*Danie J. Prinsloo**

Electronic Dictionaries viewed from South Africa

Abstract

The aim of this article is to evaluate currently available electronic dictionaries from a South African perspective for the eleven official languages of South Africa namely English, Afrikaans and the nine Bantu languages Zulu, Xhosa, Swazi, Ndebele, Northern Sotho, Southern Sotho, Tswana, Tsonga and Venda. A brief discussion of the needs and status quo for English and Afrikaans will be followed by a more detailed discussion of the unique nature and consequent electronic dictionary requirements of the Bantu languages. In the latter category the focus will be on problematic aspects of lemmatisation which can only be solved in the electronic dictionary dimension.

1. Introduction

Lexicographers increasingly acknowledge the enormous potential of electronic dictionaries (EDs) and the piling up of such virtues dominated articles on this subject in the past decade. In a state-of-the-art article, De Schryver (2003: 163-187) lists no less than 118 advantages of EDs in terms of space and speed, graphics, audio, text corpora, multimedia corpora, accessibility, user-friendliness, etc. and many of these issues are discussed in detail by Prinsloo (2001), Bolinger (1990), Nesi (1999), Atkins (1996), Geeraerts (2000), Dodd (1989) and Harley (2000) to name but a few. The great capacity and speed characteristic of electronic products, combined with enhanced query and data retrieval technology, indeed pave the way to a new generation of dictionaries unimagined in the paper-dictionary era. It will not be attempted to discuss the advantages of electronic dictionaries over paper dictionaries in detail but rather to single out the typical innovative features listed in (1) which are relevant from a South African perspective.

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

(1) a. Pop-up access
b. Bringing together of related items
c. New routes to the data
d. Less dependency on alphabetical order
e. Fuzzy spelling
f. Intelligent extrapolation of characters keyed in
g. Audible pronunciation

Such typical innovative features will simply be referred to as 'true' or 'real' electronic features.

2. Electronic dictionaries for English

As far as EDs for English is concerned the dictionary user in South Africa can benefit from the full range of electronic dictionaries internationally available such as *Macmillan English Dictionary for Advanced Learners* (MED), *Oxford English Dictionary, Second Ed*ition (OED on CD-ROM), *Oxford English Dictionary* (OED Online), *Cambridge Advanced Learner's Dictionary Online* (CALD), *Collins COBUILD on CD-ROM*, *Merriam-Webster OnLine*, etc. These dictionaries can be utilised to their full capacity in terms of true electronic features such as those given in (1). Whether online or on CD-ROM, such dictionaries present a new world of exciting electronic features. The discussion will be limited to a few outstanding features in a single online dictionary, the CALD and an ED on CD-ROM, the MED.

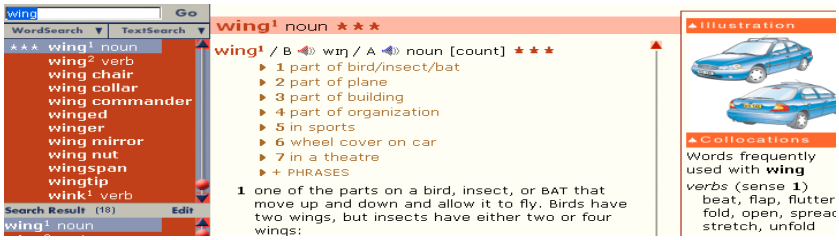
When MED is launched it immediately opens up on a random lemma which is automatically pronounced in British English and clickable options for both British and American English are provided. Audible pronunciation is an excellent example of how the ED has superseded the paper dictionary. No phonetic transcription comes close to actually *hearing*, especially problematic phonemes, such as the click sounds in Bantu languages being pronounced. Furthermore the average dictionary user in South Africa is not familiar with phonetic symbols and the IPA orthography. Adding a feature such as the self-record function that can be selected from the menu bar, MED offers the ultimate guidance in terms of pronunciation that a dictionary can give to especially learners of the language. The user's pronunciation can be recorded, played back and compared to the master recordings for British and American English.

When the user starts to type the first character(s) of the required lemma in MED, continuous intelligent extrapolation of characters is

attempted by the software. Say, for example, the user wants to look up the meaning of *intoxication*. Typing *i*, brings up the clickable lemma range **i – Iberian**, *in* triggers the range **in – inaction** while *int* returns **int. – integrity** and finally for *into*, the range **into – intoxication** is produced and the desired lemma can be clicked upon. Thus typing only 25% of the characters was required.

All words in the definitions and examples of usage are clickable and pop-up boxes appear with a definition, examples of usage and even illustrations and collocations.

Figure 1: Results of query for *wing* in MED



So-called *Smart searches* and *Sound searches* can also be performed from the menu bar, and represent excellent examples of what is referred to in (1) as ‘new routes to the data’ and ‘bringing together of related items. See Figures 2 to 4.

Figure 2: SmartSearch in MED

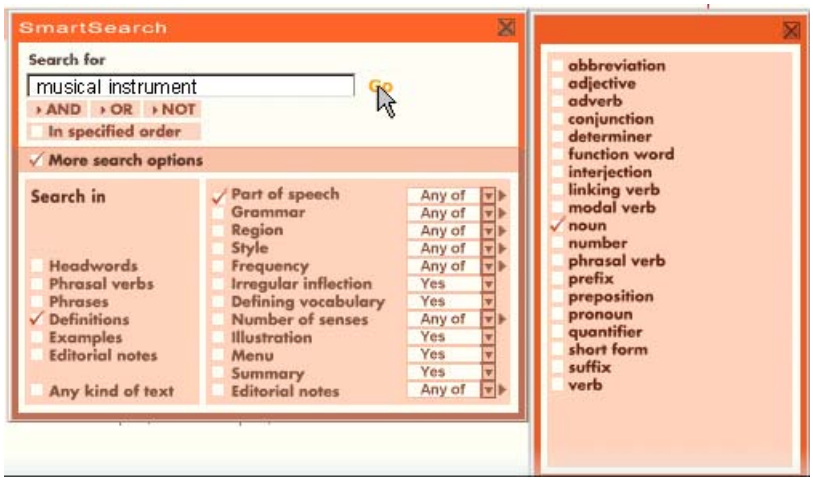
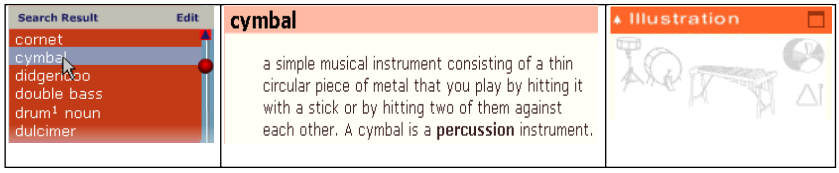


Figure 3: Result of query for *musical instrument* in MED



In Figure 3 the software response to the user's search for the unspecified item *musical instrument* is a list of musical instuments answering the user selected specified criteria including definitions and illustrations.

Figure 4: SoundSearch in MED

In Figure 4 the search is conducted on a 'sounds like' basis. As for online dictionaries for English, a simple query for *bank* in the *Cambridge Advanced Learner's Dictionary Online* returned extensive information neatly organised into 33 clickable items representing senses, homonyms, etc. related to *bank.*

Table 1: Information on *bank* in CALD

account (BANK)	bank manager	merchant bank
bank (ORGANIZATION)	the Bank of England	needle bank
bank (RAISED GROUND)	bank rate	piggy bank
bank (MASS)	bank statement	river bank
bank (MACHINES)	blood bank	savings bank
bank (TURN)	bottle bank	snow bank
state (EXPRESS)	central bank	sperm bank
bank account	clearing bank	the World Bank
bank balance	cloud bank	bank on sb/sth
bank charges	data bank	break the bank
bank holiday	fog bank	be laughing all the way to the bank

Each of these items display extensive information. Likewise, the *Merriam-Webster OnLine* offers 29 clickable entries in a pull-down menu for the lemma *bank*.

What is additionally required, for English in the South African context, however, are EDs reflecting South African English and most likely in future what is called Black South African English.

Silva (2004) states that South African English developed into a variety of English by assimilation of words and patterns from other South African languages. Dictionaries, and also EDs for English aimed at the South African market should reflect such borrowings and patterns. *A dictionary of South African English on Historical Principles* Silva (1996) represents a landmark in this regard and is a valuable source for the compilation of a true ED of South African English.

Wade (1998) lists a number of typical characteristics of Black South African English such as non-standard verb complementation, embedded questions and pronoun copying. He defines pronoun copying as instances where a noun phrase is followed immediately by a pronoun with the same referent, e.g. the parents, *they* are supposed to pay ten rands. For non-standard verb complementation he cites examples where *make* is usually followed by a 'to' infinitive rather than a bare infinitive as is illustrated in (2).

(2) Non-standard verb complementation (Wade 1998)

a. What makes them to stop that product if there are people who do come to that shop and buy them.

b. So what will we… made you to come and buy.

c. That make the meaning to be different than other countries.

d. ELS makes the second language students to be able to adapt themselves to the university.

3. True electronic dictionaries versus paper dictionaries on computer that display some electronic features

Sharpe (1995: 48), and Atkins (1996: 515-516), caution against a situation where electronic dictionaries simply use the content of printed dictionaries as their database thus not utilizing the potential of the electronic dictionary to the full.

> … dictionaries of the present … may even come to you on a CDROM rather than in book form, but underneath these superficial modernizations lurks the same old dictionary. … Will the dictionary of the future simply blip its little electronic way off into the sunset dazzling its readers with the speed which it dishes up the same old

> facts on a technicolor screen? It is up to us to take up the real challenge of the computer age, by asking not how the computer can help us to produce old-style dictionaries better, but how it can help us to create something new… Atkins (1996: 515-516)

Thus, in principle a clear decision should be made between EDs which are merely 'paper dictionaries on computer' and 'true electronic dictionaries' which utilise advanced computer technology to offer functions such as those listed in (1) that is not possible in the paper dimension.

Electronic dictionaries, for Afrikaans and the Bantu languages unfortunately fall to a large extent in the former category and much development towards the latter is still required.

For Afrikaans four electronic dictionaries, *Elektroniese WAT* (Electronic version of the Woordeboek van die Afrikaanse Taal) and *Pharos Woordeboeke Dictionaries 5-in-1* on CD-ROM and two online dictionaries *Travlangs* and *DDP Freeware* will briefly be evaluated in terms of true electronic features.

The *Pharos Woordeboeke Dictionaries 5-in-1* offers Pharos' *Major Dictionary, Bilingual Phrase Dictionary, New Words, Verklarende Afrikaanse Woordeboek* and the *Groot Tesourus van Afrikaans* on a single CD-ROM. The virtues are maximally highlighted by the publisher as follows:

> 'Whether you need guidance on spelling, meaning, synonyms, abbreviations, English and Afrikaans usage or translations, these authoritative reference sources can provide the answers. … Searches which would be time-consuming or even impossible with the printed versions can be accomplished quickly and easily in the powerful Logos Library System. … Do global searches across all five books and view the results side by side on your screen. You can find any given word in a matter of seconds. You can cross-reference easily, add your own user notes and copy-and-paste sections into your word-processor documents. Use * and ? wildcards to extend the scope of your search, to find that word on the tip of your tongue or missing from a crossword puzzle, or when you are not sure how to spell a word.'
>
> http://www.nb.co.za/Pharos/phCatalogueDisplay.asp

Even the fontsize is adjustable. All this is fine and surely offers added value but still does not offer any significant electronic features. Even the front page, title page, table of contents, etc. are exact images of the paper version. The user might still prefer to rather use the paper versions instead of 'starting-up' the computer simply to look up a few words 'on screen'.

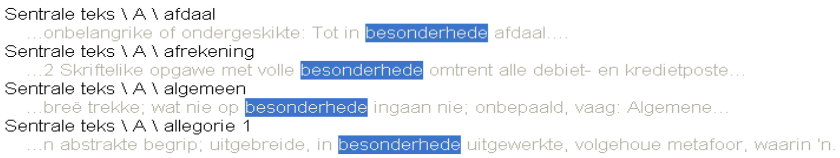
The *Elektroniese WAT* also offers certain advanced search functions and a number of cross-references, such as *oëbank* in (3) which is conveniently hyperlinked to the reference address *oogbank* that is clickable in the article of *oëbank*:

(3) Elektroniese WAT
a. oë s.nw. Selde ook, geselstaal, oge. Mv. van oog.
b. oëbank s.nw (*ongewoon*) Sien OOGBANK: *Die oëbank het 'n lys van ...*

It is good that WAT, unlike some other Afrikaanse dictionaries, did lemmatise *oë* 'eyes' which is an irregular plural for *oog* 'eye' and give a cross-reference to *oog*, where sound and elaborate treatment is offered. However, the reference address *oog* in the article of *oë*, even though it is an implicit reference, should be clickable. Since it is not, the user has to manually scroll to *oog* in some way which is not much better than paging around in the paper version. In a true electronic dictionary implicit references, in fact, all words, as in the case of MED mentioned above, should be hyperlinked to the relevant lemma.

An excellent feature in the *Elektroniese WAT* is the 'hitlist' function which generates concordance lines indicating the applicable lemma in each case.

Figure 5: Concordance lines for *besonderhede* 'particulars' in Elektroniese WAT



In Figure 5, *besonderhede* 'particulars' is given in context with 5 words of co-text on either side and it indicates that *besonderhede* occurs in the articles of lemmas such as *algemeen* 'general', *afdaal* 'descend', etc.

Elektroniese WAT overdid protection against copying by not allowing the user to copy and paste even a single word. This is nullifying one of the advantages of the electronic dictionary i.e. that users can copy and paste small sections of, or even an entire article for academic writing

purposes. Here MED is a textbook example of how it should be done namely allowing the user not only to copy an entire article but also to automatically add the source reference.

(4) **electronic** ... adjective ***
using electricity and extremely small electrical parts such as MICROCHIPS and TRANSISTORS: …

Elektroniese WAT also contains numerous untreated lemmas such as the examples given in Figure 6 reminiscent of a paper dictionary on computer. In an electronic dictionary treatment should be offered or at least clickable rerouting to the relevant lemma that is treated.

Figure 6: Untreated lemmas in *Elektroniese WAT*

onbewoë, onbewolk, onbewoon, onbewoonbaar, onbewoond, onbewus, onbewustelik, onbillik, onbloedig, onblusbaar, onbruikbaar, onbuigbaar, onbuigsaam, ondankbaar, ondeelbaar, ondefinieerbaar, ondefinitief, ondelikaat, ondenkbaar, ondenkend, ondeurdag, ondeurdringbaar, ondeurgaanbaar, ondeurgangbaar, ondeurganklik, ondeurgrondbaar, ondeurgrondelik, ondeurlaatbaar, ondeurlatend, ondeurleef, ondeursigtig,

The fact that *WAT* is currently in either paper or electronic format only completed up to the alphabetical stretch *O* in itself makes it less attractive than a full A-Z version would have been. Notwithstanding the shortcomings expressed above in terms of real electronic features, *Elektroniese WAT* remains a valuable source of information for Afrikaans.

Online dictionaries for Afrikaans generally leaves much to be desired since only a limited number of lemmas are offered and treatment is very limited. Consider (5) and (6) as typical examples.

(5) Travlang's Afrikaans-English On-line Dictionary
bank 1.bank
bankrekening 1. bank account, banking account

(6) DDP Freeware Afrikaans/English Dictionary online

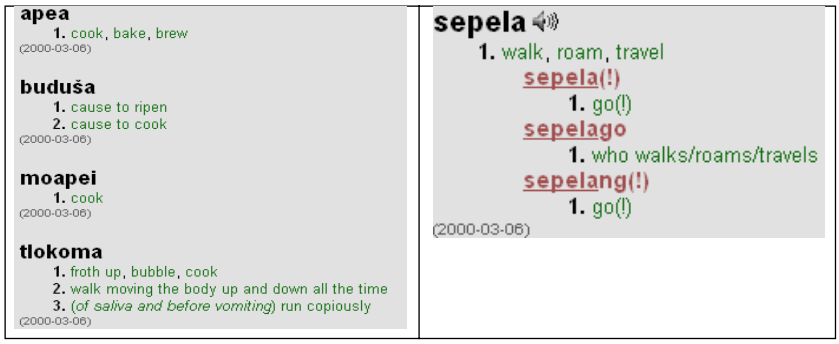
English	African
bank	oewer, bank

Compared to CALD (Table 1) and *Merriam-Webster online's* extensive treatment (5) and (6) contains very limited information, not to mention that in the latter example the name of the target language is consistently misspelt as *African* instead of *Afrikaans*.

4. Electronic dictionaries for Bantu languages – essentials or 'nice-to-haves'?

The fact that compilers of dictionaries for Bantu languages increasingly experiment with electronic and especially online dictionaries is encouraging. Unfortunately with a few exceptions, these dictionaries still offer little more than their paper counterparts or source dictionaries. Compare the following extract from the online *Sesotho sa Leboa (Northern Sotho) - English Dictionary*.

Figure 7: Online *Sesotho sa Leboa (Northern Sotho) - English Dictionary*



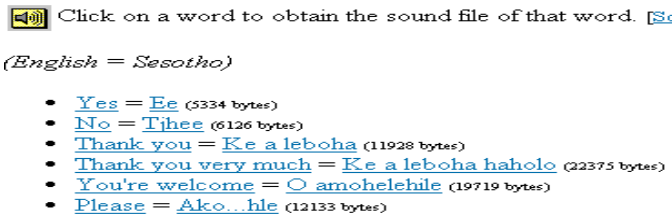
For the lemmas *apea*, *buduša*, *moapei* and *tlokoma* the dictionary offers only a number of translation equivalent paradigms. Thus no true electronic features such as those listed in (1) or added value to the paper dictionary it is based upon. However, since the paper version is mono-directional Northern Sotho → English, English words cannot be looked up. In its electronic version, English lemmas can be looked up since the software then merely collates, say, all entries containing the translation equivalent *cook* in (8). Thus a rather peculiar way of adding value, but significant for the following reasons. Firstly, the only other Northern Sotho dictionary that contains more lemmas, the *Groot Noord-Sotho Woordeboek* (Ziervogel and Mokgokong 1975) is mono-directional Northern Sotho → English/Afrikaans. Secondly, this dictionary as well as the *New English Northern Sotho dictionary*. (Kriel: 1985) is out of print for more than 10 years. Thus the online *Sesotho sa Leboa (Northern Sotho) - English Dictionary* can be regarded as the big-

gest available dictionary in the direction English → Northern Sotho, although it is a simulated direction.

For a number of words like *sepela*, in the second column of Figure 7, audible pronunciation is clickable. Ideally this option should be extended to all lemmas.

The Travlang Worldwide Travel Guides contain useful translation equivalents and phrases and are clickable for pronunciation.

Figure 8: Travlang's Worldwide Travel Guides



Consider also examples (7) and (8) for Tswana and Zulu respectively.

(7) Webster's Online Dictionary
bua speak
rata enjoy, like
robonngwe nine

(8) Zulu-English/English-Zulu online dictionary.
-thenga v. buy; purchase
njenga- prefix foll. by noun like; just as
eThekwini loc. of **iTheku** in/at/to/from Durban…

There is no doubt that the Bantu languages will benefit from all the innovative true electronic dictionary features such as those mentioned in (1) and illustrated by means of English electronic dictionaries such as MED. The real challenge for Bantu-language EDs, however, lies in a number of problematic lexicographic aspects characteristic of these languages mainly revolving around lemmatisation problems and very complicated grammatical systems. The core of the lemmatisation problem lies in a complicated derivational system in Bantu and such difficulties are multiplied if the language has a conjunctive orthography. Verbs in Bantu languages combine with numerous affixes. Van Wyk (1985: 87) calculates that a single verb in Zulu for example can have up to 18 x

19 x 6 x 2 = 4,104 combinations. Compare the following extract from a set of derivations for the verb *sebenza* (verbal root = -*sebenz*-) 'work' in Table 2 generated from the Pretoria Zulu-Corpus (PZC) and a typical example of concordance lines for Zulu verbs occurring with the prefixal cluster *wayesezo*- 'he/she would have' in Table 3.

Table 2: Derivations for the verb *sebenza* in PZC in the alphabetical sub-category a-aba

ababe**sebenz**a	aba**sebenz**ayo	abawu**sebenz**elayo
ababe**sebenz**isa	aba**sebenz**ela	abawu**sebenz**isayo
ababewa**sebenz**isa	aba**sebenz**elayo	abayi**sebenz**ayo
ababezi**sebenz**isa	aba**sebenz**i	abayi**sebenz**e
abaku**sebenz**ayo	aba**sebenz**isa	abayi**sebenz**elayo
abali**sebenz**isa	aba**sebenz**isi	abayi**sebenz**isa
abali**sebenz**ise	abasem**sebenz**ini	abayi**sebenz**isayo
abalu**sebenz**isayo	abasi**sebenz**isayo	abayi**sebenz**elayo
abanga**sebenz**i	abawa**sebenz**isayo	abayi**sebenz**isa
aba**sebenz**a	abawu**sebenz**ayo	abayi**sebenz**isayo

Table 2 lists the first 30 occurences of the alphabetically sorted derivations of the verbal root -sebenz- in PZC. Note that this list does not even go beyond the first section, *Aba*, in the alphabetical stretch *A*.

Table 3: Concordance lines for Zulu verbs occurring with the prefixal cluster *wayesezo*-

Lachamusela isu likaMjike-Joe Umona usuka esweni *Mjike-Joe's plan hatched. Jealousy lies in the eye of the beholder*	wayesezo**fika** *He would have arrived*	ekhaya Bambuyisela eGoli Leyonsebe *at home but they let him go back to Johannesburg*
khona ePrince of Wales Training College. UJabulani *there at Prince of Wales Training College. Jabulani*	wayesezo**thola** *Would have received*	izincwadi zokufundisa ekupheleni *his study material at the end of*
Sathi sehlukana noDolly wayengitshela ukuthi *Just when we said goodbye to Dolly she told me that*	wayesezo**qala** *she now began*	ukumemezela ukuthi uphethwe yisisu *to proclaim that she was pregnant*
UDlaba akafundanga okutheni, wayeka phakathi *He did not learn much and gave up in the middle*	wayesezo**sebenza** *He would by now have worked*	kwaVukusebenze. Ufike exova udaga *at Vukusebenze. He then started mixing mortar*

nje ukuthi okwakuyikhona kumphethe kabi yikuthi *in this manner, that which existed made him bad, it is because.*	wayesezo**lahl**ekelw**a** *He would have lost*	ngabantu labo ababeza kuye *those people who had come to him*
umuntu wayephumelele yini eLuhlolweni njengoba *someone was successful or not in the adjudication since*	wayesezo**qala** *He would have begun*	nje uNhlolanja. Ngazo lezozinsuku ng *in January. In those specific days*

Verb stems in Zulu for example almost always occur with one or more affixes. Traditionally Zulu dictionaries follow a stem lemmatisation strategy. This means that the lemmasign for all words in Table 2 for example will be *-sebenza* and the stems indicated in boldface in Table 3 i.e. *fika*, *thola*, *qala*, *sebenza* and *lahla*. The target users of a Zulu dictionary, especially learners of the language, are confronted with such long orthographic words and cannot look them up in Zulu dictionaries unless they know what the stem is. Isolating the stem often requires advanced knowledge of the morphological system of the language and the problem becomes critical in cases where neither the lexicographer nor the user is able to identify the stem! See Van Wyk (1985) for a detailed discussion.

Lexicographers have struggled for many decades to solve this problem by means of a variety of lemmatisation strategies. Ziervogel and Mokgokong (1975) took an approach which can be labelled an *enter-them-all-strategy* according to which they physically tried to enter all derivations of verbs. Consider the following example of the derivations actually lemmatised by them for the Northern Sotho verb *aga* 'build' which reflects 16 of the more than 30 possible suffixal clusters/derivation modules.

Table 4: Derivations of the Northern Sotho verb *aga*

1	**VR**	**aga**		VRRevtCauRecPer	agollišane
	VRPer	agile		VRRevtCauRecPas	agollišanwa
	VRPas	agwa		VRRevtCauRecPerPas	agollišanwe
	VRPerPas	agilwe	**19**	**VRAppApp**	**agelela**
5	**VRNeu-Pas**	**agega**		VRAppAppPer	ageletše
	VRNeu-PasPer	agegile		VRAppAppPas	agelelwa
6	**VRApp**	**agela**		VRAppAppPerPas	ageletšwe
	VRAppPer	agetše	**20**	**VRAppAppRec**	**agelelana**
	VRAppPas	agelwa		VRAppAppRecPer	agelelane

	VRAppPerPas	agetšwe		VRAppAppRecPas	agelelanwa
7	**VRAppRec**	**agelana**		VRAppAppRecPerPas	agelelanwe
	VRAppRecPer	agelane	**21**	**VRRevit**	**agologa**
	VRAppRecPas	agelanwa		VRRevitPer	agologile
	VRAppRecPerPas	agelanwe		VRRevitPer	agologwa
8	**VRCau**	**agiša**		VRRevitPer	agologilwe
	VRCauPer	agišitše	**28**	**VRAppAppCau**	**ageledišа**
	VRCauPas	agišwa		VRAppAppCauPer	ageledišitše
	VRCauPerPas	agišitšwe		VRAppAppCauPas	ageledišwa
9	**VRCauRec**	**agišana**		VRAppAppCauPerPas	ageledišitšwe
	VRCauRecPer	agišane	**29**	**VRAppAppCauRec**	**ageledišana**
	VRCauRecPas	agišanwa		VRAppAppCauRecPer	ageledišane
	VRCauRecPerPas	agišanwe		VRAppAppCauRecPas	ageledišanwa
13	**VRRevt**	**agolla**		VRAppAppCauRecPerPas	ageledišanwe
	VRRevtPer	agolotše	**30**	**VRAppAppAlt-Cau**	**ageletša**
	VRRevtPas	agollwa		VRAppAppAlt-CauPer	ageleditše
	VRRevtPerPas	agolotšwe		VRAppAppAlt-CauPas	ageletšwa
17	**VRRevtCau**	**agolliša**		VRAppAppAlt-CauPerPas	ageleditšwe
	VRRevtCauPer	agollišitše	**31**	**VRAppAppAlt-CaurRec**	**ageletšana**
	VRRevtCauPas	agollišwa		VRAppAppAlt-CauRecPer	ageletšane
	VRRevtCauPerPas	agollišitšwe		VRAppAppAlt-CauRecPas	ageletšanwa
18	**VRRevtCauRec**	**agollišana**		VRAppAppAlt-CauRecPerPas	ageletšanwe

VR=verbal root; Per=perfect; Pas=passive; Neu-Pas=neutro-passive; App=applicative; Rec=reciprocal; Cau=causative; Revt=reversive transitive; Revit=reversive intransitive; Alt-Cau=alternative causative

Although successful in terms of entering ‘all’ the derivations, finding the meaning of the word remains a problem for the user as is illustrated by means of *dikagollišano* in Table 5. Here the user firstly has to strip the suffixes in order to find the verb stem and its meaning and then to ‘add’ the semantic connotations in a cumulative way in order to find the meaning – thus up to 12 steps in total:

Table 5: Information retrieval process for *dikagollišano* in *Groot Noord-Sotho Woordeboek*

1	**dikagollišano**	↓	plural deverbative consisting of root + reversive transitive + causative + reciprocal + ending
2	kagollišano	↓	singular deverbative consisting of root + reversive transitive + causative + reciprocal + ending
3	agollišana	↓	verb root + reversive transitive + causative + reciprocal + ending
4	agolliša	↓	verb root + reversive transitive + causative + ending
5	agolla	↓	verb root + reversive transitive + ending
6	**aga**	↓	**verb** (stem)
7	**build**	↓	**meaning of the verb**
8	break down	↓	reverse or opposite meaning 'un-build'
9	cause to break down	↓	add causative sense of 'let/force'
10	cause each other to break down	↓	add reciprocal sense of 'each other'
11	the process of causing each other to break down	↓	nominalise: 'the process of …' (singular)
12	**the processes of causing each other to break down**		change 'the process of …' to the plural

In step 12 the user concudes that *dikagollišano* means '*the processes of causing each other to break down*' – but it is an artificially constructed meaning and (s)he is still not sure that it is the right conclusion.

A second strategy employed by Kriel and Van Wyk (1989) can be labelled the *regulate-them-in* approach. Following this approach only verb stems are lemmatised and a complicated set of rules is designed and given in the users' guide to the dictionary. In theory it means that all derivations are catered for but in practice it boils down to exactly the same process as illustrated for *dikagollišano* in Table 5. Other efforts include so-called left-expanded article structures, where an article displaying a left-expanded structure can still maintain an undisturbed alignment of the lemma sign in the vertical macrostructural ordering, as in Table 6.

Table 6:

ngingahamba	I may go
ukuhamba	to go/walk
ngangilihamba	I was traveling it
ayengasahambeli	they no longer visited
ekuhambeni	during their journey/traveling

The Zulu words in Table 6 are thus still lemmatised according to the stem principle, i.e. the root *-hamb-* in this example, but the full orthographic forms are given with vertical alignment on *h-,* within the alphabetical stretch *H* in the dictionary. Although this approach has certain advantages over strict stem lemmatisation, it does not exempt the user from the obligation to identify the stem.

Similar problematic circumstances exist for the lemmatisation of nouns. As in the case of verbs, nouns occur with affixes.

Table 7: Concordance lines for Zulu nominal cluster *nanjengomuntu*

3. (a) USean. (b) UAda. (c) UWaite njengobaba, *3. (a) Sean. (b) Ada. (c) Waite as the father*	**nanjengomuntu** *and also as a mere person*	nje. (d) UGarrick. Sebenzisa igama *Garrick. Use the name*
obusezandleni zamaphoyisa. Kodwa njengeNkosazane *which was in the hands of the police. But as the Princess*	**nanjengomuntu** ***and also as somebody***	engimethembayo ngithe angikuvezele ka *who I must trust. I thought that I should disclose it.*
kubafundi lokho akucabangile. Sekumfikele wakuloba; *to the students that he had in mind. It occurred to him to wite it down*	**nanjengomuntu** *and as somebody*	othuka inhlamba emkhandlwini. k *who uses obscene language in the assembly.*
be nguGumede onokuchaza loko njengenhloko yomuzi. *It is Gumede who is able to explain that as the head of the village*	**nanjengomuntu** *and even as a person*	obona omahlalela efika *who sees people who don't want to work*

Here the Zulu noun *umuntu* ‘a human being’ is preceded by *na-* ‘and’ plus *ngenga* ‘as, like’ and a sound change *a+u* → *o* has occurred. The user has to know that the *na*, and *njenga* should be stripped, the sound change reversed and to remove the class prefix *(u)mu-* of the noun, in order to look it up under *-ntu* and add the semantic connotations back on similar to the process in Table 5 for *dikagollišano*.

Furthermore, apart from the problem of stem identification, singularity and plurality in Bantu is indicated by prefixes. This complicates lemmatisation in alphabetically ordered dictionaries since it is extremely redundant to lemmatise each noun twice, on singular and on plural in the dictionary.

A variety of lemmatisation strategies have been attempted for nouns such as stem lemmatisation, lemmatising singular forms supplemented by rules given in the front matter of how to convert plural to singular, lemmatising both singular and plural forms, lemmatising on the third letter of the word in an attempt to avoid the noun prefix, etc. All these strategies have major disadvantages and are discussed in great detail in Prinsloo and De Schryver (1999) and De Schryver and Prinsloo (2000a and 2000b).

As a final example of a major lexicographic problem, this time on the level of complicated grammatical structures, the lemmatisation of copulatives in Northern Sotho can be cited. The English words *is*, *am*, *are* and *be* literally have hundreds of equivalents in Northern Sotho. Consider (9) as a tiny extract from the rules determining the formation of copulatives (Poulos and Louwrens 1994: 320-326) and Table 8 as an example driven table of real examples formed on the basis of such rules.

(9) ***The indicative series*** *The present tense Principal Identifying* pos lst and 2nd persons: ***SC - CB*** Classes: ***CP - CB*** neg. 1st and 2nd persons: ***ga - SC - CB*** Classes: ***ga - se - CB*** *Participial* pos. 1 st and 2nd person: ***SC - le - CB*** Classes: ***CP - le - CB*** neg. lst and 2nd person: **SC - se - *CB*** Classes: ***CP - se - CB*** *The future tense Principal* pos. 1st and 2nd person: ***SC - tlô/tla - ba + CB*** Classes: ***CP - tlô/tla - ba + CB*** neg. 1st and 2nd person: **SC - ka - se -bê + CB SC** Classes: ***CP - ka - se -bê + CB*** *Participial* pos 1st and 2nd person: ***SC - tlô/tla - ba + CB*** Classes: ***CP - tlo/tla - ba + CB*** neg 1st and 2nd person: ***SC - ka - se-bê + CB*** Classes: ***CP - ka se - be + CB*** *The past tense Principal* pos 1st and 2nd person: ***SC - bilê + CB*** Classes: ***CP - bilê + CB*** neg 1st and 2nd person: ***ga - se - SC - be + CB ga - se - SC2 - a - ba + CB ga - SC2 - a - ba + CB*** Classes: ***ga - se - CP - be + CB ga - se - SC2 - a - ba + CB1 ga - SC2 -a - ba - CB*** *Participial* pos lst and 2nd person: ***SC - bilê + CB*** Classes: ***CP - bilê + CB*** neg. lst and 2nd person: ***SC - sa - ba + CB*** Classes: ***CP - sa - ba + CB***

Table 8: Dynamic Copulatives

Column 1: MD. = MOOD, IND. = INDICATIVE, SIT. = SITUATIVE, REL. = RELATIVE, SUB. = SUBJUNCTIVE, CON. = CONSECUTIVE, INF. = INFINITIVE, IMP. = IMPERATIVE, HAB. = HABITUAL

Column 2: PRES. = PRESENT, FUT. = FUTURE, PAS. = PAST +Pot. = containing the Potential

Column3: ACT. = ACTUALITY (p. = positive, n. = negative)

MD.	**TENSE**	**ACT.**	**Common verb**	**Identifying**	**Descriptive**	**Associative**
IND.	PRES.	p.	mosadi *o reka* dipuku	*e ba* morutiši	*o ba* bohlale	*o ba* le mpša
		n.	mosadi ga *a reke* dipuku	ga *e be* morutiši	ga *a be* bohlale	ga *a be* le mpša
	+Pot.	p.	mosadi *a* ka *reka* dipuku	*e* ka *ba* morutiši	*a* ka *ba* bohlale	*a* ka *ba* le mpša
		n.	mosadi *a* ka se *reke* dipuku	*e* ka se *be* morutiši	*a* ka se *be* bohlale	*a* ka se *be* le mpša
	FUT.	p.	mosadi *o* tlo/tla *reka* dipuku	*e* tlo/tla *ba* morutiši	*o* tlo/tla *ba* bohlale	*o* tlo/tla *ba* le mpša
		n.	mosadi *a* ka se *reke* dipuku	*e* ka se *be* morutiši	*a* ka se *be* bohlale	*a* ka se *be* le mpša
	PAS.	p.	mosadi *o rekile* dipuku	*e bile* morutiši	*o bile* bohlale	*o bile* le mpša
		n.	mosadi ga se *a reka* dipuku	ga se *ya ba* morutiši	ga se *a ba* bohlale	ga se *a ba* le mpša
SIT.	PRES.	p.	ge mosadi *a reka* dipuku	*e eba* morutiši	*a eba* bohlale	*a eba* le mpša
		n.	ge mosadi *a* sa *reke* dipuku	*e* sa *be* morutiši	*a* sa *be* bohlale	*a* sa *be* le mpša
	+Pot.	p.	ge mosadi *a* ka *reka* dipuku	*e* ka *ba* morutiši	*a* ka *ba* bohlale	*a* ka *ba* le mpša
		n.	ge mosadi *a* ka se *reke* dipuku	*e* ka se *be* morutiši	*a* ka se *be* bohlale	*a* ka se *be* le mpša
	FUT.	p.	ge mosadi *a* tlo/tla *reka* dipuku	*e* tlo/tla *ba* morutiši	*a* tlo/tla *ba* bohlale	*a* tlo/tla *ba* le mpša
		n.	ge mosadi *a* ka se *reke* dipuku	*e* ka se *be* morutiši	*a* ka se *be* bohlale	*a* ka se *be* le mpša
	PAS.	p.	ge mosadi *a rekile* dipuku	*e bile* morutiši	*a bile* bohlale	*a bile* le mpša
		n.	ge mosadi *a* sa *reka* dipuku	*e* sa *ba* morutiši	*a* sa *ba* bohlale	*a* sa *ba* le mpša
REL.	PRES.	p.	mosadi yo *a rekago* dipuku	*e bago* morutiši	*a bago* bohlale	*a bago* le mpša
		n.	mosadi yo *a* sa *rekego* dipuku	*e* sa *bego* morutiši	*a* sa *bego* bohlale	*a* sa *bego* le mpša
	+Pot.	p.	mosadi yo *a* ka *rekago* dipuku	*e* ka *bago* morutiši	*a* ka *bago* bohlale	*a* ka *bago* le mpša

		n.	mosadi yo *a* ka se *rekego* dipuku	*e* ka se *bego* morutiši	*a* ka se *bego* bohlale	*a* ka se *bego* le mpša
	FUT.	p.	mosadi *yo a* tlo/tla *rekago* dipuku	*e* tlo/tla *bago* morutiši	*a* tlo/tla *bago* bohlale	*a* tlo/tla *bago* le mpša
		n.	mosadi yo *a* ka se *rekego* dipuku	*e* ka se *bego* morutiši	*a* ka se *bego* bohlale	*a* ka se *bego* le mpša
	PAS.	p.	mosadi yo *a rekilego* dipuku	*e bilego* morutiši	*a bilego* bohlale	*a bilego* le mpša
		n.	mosadi yo *a* sa *rekago* dipuku	*e* sa *bago* morutiši	*a* sa *bago* bohlale	*a* sa *bago* le mpša
SUB.		p.	(gore) mosadi *a reke* dipuku	*e be* morutiši	*a be* bohlale	*a be* le mpša
		n.	(gore) mosadi *a* se *reke* dipuku	*e* se *be* morutiši	*a* se *be* bohlale	*a* se *be* le mpša
CON.		p.	mosadi *a reka* dipuku	*ya ba* morutiši	*a ba* bohlale	*a ba* le mpša
		n.	mosadi *a* se *reke* dipuku	*ya* se *be* morutiši	*a* se *be* bohlale	*a* se *be* le mpša
INF.		p.	go *reka* dipuku	go *ba* morutiši	go *ba* bohlale	go *ba* le mpša
		n.	go se *reke* dipuku	go se *be* morutiši	go se *be* bohlale	go se *be* le mpša
IMP.		p.	*reka* dipiku!	*eba* morutiši!	*eba* bohlale!	*eba* le mpša!
		n.	se *reke* dipuku!	se *be* morutiši!	se *be* bohlale!	se *be* le mpša!
HAB.		p.	mosadi *a reke* dipuku	*e be* morutiši	*a be* bohlale	*a be* le mpša
		n.	mosadi *a* se *reke* dipuku	*e* se *be* morutiši	*a* se *be* bohlale	*a* se *be* le mpša

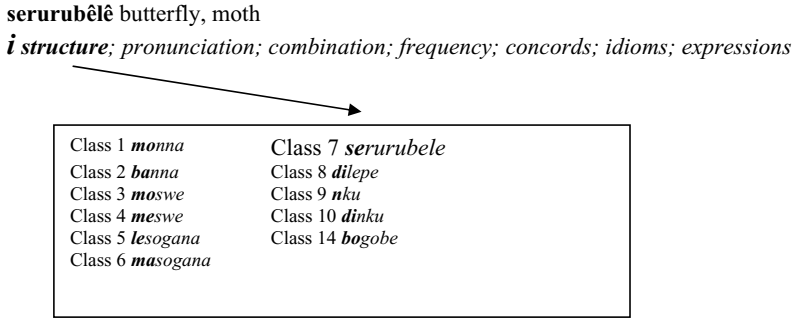
In Table 8 not less than 34 copulative forms for 3 different copulative relations were given, covering only class 1. Multiplied by the roughly 20 different sets of concords for persons and classes in Table 1, this means roughly 34 x 3 x 20 = 2,040 possible candidates for lemmatisation of the dynamic copulative.

In a good Northern Sotho dictionary the lexicographer tries to maximally utilise all available strategies and structures such as sound treatment in dictionary articles, cross-references to the back matter and even cross-references to outside sources such as grammar books in order to assist the user to understand this complicated issue in Northern Sotho.

One cannot but conclude that lemmatisation of especially nouns, verbs and copulatives cannot be solved for Bantu languages in the paper dimension especially if an accessible, user-friendly dictionary for

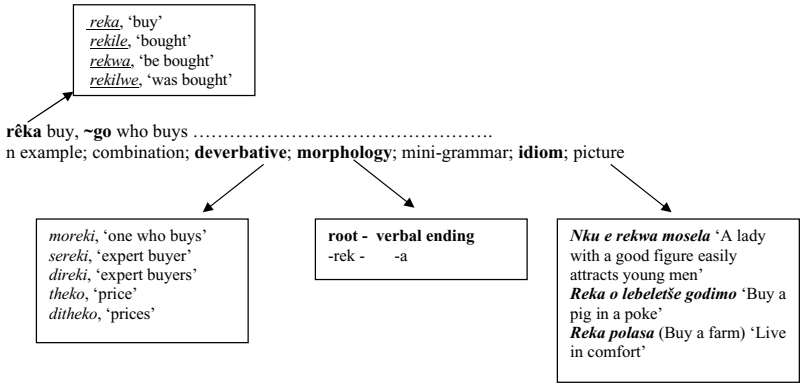
inexperienced learners of the language is the objective. The question is how can these lemmatisation problems in respect of e.g. verbs, nouns and complicated linguistic systems like the copulative be solved? The solution lies in the electronic dictionary dimension. Utilising a combination of, especially the electronic features listed in (1), i.e. pop-up access, bringing together of related items, new routes to the data, less dependency on alphabetical order, intelligent extrapolation, etc. can be the answer. In practical terms, detailed morphological analysis and parsing of nouns and verbs, annotated corpora, huge frequency lists, etc. will be the required building blocks. Hundreds of thousands of words will have to be hyperlinked to their lemma signs in order to allow intelligent extrapolation as has been illustrated above for *intoxication* in MED. Stratified/layered pop-up boxes in the case of complicated grammatical systems will have to be built as well as a complicated network of cross-referencing. Consider Figures 9 – 11 for typical suggested solutions for the lemmatisation of nouns, verbs and copulatives respectively.

Figure 9: The noun *serurubele* in an ED for Northern Sotho



In the case of nouns, the *noun class system* could be presented in an innovative but simplistic way. In Figure 9 the user looks up the word *serurubele* and finds the translation equivalents 'butterfly, moth'. If (s)he now puts the cursor on ***structure*** in the information bar, a text box opens, not only reflecting the *total scope of the noun class system*, but also *putting the word itself within its appropriate position in the noun class system*, namely class seven.

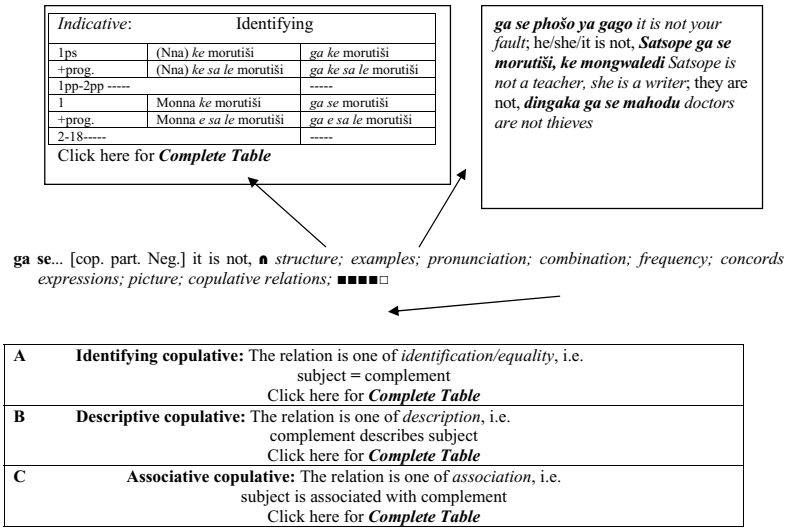
Figure 10: The verb *reka* in an ED for Northern Sotho



In the first pop-up box the user can find useful information regarding the verbal derivations of the lemma. In the left bottom box, (s)he can find all nominalizations arranged according to their nominal classification. In the right bottom box, typical occurrences of the lemma and its derivations in idioms and proverbs can be studied.

Keep in mind that all this is achieved by simply moving the mouse over different sections of the navigation bar. Thus, information boxes only appear if the user wants to see them.

Figure 11: The copulative *ga se* in an ED for Northern Sotho



For the copulative, layered, clickable options should be provided, thus presenting the user digestible sections while outlining the full scope of the complicated system.

5. Conclusion

It has been attempted in this article to give a perspective on electronic dictionaries from a South African point of view. As far as English is concerned one could conclude that South African users have the advantage of the availability of sophisticated internationally developed Eds, both on CD-ROM and online and that future developments should focus on extending the same level of sophistication to Eds catering for South African English and also for Black South African English. For Afrikaans progress has been made towards the compilation of true electronic dictionaries and it is expected that a new generation of Afrikaans Eds would include more advanced true electronic dictionary features. For the Bantu languages interest in the compilation of electronic dictionaries is picking up and the fact that successful information retrieval is so heavily dependant on the electronic dimension, provides extra motivation for the compilation of Eds for

these languages. The rate of development of Eds will also be influenced by external factors both internationally and locally. It remains to be seen how fast the presumed gradual swing from paper dictionary to electronic dictionary often advocated in publications on Eds will take place. In an African context the development and use of Eds will also be influenced by the rate of development of a dictionary culture, computational skills and access to computers and the internet. In the long run it is reasonable to expect that also in South Africa the electronic dictionary will overshadow the paper dictionary in the same way as the computer has superseded the typewriter.

References

A. Electronic dictionaries

Cambridge Advanced Learner's Dictionary Online http://dictionary.cambridge.org/

Collins COBUILD on CD-ROM. 1995. HarperCollins Publishers Ltd.

DDP Freeware Afrikaans/English Dictionary online. http://www.freedict.com/

Elektroniese WAT. Woordeboek van die Afrikaanse Taal (A-O). CD-ROM. 2003. WAT, Van Schaik.

Macmillan English Dictionary for Advanced Learners. 2002. Macmillan Publishers Limited.

Merriam-Webster OnLine http://www.m-w.com/

Oxford English Dictionary http://www.oed.com/

Oxford English Dictionary, Second Edition on Compact Disk. 1989. Oxford University Press.

Pharos Woordeboeke Dictionaries 5 in 1. 2000. Johannesburg: Pharos & Logos Information Systems.

Sesotho sa Leboa (Northern Sotho) - English Dictionary. http://africanlanguages.com/sdp/

Travlang's Afrikaans-English On-line Dictionary. http://dictionaries.travlang.com/

Travlang's Worldwide Travel Guides. http://www.travlang.com/

Webster's Online Dictionary, The Rosetta Edition.

http://www.websters-online-dictionary.org/

Zulu-English/English-Zulu online dictionary. http://www.isizulu.net/

B. Other references

Atkins, B.T. Sue. 1996. Bilingual Dictionaries: Past, Present and Future. *Proceedings of the Seventh EURALEX International Congress on Lexicography*. *Gőteborg.* 515-546.

Bolinger, D. 1990. Review of Oxford Advanced Learner's Dictionary of Current English. *International Journal of Lexicography* 3/2: 133–45.

De Schryver, Gilles-Maurice. 2003. Lexicographers' Dreams in the Electronic-Dictionary Age. *International Journal of Lexicography 16/2*: 143–199.

De Schryver, Gilles-Maurice & Daniel J. Prinsloo. 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*: 291–309.

De Schryver, Gilles-Maurice & Daniel J. Prinsloo. 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.

Dodd, W.S. 1989. Lexicomputing and the dictionary of the future. Lexicographers and their Works. James G. (Ed.) Exeter Linguistic Studies.

Geeraerts, Dirk. Euralex 2000 p75 *Proceedings of the Ninth EURALEX International Congress on Lexicography*, Stuttgart, 8-12 August 2000. (pp 75-84)

Harley, Andrew. 2000. Software Demonstration: Cambridge Dictionaries Online. *Proceedings, The Ninth Euralex International Congress*. Heid, Ulrich et al. (Eds.). Stuttgart. (pp 85-88).

Kriel, Theunis J. 1985 *New English Northern Sotho dictionary*. Johannesburg: Educum.

Kriel, Theunis J. and Van Wyk, Egidius B. 1989. *Pukuntšu woordeboek, Noord-Sotho–Afrikaans, Afrikaans–Noord-Sotho*. Pretoria: J.L. van Schaik.

Nesi, Hillary. 1999. A User's Guide to Electronic Dictionaries for Language Learners. *International Journal of Lexicography* 12/1: 55–66.

Poulos, George and Louis J. Louwrens. 1994. *A Linguistic Analysis of Northern Sotho*. Pretoria: Via Afrika.

Prinsloo, Daniel J. 2001. *The Compilation of Electronic Dictionaries for the African Languages*. *Lexikos 11*. Afrilex Series. J.C.M.D. du Plessis (Ed.). Stellenbosch. Bureau of the WAT. 139-159

Prinsloo, Daniel J. & De Schryver, Gilles-Maurice. 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.

Sharpe, P. 1995. Electronic dictionaries with particular reference to the design of an electronic bilingual dictionary for English-Speaking learners of Japanese. *International Journal of Lexicography 8/1: 39–54*.

Silva, Penny M. 1996. *A dictionary of South African English on Historical Principles*. Oxford: Oxford University Press.

Silva, Penny M. 2004 South African English: Oppressor or Liberator? Accessed at <www.ru.ac.za/affiliates/dsae/MAVEN.HTML>

Van Wyk, Egidius B. 1995. Linguistic Assumptions and Lexicographical Traditions in the African Languages. *Lexikos 5*. Afrilex Series. J.C.M.D. du Plessis (Ed.). Stellenbosch. Bureau of the WAT. 82-96

Wade, Rodrik. 1998. *Black South African English as a distinct 'new' English*. Accessed at <http://www.und.ac.za/und/ling/archive/wade-03.html>

Ziervogel, Dirk. & Pothinus C. Mokogokong. 1975. *Groot Noord-Sotho Woordeboek.* Pretoria: J.L. van Schaik.