Outomatiese lemma-identifisering en die dilemma met die lemma in Setswana

KARIEN BRITS

Adam Mickiewicz University, Poznań, Poland

Instytut Filologii Angielskiej, Uniwersytet im. Adama Mickiewicza
al. Niepodległości 4, 61-874 Poznań, Poland

karien@ifa.amu.edu.pl

RIGARDT PRETORIUS

North-West University, Potchefstroom, South Africa

Noord-Wes Universiteit
Privaatsak X6001, Potchefstroom 2520, South Africa

Rigardt.Pretorius@nwu.ac.za

Automatic lemmatisation and the dilemma with the lemma in Setswana

Abstract. Projects in human language technologies do not only imply challenges for programmers but also for grammarians. In a recent project to develop an automatic lemmatiser for Setswana, the problem arose as to what the lemma in Setswana should be, as no clear-cut definition exists in the Bantu language grammars or lexicographic studies. This article aims to determine and discuss the term "lemma" in Setswana as it should be applied in automatic lemmatisation.

Keywords: Setswana, lemmatiser, lemmatising, morphology, lexicology, lexicography

1. Inleiding

Setswana is deel van die Sothotaalgroep (of anders bekend as die Sotho-Tswanataalgroep), die tweedegrootste Bantoetaalgroep in Suid-Afrika, en is een van die amptelike tale van Suid-Afrika (Herbert, Bailey 2002: 51, 68). Dié relatiewe jong skryftaal het nou ook die arena van die taaltegnologie betree. Dit is deels daaraan toe te skryf dat Setswana grondwetlike beskerming in Suid-Afrika geniet, maar ook omdat regeringsbeleid ontwikkeling in taaltegnologie steun (Nasionale Departement van Kuns en Kultuur, 2000). Soos wat Setswana die wêreld van die taaltegnologie betree, is al hoe meer 'n behoefte aan betroubare elektroniese taalhulpmiddels soos speltoetsers, woordafbrekers, inligtingsonttrekkingsisteme, vraagbeantwoordingsisteme, outomatiese vertaalsisteme, elektroniese woordeboeke en rekenaargesteunde taalonderrigprogrammatuur. Tog om dié hulpmiddels beskikbaar te stel moet basiese hulpbronne (soos korpora en kerntegnologiemodules) eers ontwikkel word. 'n Belangrike kerntegnologiemodule wat dikwels deel uitmaak van verskeie

toepassings is 'n lemma-identifiseerder ("lemmatiser") (vergelyk Vetulani et al. 1998: 1, 50).

Ten einde so 'n lemma-identifiseerder te ontwikkel moet die term "lemma" soos wat dit in Setswana verstaan word, gedefinieer word. Daar bestaan egter tot op hede nie 'n omvattende amptelike lemmalys nie en voorts kon daar nie 'n klinkklare definisie van die term "lemma" in Setswana of in enige ander Sothotaal gevind word nie. Die algemene breë definisies gee wel leiding in wat 'n mens ongeveer onder die term kan verstaan, maar vir 'n kerntegnologiemodule soos die lemma-identifiseerder is goedgedefinieerde en gedetailleerde definisies nodig. In die eerste gedeelte van die artikel word 'n kort oorsig oor die konsep outomatiese lemma-identifisering gebied, in die tweede gedeelte word die Setwanawoordsoorte in 'n neutedop aangebied en laastens word aandag aan die ontwikkeling van 'n eie definisie vir die Setswanalemma gegee.

2. Konteks: Lemma-identifisering in natuurliketaalprosessering

Lemma-identifisering word, volgens Plisson et al. (2004), in natuurliketaalprosessering ("natural language processing") gebruik, en natuurliketaalprosessering word deur Crystal beskryf as die rekenaarmatige verwerking van teksmateriaal in natuurlike tale (2003: 309). Die doel met natuurliketaalprosessering is om prosedures te ontwikkel waarmee groot hoeveelhede teks outomaties geanaliseer kan word (Crystal 2003: 309); lemma-identifisering is een so 'n prosedure (vergelyk Manning, Schütze 1999: 132; Jurafsky, Martin 2000: 195).

Lemma-identifisering behels die redusering van woorde in 'n korpus tot hulle ooreenstemmende lekseme/lemmas, of, soos Erjavec en Džeroski dit stel: "a normalization step on textual data, where all inflected forms of a lexical word are converted or reduced to its common headword form, i.e. the lemma" (2004: 17). In aansluiting hierby definieer Hausser lemma-identifisering as die bepaling van die korrekte basisvorm van 'n woord, oftewel die verskaffing van toegang tot die ooreenstemmende lemma van 'n woord in die leksikon (1999: 125). Dit beteken dat die woord na die eenvoudigste vorm soos dit in 'n woordeboek sou voorkom (die lemma), reduseer word.

Lemma-identifisering is 'n belangrike prosedure in korpusgebaseerde navorsing (McEnery, Wilson 2001: 53). Lemma-identifisering speel 'n kernrol by die teksenkodering van 'n korpus, aangesien dit onder andere in die leksikografie die voordeel inhou dat 'n navorser al die variante van 'n lekseem uit 'n korpus kan onttrek sonder om al die moontlike variante in te sleutel (Gouws, Prinsloo 2005: 37; McEnery, Wilson 2001: 53). Volgens Plisson et al. (2004) is lemma-identifisering ook 'n belangrike voorbereidende stap vir teksontginning ("text mining") en word dit gebruik vir die skep van generiese sleutelwoorde vir soekenjins.

Opsommend kan lemma-identifisering beskryf word as 'n natuurliketaalprosesseringsprosedure wat die korrekte basisvorm/lemma van 'n toevoerwoord bepaal deur die verwydering van fleksie-affikse. Die lemma-identifiseerder het egter bepaalde reëls nodig om die regte affikse te verwyder en daarom moet die lemma noukeurig gedefinieer word.

3. Setswanawoordsoorte in 'n neutedop

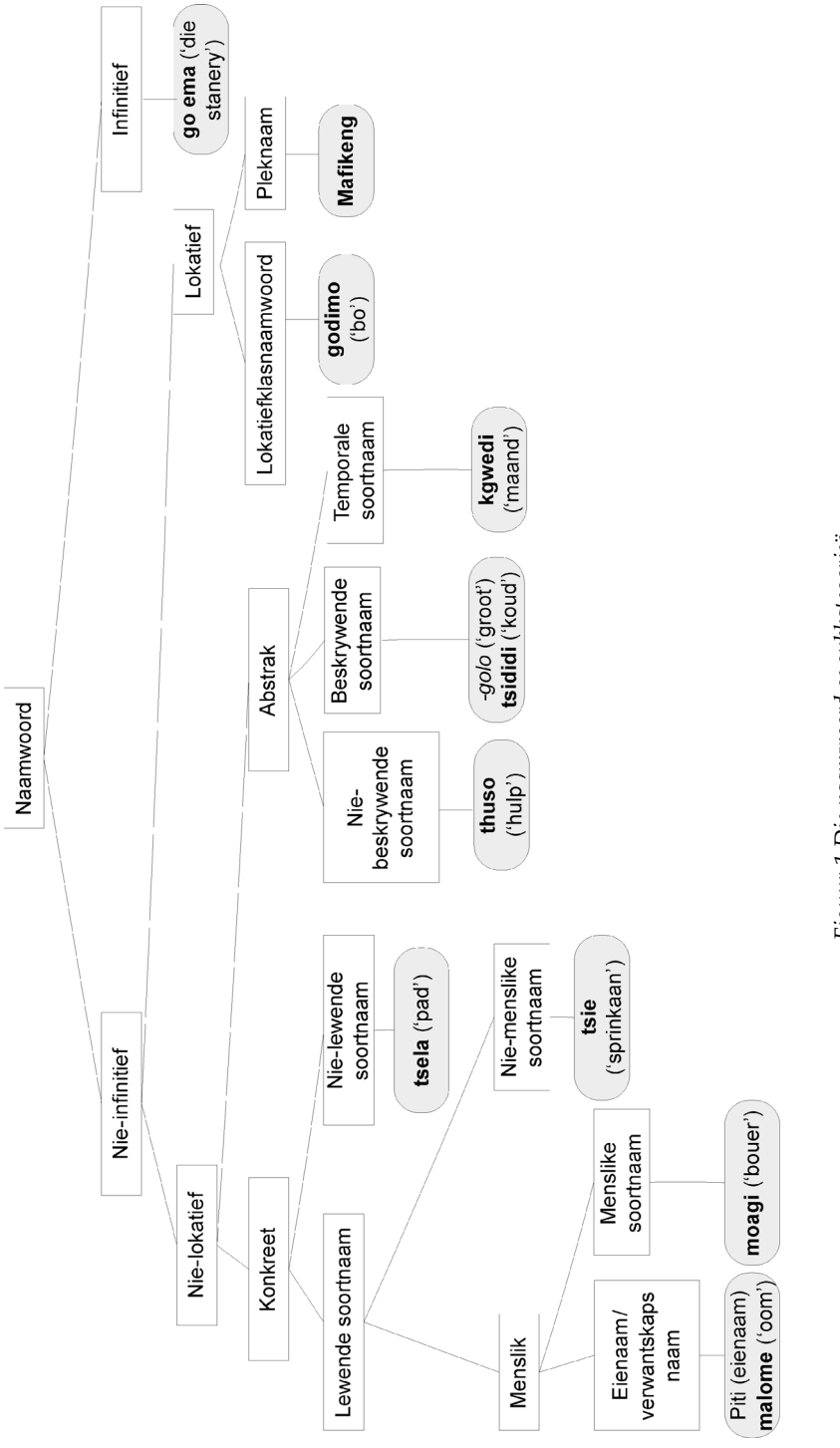
Daar is sewe woordsoorte in Setswana naamlik, die naamwoord, voornaamwoord, werkwoord, betrekkingswoord, bywoord, interjeksie en ideofoon (vergelyk Van Wyk 1966: 230, 261). Die bespreking wat volg, dek nie alles in die Setswana-grammatika nie, maar het ten doel om die leser 'n breë agtergrond van die woordsoorte te gee en met enkele voorbeelde toe te lig.

3.1 Die naamwoord

Die naamwoord word deur Krüger in subkategorieë verdeel soos wat dit in Figuur 1 voorgestel word (2006: 99). Volgens Cole (1955) kan naamwoorde – met die uitsondering van die infinitiewe en lokatiewe klasse en sekere beskrywende naamwoordwortels (byvoorbeeld *-golo*) – in pare opgedeel word om enkelvoud- en meervoudsvorms aan te dui. As daar derhalwe na klas 1 (CL1.SG of CL1.PL) verwys word, dan word daarmee bedoel die *mo-/ba*-klas; klas 2 (CL2.SG of CL2.PL) is dan die *mo-/me*-klas, ensovoorts. Ter wille van volledigheid word hierdie naamwoordklasse in Tabel 1 opgesom.

Tabel 1 Die naamwoordklasse in Setswana

klas	klasprefiks	voorbeeld	Afrikaans
1	*mo*	**monna**	'man'
	ba	**banna**	'mans'
1a	-	**mma**	'ma'
	bo	**bomma**	'ma-hulle'
2	*mo*	**motse**	'stat'
	me	**metse**	'statte'
3	*le*	**legapu**	'waatlemoen'
	ma	**magapu**	'waatlemoene'
4	*se*	**selepe**	'byl'
	di	**dilepe**	'byle'
5	*ne*	**nku**	'skaap'
	di	**dinku**	'skape'
6	*lo*	**lonaka**	'horing'
	di	**dinaka**	'horings'
7	*bo*	**boupe**	'meel'
	ma	**maupe**	'verskillende soorte meel'
8	*go*	**go ema**	'stanery'
9	*fa, go, mo*	**godimo**	'bo'



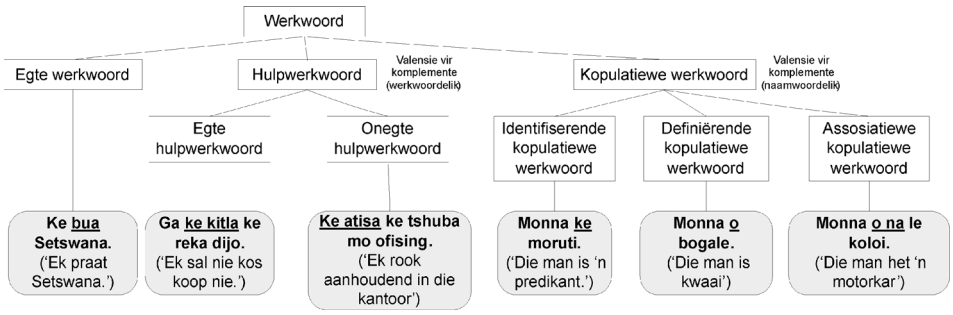
Figuur 1 Die naamwoord se subkategorieë

3.2 Die voornaamwoord

Daar is 'n sterk kongruensiesisteem in die Bantoetale, elke naamwoord het dus 'n eie voornaamwoord en hulle kan net soos die naamwoorde in klasse verdeel word: **tsona** ('hulle' verwysend na klas 4/5/6), **esi** ('alleen/slegs' verwysend na klas 2/5), **botlhe** ('almal' klas verwysend na 1/7). Voorbeeld van persoonlike naamwoorde is **nna** ('ek'), **wena** ('jy'), **rona** ('ons') en **lona** ('julle') (vergelyk Zerwick 1997).

3.3 Die werkwoord

Die werkwoord word deur Krüger (2006: 24) en Pretorius (1997) in die subkategorieë egte werkwoord, hulpwerkwoord en kopulatiewe werkwoord verdeel, soos wat dit in Figuur 2 geïllustreer word. [1]



Figuur 2 Die werkwoord se subkategorieë

3.4 Die betrekkingswoord

Daar is volgens Krüger twee soorte betrekkingswoorde, te wete die kongruerende en nie-kongruerende betrekkingswoorde (2006: 147-159). Onder die kongruerende betrekkingswoorde is daar die possessiewe betrekkingswoorde (byvoorbeeld **wa, ya**) en kwalifikatiewe betrekkingswoorde wat vormlik ooreenstem met die demonstratiewe voornaamwoord ('hierdie'), byvoorbeeld **yo**, **ba**, **tse**. Onder die nie-kongruerende betrekkingswoorde is daar voegwoordelike (byvoorbeeld **gonne –** 'want') en nie-voegwoordelike betrekkingswoorde (byvoorbeeld **ka –** 'met').

3.5 Die bywoord

Daar is min egte bywoorde in Setswana, byvoorbeeld **kae?** ('waar?'), **jang?** ('hoe?'), **jaanong** ('nou'), **jalo** ('so/soos'), **jaana** ('sodanig'), **ruri** ('regtig') (vergelyk Krüger 2006: 161-163).

[1] Die term 'egte werkwoord' word in plaas van selfstandige werkwoord, in navolging van Krüger (2006).

3.6 Interjeksies

Interjeksies of tussenwerpsels het geen vasgestelde struktuur in Setswana nie en het daarom ook geen morfeme waaruit die woord opgebou is nie. Voorbeelde van interjeksies: **Ao! Ija! Ijo! Mmalo!** ('Mensig!/Haai'), **Ehe! Oo!** ('O so!') en **X-x-x-x** ('Verskoon tog'). Dit is ook 'n beperkte groep woorde wat nie morfologies uitgebrei word nie (vergelyk Cole 1955: 395-400; Krüger 2006: 165-166; Ntsime, Krüger 1995: 177-182).

3.7 Ideofone

Soos dit met interjeksies die geval is, het die ideofoon ook nie 'n vasgestelde struktuur nie en word hierdie woordsoort nie morfologies uitgebrei nie (vergelyk Ras 1991). Byvoorbeeld: **tho tho tho** in **Pula ya-tla e-re tho tho tho** (CL5.SG.rëen AGR.SUBJ-kom AGR-sê tip tip tip -'Die reën val tip-tip-tip.').

4. Die definisie van 'n lemma in Setswana

Noudat daar 'n kort oorsig oor die begrip lemma-identifisering en die Setswana-woordsoorte gebied is, kan daar oorgegaan word na die bespreking van die term "lemma." 'n Lemma word oor die algemeen gedefinieer as die betekenisvolle basisvorm, gestroop van fleksiemorfeme wat variante vorme verteenwoordig (Choueka et al. 2000: 74; Mitkov 2003: 728; Trost 2003: 38). In die algemenetaal-wetenskapliteratuur word die term "lemma," tesame met die verbandhoudende terme "lekseem," "kanonieke vorm," "basisvorm" en "basis," dikwels verwarrend gebruik. Hierdie term word binne drie verskillende taalkundige kontekste (in die leksikografie, leksikologie en morfologie) gebruik om verskillende aspekte van dieselfde fenomeen te beskryf.

Bo en behalwe die verwarrende gebruik van dié term, is daar ook die kwessie van 'n taalspesifieke definisie, want soos wat tale verskil, kan die definisies van wat as lemma in verskillende tale beskou word, ook verskil. Een so 'n voorbeeld is 'n vertalingtoepassing om Engelse en Hebreeuse vertalings te belyn. In dié projek moes albei tale se lemmas vooraf geïdentifiseer word en moes presies beskryf word wat as lemma vir naamwoorde en werkwoorde beskou word (Choueka et al. 2000: 74). Die lemma van die werkwoordvorm in die Hebreeus is byvoorbeeld gedefinieer as die vorm in die derde persoon/enkelvoud/manlik/verlede tyd, terwyl dit in Engels gedefinieer is as die infinitiewe vorm (Choueka et al. 2000: 76). Die bepaling van die konsep "lemma" is dus, met inagneming van algemene teorieë daaroor, taalspesifiek, en alvorens 'n lemma-identifiseerder vir Setswana ontwikkel kon word, moes deeglik besin word oor wat 'n lemma in Setswana is.

Die begrip "lemma" is reeds in ander (Europese) tale wat 'n ryk geskiedenis van woordeboekopstelling het, omvattend gedefinieer. Daar is egter nie enigheid onder leksikograwe oor wat die lemma in die Bantoetale moet wees nie. Volgens De Schryver en Prinsloo (2001) debateer hulle of die wortel (wat sommige moedertaalsprekers nie eers sal herken nie) of die stam (soos wat dit in werklike taalgebruik voorkom) as lemma beskou moet word (vergelyk Gouws, Prinsloo 2005: 67-85). Ten einde 'n bevredigende definisie te ontwikkel is die term "lemma" en verbandhoudende terme binne die leksikografiese, leksikologiese en morfologiese kontekste en die implikasies daarvan op Setwana ondersoek.

4.1 "Lemma" as leksikografiese term

Die kanonieke vorm is 'n term wat in die leksikografie gebruik word, en Hartmann en James definieer dit as trefwoord ("headword") waaronder verskeie variante, woorde of frases aangehaal word (1998: 18). Die term "lemma" word in die leksikografie gebruik om na daardie leksikale elemente te verwys wat in 'n woordeboek opgeneem word (Gouws 1989: 35; Bussman 1996: 272). In aansluiting hierby beskou Hartmann die lemma as die posisie in die algehele struktuur van 'n woordeboek of naslaanwerk waar 'n inskrywing gevind kan word, gewoonlik deur die trefwoord (2001: 174).

'n Trefwoord is volgens Crystal weer die item wat aan die begin van 'n woordeboekinskrywings staan (1992: 225). Hy definieer dit as "an abstract representation [. . .] subsuming all the formal variations which may occur." 'n Voorbeeld van 'n trefwoord in Engels is "go," wat die variante woorde "goes," "going" en "went" insluit. In aansluiting by Crystal is trefwoorde volgens Jackson die basisvorms waarvan ander woordvorms afgelei word (1988: 9). Om saam te vat is 'n lemma 'n term wat na die leksikale elemente in 'n woordeboek verwys en wat, as dit as sinoniem van 'n trefwoord gebruik word, 'n abstrakte verteenwoordiging of basisvorm van ander woordvorms is.

As die leksikografiepraktyk as norm geneem word, kan 'n mens aflei wat as die lemma beskou word, deur die trefwoorde in verskillende woordeboeke van die Sothotale te bestudeer. Ter illustrasie word die naamwoorde **mosadi** ('vrou') en **selepe** ('byl'), en die werkwoorde **bula** ('oopmaak') en **araba** ('antwoord') as prototipiese voorbeelde ondersoek in die volgende woordeboeke (vergelyk Tabel 2):

- *Setswana-English Dictionary* van Brown (1988)
- *Setswana-Engels-Afrikaanse woordeboek* van Snyman et al. (1990)
- *Reader's Digest: Multi-Language Dictionary and Phrase Book* van Reynierse (1991)
- *Kompakte Setswana woordeboek* van Dent (1994)
- *Groot Noord-Sotho woordeboek* van Ziervogel en Mokgokong (1985)
- *Pukuntšu* van Kriel, Van Wyk en Makopo (1989)
- *The New Sesotho-English Dictionary* van Kriel (1958)

Tabel 2 ***Mosadi, selepe, araba*** *en* ***bula*** *in verskillende woordeboeke* [23]

		Setswana				Sesotho sa Leboa		Sesotho
		Brown (1988)	Snyman *et al.* (1990)	Reynierse (1991)	Dent (1994)	Ziervogel en Mokgokong (1985)	Kriel, Van Wyk en Makopo (1989)	Kriel (1958)
mosadi	*-sadi*		✓			✓		
	mosadi	✓		✓	✓		✓	✓
	basadi	✓			✓			✓
selepe	*-lepe*		✓			✓		
	selepe	✓		✓	✓		✓	✓
	dilepe	✓			✓			✓
araba	araba	✓	✓	✓	✓	✓	✓	✓
	arabela						✓	✓
	arajwa		✓					
bula	bula	✓	✓	✓	✓	✓	✓	✓
	bulega	✓	✓				✓	
	bulela	✓						
	budisa		✓					
	bulegileng			✓	✓			

Wat die Setswanawoordeboeke betref, is **mosadi** en **selepe** sowel as die meervoud **basadi** ('vroue') en **dilepe** ('byle') as naamwoordlemmas in die *Setswana-English Dictionary* van Brown (1988) gebruik. **Bula** ('oopmaak') word opgeneem as 'n werkwoordlemma, maar daarnaas ook **bulega** ('oopgaan/ oopmaakbaar wees') en **bulela** ('oopmaak vir'). Hoewel Brown **bulela** as 'n lemma in die woordeboek gebruik, beskou hy dit tog ook as 'n variant van **bula** as hy dit beskryf as "prep. form of *bula*" (1988: 39). Tog neem Brown (1988) net **araba** as lemma op. In Snyman *et al.* (1990) se *Setswana-Engels-Afrikaanse woordeboek* is die wortel *-sadi* en *-lepe* die lemma van **mosadi** en **selepe** onderskeidelik. Soos in die ander woordeboeke is **bula** ook weer 'n lemma, maar daarnaas ook **bulega** en **budisa** ('laat of help oopmaak'). Dieselfde gebeur met **araba**, waar **arajwa** ('word geantwoord') ook as lemma opgeneem is. Die woordeboek van Snyman *et al.* is die enigste Setswana-woordboek wat die naamwoordwortel as lemma opneem. Die naamwoordlemma van **mosadi** en **selepe** in die *Reader's Digest: Multi-Language Dictionary and Phrase Book* van Reynierse (1991) is **mosadi** en **selepe** onderskeidelik, en, anders as by Brown (1988), word die meervoud nie as 'n lemma gelys nie. **Bula** is die werkwoordlemma in dié woordeboek, en **bulegileng** ('oop/gapend') is ook as 'n lemma

[2] Sesotho sa Leboa is ook bekend as Noord-Sotho of Sepedi.
[3] Sesotho is ook bekend as Suid-Sotho.

gelys. In Dent (1994) se *Kompakte Setswana woordeboek* word **mosadi** en **basadi** albei as lemmas gelys, so ook **selepe** en **dilepe**. In hierdie woordeboek is **bula** en **bulegileng** ook albei lemmas, maar slegs **araba** word as lemma opgeneem.

Vir die naamwoordlemma in die *Groot Noord-Sotho woordeboek* van Ziervogel en Mokgokong (1985) is die wortel *-sadi* en *-lepe* gebruik. Die werkwoordlemma van **bula** en **araba** is in die infinitiefvorm opgeneem sonder die infinitiefprefiks *go-*. In 'n ander Sesotho sa Leboa–Afrikaanse woordeboek, *Pukuntšu* (Kriel, Van Wyk, Makopo, 1989), is die naamwoordlemma die naamwoord in die enkelvoud, dus **mosadi** en **selepe**. Die werkwoordlemma is in die infinitiefvorm sonder die infinitiefprefiks *go-* opgeneem, maar **bulega** ('oopgaan/oopmaakbaar wees') en **arabela** ('verantwoording doen') word ook as lemmas gelys. In Kriel (1958) se *The New Sesotho-English Dictionary* is **mosadi** en **basadi**, **selepe** en **dilepe** lemmas. **Bula** is die werkwoordlemma in dié woordeboek. **Arabela** ('inasem') word naas **araba** as lemma opgeneem.

Dit blyk uit die voorbeelde hierbo dat die naamwoordlemma wissel van 'n wortel tot enkelvoud- en meervoudvorms as aparte inskrywings. Verskeie vorms van die werkwoord word ook inkonsekwent as aparte lemmas opgeneem. Wat hier egter in ag geneem moet word, is dat die woordeboeke in 'n tyd opgestel is waar hulpmiddels, soos frekwensielyste uit korpusse gegenereer, nie beskikbaar was nie. De Schryver en Prinsloo lys soortgelyke inkonsekwenthede in woordeboeke wat nie met behulp van korpora saamgestel is nie, met verwysing na Sesotho sa Leboa-woordeboeke (2001: 376). Enkele van die inkonsekwenthede wat hulle (De Schryver, Prinsloo 2001) identifiseer, is die afwesigheid van 'n beleid hoe om produktiewe teenoor onproduktiewe suffikse te hanteer, die verskillende hanterings van prefikse (veral die wat met infleksie te doen het) en die keuse van kanonieke vorms.

Volgens De Schryver en Prinsloo is daar twee maniere waarop woorde (hier verwys hulle spesifiek na byvoeglike naamwoorde) in 'n woordeboek opgeneem kan word (d.i. hoe die kanonieke vorms gekies kan word) (2001: 379). Dit kan volgens De Schryver en Prinsloo óf in 'n sogenaamde stamgebaseerde woordeboek,[4] óf in 'n woordgebaseerde woordeboek opgeneem word. Wat hieruit afgelei kan word, is dat daar twee soorte trefwoorde is: dié wat selfstandig gebruik kan word (byvoorbeeld **bula**), en dié wat nie selfstandig gebruik kan word nie (*-sadi*) (2001).

Uit die voorafgaande ontledings en De Schryver en Prinsloo se kommentaar blyk dit duidelik dat daar meer as een lemmatiseringstradisie bestaan. Voorts vereis die leksikografiepraktyk soms, as gevolg van betekenisverskuiwing, klankveranderings of 'n hoë gebruiksfrekwensie, dat afgeleide vorme as afsonderlike inskrywings hanteer moet word. Woordeboeke bied daarom nie konsekwente

[4] Wat hulle hier as stam beskou, word in die woordmorfologie van Setswana as 'n wortel beskou.

riglyne wat as die lemma, soos van toepassing op die taaltegnologie, in die Sothotale beskou kan word nie.

4.2 "Lemma" as leksikologiese term

'n Lemma word in die konteks van die leksikologie sinoniem met 'n lekseem (of woordeskatitem/leksikale item) gebruik (vergelyk Katamba 2005: 18, 296, Van Sterkenburg 2003: 403). Ter ondersteuning hiervan wys Ooi daarop dat sommige linguiste glad nie 'n onderskeid tussen die terme lemma en lekseem tref nie (1998: 215). Tog onderskei Ooi tussen die terme "lekseem" en "lemma" en noem dat die term "lemma" gebruik word wanneer 'n mens te doen het met fleksionele variante, terwyl die term "lekseem" gebruik word vir 'n woordeboek-/leksikale inskrywing (1998: 215). In aansluiting by Ooi (1998: 215) noem Jackson lekseme die "headwords of dictionary entries" (1988: 8-9). Volgens Ooi (1998) en Jackson (1988) word die term "lemma" (anders as hierbo (3.1) aangedui) dan in die leksikologie gebruik en "lekseem" in die leksikografie. Gouws verskil van Ooi (1998) en Jackson (1988), en volgens hom beskryf 'n woordeboek nie lekseme nie: "Leksemiese klassifikasie word nie semanties gemotiveer nie, maar fleksioneel, dit wil sê vormlik" (1989:128-129).

Al word die terme "lemma" en "lekseem" uitruilbaar gebruik, is die neiging tog eerder om die term "lemma" met die leksikografie en die term "lekseem" met die leksikologie te verbind. Daarom onderskei Van Sterkenburg, in teenstelling met Ooi (1998), anders tussen die lemma en die lekseem en is die lemma die trefwoord in 'n woordeboek, terwyl die lekseem as die "smallest distinctive unit in the lexicon or vocabulary of a language, which is mostly interpreted as a combination of a form with a meaning" beskou kan word (2003: 403).

Lekseme is die basiese eenheid in die linguistiese studie van 'n woordeskat en verteenwoordig verskillende woordvorms. Indien die term "lemma" as sinoniem van die term "lekseem" gebruik word, verteenwoordig dit ook verskillende woordvorms. Tog blyk dit dat, al word die term "lemma" en "lekseem" as sinonieme gebruik, die term "lemma" eerder met die leksikografie verbind word en die term "lekseem" met die leksikologie verbind word.

Hoewel die term "lemma" glad nie in leksikologiese konteks in die Bantoetale gedefinieer word nie, bied Louwrens 'n definisie van die term "leksikale item" of "lekseem" aan as hy sê dat dit in grammatiese beskrywings van Bantoetale (spesifiek Sesotho sa Leboa) die kleinste linguistiese eenheid is wat leksikale betekenis het (1994: 94). Volgens Louwrens word die wortel van 'n woord dikwels as 'n leksikale item beskou (1994: 94). Verder skryf Louwrens dat die term soms in die semantiek in 'n wyer sin gebruik word om na 'n volledige woord te verwys (1994: 94). Indien die term "lemma" as sinoniem vir die term "lekseem" gebruik sou word, kan die lemma 'n wortel of 'n woord (stam) wees. Vervolgens word

die term "lemma" as morfologiese term bespreek en in hierdie afdeling sal die oorweging juis tussen die wortel en stam lê.

4.3 "Lemma" as morfologiese term

As die term "lemma" in die morfologie gebruik word, word daar na die basisvorm verwys. 'n Lemma kan oor die algemeen gedefinieer word as die betekenisvolle, gestroopte basisvorm ("base form") waarvan ander meer komplekse vorms (d.i. variante) afgelei word (Choueka et al. 2000: 74; Mitkov 2003: 728; Trost, 2003: 38). 'n Basisvorm is, volgens Hartmann en James met verwysing na lemma-identifisering, die kanonieke vorm ("canonical form") en met verwysing na die morfologie, die basis ("base") (1998: 12). [5]

Die definisie van 'n basis, volgens Crystal, is "[a] term used in morphology as an alternative to root or stem: it refers to any part of a word seen as a unit to which an operation can be applied, as when one adds an affix to a root or stem" (2003: 48). Hartmann en James sluit by Crystal (2003) aan as hy 'n basis definieer as 'n betekenisvolle morfologiese element wat gebruik word om woorde mee te vorm (1998: 12); hy definieer 'n wortel byvoorbeeld as "the base of a word" (Hartmann, James 1998: 120). Crystal noem egter dat sommige analiste die term "lemma" beperk deur dit aan die wortel, of die deel van 'n woord wat oorbly as al die affikse verwyder is, gelyk te stel (2003: 48).

'n Lemma kan gedefinieer word as 'n betekenisvolle morfologiese element, wat 'n wortel of stam kan wees en waarop morfologiese bewerkings gedoen kan word. Binne die konteks van lemma-identifisering word veral die morfologiese definisie gebruik, soos dit ook neerslag vind in die leksikografiese definisies (sien veral Jackson (1988), wat trefwoord/lemma gelykstel aan basisvorm). Vir doeleindes van hierdie artikel is 'n lemma die genormaliseerde basisvorm wat ook 'n lekseem is (en dus nie net die basisvorm nie). Die lemma moet dus 'n basisvorm wees waarop verdere bewerkings gedoen kan word (woorde afgelei word), maar dit moet terselfdertyd ook lekseem wees (dus selfstandig voorkom).

Die vraag oor wat in Setswana as die lemma beskou moet word, is tot dusver nog nie beantwoord word nie. Die woordmorfologie word vervolgens bespreek om daaruit te bepaal wat die lemma kan wees. Die Setswanalemma sal met die algemene definisies van lemma in gedagte, en aan die hand van die bespreking van die Setswanawoordmorfologie, gedefinieer word.

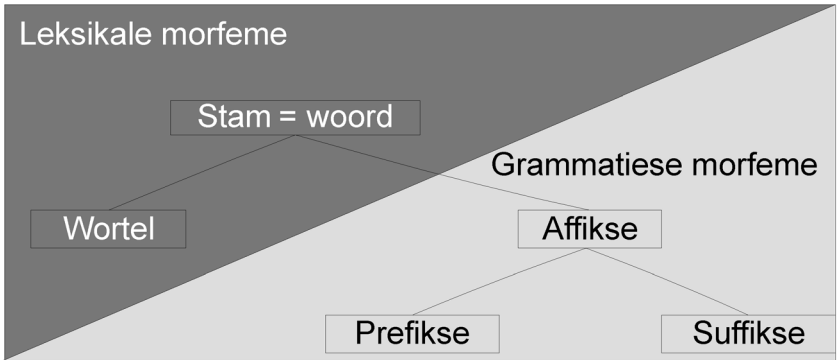
4.4 Setswanawoordmorfologie

Daar is verskillende sienings oor die woordmorfologie van die Setswanawoord. Pretorius bespreek in sy artikel oor die identifisering en beskrywing van

[5] Die terme "basis" en "basisvorm" word deur Trask as sinonieme beskou (1996:48-49).

die begrippe "stam" en "wortel" in die Sothotale die verskillende sienings (2000: 51-55). Hy noem dat morfologie of die "woordbou" in die onderrig van Bantoetale voorrang geniet, maar dat daar uiteenlopende vertolkings is van wat as 'n morfeem, stam en wortel beskou kan word (Pretorius 2000: 51). Vir doeleindes van die artikel word die beskouing van Krüger (1973, 1994, 2006) en Laas (1974) as basis vir verdere bespreking van die Setswanamorfologie geneem.

Krüger onderskei tussen leksikale morfeme (wat wortels en stamme insluit) en grammatiese morfeme (wat prefikse en suffikse insluit) (1994: 18). Na aanleiding van Krüger (1994) se beskouing kan die taksonomie in die woordmorfologie soos in Figuur 3 voorgestel word (vergelyk Pretorius, 2000: 58).



Figuur 3 'n Taksonomie vir die Setswanamorfologie.

Ten einde te bepaal wat die genormaliseerde basisvorm is, sal die terme "woord" en "morfeem" (waaronder die stam-, wortel- en grammatiese morfeme) vervolgens bespreek word en daarna met Setswanavoorbeelde toegelig word. Uit die bespreking sal afleidings gemaak word oor wat in Setswana as lemma beskou kan word.

4.4.1 Wat is 'n woord?

Dit is in die eerste plek belangrik om die konsep "woord" te definieer, omdat dit moontlik as lemma beskou kan word. 'n Woord, volgens Trask, is "a label applied in linguistics to any of several rather different conceptions of a unit which is typically larger than a morpheme but smaller than a phrase and which shows a high degree of internal coherence" (1996: 389). Die kompleksiteit van die saak word saamgevat in die woorde: "rather different conceptions of a unit." Crystal sluit by Trask aan as hy daarop wys dat daar probleme met woordgrense is (Crystal 2003: 500-501; Trask 1996: 389). Eenheid lê nie noodwendig in die ortografie nie, want dan sou "washing machine" byvoorbeeld nie as 'n woord beskou word nie. Insgelyks is die ortografie in Setswana wat die werkwoord betref disjunk

en kan woordgrense nie deur die ortografie aangedui word nie (vergelyk Creissels 1996).

Wat is die woord dan volgens die linguiste in die Bantoetale? Volgens Laas is die woord die kleinste, selfstandig-funksionerende taalsimbool en sluit daarmee by Trask se definisie van woord as 'n eenheid "which shows a high degree of internal coherence" aan (Laas 1974: 4; Trask 1996: 389). Volgens Lombard et al. word woorde "betreklik arbitrêr aangedui om hulle in skrif in die praktiese ortografie voor te stel" (1985: 9). Poulos en Louwrens sluit hierby aan as hulle verwys na die disjunktiewe skryfsisteem wat in die Sothotale (spesifiek hier Sesotho sa Leboa) gevolg word, "whereby certain elements that might belong to one and the same word category are written as 'separate' words" (1994: 7). Woorde wat op taalkundige gronde bepaal word, het volgens Lombard et al. een kenmerk in gemeen, en dit is dat die dele van dié woord altyd in "onmiddellike verband met ander dele funksioneer" (1985: 9).

Wat as outonome woorde beskou moet word en wat dele van woorde vorm kan met behulp van woordtoetse wat op woordkenmerke berus, bepaal word (Lombard et al. 1985: 9). Die algemeen aanvaarbare kenmerke van woorde is: isoleerbaarheid, skeibaarheid, omstelbaarheid en vervangbaarheid (Van Wyk 1968: 51-52, Laas 1974: 4, Krüger 2006: 13-16, Lombard et al. 1985: 11-15, Louwrens 1991: 6-9). Taalkundige woorde is "inherent stabiel" op grond van die woorddele se onskeibaarheid, onomstelbaarheid, onvervangbaarheid en onweglaatbaarheid (Lombard et al. 1985: 11). Vervolgens word dié vier toetse, soos wat dit deur Van Wyk ontwikkel is, aan die hand van voorbeelde geïllustreer (1968: 51-52). In hierdie toetse word daar ook na enkele van Zgusta se nege toetse vir multiwoord-leksikale eenhede verwys (1971: 144-151). Hierdie toetse is in verband met bepaling van wat die lemma is veral van belang by Setswana waar die werkwoord disjunk geskryf word.

Toets 1: Isoleerbaarheid

Krüger definieer "isoleerbaarheid" as die eienskap van 'n woord wat in staat is om die volledige werkwoordelike inhoud van 'n sin te konstitueer (2006: 13). Hy toets die isoleerbaarheid deur vrae rondom die werkwoordelike inhoud in van die sin (voorbeeld (1)) te vra (Krüger 2006: 13-14). Een van Zgusta se toetse bepaal dat as 'n multiwoord-leksikale eenheid (soos byvoorbeeld **di bogotse**) op sy eie betekenis het, dit 'n woord is (1971: 146). Dus: as 'n woord in terme van betekenis isoleerbaar is, is dit 'n woord.

(1) **Dintswa di bogotse bosigo.**
Di-ntswa di-bogol-il-e bo-sigo.
CL5.PL-hond AGR.SUBJ-blaf-PERF-TERM CL7.SG-nag.
'Die honde het in die nag geblaf.'

Dintswa di dirile eng?
Di-ntswa di-dir-il-e eng?
CL5.PL-hond AGR.SUBJ-doen-PERF-TERM wat?
'Wat het die honde gedoen?'

Di bogotse.
Di-bogo-il-e.
AGR.SUBJ-blaf-PERF-TERM.
'Hulle het geblaf.'

Toets 2: Skeibaarheid
Onder "skeibaarheid" word verstaan dat woorde wat in een sin direk na mekaar volg, in 'n ander sin kan verskyn met ander woorde tussen hulle (Krüger 2006: 14). Dit beteken dat as daar 'n woord tussenin gebruik word, die twee ortografiese woorde aan weerskante van die ingevoegde woord wel woorde is. Volgens Zgusta is dit onmoontlik om iets tussen dele van 'n multiwoord-leksikale eenheid soos "black market" te plaas, terwyl daar wel 'n woord tussen "illegal" en "market" geplaas kan word, soos byvoorbeeld "illegal street market" (1971: 146). In Setswana kan **yo** ('hierdie') tussen **mosadi** ('vrou') en **o tlile** ('sy het gekom') geplaas word (die woorde in voorbeeld (2) is onderstreep):

(2) **mosadi o tlile maabane**
mo-sadi o-tl-il-e maabane
CL1.SG-vrou AGR.SUBJ-kom-PERF-TERM gister
'die vrou het gister gekom'

mosadi (yo) o tlile maabane
mo-sadi yo o-tl-il-e maabane
CL1.SG-vrou hierdie AGR.SUBJ-kom-PERF-TERM gister
'hierdie vrou het gister gekom'

Toets 3: Omstelbaarheid
Die omstelbaarheidstoets hou in dat as twee konstituente plekke kan ruil en die betekenis dieselfde bly, dit bewys dat daar 'n woordgrens tussen hulle is (Krüger 2006: 15). In voorbeeld (3) is **mosadi** en **maabane** woorde omdat hulle omstelbaar is.

(3) **Mosadi o tlile maabane.**
Mo-sadi o-tl-il-e maabane.
CL1.SG-vrou AGR.SUBJ-kom-PERF-TERM gister.
'Die vrou het gister gekom.'

Maabane mosadi o tlile.
Maabane mo-sadi o-tl-il-e.
Gister CL1.SG-vrou AGR.SUBJ-kom-PERF-TERM.
'Gister het die vrou gekom.'

Toets 4: Vervangbaarheid
Wat vervangbaarheid betref, kan die woorde volgens Krüger (2006: 16) in 'n gegewe sin gewoonlik deur ander woorde of woordkombinasies vervang word (vergelyk voorbeelde (4) - (5)). Volgens Zgusta kan woorde ook deur sinonieme vervang word (1971: 148).

(4) **Mosadi o tlile maabane.**
Mo-sadi o-tl-il-e maabane.
CL1.SG-vrou AGR.SUBJ-kom-PERF-TERM gister.
'Die vrou het gister gekom.'

Mosadi kan byvoorbeeld deur **ene** vervang word:

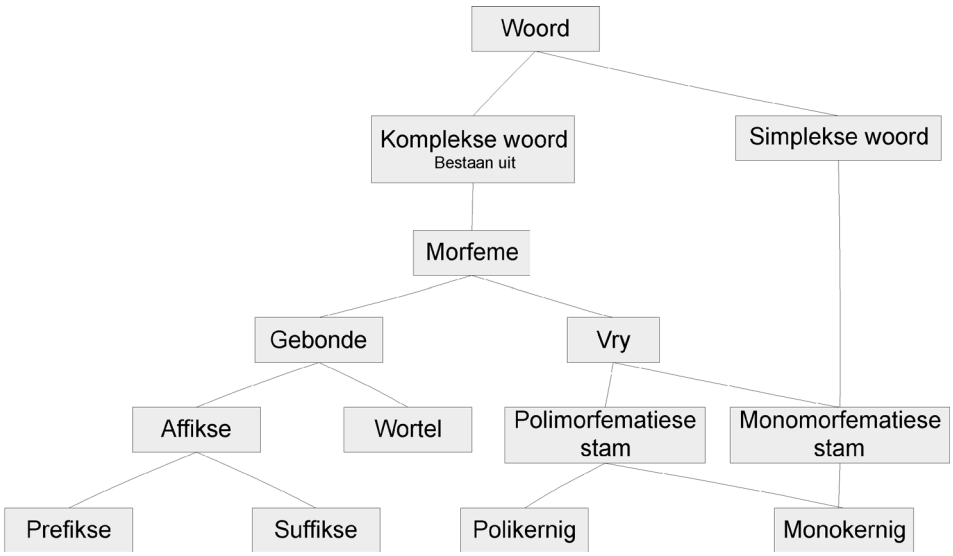
(5) **Ene o tlile maabane.**
Ene o-tl-il-e maabane.
Sy AGR.SUBJ-kom-PERF-TERM gister.
'Sy het gister gekom.'

Die *o-* in **o tlile** kan egter glad nie deur 'n ander woord vervang word sonder dat dit tot 'n ongrammatikale uitdrukking lei nie. Gestel *o-* word deur die woord **ene** ('hy/sy') vervang, dan vorm dit ***ene tlile** wat ongeldig is. In aansluiting hierby merk Zgusta op dat vervanging, met betrekking tot multiwoord-leksikale eenhede, onmoontlik is (1971: 144). Die "good" in die groetvorm "Good morning!" kan byvoorbeeld nie deur "excellent" vervang word nie, want daar is nie 'n groetvorm soos "*excellent morning!" nie.

Volgens Lombard et al. is dit voldoende indien bewys kan word dat 'n taaleenheid óf isoleerbaar óf skeibaar óf omstelbaar óf vervangbaar is, om tot die gevolgtrekking te kom dat so 'n eenheid 'n woord is (1985: 15). Indien 'n taaleenheid soos *o-* in **o tlile** nie aan die toetse voldoen nie, kan dit slegs as 'n morfeem beskryf word. Wat dan as morfeem in Setswana beskou word, word vervolgens bespreek.

4.4.2 Wat is 'n morfeem?
'n Morfeem kan oor die algemeen gedefinieer word as die minimale onderskeidende eenheid in die grammatika en is van sentrale belang in die morfologie (Crystal 2003: 300; Trask 1996: 227; Haspelmath 2002: 16). Volgens Crystal kan daar

onderskei word tussen vrye en gebonde vorms ("free and bound forms") van morfeme, waar die vrye vorms (byvoorbeeld stamme) onafhanklik kan verskyn, maar die gebondes (byvoorbeeld affikse) nie (2003: 300). Vergelyk Figuur 4 vir 'n visuele voorstelling van dié onderskeid.

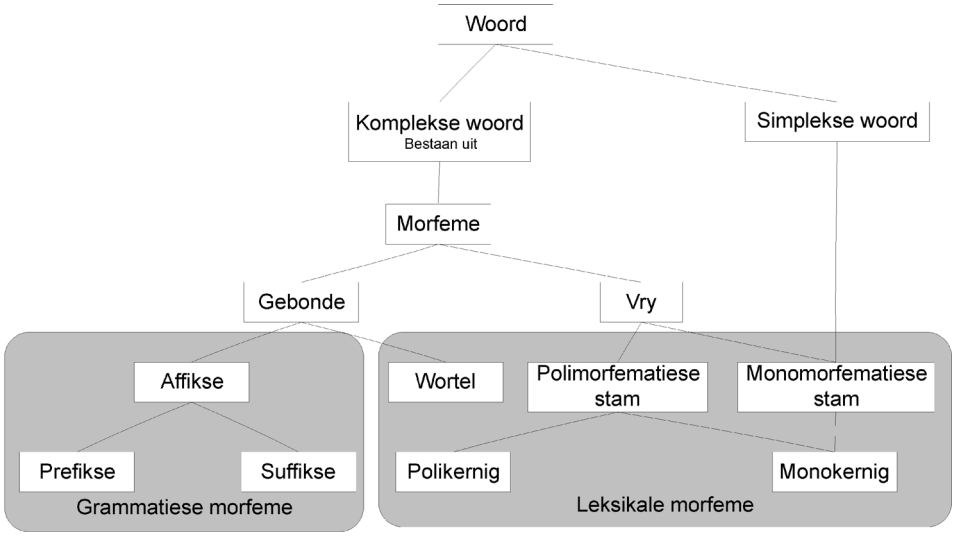


Figuur 4 Taksonomie van die Setswanawoord

Volgens Laas lei bogenoemde klassifikasie van vrye en gebonde morfeme tot verwarring (1974: 26). Wortels is, net soos affikse, gebonde morfeme omdat hulle nie selfstandig voorkom nie. Daar is dus op grond van hierdie onderskeiding nie 'n duidelike verskil tussen wortel- en grammatiese morfeme nie. Krüger verskil ook sterk van dié soort onderskeid tussen vrye en gebonde morfeme, "whereby the differences between morphemes and words are partly annulled" (1994: 15). Hy onderskei eerder tussen die begrippe "woord" en "morfeem" in navolging van die Nederlandse en Europese linguiste, asook in aansluiting by Van Wyk (1969: 35-43). Krüger (1994) onderskei daarom tussen leksikale (stamme en wortels) en grammatiese morfeme (affikse) en weer tussen die stam en wortel op grond van hulle vermoë al dan nie om op woordvlak te kan optree.

Die term "morfeem" in die Sothotale word uitvoerig deur Laas (1974) gedefinieer. Laas (1974) onderskei soos Krüger (1994) tussen grammatiese en leksikale morfeme en brei só op die konsep "leksikale morfeem" uit: "Die leksikale morfeem het direkte [. . .] korrelasie met 'n woord in die leksikon [. . .]." Hy beskou wortels en stamme as leksikale morfeme en affikse (prefikse en suffikse) as grammatiese morfeme (Laas 1974: 13-33). Krüger sluit by Laas (1974) aan en sê dat die begrip morfeem nie tot grammatiese morfeme beperk moet word nie, want

dit sluit nie al die soorte betekenisvolle woordkomponente (soos die wortels en stamme) in nie (1994: 17). Wat die onderskeiding tussen morfeme betref, stel hy eerder 'n onderskeid tussen periferale morfeme (oftewel grammatiese morfeme) en sentrale morfeme (oftewel leksikale morfeme) voor. Hierdie siening kan versoen word met bostaande, deur Figuur 4 uit te brei soos in Figuur 5, waar die onderskeid tussen leksikale en grammatiese morfeme aangetoon word.



Figuur 5 Uitbreiding van die taksonomie van die Setswanawoord

Uit Figuur 5 blyk dit dat die leksikale morfeem die wortel en stam insluit. Die wortel is die semantiese kern van 'n woord, sonder enige verdere morfeme, en het ook geen woordkorrelaat nie. Die stam, daarenteen, is die deel van 'n komplekse woord wat 'n woordkorrelaat het en een of meer grammatiese morfeme bevat (Krüger 1994: 19). In die volgende afdelings word die verskillende morfeme in meer detail verduidelik.

4.4.2.1 Stam

Volgens Trask is 'n stam "a bound form of a lexical item which typically consists of a root to which one or more affixes have been added and which serves as the immediate base for the formation of grammatic words" (1996: 334). Laas brei verder daarop uit deur dit te definieer as die morfologiese komponent wat 'n woordkorrelaat in die leksikon het en wat derhalwe op die woordvlak kan optree (teenoor die wortel wat onder die woordvlak optree) (1974: 10). Die stam kan op grond van sy betekenis selfstandig optree, maar het nogtans 'n valensie vir grammatiese morfeme, soos dit in voorbeeld (6) na vore kom (Laas 1974: 33).

(6) **ditlowana**
di-tlou-ana
CL5.PL-olifant-DIM
'olifantjies'

Tlou kan selfstandig as 'n woord gebruik word (dit het dus 'n woordkorrelaat), maar dit is terselfdertyd die stam van **tlowana**. Die stam het altyd woordstatus, al is dit die stam van 'n meerledige woord, waar die wortel altyd onder die woordvlak gerealiseer word (vergelyk Krüger 1994: 19).

Volgens Crystal kan die stam óf uit net een wortelmorfeem óf uit twee wortelmorfeme (samestelling), óf uit 'n wortelmorfeem met 'n afleidingsmorfeem bestaan (2003: 433). Die definisies van stamme in Bantoetale sluit nou by dié van Crystal aan deurdat Laas ook tussen stamme op grond van hulle morfologiese samestellings onderskei (Crystal 2003: 433; Laas 1974: 16-18). Die volgende soorte samestellings van morfeme kan geïdentifiseer word:

a) Eenledige monokernige woorde (oftewel simplekse of monomorfematiese stamme)
Dit is woorde of leksikale korrelate wat geen grammatiese morfeme bevat nie. Crystal beskou dit as woorde wat uit 'n enkele (vrye) morfeem bestaan, en dit is volgens hom die teenoorgestelde van meerledige monokernige woorde (2003: 300). Voorbeelde (7) - (9) is enkele van die simplekse in Setswana.

(7) **ruri**
'regtig'

(8) **ka**
'met'

(9) **le**
'en, asook'

b) Meerledige monokernige woorde (oftewel afgeleide of polimorfematiese stamme)
Meerledige monokernige woorde is woorde wat bestaan uit 'n wortel of eenledige woord én een of meer grammatiese morfeme. Dit is volgens Crystal die teenoorgestelde van eenledige monokernige woorde (2003: 300). Enkele meerledige monokernige woorde word geïllustreer in voorbeelde (10) en (11).

(10) **sekolong**
se-kolo-ing
CL4.SG-skool-LOC
'lokaliteit van die skool'

(11) **bomalome**
bo-malome
CL6.PL-oom
'oom-hulle'

c) Polikernige woorde
Dit is meerledige woorde wat bestaan uit twee of meer wortels of eenledige woorde. Daar word tussen samestellings (oftewel komposita) en verdubbelings (oftewel reduplikasies) onderskei. Samestellings bestaan uit twee ongelyksoortige stamme en is woorde wat tot stand kom wanneer twee outonome woorde (of 'n outonome woord en 'n woordgroep) één word (vergelyk Haspelmath 2002: 15-16). In (12) - (14) is voorbeelde van samestellings in Setswana.

(12) **mosadimogolo**
mo-sadi-mo-golo
CL1.SG-vrou-CL1.SG-groot
'ou vrou'

(13) **tautona**
tau-tona
CL5.SG.leeu-groot
'koning, groot leier'

(14) **khudutlou**
khudu-tlou
CL5.SG.skilpad-CL5.SG.olifant
'baie groot skilpad'

'n Verdubbeling is die volledige of onvolledige herhaling van 'n basis, wat weer 'n stam binne 'n meerledige woord kan vorm (vergelyk Haspelmath 2002: 24; Croft 2003: 95). Dit word in die volgende voorbeelde (15) - (17) geïllustreer.

(15) **tautau**
tau-tau
CL5.SG.leeu-CL5.SG.leeu
'kaptein/leier'

(16) **godimodimo**
go-dimo-dimo
CL9-bo-bo
'heel bo'

(17) **mogologolo**
mo-golo-golo
CL1.SG-groot-groot
'die grootste/oudste/belangrikste een'

Die vraag is nou wat die implikasies hiervan is vir lemma-identifisering en die definiëring van die konsep "lemma" in hierdie studie. Dit is by die monokernige stamme maklik om die lemma te bepaal indien die stam in sy eenvoudigste vorm as lemma beskou word. Die lemma van **sekolong** in voorbeeld (10) is eenvoudig **sekolo**.

Dit raak egter moeiliker by die polikernige stamme, veral wat die samestellings betref. Wat sou dan in voorbeeld (14) die lemma wees: **khudu** ('skilpad') of **tlou** ('olifant')? As 'n mens **khudu** ('skilpad') as lemma sou neem, is daar 'n te groot verskil in betekenis tussen die woord **khudutlou** ('baie groot skilpad') en die gekose lemma. Dieselfde geld ook as **tlou** as lemma geneem sou word. Die hele samestelling (in dié geval **khudutlou**) moet dan as lemma geneem word.

By die verdubbelings, waar die wortel volledig of onvolledig herhaal word, sou 'n mens kon redeneer dat die wortel die lemma kan wees. In voorbeeld (15) sou **tau** ('leeu') die lemma van **tautau** ('kaptein') kon wees, maar semanties sal daar dan 'n groot verskil wees tussen die woordvorm (**tautau**) en die lemma (**tau**). Soos by die samestellings, moet die verdubbeling (in dié geval **tautau**) as lemma geneem word.

Polikernige stamme kan egter op hulle beurt weer as stamme binne meerledige woorde optree, wat beteken dat hulle dus weer affikse kan neem. As 'n mens die benadering volg dat die lemma die stam sonder enige affikse is (op voorwaarde dat dit nog steeds betekenisvol is), kan die polikernige stamme, gestroop van affikse, as lemma beskou word. Polikernige stamme (soos **mosadimogolo** in voorbeeld (12)) word gevolglik, soos monokernige stamme, as lemma beskou.

4.4.2.2 Wortel

Volgens Crystal en Trask is 'n wortel daardie deel van die woord wat oorbly as al die affikse verwyder is (Crystal 2003: 402; Trask 1996: 49). Dit kan dus nie verder geanaliseer word sonder dat die betekenis verlore raak nie (Crystal 2003: 402). Indien 'n leksikale morfeem nie op die woordvlak gerealiseer word nie, is dit dus 'n wortel, "waar geen grammatiese morfeem ingesluit is nie" (Laas 1974: 13).

Volgens Laas het die wortel, al is dit die semantiese kern van 'n woord of stam, nie 'n selfstandige betekenis nie (1974: 33). Dit kom duidelik in die volgende voorbeelde na vore. In voorbeeld (18) is *-tho,* voorbeeld (19) *-kolo,* en voorbeeld (20) *-ag-* die wortels.

(18) **batho**
ba-tho
CL1.PL-mens
'mense'

(19) **sekolong**
se-kolo-ing
CL4.SG-skool-LOC
'lokaliteit van die skool'

(20) **go aga**
go-ag-a
INF-bou-TERM
'om te bou'

Die keuse van wat die lemma in Setswana moet wees, is tussen die stam en die wortel soos wat dit in die voorbeeld (21) aangetoon word.

(21) **dira, dirile, direla, dirisa**
dir-a, dir-il-e, dir-el-a, dir-is-a
werk-TERM, werk-PERF-TERM, werk-APPL-TERM, werk-CAUS-TERM
'werk, gewerk, werk vir, laat werk'

In voorbeeld (21) is die keuse vir 'n lemma tussen *dir-* (die wortel) of **dira** (die eenvoudigste stam van **dirile, direla,** en **dirisa**). Die eenvoudigste stam verwys na die stam wat die minste fleksiemorfeme moontlik het, in die stap voordat dit as 'n wortel geanaliseer word. Aangesien wortels onder die woordvlak realiseer en daarom nie selfstandig kan optree nie, kan wortels nie as lemmas beskou word nie.

4.4.2.3 Affiks

Soos reeds aangetoon, word affikse as grammatiese morfeme beskou (Laas 1974: 33; Krüger 1994: 17). Grammatiese morfeme word gebruik om grammatikale verhoudings tussen 'n woord en die konteks uit te druk (Crystal 2003: 300). Laas stel voor dat grammatiese morfeme gedefinieer word "as 'n grammatiese-relevante fonologiese segment in 'n woord of as 'n grammatiese waarde in 'n woord/ semantiese aspek in 'n woord" (1974: 18). Grammatiese morfeme is woordgebonde

in vorm en betekenis en kan nooit selfstandig optree nie. Trask (1996) sluit daarby aan en sê dat 'n affiks 'n gebonde morfeem is wat slegs voorkom as dit vas aan 'n woord of stam gebruik word. In die geval van Setswana, waar die werkwoord (wat die prefikse betref) disjunk geskryf word, word die *ke* en *a* in **ke a dira** egter as affikse beskou, ten spyte van die feit dat dit nie vas aan die woord geskryf word nie. **Ke a dira** word dus as 'n woord in Setswana beskou, ten spyte van die feit dat dit as drie ortografiese eenhede geskryf word (vergelyk voorbeeld (22)).

(22) **ke a dira**
ke-a-dir-a
AGR.SUBJ-TEMP-werk-TERM
'ek werk'

Voorbeelde van affikse is die klasprefiks *mo-* en die lokatiewe suffiks *-ing* in voorbeeld (23), en die infinitiewe prefiks *go-*, die applikatiewe suffiks *-el-* en die uitgangsmorfeem *-a* in voorbeeld (24). Hierdie affikse kan nie in isolasie voorkom nie en kan dus nie as woorde beskou word nie (vergelyk die woordtoetse onder 2.3.3.2). So byvoorbeeld kan die lokatiewe suffiks *-ing* nie sonder 'n stam of woord soos **motse** bestaan nie. Dieselfde geld ook vir die uitgangsmorfeem *-a* in voorbeeld (24).

(23) **motseng**
mo-tse-ing
CL2.SG-stat-LOC
'lokaliteit van die stat'

(24) **go direla**
go-dir-el-a
INF-werk-APPL-TERM
'om te werk vir'

Die grammatiese morfeme is nie geskik om as lemmas op te tree nie, omdat hulle woordgebonde is en nie selfstandig kan optree nie. Die grammatiese morfeme wat in Setswana nie in isolasie kan voorkom nie, is juis dié morfeme wat in die lemma-identifiseringsproses verwyder moet word.

4.5 Die lemma in Setswana

Hierbo is aangetoon dat daar twee hoofgroepe morfeme in Setswana onderskei kan word, te wete die leksikale morfeme en grammatiese morfeme. Die leksikale morfeme kan in stamme en wortels verdeel word, waar die stam op grond van

betekenis selfstandig kan optree, maar nogtans 'n valensie vir grammatiese morfeme het. Volgens Laas het die wortel, al is dit die semantiese kern van 'n woord of stam, nie 'n selfstandige betekenis nie (1974: 33). Die belangrikste verskil tussen die stam en wortel is egter dat die stam op woordvlak kan optree (d.i. selfstandig is), waar die wortel onder die woordvlak optree (d.i. onselfstandig is) (Laas 1974: 10). Die ander hoofgroep, d.i. die grammatiese morfeem, is vir bestaan en betekenis van die leksikale morfeem afhanklik (Krüger 1994: 19). Grammatiese morfeme en wortels (soos in voorbeeld (25) geïllustreer word) kan, vanweë hulle afhanklike aard, nie as lemmas geneem word nie.

(25) **lebati, lebatsana, lebating, mabati, mabatsana, mabating**
le-bati, le-bati-ana, le-bati-(i)ng, ma-bati, ma-bati-ana, ma-bat-i-(i)ng
CL3.SG-deur, CL3.SG-deur-DIM, CL3.SG-deur-LOC, CL3.PL-deur, CL3.PL-deur-DIM, CL3.PL-deur-LOC
'deur, deurtjie, lokaliteit van die deur, deure, deurtjies, lokaliteit van die deure'

Die keuse vir 'n lemma sou tussen die leksikale morfeme (die wortel *-bati* en die eenvoudigste stam **lebati** of **mabati**) lê. Die eenvoudigste stam hier verwys na die stam wat die minste fleksiemorfeme moontlik het, in die stap voordat dit as 'n wortel geanaliseer word. Weens die onafhanklike aard van die eenvoudigste stam, word dit as lemma in Setswana geneem. Die lemma in Setswana word daarom beskou as die eenvoudigste stam voordat dit as wortel geanaliseer word. Die woord moet ook nie van woordsoort verander tydens lemma-identifisering nie – wat beteken dat slegs fleksie-affikse verwyder moet word. Dit is soortgelyke beginsels wat in die ontwikkeling van 'n Poolse lemma-identifiseerder gevolg is (vergelyk Vetulani et al. 1998: 31).

Daar is sewe woordsoorte in Setswana: die naamwoord, voornaamwoord, werkwoord, betrekkingswoord, bywoord, interjeksie en ideofoon (vergelyk bespreking onder 3.1-3.7). Nie al hierdie woordsoorte is ewe groot of word ewe maklik uitgebrei nie. Handke onderskei daarom tussen oopklaswoorde ("open-class words") en gesloteklaswoorde ("closed-class words") (1995: 25). Die oop klas, bestaande uit naamwoorde, werkwoorde, bywoorde en byvoeglike naamwoorde, word maklik uitgebrei, terwyl die geslote klas, bestaande uit voorsetsels, voornaamwoorde en interjeksies, nie sommer uitgebrei word nie.

Wat Setswana betref, kan vyf van dié sewe woordsoorte (d.i. voornaamwoorde, bywoorde, betrekkingswoorde, interjeksies en ideofone) as die meer geslote kategorieë beskou word, omdat hulle nie morfologies uitgebrei kan word nie. Omrede die geslote kategorieë nie morfologies uitgebrei word nie, kan hulle net so as lemmas geneem word. Dit is derhalwe nie nodig om te bepaal watter affikse

by hierdie woordsoorte verwyder moet word ten einde die lemma te identifiseer nie. Die ander twee woordsoorte (naamwoorde en werkwoorde) vereis egter wel die implementering van morfologiese en morfo-fonologiese reëls om die lemmas (d.i. die eenvoudigste stamme) te bepaal.

3.5.1 Die naamwoordlemma

As die algemene definisie op die naamwoord van toepassing gemaak word, beteken die naamwoordlemma die vorm in die enkelvoud sonder enige suffikse (behalwe die deverbatiewe suffiks waar dit voorkom) is. Die suffikse wat nie in die lemma ingesluit word nie, is die lokatiewe, diminutiewe, feminitiewe en augmentatiewe suffikse (vergelyk Tabel 3).

Tabel 3 Die morfologie van die naamwoord

Stam		**suffikse**				
klasprefiks	wortel	deverbatief	augmentatief/ feminitief	diminutief	diminutief	lokatief
		-i/-o	*-gadi*	*-ana*	*-anyana*	*-(i)ng*

Die lemma van **koloyana** in voorbeeld (26) is dus **koloi**, d.i. die eenvoudigste stam in die enkelvoud sonder die diminutiewe suffiks *-ana*.

(26) **koloyana**
koloi-ana
CL5.SG.motorkar-DIM
'motorkarretjie'

Die uitsonderings op die definisie vir die naamwoordlemma is die beskrywende naamwoorde, plek- en eiename. Beskrywende naamwoorde neem verskeie vorms aan, omdat hulle vormlik afhanklik is van die naamwoord wat hulle beskryf. In die geval van die beskrywende naamwoorde word die wortel, en nie die eenvoudigste stam nie, as lemma gebruik (vergelyk voorbeelde (27)-(28)). Die plek- en eiename word onveranderd gelos (vergelyk voorbeeld (29)).

(27 *-golo*
'groot'

(28) **molelo o mogolo**
mo-lelo o mo-golo
CL2.SG-vuur PART.wat is CL2.SG-groot
''n groot vuur'

(29) **Mafikeng**
Ma-fika-ing
CL3.PL-fika-LOC
'Plek van die klippe'

3.5.2 Die werkwoordlemma

Die werkwoord word in die egte werkwoord, hulpwerkwoord en kopulatiewe werkwoord verdeel. Die egtewerkwoordlemma is die eenvoudigste stam in die infinitief, sonder enige suffikse (behalwe die onproduktiewe en onaktiewe suffikse). Dit is dan die stam van die werkwoord sonder die volgende affikse: die negatiewe, kongruensie-, aspek- of temporale morfeme, neutro-passiewe, iteratiewe, kousatiewe, applikatiewe, resiprokale, perfektum, passiewe suffikse en die uitgangsmorfeme **-e** en **-ng**. Dit is dus, in kort, die werkwoord in die infinitief, hoewel die *go-* in navolging van woordeboekgebruik en uit praktiese redes nie gebruik gaan word nie. Die lemma van **o agile** (voorbeeld (30)) sal dus **aga** sonder die onderwerpsmorfeem *o* en die perfektumsuffiks *-il-* wees.

(30) **o agile**
o-ag-il-e
AGR.SUBJ-bou-PERF-TERM
'jy/hy/sy het gebou'

Die hulpwerkwoord kan onderverdeel word in egte en onegte hulpwerkwoorde. Egte hulpwerkwoorde word net so as lemmas geneem (voorbeeld (31)). Beginsels wat by die bepaling van die egte werkwoordlemma geld, is ook by die onegte hulpwerkwoorde van toepassing

(31) **ketla**
'sal nie'

Hoewel die kopulatiewe werkwoord, afhangende van die tyd en modus waarin dit verskyn, verskillende vorms aanneem, kan die basiese kopulatiewe werkwoord **nna** as die lemma beskou word. Die lemma van die kopulatiewe werkwoord **ke** in voorbeeld (32) is dan eenvoudig **nna**.

(32) **Mosadi ke mooki**
Mo-sadi ke mo-ok-i
CL1.SG-vrou is CL1.SG-verpleeg-TERM
'Die vrou is 'n verpleegster.'

Die lemmas van die verskillende voornaamwoorde, betrekkingswoorde, bywoorde, interjeksies en ideofone is net soos wat dit in werklike taalgebruik voorkom. Die rede daarvoor is dat die woordsoorte redelik geslote is en dus nie soos in die geval van die naamwoord en werkwoord morfologiese uitgebrei word nie.

4. Gevolgtrekking

Na aanleiding van die bespreking is vasgestel dat die lemma die genormaliseerde basisvorm is wat ook 'n lekseem is (die lemma is daarom nie net 'n basisvorm nie). Die lemma in Setswana is, na verskeie oorwegings, met behulp van die woordmorfologie gedefinieer en daarmee is aangetoon dat daar in Setswana tussen die leksikale morfeme en grammatiese morfeme onderskei word. Die stam en die wortel is leksikale morfeme, en die affikse is grammatiese morfeme. 'n Verdere onderskeiding is tussen die wortel en die stam; die wortel is die semantiese kern van die woord, maar realiseer onder die woordvlak, terwyl die stam op die woordvlak realiseer. As die definisie in berekening gebring word dat die lemma terselfdertyd 'n basisvorm en 'n lekseem is, is 'n lemma in Setswana die stam in sy eenvoudigste vorm (met die minste affikse moontlik), voordat dit as 'n wortel geanaliseer word. Alhoewel die definisie vir die lemma in Setswana in beginsel vasgestel is, lê die uitdaging egter nou daarin om dit met behulp van 'n outomatiese lemma-identifiseerder op meer komplekse vorme, soos die polikernige naamwoorde en die kopulatiewe werkwoorde, van toepassing te maak.

Bronnelys

Brown, J.T. 1988. *Setswana-English Dictionary.* Johannesburg: Pula Press.

Bussman, H. 1996. *Routledge Dictionary of Language and Linguistics.* London: Routledge.

Choueka, Y., E.S. Conley, and I. Dagan. 2000. "A Comprehensive Bilingual Word Alignment System." *Parallel Text Processing. Alignment and Use of Translation Corpora.* Ed. J. Véronis. Dordrecht : Kluwer Academic Publishers. 69-96.

Creissels, D. 1996. "Conjunctive and Disjunctive Verb Vorms in Setswana." *South African Journal of African Languages* 16 (4): 109-115.

Croft, W. 2003. *Typology and Universals.* 2nd ed. New York: Cambridge University Press.

Crystal, D. 1992. *An Encyclopedic Dictionary of Language and Languages.* Oxford: Blackwell.

____. 2003. *A Dictionary of Linguistics and Phonetics.* 5th ed. Oxford: Blackwell.

Daelemans, W., and H. Strik, eds. 2002. *Het Nederlands in de taal- en spraaktechnologie: Prioriteiten voor basisvoorzieningen.* Final Report (01/07/2002). Dutch Language Union. 4 Des. 2005. <taalunieversum.nl/taal/technologie/docs/daelemans-strik.pdf> .

De Schryver, G-M, and D.J. Prinsloo. 2001. "Towards a Sound Lemmatisation Strategy for the Bantu Verb Through the Use of Frequency-Based Tail Slots – With Special Reference to Cilubà, Sepedi and Kiswahili." *Makala ya Kongamano la Kimataifa Kiswahili 2000 Proceedings*. Eds J.S. Mdee, and H.J.M. Mwansoko. Dar Es Salaam: Tuki, Chuo Kikuu Cha Dar Es Salaam. 216-242.

Dent, G.R. 1994. *Kompakte Setswana Woordeboek.* Pietermaritzburg: Shuter & Shooter.

Erjavec, T., and S. Džeroski. 2004. "Machine Learning of Morphosyntactic Structure: Lemmatising Unknown Slovene Words." *Applied Artificial Intelligence* 18(1): 17-40.

Gouws, R.H. 1989. *Leksikografie.* Pretoria: Akademica.

Gouws, R.H., and D.J. Prinsloo. 2005. *Principles and Practice of South African Lexicography.* Stellenbosch: Sun Press.

Handke, J. 1995. *The Structure of the Lexicon: Human Versus Machine***.** Berlin: Mouton De Gruyter.

Hartmann, R.K.K. 2001. *Teaching and Researching Lexicography***.** Harlow: Longman.

Hartmann, R.K.K., and G. James. 1998. ***Dictionary of Lexicography.*** London: Routledge.

Haspelmath, M. 2002. *Understanding Morphology.* New York: Oxford University Press.

Hausser, R. 1999. *Foundation of Computational Linguistics: Man-Machine Communication in Natural Language.* Berlin: Springer.

Herbert, R., and R. Baile. 2002. "The Bantu Language: Sociohistorical Perspectives." *Languages in South Africa*. Ed. Rajend Mesthrie. Cambridge: Cambridge University Press. 50-79.

Jackson, H. 1988. *Words and Their Meanings.* New York: Longman.

Jurafsky, D., and J.H. Martin. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* Upper Saddle River, N.J.: Prentice Hall.

Katamba, F. 2005. *English Words: Structure, History, Usage.* 2nd ed. New York: Routledge.

Knowles, G., and Z.M. Don. 2004. "The Notion of a 'Lemma': Headwords, Roots and Lexical Sets." *International Journal of Corpus Linguistics* 9(1): 69-81.

Kriel, T.J. 1958. *The New Sesotho-English Dictionary.* Johannesburg: Afrikaanse Pers-Boekhandel.

Kriel, T.J., E.B. Van Wyk, and S.A. Makopo. 1989. *Pukuntšu.* 4th ed. Pretoria: Van Schaik.

Krüger, C.J.H. 1973. "Woordanalise in Sotho." *Limi* 1(2): 1-11.

____. "Word-Based Versus Root-Based Morphology in the African Languages." *South African Journal of African Languages* 14(1): 15-23.

____. 2006. *Introduction to the Morphology of Setswana.* München: Lincom Europa.

Laas, J.A.M. 1974. "Woordbou en woordanalise in Suid-Sotho." MA Diss. Potchefstroom: PU vir CHO.

Lombard, D.P., E.B. Van Wyk, and P.C. Mokgokong. 1985. *Inleiding tot die grammatika van Noord-Sotho.* Pretoria: Van Schaik.

Louwrens, L.J. 1991. *Aspects of Northern Sotho Grammar.* Pretoria: Via Afrika.
____. 1994. *Dictionary of Northern Sotho Grammatical Terms.* Pretoria: Via Afrika.
Manning, C.D., and H. Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge: The Mit Press.
Matthews, P.H. 1997. *Oxford Concise Dictionary of Linguistics.* New York: Oxford.
Mcenery, T., and A. Wilson. 2001. *Corpus Linguistics: An Introduction.* Edinburgh: Edinburgh University Press.
Mitkov, R. 2003. *The Oxford Handbook of Computational Linguistics.* New York : Oxford University Press.
Nasionale Departement van Kuns en Kultuur: Suid-Afrika. 2005. *Strategic Plan.* 1 Feb. 2005 <http://www.dac.gov.za/publications/strategic_plan/str_plan2005_10.pdf>.
Ooi, V.B.Y. 1998. *Computer Corpus Lexicography.* Edinburgh: Edinburgh University Press.
Plisson, J., N. Lavrac, and D. Mladenic. 2004. "A Rule-Based Approach to Word Lemmatization." Proceedings C Of The 7th International Multi-Conference Information Society IS 2004, October 2004, Ljubljana. 83-86.
Poulos, G., and L.J Louwrens. 1994. *A Linguistic Analysis of Northern Sotho.* Pretoria: Via Afrika.
Pretorius, R.S. 1997. "Auxiliary Verbs as a Subcategory of the Verb in Tswana." Diss. Potchefstroom: PU vir CHO.
Pretorius, W.J. 2000. "Die identifisering en beskrywing van die begrippe stam en wortel in die Bantoetale, met besondere verwysing na die Sothotale." *Journal for Language Teaching* 34(1): 51-62.
Reynierse, C. 1991 *South African Multi-Language Dictionary and Phrase Book: English, Afrikaans, Northern Sotho, Sesotho, Tswana, Xhosa, Zulu.* Cape Town: Reader's Digest.
Snoxall, R.A. 1965. "Some Problems and Principles of Lexicography in Luganda." *African Language Studies* 6: 27-31.
Snyman, J.W., J.S. Shole, and J.C. Le Roux. 1990. *Setswana-Engels-Afrikaanse woordeboek.* Pretoria: Via Afrika.
Trask, R.L. 1996. *A Dictionary of Phonetics and Phonology.* London: Routledge.
Trost, H. 2003. "Morphology." *The Oxford Handbook of Computational Linguistics*. Ed. R. Mitkov. New York: Oxford University Press. 25-47.
Van Sterkenburg, P. 2003. *A Practical Guide to Lexicography.* Amsterdam: Benjamins.
Van Wyk, E.B. 1966. "The Word-Classes in Northern Sotho: Die woord as linguistiese eenheid." *Lingua* 17(2): 230-261.
____. 1968. "Die woord as linguistiese eenheid." *Student en dosent: Klasgids* 11 (1968).
____. 1969. "Kritiese ontleding van die morfeemopvatting." *Klasgids* 4(3): 35-43.
Vetulani, Z., J. Martinek, T. Obrębski, and G, Vetulani. 1998. *Dictionary Based Methods and Tools for Language Engineering.* Poznań: Wydawnictwo Naukowe UAM.
Zgusta, L. 1971. *Manual of Lexicography.* Paris: Mouton.
Ziervogel, D., and P.C. Mokgokong. 1985. *Groot Noord-Sotho woordeboek.* 2de uitgawe. Pretoria: Van Schaik.