with sticks.*

letswele[2] /letswêlê/ *noun 5/6* (*pl.* **matswele, mabele**) ■ breast ◆ **Matswele** a gago ke a mannyane; o tla nyantšha bjang ngwana wa gago? *Your breasts are small; how are you going to breastfeed your baby?*

For the English – Sepedi section, page allocation per alphabetical stretch in the MED as well as the 12.5 million-token University of Pretoria English Internet Corpus (PEIC), compiled by Rachelle Gauton (Taljard et al. 2007), was used as a Ruler. Once again a close correlation is observed. Table 12 reflects a comparison of the ONSD with the MED and the PEIC.

7. A brief review of additional features of the ONSD

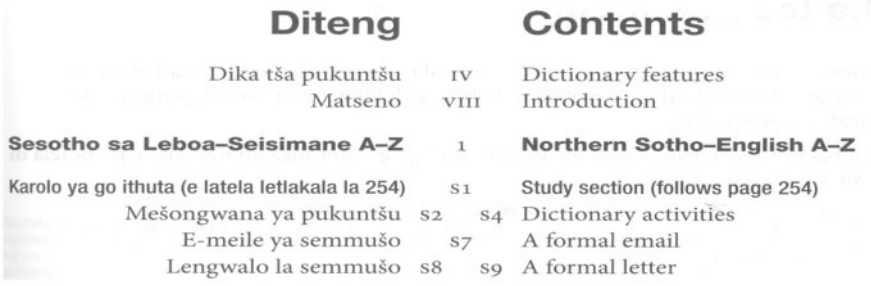
The compilers decided to use English as the metalanguage for both components of the dictionary. Using Sepedi as metalanguage could also be considered as an option in future revisions. This decision is questionable – especially in a school dictionary where all other aspects and presentations are punctiliously done on an equal basis for the two languages.

Table 12: The ONSD compared to the MED and PEIC

	Pages ONSD	% ONSD	MED pages	MED Ruler	PEIC Ruler	ONSD vs MED Ruler	ONSD vs PEIC Ruler
A	17.8	**6.2**	83	5.0	6.5	1.2	0.3
B	15.2	**5.3**	106	6.3	6.1	−1.1	0.8
C	28.6	**9.9**	156	9.3	9.1	0.6	−0.8
D	17.3	**6.0**	90	5.4	5.6	0.6	−0.4
E	12.1	**4.2**	57	3.4	3.9	0.8	−0.3
F	13.8	**4.8**	87	5.2	3.9	−0.4	−0.9
G	8.6	**3.0**	58	3.5	3.6	−0.5	0.6
H	9.6	**3.3**	69	4.1	4	−0.8	0.7
I	10.4	**3.6**	58	3.5	3.6	0.1	0
J	1.8	**0.6**	15	0.9	1.5	−0.3	0.9
K	2	**0.7**	13	0.8	1.6	−0.1	0.9
L	10.6	**3.7**	65	3.9	3.8	−0.2	0.1
M	14	**4.9**	81	4.8	6.5	0.0	1.6
N	6.2	**2.2**	34	2.0	2.4	0.1	0.2
O	8.2	**2.8**	45	2.7	2.3	0.2	−0.5
P	22.3	**7.7**	134	8.0	7.3	−0.3	−0.4
Q	1.2	**0.4**	8	0.5	0.5	−0.1	0.1
R	15.2	**5.3**	89	5.3	4.6	0.0	−0.7
S	33.5	**11.6**	210	12.5	10.4	−0.9	−1.2
T	18.6	**6.5**	95	5.7	5	0.8	−1.5
U	4.3	**1.5**	33	2.0	2.4	−0.5	0.9
V	3.9	**1.4**	18	1.1	2	0.3	0.6
W	10.7	**3.7**	62	3.7	2.7	0.0	−1
X	0.2	**0.1**	1	0.1	0.1	0.0	0
Y	1.6	**0.6**	6	0.4	0.5	0.2	−0.1
Z	0.3	**0.1**	2	0.1	0	0.0	−0.1

7.1 *Title, study section, front and back matter*

It is not clear what the exact title for reference purposes of ONSD should be: the outside cover refers to *The Oxford Sesotho sa Leboa – Seisimane English – Northern Sotho Pukuntšu ya Sekolo School Dictionary* and the first title page to *Pukuntšu ya Polelopedi ya Sekolo Sesotho sa Leboa le Seisimane E gatišitšwe ke Oxford. Oxford Bilingual School Dictionary Northern Sotho and English* and on the second title page formally with the ISBN number as *Oxford Bilingual School Dictionary: Northern Sotho and English/Pukuntšu ya Polelopedi ya Sekolo Sesotho sa Leboa le Seisimane. E gatišitšwe ke Oxford*, and the title *Oxford Bilingual School Dictionary: Northern Sotho and English* is used on the



Diteng		Contents
Dika tša pukuntšu	iv	Dictionary features
Matseno	viii	Introduction
Sesotho sa Leboa–Seisimane A–Z	1	**Northern Sotho–English A–Z**
Karolo ya go ithuta (e latela letlakala la 254)	s1	Study section (follows page 254)
Mešongwana ya pukuntšu	s2 s4	Dictionary activities
E-meile ya semmušo	s7	A formal email
Lengwalo la semmušo	s8 s9	A formal letter

Figure 2: Page references in the table of contents of ONSD.

Web site of the publisher. Listings on commercial Web sites also vary in terms of title and author reference.

The front matter of the ONSD gives a table of contents, a user-friendly explanation of the dictionary features and an introduction. The study section located between the Sepedi – English and the English – Sepedi components contains the mini-grammar[6] as well as guidance as to dictionary activities, writing of e-mails and letters, spelling and pronunciation, etc. The back matter consists of a reference section on animals, fruit and vegetables, the human body, etc. These plates and tables successfully bring together items decontextualised as an inevitable result of alphabetical ordering in dictionaries.

Reading the study section is a prerequisite for decoding certain important information when looking up words in the ONSD. So, for example, no initial easy-to-refer-to user's guide for abbreviations frequently used, such as 1p, 2p, sg., pl., PC+ Dem, etc., is given in the front matter. Thus access to sublemmata, such as [PC+] *kakanyo*, [SC+] *se kae*, [DEM+ SC+] *kgethegilego*, is subject to reading the study section.

Page references in the table of contents is somewhat confusing, cf. Figure 2.

'Dictionary features' are not found on page iv (these start on page vi) and the Introduction is not found on page viii as suggested (it is on page x). This is of course not a mistake since the intention is that the category *Dika tša pukuntšu*/Dictionary features starts on page iv and the user who wants to read the English version should page on from page iv up to where the English starts, but it is not user-friendly and is inconsistent with the approach in the S-section where a separate page indication is given, for example, the category *Mešongwana ya Pukuntšu*/Dictionary activities as S2 . . . S4. What could also be misinterpreted or be perceived as unnecessarily complex is that the Study section S1 follows page 1 if one does not note the '(follows page 254)' remark. These, however, are minor points of criticism. See De Schryver and Taljard (2007) and De Schryver (2008) for a detailed discussion of the compilation of the dictionary grammar. Marking the relevant alphabetical stretch on each page and the use of a 'single-glance' guide at the top of each page are additional user-friendly characteristics of the ONSD.

7.2 *Pronunciation*

The compilers have made a sincere effort to give pronunciation guidance by means of similar sounding English words. Some comparisons, such as 'e' versus 'i' in the guideline "**s**epela as in l**i**sten", are less successful, that is, [e] versus the common pronunciation [I]. The compilers could consider adding the IPA orthography, because it forms part of the curriculum for learners in Grade 8 and therefore will be known to many of the target users of the ONSD.

The ONSD correctly states that the circumflexed *e* and *o* are not used in everyday writing 'but should appear' in scientific texts and dictionaries... (S25). It is, however, not clear why the ONSD only indicates them in the Sepedi—English section of the dictionary and not in the English—Sepedi. Indication of circumflexes in the English—Sepedi section will support target users, especially in oral production of Sepedi.

7.3 *Text or shade(d) boxes*

This is a lexicographic device not previously used in any Sepedi dictionary and substantially enhances the quality of the treatment given in the ONSD. Shaded boxes are used to great effect in this dictionary. They give guidance in respect of lemmas treated that are not translatable, range of application, composition of multiword lemmas, spelling and word division, irregular forms, orthographic abbreviations, etc.

The series of shaded boxes highlighting the translation and use of so-called 'state of existence' (actions expressed by the past tense form of the verb continued in the present, e.g., 'sit'), however, need to be updated. In the shaded box following: *robetšego* reflecting on *robetše* it is stated: 'Although *robetše* has a perfect suffix, it is translated as a present tense verb'. However, the very example given '*ke robetše ga mogwera*...' 'I slept at my friend's...' contradicts this. Appropriate guidance to the user in this case could be given by adding another short example, for example, '*o robetše*' translated as '(s)he's asleep' or '(s)he is sleeping' to make the intended point of the shaded box clear. This is correctly done in the case of *rwele* as far as the state of existence form is concerned but no examples are given of *rwele* as a true past tense verb meaning 'carried' and also for *hloile* as 'hated'. The reason for this could be that it is less frequent and therefore omitted in terms of the policy 'gives frequently used translations only' (back cover of the dictionary). The same holds for *dutše*. However in the case of *eme*, it is translated as a present tense, but no shaded box is given. The treatment and use of shaded boxes at similar verbs, for example, *apere*, should also be checked.

Terminology used in some of the text boxes could be too difficult for the target users to interpret, for example, at *ehlwa*: 'monosyllabic auxiliary verb

stems which appear in the situative mood'. References to the moods should be supported by discussion in the mini-grammar.

7.4 *Lemmas smaller or bigger than words*

The dictionary does well to lemmatise certain multiword lemmas, such as *la ka* 'mine' *ka baka la* 'because', *ga se* '(copulative)', *ka mo go* 'here', etc. In the case of *la ka*, the rationale for lemmatisation could be found in the fact that the user should be guided against misspelling it as *laka*, which the ONSD appropriately does in the text box following *la ka* and in the inclusion of the lemma *laka* with appropriate correct-spelling-guidance to *la ka*. *Ka baka la* (1,682), *ka mo go* (345) and *ga se* (7,897), however, are apparently lemmatised, because they are frequently used, but other very frequent combinations, such as *e le* (22,314) 'being', *ka fao* (2,649) 'therefore', are not lemmatised.

No spelling errors were noted, and consistent and complete coverage of paradigms/sets of lemmas, e.g., concords, months of the year, etc., are given. The paradigm for adjectives could be extended, e.g., in the case of classes 4 and 5, *white* versus *black*. *Meso* (268) 'black' is lemmatised but not *mešweu* (31) 'white', and *leso* (101) 'black' but not *lešweu* (47) 'white'. In such cases, compilers have to make a compromise between frequency of occurrence and completion of a paradigm.

8. Conclusion

Viewed from a South African perspective Bantu language lexicography reflects a complex interplay of lexicographic traditions and lemmatisation approaches and is influenced by the orthography of the specific language. In the past decade a number of studies were undertaken to establish best practices in terms of lemmatisation, balancing of alphabetical stretches, combating inconsistencies, compilation of corpus-driven dictionaries for Sepedi, etc. The problems inherent in lemmatisation are real. These studies were performed against the background of the user-perspective. In this article it has been argued that stem lemmatisation should be avoided for nouns in disjunctively written Bantu languages such as Sepedi. An attempt was made to evaluate the ONSD on a number of these presumed best practices. School dictionaries must, by definition, be easy to use. It can be concluded that publication of the ONSD represents a new era for Sepedi — English lexicography in the sense that the latest insights, lexicographic tools, a Sepedi corpus and a state of the art dictionary writing system have been utilised. The ONSD succeeds in its aims to offer support in the key areas of helping learners choose the right translation, giving frequently used translations, showing how words are really used and the inclusion of new words from across the curriculum as well as the incorporation

of 56 pages of useful extras (a mini-grammar, activities with answers, model letters, illustrations, SMS language and more).

Notes

[1] Also referred to as Northern Sotho or Sesotho sa Leboa.
[2] Normally not done in Sepedi dictionaries but user-friendly for inexperienced users.
[3] The University of *Pretoria Sepedi Corpus* (PSC) is a collection of ca. six million running words of Northern Sotho, containing texts from different genres and domains.
[4] For an evaluation of ONSD by its editor, see De Schryver (2008).
[5] Compare De Schryver (2008: 271) and ONSD page xi for similar statistics.
[6] See De Schryver and Taljard (2007) for a detailed description.

References

A. Dictionaries

De Schryver, G-M. (Ed.). 2007. *Oxford Bilingual School Dictionary: Northern Sotho and English*. (First edition.) Cape Town: OUP Southern Africa. (ONSD).

Dent, G. R. and Nyembezi, C. L. S. 1993. *Scholar's Zulu Dictionary*. (Third edition) (First edition 1969, Second edition 1988.) Pietermaritzburg: Shuter and Shooter. (SZD).

Gouws, R. H., Stark, M. and Gouws, L. 2004. *Nuwe woordeboek sonder grense*. (First edition.) Cape Town: Maskew Miller Longman. (NWSG).

Kriel, T. J. 1976. *The New English – Northern Sotho Dictionary, English – Northern Sotho, Northern Sotho – English*. (Fourth edition.) (First edition 1950, Second edition 1958, Third edition s.d.) Johannesburg: Educum Publishers. (NEN).

Kriel, T. J. 1983. *Pukuntšu Dictionary*. (Third edition.) (First edition 1966, Second edition 1977.) Pretoria: J.L. van Schaik. (PUKU 1).

Kriel, T. J. and Van Wyk, E.B. 1989. (Fourth revised edition, cf. PUKU 1.) *Pukuntšu woordeboek, Noord-Sotho – Afrikaans, Afrikaans – Noord-Sotho*. Pretoria: J.L. van Schaik. (PUKU 2).

Kriel, T. J., Prinsloo, D. J. and Sathekge B. P. 1997. *Popular Northern Sotho Dictionary, Northern Sotho – English, English – Northern Sotho*. (Fourth edition.) (First edition 1971, Second edition 1976, Third edition 1988.) Cape Town: Pharos. (POP).

Mabille, A. and Dieterlen, H. 1988. *Southern Sotho – English Dictionary*. Revised by R.A. Paroz. Morija: Morija Sesotho Book Depot. (SSED).

Mojela, M. V., Mphahlele, M.C., Mogodi, M.P. and Selokela, M.R. 2006. *Sesotho sa Leboa | English Pukuntšu Dictionary. Cape Town: Phumelela*. (SLEPD).

Prinsloo, D. J. and Sathekge, B. P. 1996. *New Sepedi Dictionary, English – Sepedi (Northern Sotho), Sepedi (Northern Sotho) – English*. (First edition.) Pietermaritzburg: Shuter and Shooter. (NSE).

Prinsloo, D. J., Sathekge. B. P. and Kapp, L. 1997. *Nuwe Sepedi Woordeboek, Afrikaans – Sepedi (Noord Sotho), Sepedi (Noord Sotho) – Afrikaans*. (First edition.) Pietermaritzburg: Shuter and Shooter. (NSA).

Rundell, M. (Ed.). 2007. *Macmillan English Dictionary for Advanced Learners*. (Second edition.) (First edition 2002.) Oxford: Macmillan. (MED).

Rycroft, D. K. 1981. *Concise SiSwati Dictionary. SiSwati – English | English – SiSwati*. (First edition.) Pretoria: J.L. van Schaik. (CSD).

Sinclair, J. (ed.) 1995. *Collins COBUILD English Dictionary*. (First edition.) London: HarperCollins. (COBUILD).

Snyman, J. W. (Ed.). 1990. *Dikišinare ya Setswana English Afrikaans Dictionary*. (First edition.) Pretoria: Via Afrika. (DS).

Taljard, E., Gauton, R. and Gauton, L.A. 2007. Issues in the Planning and Design of a Bilingual (English – Northern Sotho) Explanatory Dictionary for Industrial Electronics. *Lexikos* 17: 1–18. (PEIC).

Ziervogel, D. and Mokgokong, P. C. M. 1975. *Comprehensive Northern Sotho Dictionary, Northern Sotho–Afrikaans/English*. (First edition.) Pretoria: J.L. van Schaik. (CNSD).

B. Other literature

De Schryver, G-M. 2008. 'Why does Africa need Sinclair?' *International Journal of Lexicography* 21.3: 267–291.

De Schryver, G-M. and Joffe, D. 2004. 'On How Electronic Dictionaries are Really Used.' In: Williams, G. and S. Vessier (eds.). 2004. *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004, Lorient, France, July 6–10, 2004*, Lorient: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud, 187–196.

De Schryver, G-M. and Prinsloo, D. J. 2000. 'The Compilation of Electronic Corpora, with Special Reference to the African Languages.' *Southern African Linguistics and Applied Language Studies* 18.1–4: 89–106.

De Schryver, G-M. and Prinsloo, D. J. 2000a. 'Electronic Corpora as a Basis for the Compilation of African-language Dictionaries, Part 1: The *Macrostructure*.' *South African Journal of African Languages* 20.4: 290–309.

De Schryver, G-M. and Prinsloo, D. J. 2000b. 'Electronic Corpora as a Basis for the Compilation of African-language Dictionaries, Part 2: The *Microstructure*.' *South African Journal of African Languages* 20.4: 310–330.

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

De Schryver, G-M. and Taljard, E 2007. 'Compiling a Corpus-based Dictionary Grammar: an Example for Northern Sotho.' *Lexikos* 17: 37–55.

Gouws, R. H. 1990. 'Information Categories in Dictionaries with Special Reference to Southern Africa.' Hartmann, R.R.K. (Ed.). *Lexicography in Africa*. Exeter: University of Exeter Press, 52–65.

Gouws, R. H. 2007. 'On the Development of Bilingual Dictionaries in South Africa: Aspects of Dictionary Culture and Government Policy.' *International Journal of Lexicography* 20.3: 313–327.

Gouws, R. H. and Prinsloo, D. J. 2005. 'Left-expanded Article Structures in Bantu with Special Reference to IsiZulu and Sepedi.' *International Journal of Lexicography* 18: 25–46.

Gouws, R. H. and Prinsloo. D. J. 2005a. Principles and Practice of South African Lexicography. (First edition.) Stellenbosch: *African Sun Media*.

Prinsloo, D. J. 1990. 'Resensie: Pukuntšu Woordeboek. (Review: Pukuntšu Dictionary)' *SA Journal of African Languages* 10, Supplement 1. Pretoria. Alasa. 109–127.

Prinsloo, D. J. 2004. 'Revising Matumo's Setswana – English – Setswana Dictionary.' *Lexikos* 14: 158–172.

Prinsloo, D. J. and De Schryver, G-M. 1999. 'The Lemmatization of Nouns in African Languages with Special Reference to Sepedi and Cilubà.' *South African Journal of African Languages* 19.4: 258–75.

Prinsloo, D. J. and De Schryver, G-M. 2002. 'Designing a Measurement Instrument for the Relative Length of Alphabetical Stretches in Dictionaries, with Special Reference to Afrikaans and English.' In: Braasch, A. and C. Povlsen (eds). *Proceedings of the Tenth EURALEX International Congress, EURALEX*. Copenhagen: Center for Sprogteknologi, Københavns Universitet, 483–494.

Prinsloo, D. J. and De Schryver, G-M. 2005. 'Managing Eleven Parallel Corpora and the Extraction of Data in all official South African languages.' In: Daelemans, W., T. du Plessis, C. Snyman and L. Teck (eds.). *Multilingualism and Electronic Language Management*. Pretoria: J.L. van Schaik, 100–122.

Prinsloo, D. J. and De Schryver, G-M. 2007. 'Crafting a Multidimensional Ruler for the Compilation of Sesotho sa Leboa Dictionaries.' *Festschrift for P.S. Groenewald*. Stellenbosch: African Sun Media, 177–201.

Prinsloo, D. J. and Gouws, R. H. 1996. 'Formulating a New Dictionary Convention for the Lemmatization of Verbs in Northern Sotho.' *South African Journal of African Languages*, 16(3): 100–107.

Van Wyk, E. B. 1995. 'Linguistic Assumptions and Lexicographical Traditions in the African Languages.' *Lexikos* 5: 82–96.