

Outomatiese Setswana Lemma-identifisering

Jeanetta Hendrina Brits

BBk. Hons.B.A.

**Verhandeling voorgelê vir die graad Magister
Artium in Setswana aan die Noordwes-Universiteit**

Studieleier: Dr. R.S. Pretorius

Medestudieleier: Prof. G.B. van Huyssteen

2006

Potchefstroomkampus

**Opgedra aan Philna Kloppers, my familie en
vriende**

Bedankings

- My studieleiers, dr. R.S. Pretorius en prof. G.B. van Huyssteen – baie, baie dankie vir julle opoffering, sinvolle leiding en begrip.
- My ouers en familie, vir die geleentheid dat ek kon studeer en vir julle ondersteuning.
- My direkteur, mnr. Theo Cloete, vir jou ondersteuning.
- My kollegas by Bemarking en Kommunikasie – dankie vir julle belangstelling.
- Prof. Attie de Lange, vir belangstelling en aanmoediging.
- Israel Matlhola, Motshediso Nyelele, Leipa Phuthi le Motheo Monaisa, ke a leboga.
- Annick van Dijkhorst – daar is nie genoeg ‘bullets’ om jou te bedank vir baie ure se geduld en hulp met die programmering nie.
- Handré Groenewald – dankie vir jou hulp met die laaste programmering.
- Ansu Berg en Suléne Pilon – vir julle raad.
- CText – vir die ure wat ek in die laboratorium kon deurbring.
- Willie Cloete – vir die taalversorging.
- Soli Deo Gloria!

SUMMARY

Key terms: Computational linguistics; Setswana grammar; Setswana morphology; Lemmatisation; Stemming; Lemma; Natural language processing, Regular expression; Finite state automata; Finite state transducer; *FSA 6*

Within the context of natural language processing, a lemmatiser is one of the most important core technology modules that has to be developed for a particular language. A lemmatiser reduces words in a corpus to the corresponding lemmas of the words in the lexicon.

A lemma is defined as the meaningful base form from which other more complex forms (i.e. variants) are derived. Before a lemmatiser can be developed for a specific language, the concept "lemma" as it applies to that specific language should first be defined clearly. This study concludes that, in Setswana, only stems (and not roots) can act independently as words; therefore, only stems should be accepted as lemmas in the context of automatic lemmatisation for Setswana.

Five of the seven parts of speech in Setswana could be viewed as closed classes, which means that these classes are not extended by means of regular morphological processes. The two other parts of speech (nouns and verbs) require the implementation of alternation rules to determine the lemma. Such alternation rules were formalised in this study, for the purpose of development of a Setswana lemmatiser. The existing Setswana grammars were used as basis for these rules. Therewith the precision of the formalisation of these existing grammars to lemmatise Setswana words could be determined.

The software developed by Van Noord (2002), *FSA 6*, is one of the best-known applications available for the development of finite state automata and transducers. Regular expressions based on the formalised morphological rules were used in *FSA 6* to create finite state transducers. The code subsequently generated by *FSA 6* was implemented in the lemmatiser.

The metric that applies to the evaluation of the lemmatiser is precision. On a test corpus of 1 000 words, the lemmatiser obtained 70,92%. In another evaluation on 500 complex nouns and 500 complex verbs separately, the lemmatiser obtained 70,96% and 70,52% respectively. Expressed in numbers the precision on 500 complex and simplex nouns was 78,45% and on complex and simplex verbs 79,59%. The quantitative achievement only gives an indication of the relative precision of the grammars. Nevertheless, it did offer analysed data with which the grammars were evaluated qualitatively. The study concludes with an overview of how these results might be improved in the future.

OPSOMMING

Sleuteltermes: Rekenaarlinguistiek; Setswanagrammatika; Setswanamorfologie; Lemma-identifisering; Stamidentifisering; Lemma; Natuurliketaalprosessering; Reëlmatige uitdrukking; Eindigestaatoutomaat; Eindigestaatoorvormer; *FSA 6*

Binne die konteks van natuurliketaalprosessering, is 'n lemma-identifiseerder ("lemmatiser") een van die belangrikste kerntegnologiemodules wat vir 'n bepaalde taal ontwikkel moet word. 'n Lemma-identifiseerder herlei woorde in 'n korpus na die ooreenstemmende lemmas van die woorde in die leksikon.

'n Lemma word gedefinieer as die betekenisvolle basisvorm waarvan ander meer komplekse vorms (d.i. variante) afgelei word. Voordat 'n lemma-identifiseerder vir 'n spesifieke taal ontwikkel kan word, moet die konsep "lemma" soos wat dit van toepassing op die spesifieke taal is, duidelik gedefinieer word. Die slotsom van dié studie is dat slegs stamme (en nie wortels nie) in Setswana selfstandig as woorde kan optree, en derhalwe moet net stamme as lemmas in die konteks van outomatiese lemma-identifisering vir Setswana aanvaar word.

Vyf van die sewe woordsoorte in Setswana kan as geslote klasse beskou word, wat beteken dat dié klasse nie deur reëlmatige morfologies prosesse uitgebrei word nie. Die ander twee woordsoorte (naamwoorde en werkwoorde), vereis die implementering van omvormingsreëls om die lemma te bepaal. Sodanige omvormingsreëls is in hierdie studie geformaliseer, met die oog op die ontwikkeling van 'n Setswana lemma-identifiseerder. Die bestaande grammatikas van Setswana is as basis vir dié reëls gebruik. Daarmee is dan ook bepaal hoe akkuraat die bestaande grammatikas geformaliseer kan word om die lemmas in Setswana te identifiseer.

Die sagteware wat deur Van Noord (2002) ontwikkel is, *FSA 6*, is een van die bekendste toepassings beskikbaar vir die ontwikkeling van eindigestaatoutomate en -oorvormers. Reëlmatige uitdrukkings wat gebaseer is op die geformaliseer-

de morfologiese reëls, is in *FSA 6* gebruik om eindigestaatoorvormers te skep. Die kode wat vervolgens deur *FSA 6* gegenereer is, is in die lemma-identifiseerder geïmplementeer.

Die kriteria wat vir die evaluering van die lemma-identifiseerder geld, is akkuraatheid ("precision"). Op 'n toetskorpus van 1 000 woorde, behaal die lemma-identifiseerder 70,92%. In 'n ander evaluering op 500 komplekse naamwoorde en 500 komplekse werkwoorde apart, behaal die lemma-identifiseerder onderskeidelik 70,96% en 70,52%. In syfers uitgedruk is die akkuraatheid op 500 komplekse en simplekse naamwoorde 78,45% en op komplekse en simplekse werkwoorde 79,59%. Die kwantitatiewe prestasie gee egter slegs 'n aanduiding van die relatiewe akkuraatheid van die grammatikas. Dit het tog geanaliseerde data gebied waarmee die grammatikas kwalitatief geëvalueer is. Die studie sluit af met 'n vooruitskouing van hoe hierdie resultate moontlik in die toekoms verbeter kan word.

INHOUDSOPGAWE

HOOFSTUK 1 – INLEIDING

1.1	INLEIDING	1
1.2	KONTEKS: WAT IS LEMMA-IDENTIFISERING?	1
1.3	PROBLEEMSTELLING	4
1.4	NAVORSINGSVRAE	10
1.5	DOELSTELLINGS	10
1.6	SENTRALE TEORETIESE STELLINGS	10
1.7	METODE VAN ONDERSOEK	11
1.7.1	Definiëring en identifisering van die lemma in Setswana	11
1.7.2	Ontwikkeling van 'n lemma-identifiseerder	12
1.7.2.1	Formalisering van bestaande grammatika	12
1.7.2.2	Die ontwikkeling van datastelle	12
1.7.2.3	Evaluering van die lemma-identifiseerder	13
1.8	Ontplooiing	13

HOOFSTUK 2 – DIE LEMMA IN SETSWANA

2.1	INLEIDING	15
2.2	DIE KONSEP “LEMMA”: ALGEMENE PERSPEKTIEF	15
2.2.1	“Lemma” as leksikografiese term	16
2.2.2	“Lemma” as leksikologiese term	17
2.2.3	“Lemma” as morfologiese term	19
2.3	DIE KONSEP LEMMA: TAALSPESIFIEKE PERSPEKTIEF	20
2.3.1	“Lemma” as leksikografiese term	21
2.3.2	“Lemma” as leksikologiese term	24
2.3.3	“Lemma” as morfologiese term	24
2.3.3.1	Algemene Setswana woordmorfologie	25
2.3.3.2	Wat is 'n woord?	29
2.3.3.3	Wat is 'n morfeem?	35
2.3.3.3.1	Stam	38

INHOUDSOPGAWE (vervolg)

2.3.3.3.2	Wortel	43
2.3.3.3.3	Affiks	44
2.3.4	Voorlopige gevolgtrekking: Die lemma in Setswana	45
2.4	MORFOLOGIE VAN WOORDSOORTE	46
2.4.1	Naamwoorde	48
2.4.2	Werkwoorde	54
2.4.2.1	Egte werkwoord	54
2.4.2.2	Hulpwerkwoord	58
2.4.2.3	Kopulatiewe werkwoorde	59
2.4.3	Voornaamwoorde	63
2.4.4	Betrekkingswoorde	65
2.4.5	Bywoorde	65
2.4.6	Interjeksies	66
2.4.7	Ideofone	67
2.5	GEVOLGTREKKING: DIE LEMMA IN SETSWANA	67

HOOFTUK 3 – LINGUISTIESE REËLS VIR LEMMA-IDENTIFISERING IN SETSWANA

3.1	INLEIDING	71
3.2	FORMALISERING VAN LINGUISTIESE REËLS: EINDIGESTAATOUTOMATE	72
3.2.1	Reëlmatige tale, reëlmatige uitdrukkings en outomata	72
3.2.2	FST en tweevlakmorfologie (“two-level morphology”)	75
3.3	EINDIGESTAATSAGTEWARE: FINITE STATE AUTOMATA UTILITIES FSA 6	77
3.3.1	Reëlmatige uitdrukkings in FSA 6	79
3.3.2	Makro’s en Prolog	81
3.4	REËLS	82
3.4.1	Morfologiese reëls vir naamwoorde	83
3.4.1.1	Die naamwoordklasse	84

	INHOUDSOPGAWE (vervolg)	
3.4.1.1.1	Lokatiewe suffiks -ng	84
3.4.1.1.2	Diminutiewe suffiks -anyana, -ana en -nyana	86
3.4.1.1.3	Augmentatiewe en feminitiewe suffikse	92
3.4.1.1.4	Meervoudsprefikse	93
3.4.1.2	Die beskrywende naamwoorde	99
3.4.2	Reëls vir werkwoorde	100
3.4.2.1	Die egte werkwoord	102
3.4.2.1.1	Prefikse	102
3.4.2.1.2	Uitgangsmorfeem -ng	107
3.4.2.1.3	Die passief in die perfektum	108
3.4.2.1.4	Passief	109
3.4.2.1.5	Perfektum	113
3.4.2.1.6	Uitgangsmorfeem -e	118
3.4.2.1.7	Resiprokalis	119
3.4.2.1.8	Applikatief	120
3.4.2.1.9	Kousatief	122
3.4.2.1.10	Iteratief	126
3.4.2.1.11	Neutro-passief	126
3.4.2.2	Die hulpwerkwoord	127
3.4.2.3	Die kopulatiewe werkwoord	129
3.4.3	Ander woordsoorte	131
3.5	GEVOLGTREKKING	131

HOOFTUK 4 – 'n LEMMA-IDENTIFISEERDER VIR SETSWANA: IMPLEMENTERING EN EVALUERING

4.1	INLEIDING	133
4.2	IMPLEMENTERING	133
4.2.1	Argitektuur van lemma-identifiseerder	133
4.2.2	Die naamwoordmodule	137
4.2.3	Die egtewerkwoordmodule	142

	INHOUDSOPGAWE (vervolg)	
4.2.4	Samevatting	152
4.3	EVALUERING	153
4.3.1	Samestelling van toetsdata	153
4.3.2	Meetinstrumente	154
4.3.3	Resultate	157
4.4	BESPREKING VAN RESULTATE	159
4.4.1	Naamwoorde	159
4.4.2	Werkwoorde	168
4.4.3	Verskillende woordsoorte	181
4.4	GEVOLGTREKKING	185
 HOOFSTUK 5 – SLOT		
5.1	INLEIDING	187
5.2	SAMEVATTING	187
5.3	GEVOLGTREKKING	189
5.4	VOORUITSKOUING	192
 BRONNELYS		 194

CD-ROM

BYLAE

Bylaag 1	1
Bylaag 2	8
Bylaag 3	46
Bylaag 4	47

LEMMALYSTE

LISENSIES

PROGRAMMATUUR

LYS VAN TABELLE EN FIGURE

Tabel 2.1 Mosadi, selepe, araba en bula in verskillende woordeboeke	22
Tabel 2.2 Van Wyk (1961) se beskouing van die werkwoordstruktuur	26
Figuur 2.1 'n Hiërargie in die woordmorfologie sover	27
Figuur 2.2 'n Hiërargiese ontleding van ba a re bona .	28
Figuur 2.3 'n Taksonomie vir die Setswanamorfologie.	29
Figuur 2.4 Taksonomie van die Setswanawoord	36
Figuur 2.5 Uitbreiding van die taksonomie van die Setswanawoord	38
Figuur 2.6 Die naamwoord se subkategorieë	48
Tabel 2.3 Die naamwoordklasse in Setswana	49
Tabel 2.4 Die morfologie van die naamwoord	50
Figuur 2.7 Die werkwoord se subkategorieë	54
Tabel 2.5 Die morfologie van die egte werkwoord (gebaseer op Krüger, 2006a:257)	55
Figuur 3.1 'n Visuele voorstelling van die outomaton	75
Figuur 3.2 Tweevlakmorfologie	76
Figuur 3.3 'n FST	76
Tabel 3.1 Basiese operators in <i>FSA 6</i>	80
Tabel 3.2 Die hiërargie vir verwydering van die suffikse van die naamwoord	83
Tabel 3.3 Die hiërargie in die lemma-identifisering van die egte werkwoord (gebaseer op Krüger, 2006a:257)	101
Figuur 4.1 'n Diagrammatiese voorstel van die sisteem	136
Figuur 4.2 'n Diagrammatiese voorstelling van die naamwoordmodule	140
Figuur 4.3 'n Diagrammatiese voorstel van die ontleding van ditaugatsaneng deur die naamwoordmodule	142
Figuur 4.4 Diagrammatiese voorstelling van die egtewerkwoordmodule	148
Figuur 4.5 'n Diagrammatiese voorstel van die ontleding van gatakisetsanweng deur die egtewerkwoordmodule	152
Tabel 4.1 Resultate van die evaluering	157
Tabel 4.2 Omvang van foute by die naamwoordmodule	158
Tabel 4.3 Omvang van foute by die werkwoordmodule	159

Tabel 4.4 Ontledings van die naamwoordmodule se resultate	163
Tabel 4.5 Ontledings van die egtewerkwoordmodule se resultate	175
Tabel 4.6 Ontledings van die volledige lemma-identifiseerder se resultate	183

AFKORTINGS EN KONVENSIES

Die morfologiese ontledings in die studie is in die internasionale styl in navolging van Croft (2003) en Haspelmath (2002) gedoen.

AGR	kongruensie
APPL	applikatief
AUG	augmentatief
CAUS	kousatief
CL	naamwoordklas
DEV	deverbatief
DIM	diminutief
FEM	feminitief
FUT	toekoms (temporale morfeem)
INF	infinitief
ITER	iteratief
LOC	lokatief
N.PASS	neutro-passief
NEG	ontkenning
OBJ	voorwerp
PASS	passief
PERF	perfektum
PL	meervoud
RECP	resiprokalis
REFL	refleksief
REL	relatief
SG	enkelvoud
SUBJ	onderwerp
TERM	uitgang

Die volgende konvensies word in hierdie verhandeling gebruik:

	Ortografie	Voorbeelde
Setswanawoorde	vetgedruk ("bold")	dintswa
Morfeme	<i>kursief ("italics")</i>	<i>-il-</i>
Non-morfeme/letters	<i>vetgedruk en kursief</i>	<i>-ile</i>
Lêername/reëlmatige uitdrukkings	Courier New	[?*, [n, g]: []]
Lemma-identifisering	kompleks > simpleks	dintswa > ntswa
Woordvorming	simpleks → kompleks	ntswa → dintswa
Engelse aanhalings	tussen dubbelaanhalingstekens ("")	"translation"
Afrikaanse vertalings	tussen enkelaanhalingstekens (')	'vertaling'

Die afkortings en ortografiese styl word só toegepas:

Setswana: **Dintswa di bogotse.**

Ontleding: *Di-ntswa di-bogol-il-e.*

Afkortings: CL5.PL-hond AGR.SUBJ-blaf-PERF-TERM.

Vertaling: 'Die honde het geblaf.'

Hoofstuk 1

Inleiding

1.1 Inleiding

Setswana, 'n relatief jong skryftaal, is in die nuwe millennium besig om 'n nuwe fase te betree. Dit is hoofsaaklik te danke aan die grondwetlike beskerming wat Setswana geniet, maar ook omdat regeringsbeleid ontwikkeling in taaltegnologie steun (Nasionale Departement van Kuns en Kultuur, 2000). In die snelveranderende tegnologiese omgewing waarin Setswanasprekers hulle bevind, is daar al hoe meer 'n behoefte aan betroubare elektroniese taalhulpmiddels soos speltoetsers, woordafbrekers, inligtingsonttrekkingsisteme, vraagbeantwoordingsisteme, outomatiese vertaalsisteme, elektroniese woordeboeke en rekenaargesteunde taalonderrigprogrammatuur. Met uitsondering van 'n speltoetser, 'n woordafbreker en rekenaargesteunde taalonderrigprogrammatuur is daar op hierdie stadium weinig sodanige hulpmiddels vir Setswana beskikbaar. Die ontwikkeling van taaltegnologiese hulpmiddels is egter nie iets wat oornag gebeur nie: basiese hulpbronne (soos korpora en kerntegnologiemodules) moet eers ontwikkel word voordat eindgebruikertoepassings in finale vorm beskikbaar is en gebruik kan word. Een sodanige kerntegnologiemodule wat dikwels deel uitmaak van verskeie toepassings is 'n lemma-identifiseerder ("lemmatiser"). Tans bestaan daar nog geen lemma-identifiseerder vir Setswana nie.

1.2 Konteks: Wat is lemma-identifisering?

Lemma-identifisering word, volgens Plisson *et al.* (2004), in natuurliketaalprosessering ("natural language processing") gebruik, en natuurliketaalprosessering word deur Crystal (2003:309) beskryf as die rekenaarmatige verwerking van teksmateriaal in natuurlike tale. 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 (2004:17) 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". In aansluiting hierby definieer Hausser (1999:125) 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. Dit beteken dat die woord na die eenvoudigste vorm soos dit in 'n woordeboek sou voorkom (die lemma), reduceer word.

Die redusering na die basisvorm behels in die meeste gevalle hoofsaaklik die verwydering van fleksie-affikse (Bussman, 1996:272). Siemens (1996) sluit hierby aan deur te sê dat lemma-identifisering die proses is waarin die fleksie- en variante vorme van 'n woord gereduseer word na hulle ooreenstemmende lemma, te wete die basisvorm of trefwoord. 'n Lemma-identifiseerder kan daarom as 'n eenvoudige morfologiese analiseerder beskou word (Daelemans & Strik, 2002:37).

"Basisvorm" is 'n term wat veral in die morfologie gebruik word (Crystal, 2003:48; Hartmann & James, 1998:12). Dié term verwys gewoonlik na enige deel van 'n woord wat as 'n eenheid beskou word en waarop 'n morfologiese bewerking toegepas kan word – soos byvoorbeeld om 'n affiks by die wortel of stam te voeg (Crystal, 2003:48). So byvoorbeeld is die enkelvoud van die selfstandige naamwoord in Engels tradisioneel die basis vir die vorming van die meervoud. In die geval van Engels sou die lemma "book" dan as basis dien om die meervoud "books" mee te vorm, en die werkwoordlemma "work" sou die basis wees vir die ander vorme "worked", "working" en "works". Die term "basisvorm" word soms as ekwivalent van die terme "woord" of "stam" gebruik (Crystal, 2003:48; Matthews, 1997:35).

'n Mens kan daarom nie van lemma-identifisering praat sonder om stamidentifisering ("stemming") in dieselfde asem te noem nie. Stamidentifisering is 'n nog eenvoudiger vorm van morfologiese analise, maar is nogtans aan die lemma-identifiseringsproses verwant (Daelemans & Strik, 2002:37). In plaas daarvan dat die toevoerwoord na 'n ander geldige woord verander word, word die affikse bloot verwyder om die stam van die toevoerwoord te bepaal (Gearailt, 2005:23; Manning & Schütze, 1999:132). Dit beteken dat "gone", "going" and "goes" se stam GO sal wees, as die *-ne*, *-ing*, en *-es* onderskeidelik verwyder word, terwyl "went" nie geanaliseer sal word nie. So ook sal "electricity", "electrify" en "electrician" na **electr-** gereduseer kon word, wat die stam is, maar nie noodwendig 'n geldige woord nie.

Stamidentifisering behels dus ongeveer dieselfde as lemma-identifisering, behalwe dat die toevoerwoord nie noodwendig tot 'n basisvorm (of geldige woord) gereduseer word nie. Daarom word dit in inligtingonttrekking gebruik, waar dit dikwels nie so belangrik is dat die stamme regte woorde is nie; dit is net belangrik dat die seekstring wel die gewenste soekresultate oplewer (Lancaster University Computing Department, 2004). Voorbeelde van die gebruik van stamidentifisering in inligtingonttrekking is byvoorbeeld in soekenjins soos *Lycos* en *Google* (Lancaster University Computing Department, 2004). Volgens Khoja (2002) kan stamidentifisering ook gebruik word vir kompressie, tekssoektogte, teksanalises en speltoetsers. In plaas daarvan om 'n volledige woord in 'n leksikon op te soek, word daar dus net 'n stam gesoek, wat tot gevolg het dat die leksikon kleiner is (Khoja, 2002).

Hoewel lemma-identifisering en stamidentifisering verwant is en die terme soms afwisselend gebruik word (Goldsmith, Higgins & Soglasnova, 2000:274), kan lemma-identifisering as 'n soort samesmeltingsprosedure ("conflation procedure") gebruik word, waar variante gereduseer word na 'n enkele kanonieke vorm (Lennon, Pierce & Willett, 1981). Hierdie genormaliseerde woordvorm is meestal die kanonieke vorm – 'n term wat in die leksikografie gebruik word en deur

Hartmann en James (1998:12,18) gedefinieer word as die trefwoord ("headword") waaronder verskeie variante aangehaal word, wat dan ook dieselfde is as 'n basisvorm met verwysing na lemma-identifisering. Met "samesmelting" word bedoel dat daar aan al die variante – al het hulle nie dieselfde stam nie – 'n gemeenskaplike lemma toegeken word. Volgens Plisson *et al.* (2004) is lemma-en stamidentifisering eenders, maar lemma-identifisering hoef nie die stam van die woord te produseer nie; dit vervang wel die suffikse om die genormaliseerde woordvorm as resultaat te lewer. In hierdie opsig kan die lemma-identifiseerder as die leksikale variant van 'n stamidentifiseerder beskou word, wat woorde, en nie net morfeme nie, as resultaat lewer.

Lemma-identifisering is 'n belangrike prosedure in korpusgebaseerde navorsing (McEnery & Wilson, 2001:53). Volgens Gouws en Prinsloo (2005:37) speel lemma-identifisering '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 (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. Bussman (1996) noem ook dat lemma-identifisering gebruik word om indekse, konkordansies en lyste uit tekskorpora op te stel.

Ter samevatting kan die volgende werksdefinisie vir lemma-identifisering aangebied word: Lemma-identifisering is 'n natuurliketaalprosesseringsprosedure wat die korrekte basisvorm/lemma van 'n toevoerwoord bepaal deur die verwydering van fleksie-affikse. As sodanig word dit as 'n eenvoudige proses van morfologiese analise beskou wat vele toepassingsmoontlikhede in korpusgebaseerde navorsings- en ontwikkelingswerk het.

1.3 Probleemstelling

Soos wat blyk uit bostaande werksdefinisie, hou lemma-identifisering verband

met die konsep "lemma" en wat as lemma beskou word, moet dus ook duidelik gedefinieer word. Volgens Gouws (1989:35) en Bussman (1996:272) word die term "lemma" gebruik om na daardie leksikale elemente te verwys wat in 'n woordeboek as trefwoorde opgeneem word. Crystal (2003:263) brei daarop uit deur te sê dat dit 'n abstrakte verteenwoordiging van leksikale variasie is: die versamelwoord "GO" verteenwoordig die vorme "went", "gone", "going" en "goes". 'n Lemma word ook algemeen gedefinieer as die betekenisvolle basisvorm, gestroop van fleksiemorfeme wat variante vorme verteenwoordig (Choueka *et al.*, 2000:74; Mitkov, 2003:728; Trost, 2003:38).

Tog, soos wat tale verskil, kan die definisies van wat as lemma in verskillende tale beskou word, ook verskil. So byvoorbeeld, in 'n vertalingtoepassing om Engelse en Hebreeuse vertalings te belyn, moes albei tale se lemmas geïdentifiseer word (Choueka *et al.*, 2000:74). Vir dié projek moes presies beskryf word wat as lemma vir naamwoorde en werkwoorde beskou word. 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 kan word, moet deeglik besin word oor wat 'n lemma in Setswana is.

Die begrip "lemma" is reeds in ander tale wat 'n ryk geskiedenis van woordeboekopstelling het, omvattend gedefinieer. Daar is egter nie enigheid oor wat die lemma in die Afrikatale moet wees nie. Volgens De Schryver en Prinsloo (2001) debatteer leksikograwe 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). Die terme "wortel" en "stam", moet derhalwe onder die loep geneem word om te bepaal wat die lemma in Setswana kan wees.

Volgens Louwrens (1994:94), met spesifieke verwysing na Sesotho sa Leboa, verwys die term “leksikale item” of “lekseem” in grammatiese beskrywings van Afrikatale na die kleinste linguistiese eenheid wat leksikale betekenis het. Dusdanig word die wortel van ’n woord dikwels as ’n leksikale item beskou (Louwrens, 1994:94). Wortels is egter nie onafhanklike eenhede nie, en daarom ook nie woorde nie. As ’n lemma-identifiseerder dan ‘regte’ woorde (of lemmas) as resultaat moet lewer, dan kan wortels nie as lemmas beskou word nie.

’n Ander moontlikheid is om stamme as lemmas te oorweeg. ’n Stam is volgens Laas (1974:10) die morfologiese komponent wat ’n leksikale woordkorrelaat in die leksikon het. Die stamme in Setswana het egter meestal fleksiemorfeme, en voldoen daarom nie aan die definisie dat ’n lemma ’n betekenisvolle basisvorm, gestroop van fleksiemorfeme, is nie.

Die volgende voorbeelde illustreer die keuse van ’n lemma tussen die wortel en die eenvoudigste stam. Wat hier met die eenvoudigste stam bedoel word, is die stam met die minste moontlike fleksiemorfeme, in die analisestap voordat dit as ’n wortel beskou kan word. In voorbeeld (1) is die keuse van wat die lemma moet wees tussen *rek-*, die wortel, en *reka*, die eenvoudigste stam van *rekile*, *rekela*, en *rekisa*.

- (1) **reka, rekile, rekela, rekisa**
 rek-a, rek-il-e, rek-el-a, rek-is-a
 koop-TERM, koop-PERF-TERM, koop-APPL-TERM, koop-CAUS-
 TERM
 ‘koop, gekoop, koop vir, laat koop/verkoop’

(2) **letsogo, matsogo, letsogwana, letsogong, matsogwana, matsogong**

le-tsogo, ma-tsogo, le-tsogo-ana, le-tsogo-(i)ng, ma-tsog-ana, ma-tsogo-(i)ng

CL3.SG-arm, CL3.PL-arm, CL3.SG-arm-DIM, CL3.SG-arm-LOC, CL3.PL-arm-DIM, CL3.PL-arm-LOC

'arm, arms, armpie, lokaliteit van die arm, armpies, lokaliteit van die arms'

Wat die naamwoord in voorbeeld (2) betref, lê die keuse tussen die wortel *-tsogo* en **letsogo** of **matsogo**, die eenvoudigste stamme van **letsogwana** en **letsogong**. Vir naamwoorde sal die keuse dus lê tussen die eenvoudigste stam in die enkelvoud en meervoud.

Uit die voorafgaande voorbeelde blyk dit dat wat as lemma in Setswana beskou word, duidelik gedefinieer moet word. Die eerste probleemstelling volg dan hieruit dat daar nie duidelike riglyne vir die definiëring van die lemma in Setswana is nie.

Die tweede probleemstelling hou verband met die feit dat dit in die ontwikkeling van 'n lemma-identifiseerder nie net gaan oor die bepaling van wat 'n lemma is nie, maar ook oor hoe die lemma geïdentifiseer moet word. Dit beteken dat die juiste benadering en tegniek gekies moet word om sodanige prosesse outomaties uit te voer. Om dit te doen, kan daar gekies word uit hoofsaaklik twee soorte benaderings wat in natuurliketaalprosessering gevolg word (Voutilainen, 1999:9). Daar was die aanvanklike linguistiese benadering in die vyftiger- en sestigjare, waar kennisgebaseerde, linguistiese reëls gebruik is in die ontwikkeling van toepassings soos woordsoortetiketteerders. Met betrekking tot tegnieke, is eindigestaatoutomate ("finite state automata") en eindigestaatoorvormers ("finite state transducers") sinoniem met so 'n linguistiese benadering. Die outomaat se ontstaan is aan Turing (1936) se model van algoritmiese berekeninge te danke,

wat gelei het tot Kleene (1951 en 1956) se werk met betrekking tot eindigestaatoutomate en reëlmatige uitdrukings (Jurafsky & Martin, 2000:10-11).

In die sewentigerjare het die datagedrewe benadering, waarin taalpatrone outomaties afgelei is uit reeds geannoteerde teks as toevoer, veld gewen. In hierdie benadering word gewoonlik van masjienleertegnieke en statistiese metodes gebruik gemaak (Voutilainen, 1999:9-11).

Die eindigestaatmodelle keer egter in die tagtigerjare terug met Kaplan en Kay (1981) se werk met betrekking tot eindigestaatmorfologie en -fonologie (Jurafsky & Martin, 2000:14) en word vandag weer vir sekere prosedures in natuurliketaalprosessering gebruik. So byvoorbeeld word die linguistiese benadering reeds gebruik in die ontwikkeling van morfologiese analiseerders vir ander Afrikatale; vergelyk byvoorbeeld die eindigestaatprogrammatuur van Xerox (d.i. *Xerox Finite State Tools*) wat gebruik word in 'n NRF-befondsde projek by UNISA (vergelyk onder andere Pretorius & Bosch, 2003).

Volgens Jurafsky en Martin (2000:65) is 'n leksikon, morfologiese reëls en morfofonologiese reëls nodig om 'n morfologiese analiseerder te bou. Alhoewel die lemma-identifiseerder nie 'n omvattende morfologiese analiseerder is soos wat in die NRF-befondsde morfologiese-analiseerderprojek by UNISA ontwikkel word nie (Pretorius & Bosch, 2003), maak dit tog ook van 'n leksikon, morfologiese reëls en morfofonologiese reëls gebruik. 'n Leksikon is die lys van stamme en affikse, saam met die basiese inligting daaroor (byvoorbeeld of dit 'n naamwoord- of 'n werkwoordstam is). Die morfologiese reëls orden die morfeme; watter morfeme voor die stam mag kom en in watter volgorde dit na die stam volg. Ortografiese/morfofonologiese reëls word gebruik om die veranderinge wat plaasvind as twee morfeme kombineer, te beskryf (Jurafsky & Martin, 2000:65; vergelyk Pretorius & Bosch, 2002:13). Die uitgangspunt en vooronderstelling van hierdie studie is dat die linguistiese benadering 'n gepaste benadering tot

lemma-identifisering vir Setswana kan wees.

Hierdie vooronderstelling berus op die feit dat die morfologiese patrone van die naamwoorde en werkwoorde in Setswana redelik reëlmatig is, terwyl die ander woordsoorte morfologies onproduktief is. So byvoorbeeld is vyf van Setswana se sewe woordsoortklasse (die voornaamwoord, bywoord, betrekkingwoord, interjeksie en ideofoon) se lemmas maklik bepaalbaar, aangesien dié woordsoortklasse in 'n hoë mate geslote is. Daar is dus net twee oop woordsoortklasse, die naamwoord en die werkwoord, waarvoor reëls gemaak moet word. Dit is dan juis ook veral die naamwoord en die werkwoord wat leksikograwe hoofbrekens besorg in die lemma-identifiseringsproses (Gouws & Prinsloo, 2005:67).

Ter verdere ondersteuning van dié vooronderstelling bestaan daar boonop 'n goeie kennisbasis vir die linguistiese benadering omdat die morfologiese struktuur van die Sothotale, en spesifiek Setswana, omvattend in handboeke en artikels beskryf is (vergelyk Cole, 1955; Krüger, 1973, 1994 & 2006a (Setswana); Lombard *et al.*, 1985; Louwrens, 1991 (Sesotho sa Leboa)). Die ontwikkeling van 'n reëlgebaseerde lemma-identifiseerder bied dan ook 'n geleentheid om te bepaal hoe goed die bestaande grammatikas sal vaar indien dit geformaliseer word (d.i. aangepas word om rekenaarmatig aangewend te word). Die beskrywings van Setswana (en ook Sesotho sa Leboa) word dus as basis vir die reëls gebruik. Op dié manier kan die kwaliteit van die grammatikas beskryf word, aangesien die werkverrigting van lemma-identifiseerder wat slegs op hierdie reëls gebaseer is eerstens 'n aanduiding gee van die reikwydte van die grammatikas en tweedens van die akkuraatheid van die grammatikas. Hierdie grammatikas is nog nie voorheen op dié manier getoets nie. Die tweede probleemstelling is dan die effektiwiteit waarmee die bestaande grammatikas geformaliseer kan word om die lemmas in Setswana te identifiseer.

Samevattend kan gestel word dat die lemma in Setswana gedefinieer moet word

soos wat dit in die konteks van outomatiese lemma-identifisering verstaan word. Ook moet 'n lemma-identifiseerder vir Setswana ontwikkel word ten einde te bepaal wat die reikwydte en korrektheid/akkuraatheid van bestaande Setswanagrammatikas (van spesifiek Cole (1955) en Krüger (2006a), maar ook met verwysing na Lombard *et al.* (1985) en Louwrens (1991) vir Sesotho sa Leboa) is.

1.4 Navorsingsvrae

Na aanleiding van bogenoemde agtergrond en probleemstelling ontstaan die volgende navorsingsvrae:

- (i) Hoe moet die lemma in Setswana gedefinieer en geïdentifiseer word met die oog op outomatiese lemma-identifisering?
- (ii) Kan bestaande Setswanagrammatikas effektief geformaliseer word om die lemmas outomaties te kan identifiseer?

1.5 Doelstellings

Na aanleiding van bogenoemde navorsingsvrae kan die volgende doelstellings geformuleer word:

- (i) Om die lemma in Setswana te definieer en te identifiseer met die oog op outomatiese lemma-identifisering; en
- (ii) Om te bepaal of bestaande Setswanagrammatikas effektief geformaliseer kan word om die lemmas outomaties te identifiseer.

1.6 Sentrale teoretiese stellings

Na aanleiding van bostaande agtergrond, probleemstelling, navorsingsvrae en doelstellings kan die onderstaande twee sentrale teoretiese stellings geformuleer word. Hierdie teoretiese stelling het eensyds betrekking op die definisie van die lemma in Setswana en andersyds op die formalisering van die Setswanagrammatika.

(i) Hoewel die wortel van 'n woord in die Afrikatale dikwels as 'n leksikale item beskou word, kan die wortel nie selfstandig gebruik word nie. Die stam, daarenteen, is 'n morfologiese komponent wat 'n leksikale woordkorrelaat in die leksikon het. Hoewel die stam en wortel dus albei as leksikale morfeme beskou kan word, is dit net die stam wat selfstandig gebruik kan word. As sodanig word 'n lemma in Setswana gesien as die eenvoudigste stam wat as 'n kanonieke vorm optree. Vir die naamwoord is dit dus die eenvoudigste stam in die enkelvoud en vir die werkwoord die eenvoudigste stam in die infinitief, sonder enige prefikse.

(ii) Die bestaande Setswanagrammatikas kan effektief geformaliseer word om 'n outomatiese lemma-identifiseerder vir Setswana te ontwikkel. Die morfologiese struktuur van die Sothotale, spesifiek Setswana, is goed gedokumenteer in handboeke en artikels van verskeie navorsers, en kan dus dien as basis vir kennisgebaseerde reëls in die volg van 'n linguistiese benadering tot die ontwikkeling van 'n lemma-identifiseerder vir Setswana. Die stelling is dus dat daar genoegsame bestaande kennis is om die linguistiese benadering suksesvol te implementeer.

1.7 Metode van ondersoek

1.7.1 Definiëring en identifisering van die lemma in Setswana

Aan die hand van 'n omvattende literatuurondersoek sal die konsep "lemma" in Setswana gedefinieer word, en kriteria waaraan 'n string karakters moet voldoen om as lemma geïdentifiseer te word, sal bepaal word. 'n Deeglike studie van die Setswana woordmorfologie, veral met betrekking tot die naamwoord en werkwoord, sal in dié fase gedoen word. Die grammatikas van Cole (1955) en Krüger (2006a) sal as basis van die bespreking van die Setswana woordmorfologie dien, aangesien dit die enigste twee gesaghebbende bronne van omvattende grammatikas vir Setswana is. Ter aanvulling sal daar ook verwys word na Lombard *et al.* (1985) en Louwrens (1991), as voorbeelde van omvattende, gesaghebbende grammatikas vir Sesotho sa Leboa.

1.7.2 Ontwikkeling van 'n lemma-identifiseerder

1.7.2.1 Formalisering van bestaande grammatika

Gegewe die definisie van die lemma in Setswana sal die bestaande linguistiese beskrywings van Cole (1955) en Krüger (2006a) in hierdie fase in reëlmatige uitdrukkings omskryf word. Die programmatuur waarmee die bestaande grammatika geformaliseer word, is *FSA 6*. Dit is deur Van Noord (2002) ontwikkel en word allerweë as een van die bekendste pakkette beskou (Piskorski & Neuman, 1999) wat beskikbaar is om eindigestaatoutomate en -oorvormers mee te ontwikkel.

1.7.2.2 Die ontwikkeling van datastelle

Met die oog op die evaluering van die reëls wat in die vorige stap geformuleer is, moet datastelle ontwikkel word. Deur middel van 'n lukrake, geoutomatiseerde steekproef word die datastelle uit die woordelys van 'n Setswanaspeltoets (NWU, 2005) met 'n grootte van 73,119 woorde onttrek. Indien daar met 'n steekproefgrootte van 'n 1,000 woorde gewerk word, is die sekerheid ("confidence") 97%; 'n sekerheid van 95% word allerweë as aanvaarbaar beskou in die bepaling van 'n steekproefgrootte (Lohr, 1999:40). Twee datastelle van onderskeidelik 1,000 woorde (500 werkwoorde en 500 naamwoorde) en 2,000 woorde (1,000 werkwoorde en 1,000 naamwoorde) word in dié fase saamgestel. Die datastel van 1,000 woorde word gebruik vir toetsing tydens die ontwikkeling, terwyl die ander 2,000 woorde gebruik word vir die evaluering van die lemma-identifiseerder. 'n Gemengde stel van verskeie woordsoorte van 1,000 woorde word ook saamgestel om die uiteindelijke lemma-identifiseerder mee te evalueer. Die lemmas van die datastelle word per hand geïdentifiseer en word dan in die evaluering gebruik om die resultate van die lemma-identifiseerder aan te meet. In hierdie fase word daar ook met behulp van die *South African multi-language dictionary and phrase book* (Reynierse, 1991) en 'n frekwensielys (Prinsloo, datum onbekend) 'n lemmalys, wat in die lemma-identifiseerder gebruik word, saamgestel.

1.7.2.3 Evaluering van die lemma-identifiseerder

In die laaste fase word die 2,000 woorde (1,000 naamwoorde en 1,000 werkwoorde) wat in 1.7.2.2 saamgestel is, asook die gemengde teks van 1,000 woorde, in die evaluering gebruik. Daarmee word bepaal hoe akkuraat die reëls vir naamwoorde en werkwoorde onderskeidelik is en hoe akkuraat die uiteindelijke lemma-identifiseerder op alle woordsoorte is. Die kriterium wat in die evaluering geld, is akkuraatheid ("precision"). Die datastelle se lemmas word in hierdie fase deur die lemma-identifiseerder geïdentifiseer. Die resultate word dan met die lemmas wat in 1.7.2.2 per hand geïdentifiseer is, vergelyk. Eerstens word die akkuraatheid van die lemma-identifiseerder op komplekse woorde (dit is die woorde wat nog omvorm moet word) bepaal. Dit word met dié formule gedoen:

$$\text{KompleksAkkuraatheid} = \frac{\text{KomplekseKorrekHanteer}}{\text{KomplekseHanteer}}$$

Tweedens word die akkuraatheid van die lemma-identifiseerder op alle woorde, komplekse en simplekse (dit is die woorde wat reeds lemmas is) bepaal. Dit word met dié formule gedoen:

$$\text{VoorspeldeAkkuraatheid} = \frac{\text{KomplekseKorrekHanteer} + \text{SimplekseKorrekHanteer}}{\text{KomplekseHanteer} + \text{SimplekseHanteer}}$$

Hoewel die resultate kwantitatief (in persentasie akkuraatheid) uitgedruk word, sal die bespreking daarvan kwalitatief van aard wees (d.i. bespreking van redes vir die resultate).

1.8 Ontplooiing

In hierdie hoofstuk is bespreek wat lemma-identifisering behels en daarmee is aangedui dat die lemma in Setswana nog nie voldoende gedefinieer is nie. In natuurliketaalprosessering is daar twee benaderings wat gevolg kan word in die ontwikkeling van 'n lemma-identifiseerder, naamlik die linguistiese en die datagedrewe benadering. Die linguistiese benadering word gevolg in die

ontwikkeling van 'n reëlgebaseerde lemma-identifiseerder en daarmee kan bepaal word hoe effektief die bestaande Setswanagrammatika geformaliseer kan word.

Na aanleiding van die navorsingsvraag in Hoofstuk 1 oor hoe die lemma in Setswana gedefinieer en geïdentifiseer moet word, word die lemma in Setswana in Hoofstuk 2 onder die loep geneem. Die definisies van wat as die lemma in Setswana beskou kan word, word aan die hand van die bespreking van die Setswanawoordmorfologie gedoen. Die keuse van wat 'n lemma is, lê in dié hoofstuk tussen die wortel en die stam. Wat as lemma beskou kan word, word in hierdie hoofstuk vir elke woordsoort bespreek.

Die navorsingsvraag of bestaande Setswanagrammatikas effektief geformaliseer kan word om die lemmas te identifiseer, word in Hoofstuk 3 en 4 beantwoord. Die formalisering van bestaande reëls kom in Hoofstuk 3 aan die beurt. Begrippe soos reëlmatige uitdrukking, eindigestaatoutomaat en eindigestaatoorvormer en die verband wat dit met die formalisering van die Setswanagrammatika hou, word daar bespreek. Daarna word die grammatikareëls, soos wat dit vir lemma-identifisering aangepas is, in FSA 6 se standaardnotasie herskryf.

Die reëls wat in Hoofstuk 3 geskryf is, word in Hoofstuk 4 in 'n sisteem geïmplementeer. Die egtewerkwoordmodule en naamwoordmodule van die lemma-identifiseerder en die uiteindelijke lemma-identifiseerder word vervolgens aan evaluering onderwerp. Laastens word die resultate en oorsake van die resultate krities bespreek.

In die laaste hoofstuk word die vorige hoofstukke saamgevat en gevolgtrekkings gemaak in antwoord op die navorsingsvrae wat in hierdie hoofstuk gestel is. Hoofstuk 5 sluit af met 'n vooruitskouing en aanbevelings vir verdere navorsing.

Hoofstuk 2

Die lemma in Setswana

2.1 Inleiding

Soos in Hoofstuk 1 bespreek, hou die begrip “lemma” ook met lemma-identifisering verband, aangesien lemma-identifisering die redusering van woorde in ’n korpus tot hulle ooreenstemmende lemmas behels. Lemma-identifisering is ook anders gestel, die bepaling van die korrekte basisvorm van ’n woord, oftewel die verskaffing van toegang tot die ooreenstemmende lemma van ’n woord in die leksikon (Hausser, 1999:125). Ten einde ’n betroubare lemma-identifiseerder te ontwikkel, moet die konsep “lemma” gedefinieer word soos wat dit in die konteks van outomatiese lemma-identifisering in Setswana verstaan word. Die definisie van ’n lemma is egter taalspesifiek, omdat dit wat as lemma beskou kan word van taal tot taal verskil (Choueika *et al.*, 2000:74; Knowles & Don, 2004:69-81).

In Hoofstuk 1 is aangetoon dat daar nie eenstemmigheid is oor wat as die lemma in Setswana beskou kan word nie; in hierdie hoofstuk word die konsep “lemma” soos wat dit in Setswana beskou kan word dus gedefinieer. Ten einde so ’n definisie te formuleer, word die konsep “lemma” soos wat dit algemeen verstaan word eerstens bespreek. Omdat daar nie met behulp van die bestaande definisies bepaal kan word wat die lemma in Setswana is nie, word aan die hand van die Setswanawoordmorfologie bepaal hoe die lemma in Setswana gedefinieer kan word. Volgende uit die definisie word die lemma vir elke woordsoort afsonderlik bespreek.

2.2 Die konsep “lemma”: Algemene perspektief

In die algenetaalwetenskapliteratuur word die term “lemma”, tesame met die verbandhoudende terme “lekseem”, “kanonieke vorm”, “basisvorm” en “basis”, dikwels verwarrend gebruik. Die term “lemma” word binne drie verskillende taalkundige kontekste (in die leksikografie, leksikologie en morfologie) gebruik om verskillende aspekte van dieselfde fenomeen te beskryf. In hierdie afdeling word die term “lemma”, soos wat dit in die verskillende kontekste voorkom, onder die loep geneem. Die verbande

tussen die term "lemma" en die terme "lekseem", "kanonieke vorm", "basisvorm" en "basis" word bespreek soos wat dit in leksikografiese, leksikologiese en morfologiese kontekste voorkom.

2.2.1 "Lemma" as leksikografiese term

Die kanonieke vorm is 'n term wat in die leksikografie gebruik word, en Hartmann en James (1998:18) definieer dit as trefwoord ("headword") waaronder verskeie variante, woorde of frases aangehaal word. 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 (2001:174) die lemma as die posisie in die algehele struktuur van 'n woordeboek of naslaanwerk waar 'n inskrywing gevind kan word, gewoonlik deur die trefwoord. In teenstelling met Hartmann (2001) is Gouws (1989:36) nie ten gunste daarvan dat die terme "lemma" en "trefwoord" as sinonieme gebruik word nie – dit verskraal vir hom die "omvang van dit wat as lemmas kan optree", want volgens hom kan lemmas "ook subwoordelike of multileksikale eenhede wees" (Gouws, 1989:36). Die lemma is dus volgens Gouws (1989:37) 'n woordeboekinskrywing.

'n Trefwoord is volgens Crystal (1992:225) weer die item wat aan die begin van 'n woordeboekinskrywings staan. 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 (1992:225) is trefwoorde volgens Jackson (1988:9) die basisvorms waarvan ander woordvorms afgelei word.

'n Lemma is '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. Vervolgens word die term "lemma", soos wat dit in die leksikologie gebruik word, bespreek.

2.2.2 “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 (1998:215) daarop dat sommige linguiste glad nie ’n onderskeid tussen die terme lemma en lekseem tref nie. Tog onderskei Ooi (1998:215) 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. In aansluiting by Ooi (1998:215) noem Jackson (1988:8-9) lekseme die “headwords of dictionary entries”. Volgens Ooi (1998) en Jackson (1988) word die term “lemma” (anders as hierbo (2.2.1) aangedui) dan in die leksikografie gebruik en “lekseem” in die leksikografie. Gouws (1989:128-129) 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.”

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 (2003:403), 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. Crystal (2003:265) sluit by Van Sterkenburg (2003) aan as hy die term “lekseem” definieer as ’n term wat deur sommige linguiste gebruik word om te verwys na die minimale onderskeidende eenheid in die semantiese sisteem van ’n taal: “Its original motivation was to reduce the ambiguity of the term word, which applied to orthographic/phonological and lexical levels, and to devise a more appropriate term for use in the context of discussing a language’s vocabulary.” (Crystal, 2003:265.) Volgens Crystal (2003:263) is die lemma die item wat aan die begin van ’n woordeboekinskrywing staan of die trefwoord wat ’n abstrakte verteenwoordiging van variante is (vergelyk die bespreking onder 2.2.1).

Hartmann en James (1998:83) definieer lekseme in aansluiting by Van Sterkenburg (2003) en Crystal (2003:265) as die basiese eenheid in die linguistiese studie van 'n woordeskat ("vocabulary"). Voorts stel Katamba (2005:18,296) en Van Sterkenburg (2003:403) lekseme gelyk aan woordeskatitems ("vocabulary items"), oftewel leksikale items. 'n Leksikale item is 'n woord wat in die groot woordsoorte (d.i. naamwoord, byvoeglike (of beskrywende) naamwoord, werkwoord, bywoord) voorkom, wat gelys kan word in die woordeboek met 'n identifiseerbare betekenis en wat onafhanklik kan voorkom (Katamba, 2005:296). Dit blyk uit Katamba (2005:296) se beskrywing van 'n leksikale item dat 'n Engelse woord soos "he" ('n voornaamwoord) dan nie 'n leksikale item is nie, maar tog het dit 'n identifiseerbare betekenis en kan dit onafhanklik voorkom. Voorts definieer Katamba (2005:296) 'n leksikale item dan ook as "a word in a sense or an item of vocabulary that can be listed in the dictionary" (Katamba, 2005:269). 'n Lekseem blyk daarom iets te wees waarvan die betekenis beskryf kan word, en verwys nie na die fenomeen "trefwoord" nie. Gouws (1989:84) sluit by Katamba (2005) aan as hy leksikale items as die elemente beskou "wat die leksikon van 'n gegewe taal konstitueer" en wat ook elemente wat groter en kleiner as woorde is, insluit.

Volgens Handke (1995:25) is die term "woord" in die morfologie die eintlike vergestaltung van 'n lekseem wat die fundamentele eenheid van die leksikon van 'n taal is. Hy gee die voorbeeld in Engels, waar "dies", "died" en "dying" woordvorme (word hieronder verder bespreek) van die lekseem "die" is. Matthews (1997:204) sluit by Handke (1995) aan en definieer 'n lekseem as 'n woord wat beskou kan word as 'n abstrakte leksikale eenheid wat 'n spesifieke vorm aanneem in spesifieke konstruksies, byvoorbeeld **THUSA** ('help') as abstrakte vorm van die verskillende woordvorme **thusitse** ('het gehelp'), **ithusa** ('self help') en **nthusa** ('my help'). 'n Lekseem is in aansluiting by Matthews (1997) volgens Bussman (1996:273) 'n basiese abstrakte eenheid van die leksikon in die dieptestruktuur of op die *langue*-vlak (dit is die *langue* en *parole* van De Saussure (1916)). Hy brei verder op die term "lekseem" uit en noem dat dit in die breë as sinoniem vir 'woord' gebruik kan word om 'n leksikale eenheid of element van die woordeskat aan te dui.

Die bespreking van die lekseem lei tot 'n ander begrip, naamlik woordvorms. Volgens Matthews (1997:405) is woordvorms die verskillende vorms van 'n spesifieke woord, hetsy 'n fonetiese of ortografiese woord. Hierdie woordvorms is spesifiek die verskillende vorms van 'n lekseem. Bussman (1996:522) sluit daarby aan as hy woordvorm definieer as "the concretely realized grammatical form of a word in the context of a sentence". Dit is met ander woorde die woord in die oppervlakstruktuur wat ooreenkom met die lekseem en realiseer in ooreenstemming met grammatiese kategorieë, soos byvoorbeeld tyd, getal en persoon.

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.

2.2.3 "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 (1998:12) met verwysing na lemma-identifisering, die kanonieke vorm ("canonical form") en met verwysing na die morfologie, die basis ("base")¹.

Die definisie van 'n basis, volgens Crystal (2003:48), 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". Hartmann en James (1998:12) sluit by Crystal (2003)

¹ Die terme "basis" en "basisvorm" word deur Trask (1996:48-49) as sinonieme beskou.

aan as hy 'n basis definieer as 'n betekenisvolle morfologiese element wat gebruik word om woorde mee te vorm; hy definieer 'n wortel byvoorbeeld as "the base of a word" (Hartmann & James, 1998:120). Crystal (2003:48) 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.

Uit bostaande is dit duidelik dat die term "lemma" in verskeie kontekste voorkom. 'n Lemma word soms in die leksikografie aan 'n trefwoord of woordeboekinskrywing gelykgestel, in die leksikologie aan 'n lekseem, en in die morfologie aan 'n basisvorm.

Om saam te vat, kan 'n basis, en daarom ook 'n lemma, 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 studie 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).

Noudat die begrip "lemma" bespreek is, kan daar oorgegaan word na die beskouing daarvan in die Afrikatale, aangesien dit wat as lemma beskou kan word, taalspesifiek is (Choueka *et al.*, 2000:74; Knowles & Don, 2004:69-81).

2.3 Die konsep "lemma": Taalspesifieke perspektief

In tale met 'n lang leksikografiese tradisie is die konsep "lemma" goed gevestig, maar wat die Afrikatale betref, is oor die algemeen nog geen konsensus oor wat die term presies behels nie. De Schryver en Prinsloo (2001:217) stel dit dan ook dat "[a] sound presentation of Bantu lexica has remained a bone of contention ever since." Volgens De Schryver en Prinsloo (2001:216-218) debatteer leksikograwe van die Afrikatale nog of die stam (soos wat dit in werklike taalgebruik voorkom) of die wortel (wat sommige moedertaalsprekers

nie eers sal herken nie) as lemma beskou moet word. De Schryver en Prinsloo (2001:216) haal Snoxall (1965: 27-28) aan waar hy in 'n bespreking oor 'n Luganda-Engelse woordeboek sê: "even many Baganda would have little idea under what root form they should look up many of the commonest words which they use.[...] The general principle of entering words in a dictionary under roots, though it was to an extent followed by some of the earlier compilers, could never be of great assistance [...] It would seem therefore that, although disappointing perhaps to etymologists, a decision to enter headwords in the form in which they are used in actual speech, as words possessing meaning, [...] will be welcomed by the great majority of the users of the dictionary". In die volgende afdelings word die term "lemma" soos wat dit as leksikografiese, leksikologiese en morfologiese term in die Sothotale beskou word, bespreek.

2.3.1 "Lemma" as leksikografiese term

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 2.1):

- *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)

		Setswana				Sesotho sa Leboa		Sesotho
		Brown (1988)	Snyman <i>et al.</i> (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			✓	✓			

Tabel 2.1 **Mosadi, selepe, araba en bula** in verskillende woordeboeke

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 (1988:39) **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*". 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 Setswanawoordeboek 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 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 Zivvogel 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.

De Schryver en Prinsloo (2001:376) lys soortgelyke inkonsekwenthede in woordeboeke wat nie met behulp van korpora saamgestel is nie, met verwysing na Sesotho sa Leboa-woordeboeke. 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 (d.i. die lemma of trefwoord, vergelyk die bespreking onder 2.2.1).

Volgens De Schryver en Prinsloo (2001:379) 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). Dit kan volgens hulle (De Schryver & Prinsloo, 2001) óf in 'n sogenaamde stamgebaseerde woordeboek,² 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*). Die woordeboeke bied daarom nie konsekwente riglyne is vir wat as lemma in die Sothotale beskou kan word nie.

2.3.2 “Lemma” as leksikologiese term

Hoewel die term “lemma” glad nie in leksikologiese konteks in die Afrikatale gedefinieer word nie, bied Louwrens (1994:94) 'n definisie van die term “leksikale item” of “lekseem” aan as hy sê dat dit in grammatiese beskrywings van Afrikatale (spesifiek Sesotho sa Leboa) die kleinste linguistiese eenheid is wat leksikale betekenis het. Volgens Louwrens (1994:94) word die wortel van 'n woord dikwels as 'n leksikale item beskou. Verder skryf Louwrens (1994:94) dat die term soms in die semantiek in 'n wyer sin gebruik word om na 'n volledige woord te verwys. Indien die term “lemma” as sinoniem vir die term “lekseem” gebruik sou word (vergelyk die bespreking onder 2.2.2), kan die lemma 'n wortel of 'n woord (stam) wees.

2.3.3 “Lemma” as morfologiese term

Die vraag oor wat in Setswana as die lemma beskou moet word, kon tot dusver nog nie beantwoord word nie. Daar sal dus na die woordmorfologie gewend moet word om te bepaal wat die lemma kan wees. Die

² Wat hulle hier as stam beskou, word in die woordmorfologie van Setswana as 'n wortel beskou (vergelyk die bespreking onder 2.3).

Setswanalemma sal met die algemene definisies van lemma in gedagte, en aan die hand van die bespreking van die Setswanawoordmorfologie, gedefinieer word.

2.3.3.1 Algemene Setswanawoordmorfologie

Daar is verskillende sienings oor die woordmorfologie van die Setswanawoord. Pretorius (2000:51-55) bespreek in sy artikel oor die identifisering en beskrywing van die begrippe “stam” en “wortel” in die Sothotale die verskillende sienings. Hy noem dat morfologie of die ‘woordbou’ in die onderrig van Afrikatale voorrang geniet, maar dat daar uiteenlopende vertolkings is van wat as ’n morfeem, stam en wortel beskou kan word (Pretorius, 2000:51). In een van die eerste Sothogrammatikas beskou Endemann (1876) byvoorbeeld ’n verbale element as ’n wortel of stam, wat dus ’n basiese werkwoord sonder ’n slotvokaal is. Hy tref glad nie onderskeid tussen die wortel en stam nie, en gebruik dit as sinonieme. **Dira** (‘werk’) sou byvoorbeeld ’n basiese werkwoord wees, terwyl die vorm *-dir-* die stam of wortel van dié basiese werkwoord sou wees.

’n Paar dekades later beskou Beyer (1920:5) ’n taalvorm soos **Thibela!** (‘Keer!’) as eenvoudige woordstam, en ’n struktuur soos **o thibela**, wat in die konteks **Mosimane o thibela dinku** (‘Die seun keer die skape’) optree, as ’n basiese werkwoordvorm. Die onderskeid tussen woordstam en woordvorm word dus baseer op die teenwoordigheid al dan nie van ’n onderwerpsmorfeem (soos *o* in bostaande voorbeeld). Beyer (1920) onderskei, net soos Endemann (1876), nie tussen ’n wortel en ’n stam nie.

Doke (1931), wat ’n belangrike invloed gehad het op die beskrywing van Setswanamorfologie, het ook nie tussen ’n wortel en ’n stam onderskei nie; daar was vir hom nie ’n verskil in rang wat morfologiese hiërargie betref nie. In navolging van Doke (1931), tref Cole (1955) in sy Setswanagrammatika ook geen onderskeid tussen ’n wortel en ’n stam nie.

Van Eeden (1956:212) breek in sy beskrywing van die isiZulu-grammatika weg van ander grammatici (vergelyk Pretorius, 2000:52). Hy beskryf die

wortel as 'n gereduseerde stam, gestroop van suffikse; daardie deel van die woord wat nie verander kan word nie. In die sestigjare gaan Van Wyk (1961) verder en onderskei vir die Sothotale byvoorbeeld in die werkwoordstruktuur tussen drie lae, naamlik kern, basis en stam, waar die kern en die basis somtyds ooreenkom. Die kern stem ooreen met wat vandag as wortel beskou kan word. Die basis is die kern plus een of meer uitbreidings, sonder die kategoriale uitgangsmorfeem (-a), waar die stam die kategoriale uitgangsmorfeem het (vergelyk Lombard *et al.*, 1985:19). Vergelyk die voorbeelde in Tabel 2.2.

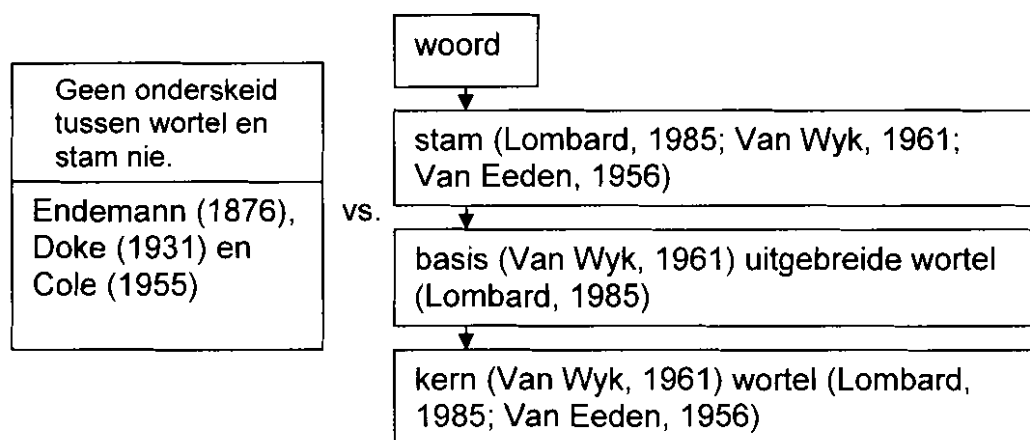
Woord	Stam	Basis	Kern
Go dira <i>Go-dir-a</i> INF-werk-TERM (‘om te werk’)	<i>-dira</i>	<i>-dir-</i>	<i>-dir-</i>
Go dirisa <i>Go-dir-is-a</i> INF-werk-CAUS-TERM (‘om te laat werk/om te gebruik’)	<i>-dirisa</i>	<i>-dir-is-</i>	<i>-dir-</i>
Go dirisana <i>Go-dir-is-an-a</i> INF-werk-CAUS-RECP-TERM (‘om vir mekaar te laat werk/om mekaar te gebruik’)	<i>-dirisana</i>	<i>-dir-is-an-</i>	<i>-dir-</i>

Tabel 2.2 Van Wyk (1961) se beskouing van die werkwoordstruktuur.

Lombard *et al.* (1985:17,18,109,135) brei in die tagtigerjare op Van Wyk (1961) se beskouings uit deur tussen 'n wortel (kern), 'n uitgebreide wortel (basis) en 'n stam te onderskei. Hulle beskryf die wortel nie in terme van die redusering van affikse nie, maar as 'n gemene woordaspek met bepaalde semantiese kenmerke (Lombard *et al.*, 1985:19). Lombard *et al.* (1985:24) onderskei tussen twee struktuurlae in polimorfematiese woorde waarvan die wortel (kern) die eerste laag is, terwyl die stam as die tweede struktuurlaag

beskou kan word. Dit word as 'n struktuurlaag onderskei omdat dit volgens Lombard *et al.* (1985:24) as 'n tonologiese eenheid funksioneer. 'n Stam bestaan dus volgens Lombard *et al.* (1985:24) uit 'n wortel plus een of meer suffigale morfeme.

Pretorius (2000:53) maak, na aanleiding van sy bespreking van die ontwikkeling in morfologie, die afleiding dat die stam 'n rangeenheid in eie reg is, wat ten opsigte van die hiërargie tussen die woord en morfeem geplaas moet word (Pretorius, 2000:53). Die hiërargie tot dusver bespreek, kan só uitgebeeld word:

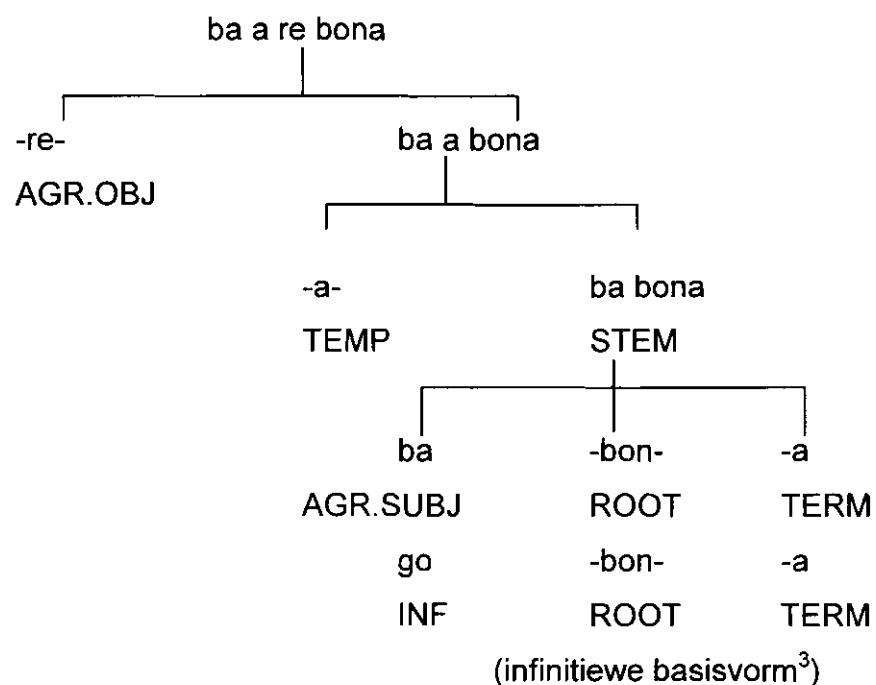


Figuur 2.1 'n Hiërargie in die woordmorfologie sover.

In die sewentigerjare het Krüger (1973) 'n nuwe denkrigting ingelei met die artikel "Woordanalise in Sotho". Tot op daardie stadium was woordanalise in die Afrikatale gekenmerk deur 'n analise van eindkomponente of 'n neutrale analise. Dit beteken dat 'n meerledige woord net sintagmaties in die kleinste eindkonstituente ontleed is (vergelyk voorbeeld (3)).

- (3) **ba a re bona**
 ba-a-re-bon-a
 AGR.SUBJ-TEMP-AGR.OBJ-sien-TERM
 'hulle sien ons'

Só 'n analise neem egter nie die feit in ag dat daar 'n onderlinge verhouding tussen die morfeme bestaan nie. Die probleem wat Krüger (1973) daarmee gehad het, is dat die eindkomponente wat tot groter eenhede binne 'n hiërgiesse struktuur gegroepeer is, nie in ag geneem word nie. Daar is volgens Krüger (1973) 'n bepaalde rangorde waarvolgens die morfeme binne die woord gerangskik is. Dit beteken dat as 'n woord ontleed word, die affikse in 'n bepaalde volgorde afgehaal word, soos dit ook in Figuur 2.2 geïllustreer word. Dié tipe ontleding word onder meer gekenmerk deur 'n proses van redusering, aangesien die affikse verwyder word (Pretorius, 2000:54-55).

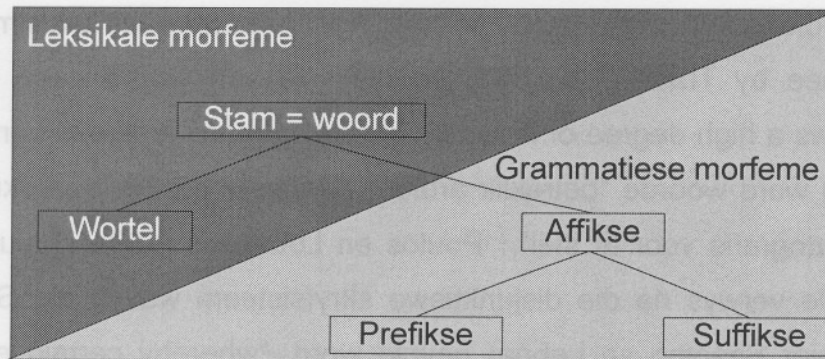


*Figuur 2.2 'n Hiërgiesse ontleding van **ba a re bona**.*

Posthumus (1994) sluit in die negentigerjare aan by Krüger (1973; 1994) en Laas (1974) se uitgangspunte wat die woordanalitiese prosesse betref. Hy verskil wel van Krüger (1973; 1994) as hy die wortel definieer as die leksikale morfeem van die woord wat morfologies onanaliseerbaar is en waaraan affikse bygevoeg kan word. Krüger (1994:18) onderskei tussen leksikale morfeme (wat wortels en stamme insluit) en grammatiese morfeme (wat

³ In die morfologie van die Afrikatale is daar verskillende benaderings ten opsigte van die basisvorm. Daar is die een kant die woordgebaseerde benadering en aan die ander kant die wortelgebaseerde benadering (Posthumus, 1994). Krüger (1994) staan die woordgebaseerde morfologie voor, waarin die stam 'n belangrike rol speel.

prefikse en suffikse insluit). Na aanleiding van Krüger (1994) se beskouing kan die taksonomie in die woordmorfologie soos in Figuur 2.3 voorgestel word (vergelyk Pretorius, 2000:58).



Figuur 2.3 '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.

2.3.3.2 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 (1996:389), 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”. Die kompleksiteit van die saak word saamgevat in die woorde: “rather different conceptions of a unit”. Crystal (2003:500-501) sluit by Trask (1996:389) aan as hy daarop wys dat daar probleme met woordgrense is. 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). Daar is nie net onduidelikheid wat woordgrense betref nie, maar ook waar die term gebruik word, want Hartmann (2001:180)

beskou 'n woord as die basiese eenheid van woordeskat, en Ooi (1998:9-10) beskou dit weer as die basiese eenheid in die linguistiek.

Wat is die woord dan volgens die linguiste in die Afrikatale? Volgens Laas (1974:4) is die woord die kleinste, selfstandig-funksionerende taalsimbool en sluit daarmee by Trask (1996:389) se definisie van woord as 'n eenheid "which shows a high degree of internal coherence" aan. Volgens Lombard *et al.* (1985:9) word woorde "betreklik arbitrêr aangedui om hulle in skrif in die praktiese ortografie voor te stel". Poulos en Louwrens (1994:7) 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". Woorde wat op taalkundige gronde bepaal word, het volgens Lombard *et al.* (1985:9) een kenmerk in gemeen, en dit is dat die dele van dié woord altyd in "onmiddellike verband met ander dele funksioneer".

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, skeikbaarheid, omstelbaarheid en vervangbaarheid (Van Wyk, 1968:51-52; Laas, 1974:4; Krüger, 2006a:13-16; Lombard *et al.*, 1985:11-15; Louwrens, 1991:6-9). Taalkundige woorde is "inherent stabiel" op grond van die woorddele se onskeikbaarheid, onomstelbaarheid, onvervangbaarheid en onweglaatbaarheid (Lombard *et al.*, 1985:11). Vervolgens word dié vier toetse, soos wat dit deur Van Wyk (1968:51-52) ontwikkel is, aan die hand van voorbeelde bespreek. In hierdie toetse word daar ook na 'n paar van Zgusta (1971:144-151) se nege toetse vir multiwoordleksikale eenhede verwys. 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 (2006a:13) definieer "isoleerbaarheid" as die eienskap van 'n woord wat in staat is om die volledige werkwoordelike inhoud van 'n sin te

konstitueer. Dit beteken dat sommige woorde isoleerbaar is, wat impliseer dat hulle hulle betekenis behou as hulle as die enigste woorde in 'n sin verskyn, soos byvoorbeeld **Mang?** ('Wie?') (Van Wyk, 1968:51-52). Krüger (2006a:13-14) toets die isoleerbaarheid deur vrae rondom die werkwoordelike inhoud van die sin te vra (voorbeeld (4)):

(4) **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.'

Di bogotse leng?

Di-bogo-il-e leng?

AGR.SUBJ-blaf-PERF-TERM wanneer?

'Wanneer het hulle geblaf?'

Bosigo

Bo-sigo

CL7.SG-nag

'In die nag'

Ke bomang ba ba bogotseng bosigo?

Ke bo-mang ba ba-bogol-il-e-ng bo-sigo?

Dit is CL1.PL-wie wat AGR.SUBJ-blaf-PERF-TERM-TERM.REL

CL7.SG-nag?

‘Wie het in die nag geblaf?’

Dintswa

Di-ntswa

CL5.PL-hond

‘Honde’

Konstituente soos **di bogotse** (‘hulle blaf’), **bosigo** (‘nag’) en **dintswa** (‘honde’) kan in terme van isoleerbaarheid as woorde gesien word. Isoleerbaarheid is volgens Louwrens (1991:7) ’n belangrike toets waarvolgens woorde van woordaspekte, of morfeme van mekaar, onderskei kan word. In die voorbeeld **Dintswa di bogotse bosigo**, kan *di-* en *-bogotse* nie net so as sinne gebruik word nie, en is daarom morfeme van ’n taalkundige woord. Hierdie toets bepaal dus hoe isoleerbaar ’n woord is, en of dit as die enigste woord in ’n sin gebruik kan word. Een van Zgusta (1971:146) se toetse bepaal dat as ’n multiwoord-leksikale eenheid (soos byvoorbeeld **di bogotse**) op sy eie beteken is, dit ’n woord is. Dus: as ’n woord in terme van beteken is isoleerbaar is, is dit ’n woord. Dié toets is egter beperk, omdat nie alle outonome woorde isoleerbaar is nie (Lombard *et al.*, 1985:11-12). Daarom is dit nodig om ook die volgende woordtoetse te doen.

Toets 2: Skeikbaarheid

Onder “skeikbaarheid” word verstaan dat woorde wat in een sin direk na mekaar volg, in ’n ander sin kan verskyn met ander woorde tussen hulle (Krüger, 2006a: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 (1971:146) 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”. In Afrikaans kan die woord *mooi* tussen *die* en *vrou* in

die sin *die vrou gaan huis toe* gevoeg word: *die (mooi) vrou gaan huis toe*. Hieruit kan dan afgelei word dat *die* en *vrou* woorde is. In Setswana kan *yo* ('hierdie') tussen **mosadi** ('vrou') en **o tlike** ('sy het gekom') geplaas word (die woorde in voorbeeld (5) is onderstreep):

(5) **mosadi o tlike 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 tlike maabane

mo-sadi yo o-tl-il-e maabane

CL1.SG-vrou hierdie AGR.SUBJ-kom-PERF-TERM gister

'hierdie vrou het gister gekom'

Die feit dat 'n woord tussenin ingevoeg kan word, bewys dus dat **mosadi** en **o tlike** twee woorde is. Hoe weet 'n mens dan dat *o-* en **tlile** nie dalk verskillende woorde is nie? As daar nie 'n woord tussen *o-* en **tlile** gevoeg kan word nie, is dit nie verskillende woorde nie. **Yo** kan byvoorbeeld nie tussen *o-* en **-tlile** gevoeg word nie en ***o yo tlike** sou dus ongrammatikaal wees. Dit kan ook met behulp van die volgende toets bepaal word.

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, 2006a:15). In voorbeeld (6) is **mosadi** en **maabane** woorde omdat hulle omstelbaar is.

(6) **Mosadi o tlike 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 tlike.

Maabane mo-sadi o-tl-il-e.

Gister CL1.SG-vrou AGR.SUBJ-kom-PERF-TERM.

'Gister het die vrou gekom.'

Met die voorafgaande vraag in gedagte (d.i. hoe 'n mens weet of o- en **-tlike** afsonderlike woorde is), beteken dit dat as 'n mens die o- en **-tlike** nie kan omruil nie, **o tlike** 'n woord is. As o- en **-tlike** omgeruil word (***tlike o**) lei dit tot 'n ongrammatikale uitdrukking. Volgens Krüger (2006a:15) kan wel bewys word dat die konstituent o- morfeemstatus het. Morfeme is nie omstelbaar nie, omdat hulle 'n vasgestelde posisie in die woordstruktuur beklee.

Toets 4: Vervangbaarheid

Wat vervangbaarheid betref, kan die woorde volgens Krüger (2006a:16) in 'n gegewe sin gewoonlik deur ander woorde of woordkombinasies vervang word (vergelyk voorbeelde (7) - (10)). Volgens Zgusta (1971:148) kan woorde ook deur sinonieme vervang word.

(7) **Mosadi o tlike 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:

(8) **Ene o tlike maabane.**

Ene o-tl-il-e maabane.

Sy AGR.SUBJ-kom-PERF-TERM gister.

'Sy het gister gekom.'

O tlike kan byvoorbeeld deur **ga a a tla** vervang word:

(9) **Mosadi ga a a tla maabane.**

Mo-sadi ga-a-a-tl-a maabane.

CL1.SG-vrou NEG-AGR.SUBJ-TEMP-kom-TERM gister.

'Die vrou het nie gister gekom nie.'

Maabane kan byvoorbeeld deur **bogologolo** vervang word:

(10) **Ene o tlele bogologolo.**

Ene o-tl-il-e bo-golo-golo.

Sy AGR.SUBJ-kom-PERF-TERM CL7.SG-groot-groot

‘Sy het lank terug al gekom.’

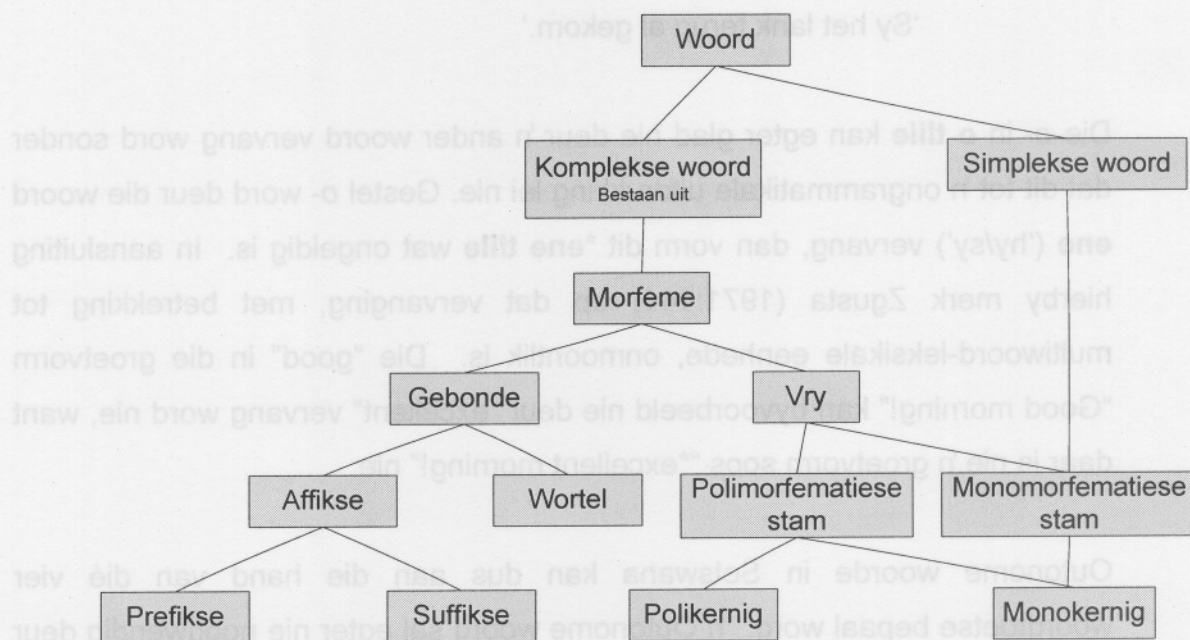
Die *o-* in **o tlele** kan egter glad nie deur ’n ander woord vervang word sonder dat dit tot ’n ongrammatikale uitdrukking lei nie. Gestel *o-* word deur die woord **ene** (‘hy/sy’) vervang, dan vorm dit ***ene tlele** wat ongeldig is. In aansluiting hierby merk Zgusta (1971:144) op dat vervanging, met betrekking tot multiwoord-leksikale eenhede, onmoontlik is. 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.

Outonome woorde in Setswana kan dus aan die hand van dié vier woordtoetse bepaal word. ’n Outonome woord sal egter nie noodwendig deur al vier toetse bewys kan word nie, omdat die toetse beperk is. Daar is wel woorde, soos **mosadi**, wat isoleerbaar, skeibaar, omstelbaar en vervangbaar is. ’n Woord soos **ke** (‘dit is’) in **ke mosadi** (‘dit is ’n vrou’) is nie isoleerbaar nie (kan dus nie alleen in ’n sin staan nie) en is nie omstelbaar nie (***mosadi ke**), maar dit is wel skeibaar (**ke (ene) mosadi** – ‘dit is sy, die vrou’) en vervangbaar (**ga se mosadi** – dit is nie ’n vrou nie’). Volgens Lombard *et al.* (1985:15) 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. Indien ’n taaleenheid soos *o-* in **o tlele** 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.

2.3.3.3 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 (2003:300) 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. Vergelyk Figuur 2.4 vir 'n visuele voorstelling van dié onderskeid.

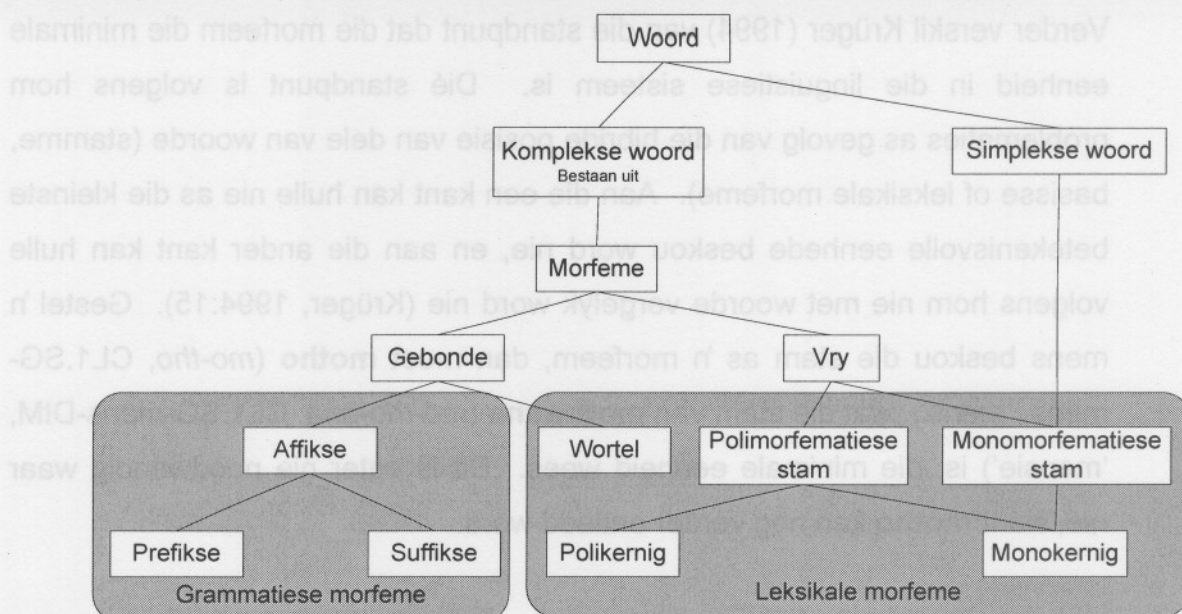


Figuur 2.4 Taksonomie van die Setswanawoord

Volgens Laas (1974:26) lei bogenoemde klassifikasie van vrye en gebonde morfeme tot verwarring. 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 (1994:15) verskil ook sterk van dié soort onderskeid tussen vrye en gebonde morfeme, "whereby the differences between morphemes and words are partly annulled". Hy onderskei eerder tussen die begrippe 'woord' en 'morfeem' in navolging van die Nederlandse en Europese linguïste, 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.

Verder verskil Krüger (1994) van die standpunt dat die morfeem die minimale eenheid in die linguistiese sisteem is. Dié standpunt is volgens hom problematies as gevolg van die hibride posisie van dele van woorde (stamme, basisse of leksikale morfeme). Aan die een kant kan hulle nie as die kleinste betekenisvolle eenhede beskou word nie, en aan die ander kant kan hulle volgens hom nie met woorde vergelyk word nie (Krüger, 1994:15). Gestel 'n mens beskou die stam as 'n morfeem, dan moet **motho** (*mo-tho*, CL1.SG-mens, 'mens'), wat die stam van **mothwana** (*mo-tho-ana*, CL1.SG-mens-DIM, 'mensie') is, die minimale eenheid wees. Dit is egter nie noodwendig waar nie, want **motho** kan nog verder ontleed word.

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 (1994:17) 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. Wat die onderskeiding tussen morfeme betref, stel hy eerder 'n onderskeid tussen perifere morfeme (oftewel grammatiese morfeme) en sentrale morfeme (oftewel leksikale morfeme) voor. Hierdie siening kan versoen word met bostaande, deur Figuur 2.4 uit te brei soos in Figuur 2.5, waar die onderskeid tussen leksikale en grammatiese morfeme aangetoon word.



Figuur 2.5 Uitbreiding van die taksonomie van die Setswanawoord

Uit Figuur 2.5 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.

2.3.3.3.1 Stam

Volgens Trask (1996:334) 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."

'n Grammatiese woord⁴ is volgens Crystal (1992:160) 'n woord met geen leksikale betekenis nie, en het slegs 'n funksie om grammatiese verhoudings uit te druk (byvoorbeeld voorsetsels en voornaamwoorde).

Volgens Laas (1974:10) is 'n stam die morfologiese komponent wat 'n woordkorrelaat in die leksikon het en wat derhalwe op die woordvlak kan optree

⁴ Die definisie wat hier van grammatiese woord gegee word, stem ooreen met wat in Setswana onder grammatiese morfeem verstaan word, sodat dit hier eintlik as *morfeem* gelees kan word.

(teenoor die wortel wat onder die woordvlak optree). Die stam kan op grond van sy betekenis selfstandig optree, maar het nogtans 'n valensie vir grammatiese morfeme (Laas, 1974:33), soos dit in voorbeeld (11) na vore kom.

- (11) **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 (2003:433) kan die stam óf uit net een wortelmorfeem óf uit twee wortelmorfeme (samestelling), óf uit 'n wortelmorfeem met 'n afleidingsmorfeem bestaan. Die definisies van stamme in Afrikatale sluit nou by dié van Crystal (2003:433) aan deurdat Laas (1974:16-18) ook tussen stamme op grond van hulle morfologiese samestellings onderskei. 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 (2003:300) beskou dit as woorde wat uit 'n enkele (vrye⁵) morfeem bestaan, en dit is volgens hom die teenoorgestelde van meerledige monokernige woorde. Voorbeelde (12) - (15) is enkele van die simplekse in Setswana.

- (12) **malome**
CL1.SG.oom
'oom'

⁵ Vergelyk met bespreking by 2.3.2.

(13) **ruri**
'regtig'

(14) **ka**
'met'

(15) **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 (2003:300) die teenoorgestelde van eenledige monokernige woorde.

Enkele meerledige monokernige woorde word geïllustreer in voorbeelde (16) en (17).

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

(17) **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 (18) - (20) is voorbeelde van samestellings in Setswana.

- (18) **mosadimogolo**
mo-sadi-mo-golo
CL1.SG.vrou-CL1.SG-groot
'ou vrou'
- (19) **tautona**
tau-tona
CL5.SG.leeu-groot
'koning, groot leier'
- (20) **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 (21) - (23) geïllustreer.

- (21) **tautau**
tau-tau
CL5.SG.leeu-CL5.SG.leeu
'kaptein/leier'
- (22) **godimodimo**
go-dimo-dimo
CL9-bo-bo⁶
'heel bo'

⁶ Klas 9 is getalloos, daar is nie 'n meervoud of enkelvoud nie.

(23) **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 (16) is eenvoudig **sekolo**.

Dit raak egter moeiliker by die polikernige stamme, veral wat die samestellings betref. Wat sou dan in voorbeeld (20) 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 (21) 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 meerlede 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 (18)) word gevolglik, soos monokernige stamme, as lemma beskou.

2.3.3.3.2 Wortel

Volgens Crystal (2003: 402) en Trask (1996: 49) is 'n wortel daardie deel van die woord wat oorbly as al die affikse verwyder is. 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 (1974:33) het die wortel, al is dit die semantiese kern van 'n woord of stam, nie 'n selfstandige betekenis nie. Dit kom duidelik in die volgende voorbeelde na vore. In voorbeeld (24) is *-tho*, voorbeeld (25) *-kolo*, en voorbeeld (26) *-ag-* die wortels.

(24) **batho**

ba-tho

CL1.PL-mens

'mense'

(25) **sekolong**

se-kolo-ing

CL4.SG-skool-LOC

'lokaliteit van die skool'

(26) **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 (27) aangetoon word.

(27) **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 (27) 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. In Hoofstuk 1.2 is genoem dat 'n lemma-identifiseerder 'regte' woorde (of lemmas) as resultaat moet lewer. Aangesien wortels onder die woordvlak realiseer en daarom nie selfstandig kan optree nie, kan wortels nie as lemmas beskou word nie.

2.3.3.3.3 Affiks

Soos hierbo aangetoon (2.3.3.3.2), 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 (1974:18) 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." 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 (28)).

- (28) **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 (29), en die infinitiewe prefiks *go-*, die applikatiewe suffiks *-el-* en die uitgangsmorfeem *-a* in voorbeeld (30). 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 (30).

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

- (30) **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.

2.3.4 Voorlopige gevolgtrekking: 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 (vergelyk die bespreking onder 2.3.3.3.1). Volgens Laas (1974:33) het die wortel, al is dit die semantiese kern van 'n woord of stam, nie 'n selfstandige betekenis nie. 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 (31) geïllustreer word) kan, vanweë hulle afhanklike aard, nie as lemmas geneem word nie.

- (31) **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. Die woorde van die verskillende woordsoorte het verskillende kenmerke en affikse. Wat dan presies as lemmas beskou kan word, moet volgens die verskillende woordsoorte bespreek word.

2.4 Morfologie van woordsoorte

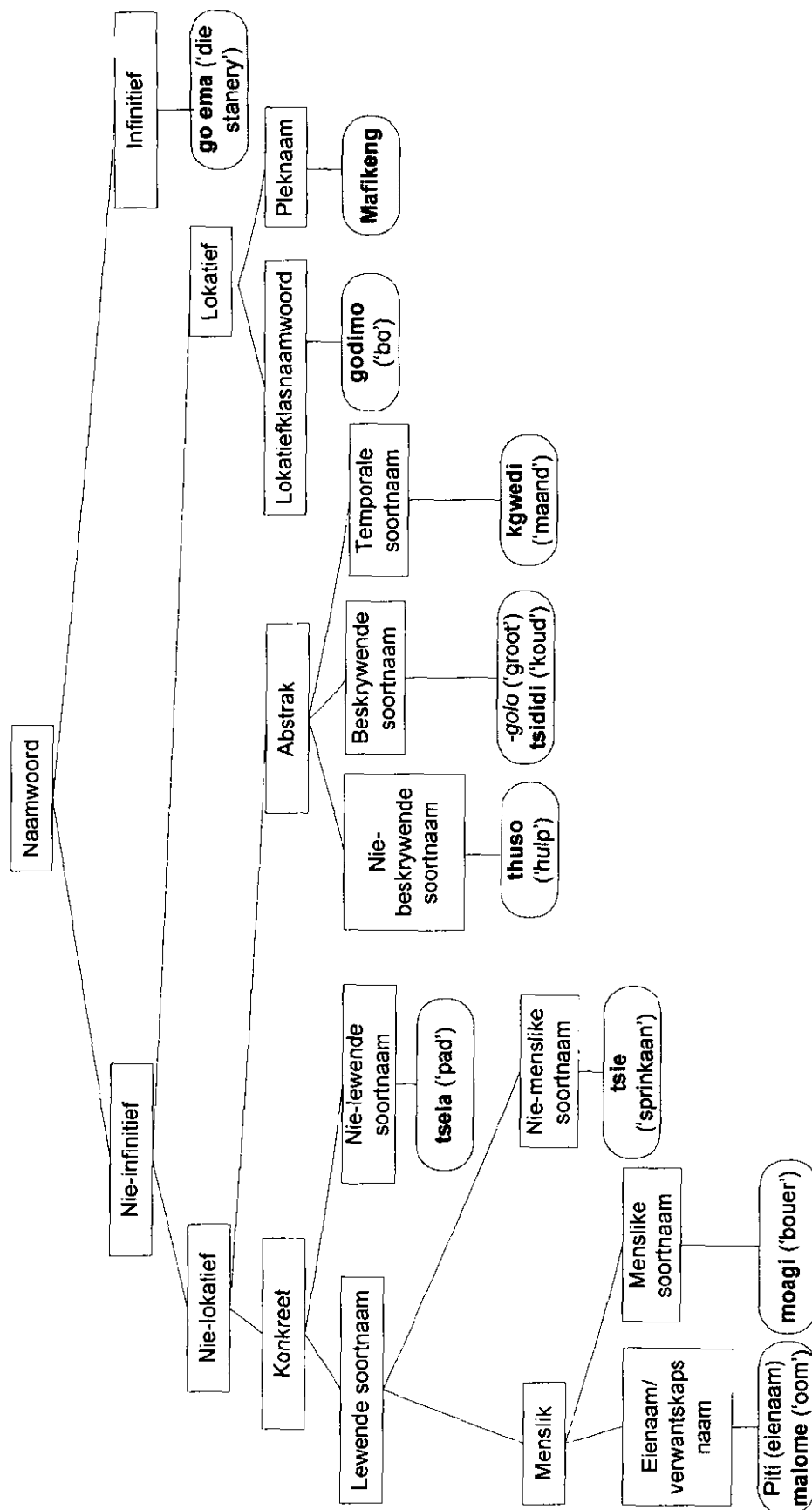
Daar is sewe woordsoorte in Setswana: die naamwoord, voornaamwoord, werkwoord, betrekkingwoord, bywoord, interjeksie en ideofon (vergelyk Van Wyk, 1966:230, 261). Nie al hierdie woordsoorte is ewe groot of word ewe maklik uitgebrei nie. Handke (1995:25) onderskei daarom tussen oopklaswoorde ("open-class words") en gesloteklaswoorde ("closed-class words"). 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, betrekkingwoorde, 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 dit 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.

2.4.1 Naamwoorde

Die naamwoord word deur Krüger (2006a:99) in subkategorieë verdeel soos wat dit in Figuur 2.6 voorgestel word.



Figuur 2.6 Die naamwoord se subkategorieë

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 (vergelyk Figuur 2.6). As daar derhalwe na klas 1 verwys word, dan word daarmee bedoel die *mo-/ba*-klas; klas 2 is dan die *mo-/me*-klas, ensovoorts. Ter wille van volledigheid word hierdie naamwoordklasse in Tabel 2.3 opgesom.

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

Tabel 2.3 Die naamwoordklasse in Setswana

Die keuse van die naamwoordlemma blyk tussen die eenvoudigste stam in die enkelvoud- en die meervoudsvorm te wees, behalwe klas 8 en 9 wat nie 'n enkelvoud of meervoud het nie (vergelyk Tabel 2.3). Die eenvoudigste stam waarna verwys word, is die stam met die minste affikse voordat dit as 'n wortel

geanaliseer kan word. Die deverbatiëwe suffiks sal nie verwyder word nie, omdat die woord van woordsoort sal verander (vergelyk die bespreking onder 2.3.4)

Die naamwoord kan deverbatiëwe, augmentatiëwe, feminitiewe, diminutiewe en lokatiëwe suffikse neem, soos wat dit in Tabel 2.4 uiteengesit is (vergelyk Krüger, 2006a).

stam		suffikse				
klasprefiks	wortel	deverbatiëf	augmentatiëf/feminitief	diminutief	diminutief	lokatiëf
		-i/-o	-gadi	-ana	-anyana	-(i)ng

Tabel 2.4 Die morfologie van die naamwoord

Voorbeelde (32) - (36) illustreer hoe die suffikse in Tabel 2.4 gebruik word. Omdat die naamwoord in voorbeeld (31) van woordsoort sal verander indien die deverbatiëwe suffiks verwyder word, sal dit vir die doeleindes van lemma-identifisering nie verwyder word nie. Die suffikse wat wel nie in die lemma ingesluit kan word nie, is die augmentatiëwe/feminitiewe (32) - (33), diminutiewe (34) en lokatiëwe (35) suffikse.

(32) **modumedi**
mo-dumel-a-i
 CL1.SG-glo-TERM-DEV.TERM
 'gelowige'

(33) **podigadi**
podi-gadi
 CL5S.G.bok-AUG
 'bok-ooi/n baie groot bok'

- (34) **kgomogadi**
kgomo-gadi
 CL5.SG.bees-FEM
 'koei/uitstaande koei'
- (35) **tawana**
tau-ana
 CL5.SG.leeu-DIM
 'leeutjie'
- (36) **kerekeng**
kereke-ing
 CL5.SG.kerk-LOC
 'lokaliteit van die kerk'

Die infinitief (klas 8) word grammaties onder die naamwoord gelys (vergelyk Figuur 2.6), omdat dit ook die vermoë het om kongruensie teweeg te bring (soos geïllustreer in voorbeeld (37)). Dit stem egter ook vormlik ooreen met die werkwoord. Die onderskeid wat tussen die infinitief en die werkwoord getref word, is sintakties van aard en daarom kan die infinitief vir praktiese doeleindes onder die werkwoord hanteer word.

- (37) **Go ema go tla re lapisa.**
Go-em-a go-tla-re-lap-is-a.
 INF-staan-TERM AGR.SUBJ-FUT-AGR.SUBJ-moeg-CAUS-
 TERM
 'Die stanery sal ons moeg maak.'

Behalwe die naamwoordklasse word die beskrywende naamwoorde (in ander tale gewoonlik adjektiewe of byvoeglike naamwoorde) ook as naamwoorde behandel (vergelyk Vermeulen & Pretorius, 1999:88-89, Van Rooy & Pretorius, 2003). Hieronder volg voorbeelde van die wortels (38) - (39) en stamme (40) - (41) van beskrywende naamwoorde. In voorbeelde (42) - (43)

blyk dit dat die stamme se vorm afhanklik is van die naamwoord wat hulle beskryf.

(38) *-golo*
'groot'

(39) *-bedi*
'twee'

(40) **kgolo**
(ne)-golo
CL5.SG-groot
'groot'

(41) **mebedi**
me-bedi
CL2.PL-twee
'twee'

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

(43) **melelo e mebedi**
me-lelo e me-bedi
CL2.PL-vuur PART.wat is CL2.PL-twee
'twee vure'

Alhoewel die algemene riglyn vir die bepaling van die lemma die eenvoudigste stam is, sal dit in hierdie geval beter wees om die wortelvorm van die beskrywende naamwoorde te gebruik. Die verskillende stamme **mogolo, bagolo, megolo, segolo, kgolo, bogolo**, ensovoorts kan onder een wortel, **-golo**, opgeneem word. Dit is dan ook hoe hulle in meeste

woordeboeke opgeteken word (vergelyk Dent, 1994; Kriel, 1958; Kriel, Van Wyk & Makopo, 1989; Reynierse, 1991; Snyman *et al.*, 1990; Ziervogel & Mokgokong, 1985).

Die temporale soortname is afgelei van gewone naamwoorde en bywoorde en sal as sodanig beskou word (vergelyk Figuur 2.6). Plekname en eiename kan – omdat hulle nie morfologies uitgebrei word nie – net so as lemma geneem word. Die temporale soortnaam in voorbeeld (44) is afgelei van 'n naamwoord in klas 9, **Mafikeng** in voorbeeld (45) is 'n pleknaam en **Lerato** in voorbeeld (46) is 'n eienaam. Die plek- en eiename het gewoonlik 'n betekenis en kan daarom streng gesproke ontleed word, maar word as plek- en eiename herken as die woorde met 'n hoofletter begin (vergelyk Krüger, 2006a:93-99).

(44) **morago**

CL9.agter

'daarna'

(45) **Mafikeng**

Ma-fika-ing

CL3.PL-fika-LOC

'Plek van die klippe'

(46) **Lerato**

Le-rat-a-o

CL3.SG-liefhê-TERM-DEV.TERM

'Liefde'

Samevattend kan gestel word dat die lemma van die naamwoord beskou word as daardie vorm in die enkelvoud sonder lokatiewe, diminutiewe, feminitiewe en augmentatiewe suffikse. Klas 8 en 9, wat nie 'n enkel- of meervoudvorm het nie, word met hulle klasprefikse sonder bogenoemde suffikse geneem. Plek- en eiename word net so as lemmas geneem.

'n Uitsondering by die naamwoorde is die beskrywende naamwoorde, waar die wortel as die lemma geneem word. Die lemma van die beskrywende naamwoord **segolo**, soos wat dit in voorbeeld (47) voorkom, is dus **-golo**.

(47) **Monna o reka sejanaga se segolo.**

Mo-nna o-rek-a se-ja-naga se se-golo

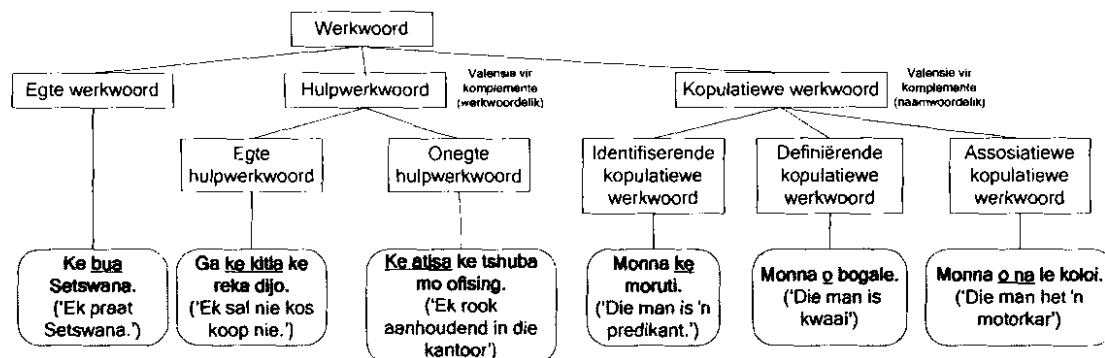
CL1.SG-man AGR.SUBJ-koop-TERM CL4.SG-eet-CL5.SG.veld

wat is CL4.SG-groot

'Die man koop 'n groot motorkar.'

2.4.2 Werkwoorde

Die werkwoord word deur Krüger (2006a:24) en Pretorius (1997) in die subkategorieë egte werkwoord⁷, hulpwerkwoord en kopulatiewe werkwoord verdeel, soos wat dit in Figuur 2.7 geïllustreer word.



Figuur 2.7 Die werkwoord se subkategorieë

Vervolgens word elk van hierdie subkategorieë bespreek om te bepaal wat as lemma onder elkeen van dié werkwoorde beskou kan word.

2.4.2.1 Egte werkwoorde

Krüger (2006a) verdeel die suffikse van die egte werkwoord in semi-produktiewe, produktiewe en uitgangsmorfeme. Die prefikse word verdeel in die infinitiewe prefiks, negatiewe, onderwerps- aspektuele, temporale en voorwerpsmorfeme soos wat dit in Tabel 2.5 voorgestel word.

⁷ Die term ‘egte werkwoord’ word in plaas van selfstandige werkwoord, in navolging van Krüger (2006a) gebruik.

Prefikse							Wortel	Semi-produktiewe suffikse					Produktiewe suffikse					Uitgang
INF	Neg	Om	Neg	Asp	Temp	Vm		Rev	Rev	Denom	Neutro- passief	Iter	Kous	Appl	Resipr	Perf	Pass	Neg/ Modale
go	ga	nie- lok. (klas) lok. go inf. go klas e	-se- -sa- -a-	-a- -sa- -ka-	-tia-	nie- lok. (klas) lok. -go- inf. -go- refl. -i-		-og- -olog-	-ol- -olol-	-f- -fal-	-eg- -al- -agal- -eseg- -agan-	-ak-	-is- -y-	-el- -ets-	-an-	-il-	-iw- +perf -w-	-a (mod) -e (neg) -ê (mod) -ng (mod)

Tabel 2.5 Die morfologie van die egte werkwoord (gebaseer op Krüger, 2006a:257)

Die egtewerkwoordlemma is die stam in die infinitiefvorm, aangesien dit gesien kan word as die eenvoudigste stam van die werkwoord, hoewel die *go-* in navolging van die gebruik in die woordeboeke (vergelyk Tabel 2.1) en uit praktiese oorwegings (alfabetisering) nie gebruik word nie. Dit is dus die werkwoord sonder die volgende prefikse: die negatiewe, kongruensie-, aspek- of temporale morfeme (vergelyk voorbeeld (48)), of suffikse soos die neutropassiewe, iteratiewe, kousatiewe, applikatiewe, resiproke, perfektum-, passiewe suffikse of die uitgangsmorfeme *-e* en *-ng* (vergelyk voorbeelde (49) - (56)). Indien dié affikse verwyder word, sal die werkwoordlemma oorbly, soos dit in die voorbeelde geïllustreer word.

- (48) **ke tla go thusa**
ke-tla-go-thus-a
 AGR.SUBJ-FUT-AGR.OBJ-help-TERM
 ‘ek sal jou help’
 Lemma: **thusa**

- (49) **robega**
rob-eg-a
 breek-N.PASS-TERM
 ‘breek/breekbaar’
 Lemma: **roba**

- (50) **ragaka**
rag-ak-a
 skop-ITER-TERM
 ‘herhaaldelik skop’
 Lemma: **raga**

- (51) **kwadisa**
kwaf-is-a
skryf-CAUS-TERM
'laat skryf'
Lemma: **kwala**
- (52) **rekela**
rek-el-a
koop-APPL-TERM
'koop vir'
Lemma: **reka**
- (53) **thusana**
thus-an-a
help-RECP-TERM
'mekaar help'
Lemma: **thusa**
- (54) **badiwa**
bal-iw-a
lees/tel-PASS-TERM
'word gelees/getel'
Lemma: **bala**
- (55) **kopile**
kop-il-e
vra-PERF-TERM
'gevra'
Lemma: **kopa**

- (56) **bitsang**
 bits-a-ng
 roep-TERM-TERM.REL
 'wat roep'
 Lemma: **bitsa**

Die onproduktiewe of onaktiewe suffikse, te wete die kontaktiewe, dispersiewe en posisionele suffikse, asook sommige van die semi-produktiewe suffikse, te wete die denominatiewe en reversiewe suffikse, word as deel van die lemma beskou. Hierdie suffikse is nie meer so aktief in Setswana nie en sommige vorms het nie meer ooreenstemmende eenvoudige stamme nie (vergelyk Cole, 1955:211-221; Krüger, 2006a:198; Lombard *et al.*, 1985:101-164; Pretorius, 1997:61-76).

2.4.2.2 Hulpwerkwoorde

Die hulpwerkwoord kan onderverdeel word in egte en onegte hulpwerkwoorde (vergelyk Figuur 2.7). Daar is egter min egte hulpwerkwoorde, en hulle kan daarom net soos wat hulle voorkom as lemmas gebruik word (vergelyk voorbeeld (57) - (59)). Die onegte hulpwerkwoorde is volgens Louwrens (1991:50-60) en Pretorius (1997:76-100) egte werkwoorde wat as hulpwerkwoorde gebruik word, en daarom geld dieselfde beginsels vir egte werkwoorde ook vir die onegte hulpwerkwoorde.

- (57) **nke**
 'sal nie'
- (58) **atisa**
 'herhaaldelik'
- (59) **setse**
 'reeds'

Die hulpwerkwoord kan nie alleen gebruik word nie en word altyd deur 'n egte werkwoord of kopulatiewe werkwoord gevolg. Die sin in (60) is 'n voorbeeld

van die hulpwerkwoord en egte werkwoord wat saam gebruik word. In hierdie sin is die hulpwerkwoord **o setse** se lemma **setse**, en die werkwoord **o segile** se lemma is **sega**.

(60) **Monna o setse o segile nama.**

Mo-nna o-sets-e o-seg-il-e nama

CL1.SG-man AGR.SUBJ-reeds-TERM AGR.SUBJ-sny-PERF-
TERM CL5.SG.vleis

‘Die man het reeds die vleis gesny.’

2.4.2.3 Kopulatiewe werkwoorde

Daar is drie soorte kopulatiewe werkwoorde, te wete die identifiserende, definiërende en assosiatiewe kopulatiewe werkwoorde (vergelyk Figuur 2.7). Die kopulatiewe werkwoord word modaal verbuig en kom in verskillende tye en aspekte, positief en negatief voor⁸. Dit bring mee dat die identifisering van die kopulatiewe werkwoord se lemma kompleks is. Dit word aan die hand van die identifiserende kopulatiewe werkwoord geïllustreer.

Die identifiserende kopulatiewe werkwoord is in die positief, teenwoordige tyd **ke** (‘dit is’), en in ’n negatiewe stelling **ga se** (dit is nie), soos geïllustreer in voorbeelde (61) - (62). In die indikatiewe modus se progressiewe aspek (positief) is die identifiserende kopulatiewe werkwoord **e sa le** (voorbeeld (63)). In die negatief van die progressiewe aspek is die kopulatiewe werkwoord ’n komplement van die hulpwerkwoorde **e sa ntse** en **ga e tihole** en dan is die vorm **e le** (voorbeelde (64) en (65)). Die identifiserende kopulatiewe werkwoord in die indikatiewe modus, toekomstige tyd, positief is **e tla nna** en in die negatief is dit **e nna** (voorbeelde (66) en (67)).

⁸ Nie alle grammatici beskou alle modale vorme van kopulatiewe as werkwoorde nie, maar verwys soms daarna as kopulatiewe betrekkingswoorde/partikels (vgl. Lombard *et al.*, 1985:173, 194-200; Louwrens, 1991:64-68).

- (61) **Monna ke moruti.**
Mo-nna ke mo-rut-i
 CL1.SG-man dit is CL1.SG.onderrig-TERM
 'Die man is 'n predikant.'
- (62) **Monna ga se moruti.**
Mo-nna ga-se mo-rut-i
 CL1.SG-man NEG-is CL1.SG.onderrig-TERM
 'Die man is nie 'n predikant nie.'
- (63) **Monna e sa le moruti.**
Mo-nna e-sa-le mo-rut-i.
 CL1.SG-man AGR.SUBJ-PROG-is CL1.SG.onderrig-TERM
 'Die man is nog steeds 'n predikant.'
- (64) **Monna e sa ntse e le moruti.**
Mo-nna e-sa-nna-il-e e-le mo-rut-i.
 CL1.SG-man AGR.SUBJ-PROG-steeds-PERF-TERM
 AGR.SUBJ-is CL1.SG-onderrig-TERM
 'Die man is nog steeds 'n predikant.'
- (65) **Monna ga e tlhole e le moruti.**
Mo-nna ga-e-tlhol-e e-le mo-rut-i.
 CL1.SG-man NEG-AGR.OBJ -aanhoudend-TERM
 AGR.SUBJ-is CL1.SG.onderrig-TERM
 'Die man is nie meer 'n predikant nie.'
- (66) **Monna e tla nna moruti.**
Mo-nna e-tla-nna mo-ruti.
 CL1.SG-man AGR.SUBJ-FUT-is CL1SG-onderrig-TERM
 'Die man sal 'n predikant wees/word.'

(67) **Monna ga e ketla e nna moruti.**

Mo-nna ga-e-ketl-a e-nna mo-rut-i.

CL1.SG-man NEG-AGR.SUBJ-FUT-sal-TERM

AGR.SUBJ-is CL1.SG-onderrig-TERM

‘Die man sal nie ’n predikant wees/word nie.’

In hierdie enkele voorbeelde (wat slegs in die indikatiewe modus aangebied is) het die identifiserende kopulatiewe werkwoord die volgende vorms aange-
neem: **ke, ga se, e sa le, e le, e tla nna, e nna**. Daar is dan ook nog ander
vorms in die ander modi, soos byvoorbeeld **ya nna, ya se nne, nna, se nne,
e nne, e se nne, e se**. Dit blyk uit bostaande voorbeelde feitlik onmoontlik
om ’n lemma met ooreenstemmende kenmerke vir al die verskillende vorms
van die kopulatiewe werkwoorde te identifiseer.

Die lemma kan egter wel op een van drie maniere bepaal word. Dit kan by-
voorbeeld elke modale vorm wees (byvoorbeeld **ya nna, ya se nne, nna, se
nne, e nne, e se nne, e se**). Andersins kan dit ook, na analogie van die
hantering van die selfstandige werkwoord, die vorm soos dit in die indikatiewe
teenwoordige tyd voorkom, wees. Indien hierdie tweede manier gevolg word,
sal die indikatiewe kopulatiewe werkwoordlemma **ke** en die assosiatiewe
kopulatiewe werkwoordlemma **na** (voorbeeld (68)) wees. Die definiërende
kopulatiewe werkwoord kom vormlik ooreen met die onderwerpsmorfeem.
Daarom sal die lemma dan **o, ba, e, le, a, se, di, lo, bo** of **go** kan wees
(voorbeelde (69)-(70)).

(68) **Monna o na le koloi.**

Mo-nna o-na le koloi

CL1.SG-man AGR.SUBJ-is saam met CL5.SG.motor.

‘Die man het ’n motor.’

Lemma: **na**

- (69) **Monna o bogale.**

Mo-nna o bo-gale

CL1.SG-man is CL7.SG-kwaaiheid.

'Die man is kwaai.'

Lemma: **o**

- (70) **Dikgomo di bogale.**

Di-kgomo di bo-gale.

CL5.PL-bees is CL7.SG-kwaaiheid.

'Die beeste is kwaai.'

Lemma: **di**

Die laaste en, myns insiens, mees sinvolle manier is om die basiese infinitiewe kopulatiewe werkwoord **go nna** ('om te wees') as lemma te neem (vergelyk Cole, 1955:336-339; Krüger, 2006a:275-278). Dit kom ook vormlik gedeeltelik ooreen met die werkwoordvorms van die verskillende kopulatiewe soos dit in die toekomstige tyd (indikatiwe modus) voorkom (vergelyk voorbeeld (71) - (73)). Die voorstel hier is dus dat die lemma van die kopulatiewe werkwoord die basiese kopulatiewe werkwoord **nnā** (sonder *go*, in navolging van die egte werkwoord) moet wees; **o tla nna** (voorbeeld (71)), **o tla nna** (voorbeeld (72)) en **di tla nna** (voorbeeld (73)) moet dus almal in die lemma-identifiseringsproses na **nnā** verander.

- (71) **Monna o tla nna morutabana.**

Mo-nna e-tla-nna mo-ruta-ba-ngwana.

CL1.SG-man AGR.SUBJ-FUT-is CL1SG-onderig-CL1PL-kind

'Die man is sal 'n onderwyser wees/word.'

Lemma: **nnā**

- (72) **Monna o tla nna le koloi.**

Mo-nna o-tla-nna le koloi.

CL1.SG-man AGR.SUBJ-FUT-is saam met CL5.SG.motor.

'Die man sal 'n motor hê.'

Lemma: **nnā**

(73) **Dikgomo di tla nna bogale.**

Di-kgomo di-tla-nna bo-gale.

CL5.PL-bees AGR.SUBJ-FUT-is CL7.SG-kwaaiheid.

‘Die beeste sal kwaai wees/word.’

Lemma: **nna**

2.4.3 Voornaamwoorde

Die verskillende voornaamwoorde kan net soos hulle voorkom as lemmas gebruik word. Hieronder volg voorbeelde (74) - (87) van die verskillende voornaamwoorde soos dit in Setswana voorkom (vergelyk Zerwick, 1997).

- Persoonlike voornaamwoord

(74) **nna**

‘ek’

(75) **rona**

‘ons’

- Absolute voornaamwoord

(76) **bone**

‘hulle’ (klas 1/7)

(77) **(w)one**

‘hy, sy, dit’ (klas 2)

- Aanwysende voornaamwoorde (demonstratiewe)

(78) **yo**

‘hierdie’ (klas 1)

(79) **yoo**
'daardie' (klas 1)

(80) **yole**
'doerdie' (klas 1)

- Kwantitatiewe voornaamwoorde

(81) **esi**
'alleen/slegs' (klas 2/5)

(82) **botlhe**
'almal' (klas 1/7)

- Enumeratiewe voornaamwoorde

(83) **ofe**
'watter' (klas 1/2)

(84) **ope**
'geeneen' (klas 1/2)

(85) **osele**
'ander/vreemde' (klas 1/2)

- Possessiewe voornaamwoorde

(86) **me**
'(van) my'

(87) **rona**
'(van) ons'

(Sien 1.1.1 - 1.1.6 in Bylaag 1 vir 'n volledige tabel van al die voornaamwoorde.)

2.4.4 Betrekkingswoorde

Daar is volgens Krüger (2006a:147-159) twee soorte betrekkingswoorde, te wete die kongruerende en nie-kongruerende betrekkingswoorde en albei kan net so as lemma gebruik word. 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 en nie-voegwoordelike betrekkingswoorde (vergelyk Krüger, 2006a:147-159). Hieronder volg enkele voorbeelde (88) - (89) van betrekkingswoorde (kursief) in sinsverband.

(88) **Ke batla go ya gae *gonne* ke lapile.**

Ke-batl-a go-y-a gae gonne ke-lap-il-e.

AGR.SUBJ-soek-TERM INF-gaan-TERM huis, want

AGR.SUBJ-moeg-PERF-TERM.

'Ek wil huis toe gaan, want ek is moeg.'

(89) **Ke rema *ka* selepe.**

Ke-rem-a ka se-lepe.

AGR.SUBJ-kap-TERM met CL4.SG-byl

'Ek kap met die byl.'

(Sien 1.2 in Bylaag 1 vir 'n volledige lys van betrekkingswoorde.)

2.4.5 Bywoorde

Egte bywoorde in Setswana is volgens Krüger (2006a:161-163) baie min en word daarom net soos hulle is as lemmas beskou. Hieronder is voorbeelde (90) - (95) van die bywoorde, terwyl 'n uitgebreide lys in Bylaag 2 onder 2.7 opgeneem word.

- (90) **kae?**
'waar?'
- (91) **jang?**
'hoe?'
- (92) **jaanong**
'nou'
- (93) **jalo**
'so/soos'
- (94) **jaana**
'sodanig'
- (95) **ruri**
'regtig'

2.4.6 Interjeksies

Interjeksies of tussenwerpsels het geen vasgestelde struktuur in Setswana nie en het daarom ook geen morfeme waaruit die woord opgebou is nie. Dit is ook 'n beperkte groep woorde wat nie morfologies uitgebrei word nie (vergelyk Cole, 1955: 395-400; Krüger, 2006a:165-166; Ntsime & Krüger, 1995: 177-182). Die interjeksies word net soos die voornaamwoord, betrekkingwoord en bywoord as lemma geneem (vergelyk voorbeelde (96) - (98)).

- (96) **Ao! Ija! Ijo! Mmalolo!**
'Mensig!/Haai'
- (97) **Ehe! Oo!**
'O so!'

(98) **X-x-x-x**

'Verskoon tog'

2.4.7 Ideofone

Soos dit met interjeksies die geval is, het die ideofoon ook nie 'n vasgestelde struktuur nie (vergelyk voorbeeld (99)). Omrede hierdie woordsoort nie morfologies uitgebrei word nie, word hulle net soos hulle in die taal voorkom, as lemmas geneem (Ras, 1991).

(99) **Pula ya tla e re tho tho tho.**

Pula ya-tla e-re tho tho tho.

CL5.SG.reën AGR.SUBJ-kom AGR-sê tip tip tip.

'Die reën val tip-tip-tip.'

2.5 Gevolgtrekking: Die lemma in Setswana

Die doel van hierdie hoofstuk was om die konsep "lemma" te definieer soos wat dit in die konteks van outomatiese lemma-identifisering in Setswana verstaan moet word. Ten einde so 'n definisie te formuleer, is die konsep "lemma", soos wat dit algemeen verstaan word, bespreek. Daar is vasgestel dat die lemma die genormaliseerde basisvorm is wat ook 'n lekseem is (die lemma is daarom nie net 'n basisvorm nie). Daarop het 'n bespreking van die Setswanawoordmorfologie gevolg aan die hand waarvan bepaal is hoe die lemma in Setswana gedefinieer kan word.

Daar 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. Daar moes bepaal word watter leksikale morfeem – die stam of die wortel – onafhanklik kan optree. 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 kan word.

Dit wat as lemma beskou kan word, is ook vir elke woordsoort afsonderlik bespreek. Dit beteken vir die naamwoord dat die lemma die vorm in die enkelvoud sonder enige suffikse (behalwe die deverbatiwiese suffiks waar dit voorkom) is. Die suffikse wat nie in die lemma ingesluit word nie, is die lokatiewe, diminutiewe, feminitiewe en augmentatiewe suffikse. Die lemma van **koloyana** in voorbeeld (100) is dus **koloi**, d.i. die eenvoudigste stam in die enkelvoud sonder die diminutiewe suffiks *-ana*.

(100) **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. Die plek- en eiename word onveranderd gelos.

Die werkwoord word in die egte werkwoord, hulpwerkwoord en kopulatiewe werkwoord verdeel (vergelyk die bespreking onder 2.4.2). 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 (101)) sal dus **aga** sonder die onderwerpsmorfem *o* en die perfektsuffiks *-il-* wees.

- (101) **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 (vergelyk die bespreking onder 2.4.2.2). Egte hulpwerkwoorde word net so as lemmas geneem (voorbeeld (102)). Beginself wat by die bepaling van die egte werkwoordlemma geld, is ook by die onegte hulpwerkwoorde van toepassing.

- (102) **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 (103) is dan eenvoudig **nna**.

- (103) **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, betrekkingwoorde, 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. (Sien 2.6 - 2.10 in Bylaag 2 vir die volledige lysie van dié woordsoorte.)

Die eerste doelstelling in Hoofstuk 1, om die lemma in Setswana te definieer en te identifiseer met die oog op outomatiese lemma-identifisering, is in hierdie hoofstuk bereik. Nou kan daar oorgegaan word om te bepaal of bestaande Setswanagrammatikas effektief geformaliseer kan word om die lemmas outomaties te identifiseer (die tweede doelstelling van die studie). In Hoofstuk 3 word die bestaande grammatika geformaliseer sodat die affikse in die lemma-identifiseringsproses verwyder kan word om die regte stam, oftewel die lemma, te kry.

Hoofstuk 3

Linguistiese reëls vir lemma-identifisering in Setswana

3.1 Inleiding

In Hoofstuk 1 is die begrip lemma-identifisering bespreek en is daarna in Hoofstuk 2 bepaal wat as lemma in Setswana beskou kan word. Dit volg uit Hoofstuk 2 dat 'n lemma in Setswana die stam in sy eenvoudigste vorm is voordat dit as 'n wortel beskou kan word. Dit beteken dan dat die naamwoordlemma in die enkelvoud sonder enige suffikse (behalwe die deverbatiëwewortel suffiks) is. Die egtewerkwoorde is die eenvoudigste stam in die infinitief, sonder enige suffikse (behalwe die onproduktiewe en onaktiewe suffikse).

Dit gaan in die ontwikkeling van 'n lemma-identifiseerder nie net oor die bepaling van wat 'n lemma is nie, maar ook oor hoe die lemma geïdentifiseer moet word. Dit hou in dat die juiste benadering en tegniek gekies moet word om die proses outomaties uit te voer. In Hoofstuk 1 is daar op die linguistiese benadering besluit omdat die morfologiese patrone van die naamwoorde en werkwoorde in Setswana redelik reëlmatig is, terwyl die ander woordsoorte morfologies onproduktief is. Daar is ook 'n goeie kennisbasis vir die linguistiese benadering omdat die morfologiese struktuur van die Sothotale, en spesifiek Setswana, omvattend in handboeke en artikels beskryf is (vergelyk onder andere Cole, 1955; Krüger, 1967, 1994 & 2006a (Setswana); Lombard *et al.*, 1985; Louwrens, 1991 (Sesotho sa Leboa)). Met die ontwikkeling van 'n reëlgebaseerde lemma-identifiseerder kan ook bepaal word hoe effektief bestaande Setswanagrammatikas geformaliseer kan word om die lemmas te identifiseer.

Die formalisering van die lemma-identifiseringsreëls in Setswana word in die notasie van die sagteware (FSA 6) gedoen. Voordat die bestaande grammatikareëls dus geformaliseer kan word, moet begrippe soos "reëlmatige uitdrukking", "eindigestaatvormer" en "eindigestaatoutomaat" bespreek word. Daarna word die notasieformaat van die reëlmatige uitdrukkings in FSA 6 bespreek, waarna die reëls vir die bepaling van die naamwoordlemma aan

die beurt kom, gevolg deur die bespreking van die formalisering van die werkwoordreëls en die reëls vir die ander woordsoorte.

3.2 Formalisering van linguistiese reëls: Eindigestaatoutomate

Daar is volgens Voutilainen (1999:9) hoofsaaklik twee soorte benaderings wat in natuurliketaalverwerking gevolg word. In die vyftiger- en sestigerjare is die linguistiese benadering gevolg, waarin linguistiese reëls gebruik is in die ontwikkeling van toepassings soos woordsoortetiketteerders. Die eindigestaatoutomate wat ook sinoniem met die linguistiese benadering is, se ontstaan is aan Turing (1936) se model van algoritmiese berekeninge te danke. Turing se model lei tot Kleene (1951 en 1956) se werk op eindigestaatoutomate en reëlmatige uitdrukkings (Jurafsky & Martin, 2000:10-11). In die sewentigerjare word daar van die linguistiese benadering wegbeweeg en ontstaan die datagedrewe benadering waarin taalpatrone outomaties afgelei is uit reeds geannoteerde teks as toevoer. In hierdie benadering word gewoonlik van masjienleertegnieke en statistiese metodes gebruik gemaak (Voutilainen, 1999:9-11). Die eindigestaatmodelle kry in die tagtigerjare met Kaplan en Kay (1981) se werk op eindigestaatmorfologie en -fonologie weer prominensie in natuurliketaalprosessering (Jurafsky & Martin, 2000:14).

In die linguistiese benadering word begrippe soos “reëlmatige uitdrukkings” en “reëlmatige tale” in noue verband met eindigestaatoutomate gebruik. Reëlmatige uitdrukkings is die formalisme wat gebruik word om die outomata te beskryf en dit beskryf terselfdertyd ook reëlmatige tale. Die begrippe “reëlmatige tale” en “reëlmatige uitdrukkings” word dus eers bespreek voordat daar oorgegaan word na die bespreking van die outomaat.

3.2.1 Reëlmatige tale, reëlmatige uitdrukkings en outomata

Reëlmatige tale bestaan uit 'n stel stringe (soms sinne genoem) wat op hulle beurt uit aaneengerygde simbole (karakters of woorde) bestaan. Hierdie simbole vorm deel van 'n eindige alfabet of woordeskat (Beesley & Karttunen, 2003:261; Jurafsky & Martin, 2000:72; Kaplan, 1996). So byvoorbeeld kan 'n reëlmatige taal uit die letters *a* en *b* bestaan. Stringe soos *aab*, *abb* en *aba*

kan dan geldige 'woorde' in die reëlmatige taal wees, gegewe dat 'n woord net met 'n **a** mag begin en net uit 'n **a** en **b** mag bestaan. 'Woorde' soos **baab** en **abc** sou dan in hierdie taal ongeldig wees.

Natuurlike tale is ook tot 'n bepaalde mate reëlmatig (vergelyk Hoard, 1998:119). 'n Natuurlike taal soos Setswana bestaan ook uit aaneengerygde simbole (wat in hierdie geval van die Latynse alfabet afkomstig is). Die werkwoorde in Setswana eindig byvoorbeeld in bepaalde sintaktiese kontekste (teenwoordige tyd, positief) op 'n **-a** of **-ang**, soos dit in **dula** ('sit') en **dulang** ('wat sit') geïllustreer word. Tog is daar, anders as by reëlmatige tale, uitsonderings; so byvoorbeeld eindig die werkwoord **itse** ('weet') in die teenwoordige tyd, positief op 'n **-e**. Tog kan natuurlike tale, ten spyte van die uitsonderings, beskryf word deur middel van reëlmatige uitdrukkings.

Die reëlmatige uitdrukking is die standaardnotasie (d.i. formalisme) wat gebruik word vir die beskrywing van reëlmatige teksstringe. Aan die hand van die reëlmatige taal wat hierbo beskryf is, kan die volgende reëlmatige uitdrukking geformaliseer word: $[a, ?^*]$. Die vierkantige hakie $[$ dui die begin van 'n woord/string aan en in hierdie geval begin die string met 'n **a**. Hierdie **a** kan gevolg word deur enige simbool (?) in die taal, wat in hierdie geval **a** en **b** is. Die sterretjie (*), oftewel Kleene-ster, beteken dat die simbool nul of meer kere kan voorkom en die $]$ dui die einde van die woord/string aan. Dit kan ook op Setswana van toepassing gemaak word. Indien 'n mens die werkwoord in die teenwoordige tyd positief wil beskryf, kan dit só lyk: $[?^*, \{a, [a, n, g]\}]$. 'n Werkwoord kan dus met enige letter begin en op 'n **-a** of **-ang** eindig.

'n Ander voorbeeld van 'n reëlmatige uitdrukking is $[m, o, n, n, a]$. In hierdie geval stel dit 'n blote opeenvolging van 'n string karakters voor. In enige string, soos voorbeeld (104), sal die reëlmatige uitdrukking slegs die substring **monna** pas.

(104) **Mosadi o bona monna yo.**

Mo-sadi o-bon-a mo-nna yo.

CL1.SG-vrou AGR.SUBJ-sien-TERM CL1.SG-man hierdie.

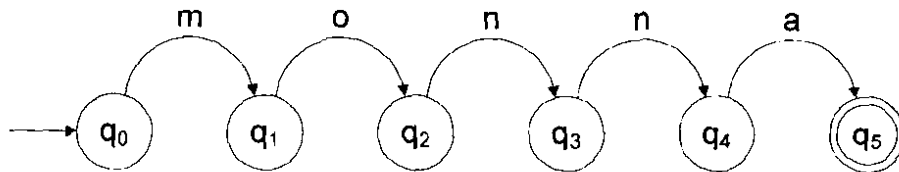
‘Die vrou sien hierdie man.’

Reëlmatige uitdrukkings word dikwels gebruik om soekpatrone in woordverwerkers en bedryfstelsels te spesifiseer, soos byvoorbeeld Unix se *Grep* (Hausser, 1999:34). Die reëlmatige uitdrukkings wat as een van die suksesse in die standaardisering van rekenaarwetenskap beskou word, word via die eindigestaatoutomaat (“finite state automaton”; voortaan FSA) en variasies daarvan, soos die eindigestaatoorvormer (“finite state transducer”; voortaan FST) geïmplementeer om in websoektogte en woordverwerking gebruik te word (Jurafsky & Martin, 2000:21-22). Kaplan (1996) sluit daarby aan en stel dat reëlmatige uitdrukkings en eindigestaatmasjiene wiskundig eenvoudig en maklik programmeerbaar is, en gevolglik in vele inligtingsprosesseringstake gebruik word.

Eindigestaatmasjiene word nie net vir inligtingsprosesseringstake gebruik nie, maar FSA’s word ook ingespan vir morfologiese herkenning. Volgens Jurafsky en Martin (2000:57-58) is morfologiese ontleding nodig vir die herkenning van stringe, en die hoofkomponent van ’n belangrike algoritme vir morfologiese ontleding is die eindigestaatoorvormer. Voorts word FSA’s gebruik om ’n leksikon te verteenwoordig en is daar toenemende belangstelling om eindigestaatmasjiene te gebruik vir die stoor en gebruik van natuurliketaalwoordeboeke (Jurafsky & Martin, 2000:71; Kaplan, 1996; Ritchie *et al.*, 1987:253). Hoe FSA’s gebruik word, sal aan die hand van verduidelikings hieronder duideliker word.

Die outomaton (of FSA) wat volgens Jurafsky en Martin (2000:33) deur reëlmatige uitdrukkings beskryf word, bestaan uit ’n eindige aantal state en ’n funksie wat die oorgange van die een staat na ’n ander bepaal soos wat die toevoer gelees word. Die outomaton begin by die eerste staat met die eerste simbool van ’n bepaalde string. Die outomaton beweeg dan van staat na staat soos dit die toevoer lees, totdat dit by die einde van die string kom. Hier is die

automaton in die eindstaat en het dit die string aanvaar wat tot die taal behoort. Die proses word gewoonlik in 'n diagram voorgestel waar die sirkels die state voorstel en verbind word deur pyle om die oorgange aan te dui. Die laaste en eindstaat word verteenwoordig deur die dubbelsirkel (Kaplan, 1996). Figuur 3.1 is 'n visuele voorstelling van die automaton van die reëlmatige uitdrukking $[m, o, n, n, a]$:

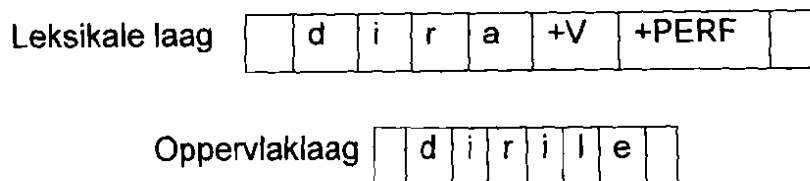


Figuur 3.1 'n Visuele voorstelling van die automaton

Soos reeds genoem, word reëlmatige uitdrukkings nie net deur middel van die FSA implementeer nie, maar ook via variasies daarvan, soos die FST (Jurafsky & Martin, 2000:21-22). Om die verskil tussen 'n FSA en 'n FST te verduidelik, word die eindigestaatoorvormer vervolgens aan die hand van tweevlakmorfologie bespreek.

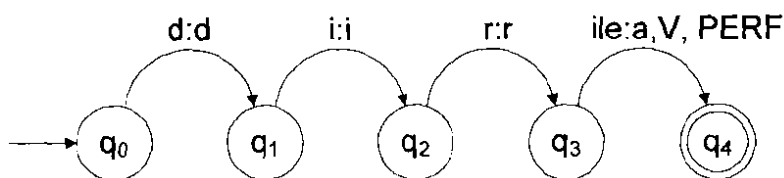
3.2.2 FST en tweevlakmorfologie ("two-level morphology")

Volgens Jurafsky en Martin (2000:71) kan die verskil tussen 'n FSA en 'n FST aan die hand van tweevlakmorfologie (two-level morphology), soos eerste deur Koskeniemi (1983) voorgestel, verduidelik word. In tweevlakmorfologie word 'n woord verteenwoordig as 'n ooreenkoms tussen 'n **leksikale laag** wat die aaneenskakeling van morfeme om 'n woord te vorm verteenwoordig, en die **oppervlaklaag** wat die werklike spelling van die finale woord verteenwoordig. Morfologiese ontleding vind plaas deur passings ("mapping rules") wat stringe in die oppervlaklaag verander na die morfeem- en eienskapstringe, in die leksikale laag. Dit word grafies in Figuur 3.2 voorgestel, waar die string **dirile** ('het gewerk') in die oppervlaklaag na **dira +V +PERF** in die leksikale laag verander word.



Figuur 3.2 Tweevlakmorfologie

Die outomaat wat gebruik word om die verandering tot stand te bring, is die FST. 'n FST verander die een stel simbole in 'n ander stel via die eindige outomaat. Dieselfde voorbeeld in Figuur 3.2 word in Figuur 3.3 voorgestel as 'n spesiale outomaat, naamlik die FST. Die dubbelpunt dui aan dat die een simbool na 'n ander verander moet word.



Figuur 3.3 'n FST

'n FST is dus 'n spesiale soort outomaton wat een string lees en 'n ander genereer (Jurafsky & Martin, 2000:71-72). Die FST herken of genereer pare stringe en het daarmee 'n meer algemene funksie as 'n FSA. FSA's beskryf reëlmatige tale en FST's beskryf weer reëlmatige verhoudings. Over geset synde definieer 'n FSA stelle stringe en 'n FST die verhouding tussen stelle stringe. Waar FSA's verband hou met reëlmatige tale ("regular languages"), hou FST's verband met reëlmatige verhoudings ("regular relations"). Soos wat reëlmatige tale bestaan uit 'n stel stringe, so bestaan reëlmatige verhoudings uit 'n stel van stringpare (Jurafsky & Martin, 2000:72).

In Hoofstuk 1 is reeds aangetoon dat morfologiese en morfo-fonologiese reëls die onderbou van 'n morfologiese analiseerder/lemma-identifiseerder (wat uit FST's bestaan) is. Die morfologiese reëls orden die morfeme – watter morfeme voor die stam mag kom en in watter volgorde dit na die stam volg. Ortografiese/morfo-fonologiese reëls word gebruik om die veranderings wat

plaasvind as twee morfeme kombineer, te beskryf (Jurafsky & Martin, 2000:65. Die klankveranderings soos wat dit in die ortografie van Setswana gereflekteer word as sekere morfeme kombineer, word breedvoerig deur Krüger en Snyman (1986) bespreek. Hierdie reëls kom in 3.4, soos wat dit vir 'n lemma-identifiseerder van toepassings sal wees, aan die beurt.

Ten einde FSA's en FST's in 'n lemma-identifiseerder te implementeer, moet daar van eindigestaatsagteware gebruik gemaak word. Reëlmatige uitdrukkings word in eindigestaatsagteware gebruik om FSA's en FST's te implementeer. Die algemene manier om 'n reëlmatige taal te beskryf, is deur middel van die reëlmatige uitdrukkings, wat 'n standaardnotasie is wat aandui in watter volgorde hierdie simbole aanmekaar skakel, watter alterniewe by sekere punte moontlik is en of substringe herhaal kan word (Kaplan, 1996; vergelyk Beesley & Karttunen, 2003:44). Vervolgens word die eindigestaatsagteware *FSA 6* en die notasie wat in *FSA 6* gebruik word, beskryf.

3.3 Eindigestaatsagteware: Finite State Automata Utilities *FSA 6*

Daar word heelwat eindigestaatsagteware (meer as dertig produkte/sagteware) op *Grail+*, 'n webwerf van die rekenaarwetenskapdepartement van die "University of Western Ontario" in Kanada, gelys (University of Western Ontario Department of Computer Science, 2002)⁹. Van die programmatuur wat beskikbaar is, is die bekendste sekerlik die *Xerox Lexical Technology*-produkte, wat ook in 'n NRF-befondsde projek by UNISA (vergelyk onder andere Pretorius & Bosch, 2003) gebruik word. *Xfst* (Xerox se eindigestaatoorvormer) kan gratis met 'n nie-kommersiële lisensie gebruik word, maar indien dit kommersieel gebruik sou word, behoort 50% van die lisensie aan Xerox (Beesley & Karttunen, 2003:508). Die bruikbaarheid van hierdie sagteware in terme van praktiese

⁹ Na afhandeling van hierdie verhandeling is nuwe eindigestaatoutomate ("finite-state registered automata" - FSRA) vir nie-agglutinerende morfologie deur Cohen-Sygal en Wintner (2006) bekendgemaak. Hierdie insigte kon dus ongelukkig nie hier verreken word nie, maar moet in verdere navorsing ondersoek word.

implementering in kommersiële produkte verlaag dus aangesien dit 'n duur proses kan word.

Die programmatuur wat deur Van Noord (2002) ontwikkel is, naamlik *Finite State Automata Utilities Version 6 (FSA 6)*, word ook allerweë beskou as een van die bekendste pakkette wat beskikbaar is om eindigestaatoutomate en -oortvormers mee te ontwikkel (Piskorski & Neuman, 1999). Die “FSA Utilities toolbox” is 'n versameling van nutsprogramme om reëlmatige uitdrukkings, eindigestaatoutomate en eindigestaatoortvormers te manipuleer. *FSA 6* kan dus eindigestaatoortvormers hanteer, wat in die geval van 'n reëlgebaseerde lemma-identifiseerder van kardinale belang is (Van Noord, 2002b). In 'n e-pos aan die outeur het Van Noord (2004) bevestig dat dit wel moontlik is om 'n lemma-identifiseerder met sy nutsprogramme te bou, solank dit nie 'n groot leksikon insluit nie.

Die programmatuur, *FSA 6*, is goed gedokumenteerd, kan gratis afgelaai word, en die kode wat daarmee genereer word, kan ook in ander produkte gebruik word. Die volgende is nodig vir die installering van *FSA 6* (Van Noord, 2002):

- SICStus Prolog (weergawe 3.12.5), insluitend 'n Tcl/Tk-biblioteek wat in die grafiese gebruikerskoppelvlak (sien Bylaag 3) gebruik word (dit is vinnig en robuust); of
- Yap, wat geen grafiese gebruikerskoppelvlak het nie, maar baie vinnig is; of
- SWI, wat eweneens geen grafiese gebruikerskoppelvlak het nie, en stadig is. Die gebruikershandleiding noem dat die grafiese koppelvlak net met SICStus Prolog gebruik kan word, maar dit is wel moontlik om die grafiese koppelvlak met SWI-Prolog te gebruik.

Behalwe die hulplêers vir die installering van *FSA 6*, is daar 'n verwysingshandleiding met voorbeelde en studiemateriaal beskikbaar. In die installeringslêer word aanwysings gegee vir die installering in Unix en Windows. *FSA 6* is ook versoenbaar met AT&T se “FSM toolset”. Een van die sterkste punte wat in die handleiding na vore gekom het, is dat *FSA 6*

kodes (Prolog, C, C++ en Java) van reëlmatige uitdrukkings genereer (Van Noord, 2002).

Die kodes wat met *FSA 6* gegenereer word, word in Hoofstuk 4 in 'n lemma-identifiseerder geïmplementeer om te bepaal tot watter mate die morfologiese reëls geformaliseer kan word. Die bespreking van die notasie van reëlmatige uitdrukkings kom volgende onder 3.3.1 en 3.3.2 aan die beurt.

3.3.1 Reëlmatige uitdrukkings in *FSA 6*

FSA 6 gebruik dieselfde beginsels as Xerox se *Xfst* vir reëlmatige uitdrukkings, alhoewel die sintaksis verskil (vergelyk Beesley & Karttunen, 2003; Xerox, 2004). Die basiese operators van *FSA 6* se reëlmatige uitdrukkings kan soos in Tabel 3.1 opgesom word (Bouma, 2004; Van Noord, 2000).

Operator	Betekenis	Voorbeeld
[]	leë string	{ [n,g]:[] } (Die string ng word deur niks – die leë string – vervang.)
[E1,E2,...En]	aaneenskakeling van E1, E2,...En	{ [g,a,d,i] } (Die letters g-a-d-i word aaneengeskakel om die string gadi te vorm.)
{ }	leë taal	{ }
{E1,E2,...En}	vereniging van E1,E2,...En	{ [e:a],[b:c] } (Indien van toepassing, word een of meer van die elemente in die krulhakies uitgevoer - dit string dus 'opdragte' aanmekaar. Die string eb sal in hierdie geval ac word.)
E*	Kleene star, nul of 'n oneindige reeks van E	a* (Daar kan geen of 'n oneindige getal a 's wees. [a*,b*] : definieer die taal wat bestaan uit 'n willekeurige aantal a 's gevolg deur 'n willekeurige aantal b 's. Die stringe aabbbb , aaa , bbb word

		deur hierdie uitdrukking herken, maar aba, bbaa, abc nie. $[a,b]^*$: definieer die taal wat uit 'n willekeurige aantal ab's bestaan; dus word ab en abababab herken, maar nie abbba, aab of bb nie.)
E^*	opsionaliteit	$\{[b^*,a], [a,b^*]\}$ (Die b 's is in hierdie geval opsioneel en die stringe bo , ba , o en a sal in hierdie geval herken word.)
$\sim E$	komplement van E	$\{[\sim \quad \$[n,g]], [w,e]:e\}$ (In hierdie geval sal die we net e word as dit nie deur 'n ng voorafgegaan is nie.)
$E_1 - E_2$	verskil	$\{[a,b,c,d] - [a,b,d]\}$ (Die verskil tussen die twee uitdrukkings is c , dus sal net die c aanvaar word.)
$\$ E$	bevat E	$\$[n,g]$ (Die string bevat 'n ng .)
$E_1 \& E_2$	snyding	$\$[i,l,e] \& \$[n,g]$ (Daar moet dus iewers in die string 'n ile en ng wees - sonder 'n spesifieke volgorde.)
?	enige simbool	? (dit kan 'n a , b , c , d , e ens. wees)
$A:B$	paar (eindigestaatoorvoerder)	$[a:b]$ (enige plek waar daar 'n a in 'n string voorkom, sal dit na 'n b verander word)
$E_1 \times E_2$	kruisproduk	$\{[a,b] \times [c]\} = a:c \text{ en } b:[]$
$A \circ B$	samestelling	$\{[?^*, [n,g]:[]]\} \circ \{[?^*, [i,l,e]:a]\}$ (Die resultaat word die toevoer van die volgende uitdrukking. Die string rekileng ('wat gekoop het') word rekile ('gekoop

		het') en rekile word weer reka ('koop').)
--	--	---

Tabel 3.1 Basiese operators in FSA 6

Die operators wat in Tabel 3.1 gelys is, kan gekombineer word in reëlmatige uitdrukkings en in Setswana toegepas word om lemmas te identifiseer. Indien die diminutiewe suffiks van byvoorbeeld die woord **tsejwana** ('oortjie') verwyder moet word om die lemma **tsebe** ('oor) te bepaal, kan die reël **-jwana** > **-be** só geskryf word: $[?^*, [j, w, a, n, a] : [b, e]]$. In hierdie geval kan die string **jwana** wat na **be** verander, voorafgegaan word deur 'n willekeurige aantal (*) simbole (?).

Daar is byvoorbeeld ook twee reëls waarmee neutro-passiewe suffikse verwyder kan word: **-esega** kan deur **-a** vervang word (oftewel $[?^*, [e, s, e, g, a] : a]$) en **-ega** word deur **-a** vervang in die geval waar dit nie deur **-es-** voorafgegaan word nie (oftewel $[?^*, ? -e, ? -s, [e, g, a] : a]$). Die laaste twee simbole (?) wat die string **ega** voorafgaan kan dus enige simbole behalwe (-) die simbole **e** of **s** wees.

Onder 3.4 word die grammatiese reëls wat in reëlmatige uitdrukkings omskryf is met behulp van die operators in FSA 6 bespreek. Die operators in FSA 6 stel die gebruiker in staat om reëlmatige uitdrukkings vir 'n taal soos Setswana te skep, wat in makro's (wat hieronder bespreek word) en uiteindelik ook in 'n lemma-identifiseerder gebruik kan word.

3.3.2 Makro's en Prolog

FSA 6 is op die programmeertaal Prolog gebaseer en dit vereis dus dat Prolog ook geïnstalleer moet wees. Die voordeel hiervan is dat komplekse reëlmatige uitdrukkings in makro's gedefinieer kan word. Hierdie makro's word buite FSA 6 in Prolog geskryf.

Makro's kan byvoorbeeld gebruik word om klinkers en medeklinkers te definieer. In die voorbeeld hieronder kry die veranderlike *klinker* dan die waarde a, e, i, o of u en die veranderlike *medeklinker* is dan die res van die alfabet (d.i. – klinker). So 'n makro lyk soos volg: :- multifile macro/2.

```
macro(klinker,{a,e,i,o,u}).
```

```
macro(medeklinker, a..z – klinker)
```

Die makro word soos volg in FSA 6 gebruik: Gaan na File > LoadAux en selekteer die betrokke Prolog-lêer. Die veranderlikes wat in die makro gedefinieer is, in hierdie geval *klinker* of *medeklinker*, kan op hulle eie as reëlmatige uitdrukkings of as deel van 'n reëlmatige uitdrukking gebruik word. As die veranderlike *klinker* byvoorbeeld ingetik word, interpreteer FSA 6 dit as 'n FSA wat spesifiseer dat net 'n **a**, **e**, **i**, **o** of **u** aanvaar kan word. Die veranderlike kan ook as deel van 'n reëlmatige uitdrukking gebruik word. *Klinker* kom byvoorbeeld voor in die reëlmatige uitdrukking waarin die voorwerpsmorfeem *n*- en die refleksief *i*- voor 'n werkwoord wat met 'n klinker begin, verwyder word. Dit is dus makliker om die reëlmatige uitdrukking eerder só te gebruik: `[[{n,i},k]:[],klinker,?*` as om dit só te gebruik: `[[{n,i},k]:[],{a,e,o,u,i},?*`

Lemma-identifisering behels die verwydering/vervanging van affikse en die grammatiese reëls wat gebruik word om die vorming van woorde te beskryf, word dus in der waarheid herskryf om die affikse te kan verwyder. Die reëlmatige uitdrukkings en makro's wat tot dusver bespreek is, word vervolgens gebruik om dié reëls in Setswana te formaliseer.

3.4 Reëls

In hierdie afdeling word die reëls, soos wat dit vir Setswana geformaliseer kan word, bespreek. Dit gaan egter nie net oor die reëls as sulks nie, maar ook oor die volgorde waarin die reëls geïmplementeer moet word. Hierdie ordening sal derhalwe in hierdie afdeling en ook in Hoofstuk 4 aan die orde kom. Die notasie soos wat dit in FSA 6 van toepassing is (vergelyk die bespreking onder 3.3.1), word in die formalisering van die reëls gebruik. Die

reëls vir die bepaling van naamwoordlemmas kom eerste aan die beurt, waarna die reëls vir die bepaling van die werkwoordlemmas bespreek word. Hierdie afdeling word met die bespreking van die geslote woordsoorte, naamlik die voornaamwoord, betrekkingwoord, bywoord, interjeksie en ideofoon, afgesluit.

3.4.1 Morfologiese reëls vir naamwoorde

In Hoofstuk 2 (2.4.1) is daar aangedui dat, met betrekking tot naamwoorde, die lemma daardie vorm is wat in die enkelvoud sonder enige suffiks (behalwe die deverbative suffiks) is.

Die ordening van die reëls kan dieselfde hiërargiese orde volg as wat Krüger (1994:20-22) in morfologiese analyses gebruik. Wat die naamwoord betref, moet die lokatiewe suffiks eerste verwyder word, gevolg deur die diminutiewe suffiks en laastens feminitiewe/augmentatiewe suffikse. In Tabel 3.2 hieronder word die hiërargie vir die verwydering van die suffikse van die naamwoord uiteengesit.

stam		suffikse			
klasprefiks	wortel	augmentatief/feminitief	diminutief	diminutief	lokatief
verander na enkelvoudsprefiks		<i>-gadi</i>	<i>-ana</i>	<i>-anyana</i>	<i>-(i)ng</i>
Volgorde		4	3	2	1

Tabel 3.2 Die hiërargie vir verwydering van die suffikse van die naamwoord

Behalwe die ordening van die morfologiese reëls, speel die morfo-fonologiese veranderings, soos onder 3.2.2 genoem, ook 'n belangrike rol. Die morfologiese reëls vir byvoorbeeld die vorming van **dikolojwana** ('varkies') sal *dikolobe-* + *-ana* wees. As gevolg van die morfo-fonologiese veranderings kan die *-ana* nie net bloot verwyder word nie, maar **-jwana** moet na **-be** verander (vergelyk Krüger & Snyman, 1986:129-130).

Vervolgens word die bestaande reëls vir naamwoorde bespreek soos wat dit vir die ontwikkeling van die lemma-identifiseerder geformaliseer is. Eerstens word die naamwoordklasse hanteer waar die lokatiewe, diminutiewe en feminitiewe/augmentatiewe suffikse verwyder moet word, en ook die meervoudsklasprefikse deur die enkelvoudsprefikse vervang moet word (3.4.1.1). Ten slotte word die reëls vir die beskrywende naamwoorde bespreek (3.4.1.2). In Bylaag 2.1 word die reëls vir die bepaling van die naamwoordlemmas volledig weergegee.

3.4.1.1 Die naamwoordklasse

In hierdie afdeling word die morfologiese reëls en morfo-fonologiese reëls van die naamwoordklasse bespreek. In lemma-identifisering word dieselfde reëls gebruik – nie om woorde te vorm nie, maar om die lemma van die woord te bepaal. Die reëls word dan in der waarheid in trurat gebruik. Die reëls wat gebruik word om die affikse te verwyder, word in dieselfde notasie gegee as wat dit in FSA 6 gebruik is.

3.4.1.1.1 Lokatiewe suffiks -ng

As die naamwoord op die vokale *o*, *e*, *u* of *i* uitgaan, word -ng sonder klankveranderinge aan die naamwoord gesuffigeer (Cole, 1955:342-343; Krüger, 2006a:88; Lombard *et al.*, 1985:73). Vergelyk voorbeelde (105) - (108).

(105) **kereke → kerekeng**

kereke → *kereke-ng*

CL5.SG.kerk → CL5.SG.kerk-LOC

'kerk' → 'lokaliteit van die kerk'

(106) **lebati → lebating**

le-bati → *le-bati-ng*

CL3.SG-deur → CL3.SG-deur-LOC

' deur' → 'lokaliteit van die deur'

(107) **legodimo → legodimong**

le-godimo → le-godimo-ng

CL3.SG-hemel → CL3.SG-hemel-LOC

'hemel' → 'lokaliteit van die hemel'

(108) **leru → lerung**

le-ru → le-ru-ng

CL3.SG-wolk → CL3.SG-wolk-LOC

'wolk' → 'lokaliteit van die wolk'

Ten einde die lokatiewe suffiks te verwyder, moet *-ng* deur die leë karakter vervang word: $[?*, \{o, e, u, i\}, [n, g]: []]$. As hierdie reël geïmplementeer word, verander dit dus **kerekeng** na **kereke**, **lebating** na **lebati**, **legodimong** na **legodimo**, en **lerung** na **leru**.

Wat die naamwoorde wat op *-a* uitgaan, betref, word die suffiks *-ing* aangevoeg (vergelyk voorbeeld (109)). Assimilasie vind plaas tussen die *-a* en die *-i-* van *-ing*, en dit verander na *-e-* (Cole, 1955:342; Krüger, 2006a:88; Krüger & Snyman, 1986:108).

(109) **legora → legoreng**

le-gora → le-gora-(i)ng

CL3.SG-heining → CL3.SG-heining-LOC

'heining' → 'lokaliteit van die heining'

Om die lokatiewe suffikse te verwyder moet *-eng* deur *-a* vervang word: $[?*, [e, n, g]: a]$. **Legoreng** se lemma is dan **legora**, nadat die proses voltooi is.

Die reël hierbo kan egter veroorsaak dat 'n woord soos **kerekeng** (voorbeeld (105)) na ***kereka** verander word. Die oplossing sou wees om die reëlmatige uitdrukkings $[?*, \{o, e, u, i\}, [n, g]: []]$ en $[?*, [e, n, g]: a]$ afsonderlik uit te voer en die resultaat aan 'n lemmalys te toets om te bepaal

of dit 'n geldig woord is. Hierdie proses om meervoudige resultate te hanteer, word in die volgende hoofstuk onder die naamwoordmodule (4.2.2) bespreek.

Die naamwoord wat reeds op 'n **-ng** eindig neem **-nye** of **-nyeng** in die plek van die **-ng** (Cole, 1955:343; Krüger, 2006a:88; Lombard *et al.*, 1985:74), soos dit uit voorbeeld (110) blyk.

(110) **bojang** → **bojannye/bojannyeng**

bo-jang → *bo-jang-nye/bo-jang-nyeng*

CL7.SG.gras → CL7.SG.gras-DIM/ CL7.SG.gras-DIM

'gras' → 'lokaliteit van die gras'

Daarom om die lokatiewe suffiks in hierdie geval te verwyder moet **-nnyeng** of **-nnye** deur **-ng** vervang word: $[?*, [n, n, y, e, \{ [n, g], [] \}]: [n, g]]$. Die reël het tot gevolg dat **bonjannye** verander word na **bojang**.

3.4.1.1.2 Diminutiewe suffiks **-anyana**, **-ana** en **-nyana**

Die diminutief in Setswana word gevorm deur die suffigering van 'n suffiks, hetsy **-ana**, **-nyana** of **-anyana** (Cole, 1955:105-109; Krüger, 2006a:74-87). Die suffiks **-anyana** dui verdere verkleining aan wanneer dit aan 'n diminutiewe naamwoord gesuffigeer word (Cole, 1955:109; Krüger, 2006a:74, 81-82). Hierdie suffiks moet derhalwe eerste verwyder word, voor die gewone diminutiewe suffikse **-ana** of **-nyana** afgehaal kan word (vergelyk voorbeeld (111)).

(111) **katsana** → **katsananyana**

katse-ana → *katse-ana-anyana*

CL5.SG.kat-DIM → CL5.SG.kat-DIM-DIM

'katjie' → 'klein katjie'

Ten einde die suffiks te verwyder, moet **-anyana** deur **-a** vervang word: $[?*, [a, n, y, a, n, a]: [a]]$. Dit beteken dat as die reël geïmplementeer word, **katsananyana** na **katsana** (wat nog nie 'n lemma is nie) verander. Die

woord moet daarom deur nog 'n stap gaan voordat die lemma geïdentifiseer kan word. Vervolgens kom die volgende reëls aan die beurt, waar die ander diminutiewe suffikse *-ana* en *-nyana* verwyder word.

Wat die suffikse *-ana* betref vind daar heelwat klankveranderings plaas as dit aan die naamwoord gesuffigeer word (Cole, 1955:105-109; Krüger, 2006a:74-87; Lombard *et al.*, 1985:74-82). As die slotvokaal *-a* (voorbeeld (112)) is, of as die slotvokaal *-e*, of *-i* (voorbeeld (113)) is, en nie deur *t*, *r*, *d*, *l* of *n* voorafgegaan word nie, vind daar elisie plaas (Krüger & Snyman, 1986:107-108).

(112) **tsela** → **tselana**

tsela → *tsela-ana*

CL5.SG.pad → CL.SG.pad-DIM

'pad' → 'paadjie'

(113) **kgosi** → **kgosana**

kgosi → *kgosi-ana*

CL5.SG.koning → CL5.SG.koning-DIM

'koning' → 'koninkie'

Om die diminutiewe suffikse wat aan 'n naamwoord eindigend op 'n *-a* gesuffigeer was, te verwyder, moet *-ana* dus deur *-a* vervang word: $[?*, [a, n, a]:a]$. 'n Woord soos **thipana** ('messie') sal dan verander word na **thipa** ('mes'), nadat die reëlmatige uitdrukking uitgevoer is.

Insgelyks, om 'n woord soos **metsana** ('n bietjie water') se lemma **metsi** ('water') te identifiseer, moet die diminutiewe suffikse vervang word deur *-i* as dit nie deur *t*, *r*, *d*, *l* of *n* voorafgegaan word nie: $[?*, ? - \{t, r, d, l, n\}, [a, n, a]:i]$. Dieselfde geld ook vir **motsana** ('statjie'), waar die *-ana* deur *-e* vervang moet word om die lemma, **motse** ('stat'), te identifiseer: $[?*, ? - \{t, r, d, l, n\}, [a, n, a]:i]$.

As die slotvokaal van die naamwoord **-u** of **-o** is, verander dit na die semivokaal **-w-** en die **-ana** word gesuffigeer, behalwe as die voorafgaande konsonant **b**, **p**, **ph**, **f** of **d** is (Krüger & Snyman, 1986:116). Vergelyk voorbeeld (114).

- (114) **tau** → **tawana**
tau → *tau-ana*
 CL5.SG.leeu → CL5.SG.leeu-DIM
 'leeu' → 'leeutjie'

Hier word die diminutiewe suffiks vervang deur **-i** as dit nie voorafgegaan is deur **b**, **p**, **ph**, **f** of **d** nie: [$?^*, ? - \{b, p, [p, h], f, d\}, [w, a, n, a] : u$]. As hierdie reël geïmplementeer word, verander dit dus **tawana** na **tau**. Die reëlmatige uitdrukking [$?^*, ? - \{b, p, [p, h], f, d\}, [w, a, n, a] : u$], wat voorsiening maak vir die naamwoorde wat op **-o** eindig, verander **ntlwana** ('huisie') na **ntlo** ('huis') en **powana** ('bulletjie') na **poo** ('bul').

Net soos by die lokatiewe suffiks (3.4.1.1.2) kan die reëls hierbo meervoudige resultate lewer. Die oplossing sou wees, om soos by die lokatiewe suffiks, die reëlmatige uitdrukkings afsonderlik uit te voer en die resultaat aan 'n lemmalys te toets.

Klankverandering vind ook plaas indien die konsonant van die laaste sillabe 'n **b**, **p**, **ph**, **f** of **m** is en die **-ana** gesuffigeer word (vergelyk Krüger & Snyman, 1986:129-130). So byvoorbeeld sal die **-b-** in 'n **-jw-** verander in 'n naamwoord wat op **-bo** eindig – vergelyk voorbeeld (115).

- (115) **kgabo** → **kgajwana**
kgabo → *kgabo-ana*
 CL5.SG.aap → CL5.SG.aap-DIM
 'aap' → 'apie'

Ten einde die diminutiewe suffiks te verwyder, moet **-jwana** deur **-bo** vervang word: [ʔ*, [j, w, a, n, a]: [b, o]]. **Kgajwana** sal dan verander word na **kgabo**, nadat die reëlmatige uitdrukking uitgevoer is.

Die ander klankveranderings word kortliks in voorbeelde (116) - (119) geïllustreer.

-p- → -tšw-

(116) **selepe → seletšwana**

se-lepe → se-lepe-ana

CL4.SG-byl → CL4.SG-byl-DIM

'byl' → 'byltjie'

-ph- → -tšhw-

(117) **tshephe → tshetšhwana**

tshephe → tshephe-ana

CL5.SG.springbok → CL5.SG.springbok-DIM

'springbok' → 'springbokkie'

-f- → -tšhw-

(118) **phefo → phetšhwana**

phefo → phefo-ana

CL5.SG.wind → CL5.SG.wind-DIM

'wind' → 'briesie'

-m- → -ngw-

(119) **kgomo → kgongwana**

kgomo → kgomo-ana

CL5.SG.bees → CL5.SG.bees-DIM

'bees' → 'jong bulletjie of koei'

As die konsonant in die laaste sillabe 'n **t**, **r**, **d**, **l**, **n** of **ng** is en as die slotvokaal **-i** of **-e** is, vind daar ook klankveranderings plaas (vergelyk Krüger

& Snyman, 1986:128). Die **d** kan ook klankverandering ondergaan as die slotvokaal 'n **u** is (Cole, 1955:108; Krüger, 2006a:86). So byvoorbeeld om die woord **lebatsana** na die lemma **lebati** te verander, moet **-tšana** of **-tsana** deur **-ti** vervang word: [?* , [{s, š}, a, n, a]:[i]]. Vergelyk voorbeeld (120).

-t- → -ts of -tš

- (120) **lebati → lebatsana/lebatšana**
le-bati → le-bati-ana
 CL3.SG-deur → CL3.SG-deur-DIM
 ' deur' → 'deurtjie'

Die ander klankverandering is kortliks in voorbeelde (121) - (126) gelys (Cole, 1955:108-109; Krüger, 2006a:86):

-r- → -tsh

- (121) **nare → natshana**
nare → nare-ana
 CL5.SG.buffel → CL5.SG.buffel-DIM
 'buffel' → 'buffeltjie'

-di → -ts-

- (122) **podī → potsana**
podī → podī-ana
 CL5.SG.bok → CL5.SG.bok-DIM
 'bok' → 'bokkie'

-du → -tšw-

- (123) **legodu → legotšwana**
le-godu → le-godi-ana
 CL3.SG-dief → CL3.SG-dief-DIM
 'dief' → 'diefie'

-l- → -j-

(124) **mmele → mmejana**

mmele → mmele-ana

CL2.SG.liggaam → CL2.SG.liggaam-DIM

'liggaam' → 'liggaampie'

-n- → -nyana

(125) **mosimane → mosimanyana**

mo-simane → mo-simane-ana

CL1.SG-seun → CL1.SG-seun-DIM

'seun' → 'seuntjie'

-ng > -ny

(126) **legong → legonyana**

le-gong → legong-ana

CL3.SG-stuk hout → CL3.SG-stuk hout-DIM

'stuk hout' → 'stukkies hout'

Indien 'n *i* of *e* volg op 'n konsonant anders as hierbo genoem, of 'n ander vokaal of affrikaat, dan vorm *e*, *i* + *-ana* > **-yana** (vergelyk Krüger & Snyman, 1986:115). Vergelyk voorbeeld (127).

(127) **koloi → koloyana**

koloi → koloi-ana

CL5.SG.wa → CL5.SG.wa-DIM

'wa' → 'waentjie'

Om die woord **koloyana** na die regte lemma **koloi** te herlei, moet **-yana** deur **-i** vervang word: [?*, [y, a, n, a] : [i]].

Die laaste diminutiewe suffiks wat bespreek word is **-nyana**. Daar vind geen klankverandering plaas as **-nyana** aan naamwoorde gevoeg word nie

(voorbeeld (128)). As die woord egter op 'n sillabiese **-ng** eindig (voorbeeld (129)), vind daar palatalisasie plaas (vergelyk Krüger & Snyman, 1986:128).

(128) **pene → penenyana**

pene → pene-nyana

CL5.SG.pen → CL5.SG.pen-DIM

'pen' → 'pennetjie'

(129) **bojang → bojannyana**

bo-jang → bo-jang-nyana

CL7.SG-gras → CL7.SG-gras-DIM

'gras' → 'bietjie gras'

Om die diminutiewe suffiks in voorbeeld (128) te verwyder, moet *-nyana* net deur die leë karakter vervang word: $[?*, [n, y, a, n, a]: []]$. So sal ook **bakinyana** ('baadjie') dus na **baki** verander word nadat die reëlmatige uitdrukking uitgevoer is.

Insgelyks, in hierdie geval moet die diminutiewe suffiks *-nnyana* deur **-ng** vervang word: $[?*, [n, n, y, a, n, a]: [n, g]]$. As hierdie reël geïmplementeer word, verander dit 'n woord soos **legonnyana** ('n houtjie') na **legong** ('hout'). Daar is in die geval van die suffigering van *-nyana* nie enige ander klankveranderings of uitsonderings bekend nie.

3.4.1.1.3 Augmentatiewe en feminitiewe suffikse

Wat die augmentatiewe en feminitiewe suffikse (*-gadi*) betref, word hulle sonder enige klankveranderings aan die naamwoord gesuffigeer (Cole, 1955:110; Krüger, 2006a:73; Lombard *et al.*, 1985:83). Vergelyk voorbeeld (130).

- (130) **tau** → **taugadi**
tau → *tau-gadi*
 CL5.SG.leeu → CL5.SG.leeu-FEM
 'leeu' → 'leeuwylie'

Hier word *-gadi* deur die *leë* karakter vervang: $\{[?*, [g, a, d, i] : []]\}$. Dit beteken dat 'n woord soos **podigadi** ('n groot bok') na **pod** ('bok') sal verander as die reël toegepas word. Die verwydering van dié suffiks is onproblematis omdat daar geen klankveranderings plaasvind nie.

3.4.1.1.4 Meervoudsprefikse

Die bestaande grammatikareëls wat die prefikse betref, word in die afdeling met behulp van *FSA* 6 geformaliseer. Soos reeds onder 3.4.1 en 2.4.1 genoem, moet die naamwoordlemma in die enkelvoud wees, aangesien dit die kanonieke vorm van die naamwoord is. Die klasprefikse moet dus van die meervoud na die enkelvoud herlei word. Dit beteken dat 'n woord soos **basadi** ('vroue') na **mosadi** ('vrou') verander sal word (vergelyk voorbeeld (131)).

- (131) **mosadi** → **basadi**
mo-sadi → *ba-sadi*
 CL1.SG-vrou → CL1.PL-vrou
 'vrou' → 'vrouens'

Die meervoudsprefiks *ba-* in voorbeeld (131) moet dus deur *mo-* vervang word: $[[b, a] : [m, o], ? -b, ?*]$. Woorde soos **banna** ('mans'), **badiri** ('werkers') en **baeti** ('kuiergaste') sal onderskeidelik na **monna** ('man'), **modiri** ('werker') en **moeti** ('kuiergas') verander.

Daar is egter ook, soos by die suffikse van die naamwoorde, klankveranderings om in gedagte te hou (Cole, 1955:69-75; Krüger, 2006a:57-59; Lombard *et al.*, 1985:29-50). Hierdie klankveranderings word vervolgens kortliks bespreek.

(a) Klas 1 (mo/ba)

Die variante prefiks *m-* (in plaas van *mo-*) kom voor waar die wortel (afgelei van 'n werkwoord) met 'n *b* begin. Die *b* word genasaleer na 'n *m*, nadat die *o* weggeval het (voorbeeld (132)) (Cole, 1955:71; Krüger, 2006a:58).

(132) **mmusi** → **babusi**

mo-busa-i → *ba-busa-i*

CL1.SG-regeer-TERM → CL1.PL-regeer-TERM

'regeerder' → 'regeerders'

Mmusi (**mo-bus-i*) is 'n deverbatief wat van **go busa** ('om te regeer') afgelei is. In die proses het die volgende klankveranderings plaasgevind:

affiksering : *mo* + **go busa** + *i* > *mo-bus-i*

elisie van *o* : *mo-bus-i* > *m-bus-i*

nasalering van die *b* in kombinasie met die *m*: : *m-bus-i* > **mmusi**

In die meervoud keer die *b* weer terug as die variante prefiks *m-* deur die meervoudsprefiks *ba-* vervang word (Cole, 1955:71,113-114; Krüger, 2006a:59). Vir die lemma-identifiseringsproses beteken dit dat, as daar 'n *b* na die *ba-* verskyn, dit weer 'n **mm-** moet word: $[[b, a, b] : [m, m], ?*]$. Vergelyk die verandering wat die reël in voorbeeld (133) tot gevolg het.

(133) **babui** > **mmui**

ba-bua-i > *mo-bua-i*

CL1.PL-praat-TERM > CL1.SG-praat-TERM

'sprekers' > 'spreker'

Die uitsonderings wat nie met reëls bepaal kan word nie, is volgens Cole (1955:72) en Krüger (2006a:59) **beng** ('eienaars') in voorbeeld (134) en **baba** ('vyande') in voorbeeld (135).

(134) **mong** → **beng**

mo-eng → *ba-eng*

CL1.SG-eienaar → CL1.PL-eienaar

'eienaar' → 'eienaars'

(135) **mmaba** → **baba**

mo-baba → *(ba)-baba*

CL1.SG-vyand → CL1.PL-vyand

'vyand' → 'vyande'

(b) Klas 2 (mo/me)

Hier is ook 'n variante prefiks *m-* wat voor die oorspronklike wortels staan wat met **b** begin. Dieselfde klankveranderings as by klas 1 se variante prefiks *-m* vind ook hier plaas. Die **b** word genasaleer na **m**, nadat die **o** weggeval het (Cole, 1955:115; Krüger, 2006a:60). In die meervoud keer die **b** weer terug as die *m-* deur die meervoudsprefiks *me-* vervang word (Cole, 1955:77,115; Krüger, 2006a:61). Vir die lemma-identifiseringsproses beteken dit dat as daar 'n **b** na die *me-* verskyn, dit weer 'n **mm-** moet word: $[[m, e, b] : [m, m], ?^*]$. As hierdie reël geïmplementeer word, word **mebala** ('kleure') **mmala** ('kleur') in voorbeeld (136).

(136) **mebala** > **mmala**

me-bala > *mo-bala*

CL2.PL-kleur > CL2.SG-kleur

'kleure' > 'kleur'

Ander klankveranderings in klas 2 kom voor waar die *m-* voor 'n wortel wat met 'n **h** as 'n foneemvariant van **f** begin, verskyn (Cole, 1955:77; Krüger, 2006a:61). Vergelyk voorbeeld (137), waar **mefero** ('grane/onkruid') se enkelvoud **mhero** ('graan/onkruid') is.

(137) **mhero** → **mefero**

mo-hero → *me-fero*

CL2.SG-graan/onkruid → CL2.PL-graan/onkruid

'graan/onkruid' → 'grane/onkruid'

Die reël wat dan hier gebruik word, is: $[[m, e, f] : [m, h], ?^*]$. **Mefero** sal dus na **mhero** toe verander as die reël toegepas word (vergelyk Cole, 1955:77; Krüger, 2006a:61).

(c) **Klas 3 (le/ma)**

Daar is nie 'n variante prefiks vir *le-* nie (Cole, 1955:79; Krüger, 2006a:61); die meervoudsprefiks *ma-* verander dus gewoon na *le-*: $[[m, a] : [l, e], ?^*]$.

Magapu ('pampoene/waatlemoene') verander dan na **legapu** as die reëlmatige uitdrukking uitgevoer word.

Daar is egter 'n paar uitsonderings wat die *ma-* of die variante *m-*prefiks betref. In voorbeeld (138) is **matlho** se enkelvoud is **leitho** en in voorbeeld (139) is **meno** se enkelvoud is **leino**. **Metsi** bly dieselfde aangesien dit nie in die enkelvoud voorkom nie (voorbeeld (140)) (Cole, 1955:79-84; Krüger, 2006a:61-65).

(138) **matlho** > **leitho**

ma-itho > *le-itho*

CL3.PL-oog > CL3.SG-oog

'oë' > 'oog'

(139) **meno** > **leino**

ma-ino > *le-ino*

CL3.PL-tand > CL3.SG-tand

'tande' > 'tand'

- (140) **metsi**
ma-itsi
 CL3.PL-water
 'water'

(d) Klas 4 (se/di)

Hierdie klas word as die 'skoon klas' beskou, omdat daar geen fonologiese veranderings in die klas plaasvind nie (Cole, 1955:84-86; Krüger, 2006a:65-66). Die reël is dus eenvoudig: $[[d,i]:[s,e],?^*]$. As hierdie reël geïmplementeer word, verander dit 'n woord soos **dilepe** ('byle') na **selepe** ('hout').

(e) Klas 5 (ne/di)

In klas 5 vind daar nie enige klankveranderings plaas nie. Die enkelvoudsprefiks word morfologies verteenwoordig deur *ne-*, maar die veranderings wat hierdeur veroorsaak word, is histories van aard. Dit beteken dat die klasprefiks *ne-* nie soos in die ander klasse deur die meervoudsprefiks vervang word nie. Die *di-* word vooraan die enkelvoudvorm gevoeg en daar vind geen klankverandering plaas nie. Vir die lemma-identifisering word die *di-* net verwyder: $[[d,i]:[],?^*]$. Die meervoud **dinku** ('skape') word na die toepassing van die reël na **nku** ('skaap') verander (Cole, 1955:86-90; Krüger, 2006a:66-68).

(f) Klas 6 (lo/di)

In hierdie klas kom daar heelwat klankveranderings by die meervoudsprefiks *di-* voor (Cole, 1955:91-93; Krüger, 2006a:68-69). Vergelyk voorbeelde (141) - (144). Die volgende reëls maak voorsiening vir klankverandering wat in klas 6 plaasvind: $[[d,i,t,h]:[l,o,r],?^*]$ vir voorbeeld (141), $[[d,i,p,h]:[l,o,f],?^*]$ vir voorbeeld (142), en $[[d,i,k,g]:[l,o,g],?^*]$ vir voorbeeld (143). In die laaste voorbeeld (144) is dié reëlmatige uitdrukking van toepassing: $[[d,i,t]:[l,o,l],? - \{h,s\},?^*]$. Hier mag die **t** nie deur 'n **h** of **s** gevolg word nie om die generering van verkeerde lemmas te voorkom. Andersins sou **dithole**

byvoorbeeld verkeerdelik ***lolhole** geword het (Cole, 1955:91-93; Krüger, 2006a:68-69).

(141) **dithole > lorole**

di-role > lo-role

CL6.PL-stof > CL6.SG-stof

'stof'

(142) **diphafa > lofafa**

di-fafa > lo-fafa

CL6.PL-stertveer > CL6.SG-stertveer

'stertvere' > 'stertveer'

(143) **dikgong > logong**

di-gong > lo-gong

CL6.PL-hout > CL6.SG-hout

'hout'

(144) **diteme > loleme**

di-leme > lo-leme

CL6.PL-tong > CL6.SG-tong

'tonge/tale' > 'tong/taal'

(g) Klas 7 (bo/ma)

Net soos by klas 4 is daar nie klankveranderings as die meervoudsprefiks *ma-* deur die enkelvoudsprefiks *bo-* vervang word nie. Die reëlmatige uitdrukking vir die prefiksvervanging is: $[[m, a] : [b, o], ?^*]$. 'n Woord soos **maupi** ('soorte meel') sal verander word na **boupi** ('meel') nadat die reël toegepas is (Cole, 1955:93-95; Krüger, 2006a:70-71).

(h) Klas 8

Klas 8 is die infinitiefklas en dié klas het soos die ander klasse die vermoë om kongruensie teweeg te bring (soos geïllustreer in voorbeeld (145)) (vergelyk

die bespreking onder 2.4.1). Dit stem egter ook vormlik ooreen met die werkwoord en word derhalwe onder werkwoorde (3.4.2) hanteer.

(145) **Go ema go tla re lapisa.**

Go-em-a go-tla-re-lap-is-a.

INF-staan-TERM AGR.OBJ-FUT-AGR.SUBJ-moeg-CAUS-
TERM

‘Die stanery sal ons moeg maak.’

(i) Klas 9 Lokatiefnaamwoorde

Die term lokatiefklasnaamwoorde is soms verwarrend, maar hierdie is gewone naamwoorde met ’n lokatiewe betekenis wat in klas 9 voorkom en nie ’n meervoudsprefiks het nie. Dit is ’n geslote groep woorde en kan daarom in ’n lys opgeneem word, dit wil sê as outomate geïmplementeer word. Cole (1955:97-99) en Krüger (2006a:93-98) het dertien lokatiefnaamwoorde geïdentifiseer en die volledige lys is in Bylaag 2.3 opgeneem.

3.4.1.2 Die beskrywende naamwoorde

Soos in 2.4.1 aangetoon is, is adjektiewe in Setswana deel van die naamwoord en word beskrywende naamwoorde genoem (vergelyk Vermeulen & Pretorius, 1999:88-89; Van Rooy & Pretorius, 2003). Aldaar is aangevoer dat die wortels van beskrywende naamwoorde soos dit in die woordeboek voorkom (dus sonder prefikse) as die lemmas beskou moet word. Aangesien die aantal wortels wat as beskrywende naamwoorde kan funksioneer betreklik min is, word dit daarom maklik in ’n lys hanteer (vergelyk Cole 1955:138-147; Krüger, 2006a:98; Lombard *et al.*, 1985:57-59).

Om aan te dui dat die lemma in hierdie geval ’n wortel is, word die prefiks deur die koppelteken (‘-’) vervang as dit deur een van die adjektiefwortels gevolg word. ’n Beskrywende naamwoord soos **mogolo** (‘groot’) se prefiks *mo-* sal met behulp van die reëlmatige uitdrukking $\{[m, o] : [-], [g, o, l, o]\}$ deur ’n koppelteken vervang word en die resultaat sal *-golo* wees. Die ander

beskrywende naamwoorde word soortgelyk hanteer, en die volledige lys van beskrywende naamwoorde word in Bylaag 2.2 in die notasievorm weergegee.

3.4.2 Reëls vir werkwoorde

In hierdie afdeling gaan daar, soos by die naamwoorde, van die bestaande grammatika gebruik gemaak word om reëls vir die bepaling van die werkwoordlemma te skryf. In Hoofstuk 2 (2.4.2.1) is daar aangedui dat die werkwoordlemma die stam in die infinitiewe vorm is. Die lemma is dus die stam sonder enige prefikse (d.i. negatiewe, kongruensie-, aspektuele of temporale morfeme) of suffikse (d.i. die neutro-passiewe, iteratiewe, kousatiewe, applikatiewe, resiprokale, perfektum, passiewe suffikse en die uitgangsmorfeme *-e* en *-ng*). Die onproduktiewe suffikse (d.i. die kontaktiewe, dispersiewe en posisionele suffikse) en sommige van die semi-produktiewe suffikse (d.i. die denominatiewe en reversiewe suffikse) sal net so gelaat word. Hierdie suffikse is nie meer so produktief in Setswana nie en sommige woorde met dié suffikse het nie meer ooreenstemmende eenvoudige stamme nie (vergelyk Cole, 1955:211-221; Krüger, 2006a:171-255).

Die ordening van die reëls vir lemma-identifisering volg 'n hiërargiese orde wat effens verskil van die een wat Krüger (2006a:257) in morfologiese analises gebruik ('n vergelyking tussen die aangepaste en gewone hiërargie is in Tabel 3.3). Dit is om praktiese redes dat die volgorde effe aangepas is. In morfologiese analise word die prefikse in kombinasie met bepaalde suffikse uitgehaal, maar by lemma-identifisering word die prefikse wat ortografies disjunktief is almal tegelyk verwyder. Die uitgangsmorfeem *-ng* word ook eerste van die suffikse verwyder, omdat dit dan makliker is om die ander suffikse te herken. Die uitgangsmorfeem *-e* word, om die perfektum te herken, eers nadat die perfektumuitgang *-ile* verwyder is, na *-a* verander (vergelyk Tabel 3.3). Vir praktiese doeleindes sal die stam sonder die infinitiewe prefiks *go-* gebruik word.

Die morfologie van die egte werkwoord (A = liniêre indeling, B = aangepaste hiërargiese indeling, C=Krüger (2006a) se hiërargie)

Prefikse								Wortel	Semi-produktiewe suffikse					Produktiewe suffikse					Uitgang
A	Inf	Neg	Om	Neg	Asp	Temp	Vm		Rev Onoorg	Rev Oorg	Denom	Neutro- passief	Iter	Kous	Appl	Resipr	Perf	Pass	Neg/ Modale
	go	ga	nie- lok. (klas) lok. go inf. go klas e	-se- -sa- -a-	-a- -sa- -ka-	-tla-	nie- lok. (klas) lok. -go- inf. -go- refl. -i-		-og- -olog-	-ol- -olol-	-f- -fal-	-eg- -al- -agal- -eseg- -agan-	-ak-	-is- -y-	-el- -ets-	-an-	-il-	-iw- +perf -w-	-a (mod) -e (neg) -ê (mod) -ng (mod)
B	1	1	1	1	1	1	2-10		nvt	nvt	nvt	12	11	10	9	8	6	4 (+perf) 5 (-w-)	7 (e/ê) 3 (-ng)
C	6a	1	6>6a	1	2	3	4		15	14	13	12	11	10	9	8	5	7	6 (mod-ê) 1 (neg-e)

Tabel 3.3 Die hiërargie in die lemma-identifisering van die egte werkwoord (gebaseer op Krüger, 2006a:257)

Vervolgens word die reëls vir die werkwoorde soos dit vir die lemma-identifiseerder geformaliseer is, bespreek. Die egte werkwoord (3.4.2.1) se reëls word bespreek aan die hand van die affikse wat verwyder word. Daarop volg die bespreking van die kopulatiewe (3.4.2.2) en hulpwerkwoord (3.4.2.3).

3.4.2.1 Die egte werkwoord

In hierdie afdeling word die morfologiese reëls en ook van die klankveranderinge of morfo-fonologiese veranderinge wat as gevolg van affigering plaasvind, bespreek. In die bepaling van die werkwoordlemma word dieselfde reëls wat vir woordvorming geld, gebruik, maar soos wat by die naamwoorde ook die geval was, word dit net omgekeerd gebruik. (Die volledige reëls vir die lemma-identifisering van die egte werkwoord is in Bylaag 2.12 opgeneem.)

3.4.2.1.1 Prefikse

Daar is enkele prefikse wat vas aan die werkwoordstam geskryf word, te wete (Cole, 1955:229-234; Krüger, 2006a:178-181):

- klas 1 se voorwerpsmorfeem *mo-* as dit voor 'n werkwoordstam wat met 'n **b** begin, gebruik word;
- die refleksiewe morfeem *i-*; en
- die eerstepersoon enkelvoud se voorwerpsmorfeem *n-*.

In die vorming van werkwoordstamme met die voorwerpsmorfeem *mo-* word die **b** genasaleer na **m**, nadat die **o** weggeval het (Krüger & Snyman, 1986:139). In voorbeeld (146) verander **bona** ('sien') na **mmona** ('sien hom/haar') indien die voorwerpsmorfeem *mo-* gebruik word. Ten einde die lemma te identifiseer, moet die reël omgekeer word $[[m,m]:b, [?*]]$. Indien die reël op **mmitisa** ('roep hom/haar') toegepas word, verander dit na **bitsa** ('roep') (vergelyk voorbeeld (147))

- (146) **mmona**
mo-bon-a
 AGR.OBJ-sien-TERM
 'sien hom/haar'

In die proses het die volgende klankveranderings plaasgevind:

affiksering : *mo + bon-a*
 elisie van *o* : *mo-bon-a > m-bon-a*
 nasalering van die *b* in kombinasie met die *m*: : *m-bon-a > mmona*

- (147) **mmitsa > bitsa**
mo-bits-a > bits-a
 AGR.OBJ-roep-TERM > roep-TERM
 'roep hom/haar' > 'roep'

Die voorwerpsmorfeem *n-* en die refleksief *i-* kan, waar dit nie klankveranderings meebring nie, met dié reël verwyder word: { [[*i, n*]: [], ?*] }. In voorbeelde (148) - (149) word geïllustreer wat die resultate van die reëls is.

- (148) **nthusa > thusa**
n-thus-a > thus-a
 AGR.OBJ-help-TERM > help-TERM
 'help my' > 'help'

- (149) **ithusa > thusa**
i-thus-a > thus-a
 REFL-help-TERM > help-TERM
 'self help' > 'help'

Ander klankveranderings vind by dié twee morfeme plaas op die volgende plekke (vergelyk Cole, 1955:232-233; Krüger, 2006a:178-181; Krüger & Snyman, 1986:120):

- Voor vokale

Indien die *n-* of *i-* voor 'n stam wat met 'n vokaal begin, gebruik word, word daar 'n **-k-** tussen die *n-* of *i-* en die vokaal gevoeg (vergelyk voorbeelde (150) - (151)).

(150) **nkaraba > araba**

n-arab-a > arab-a

AGR.OBJ-antwoord-TERM > antwoord-TERM

'antwoord my' > 'antwoord'

(151) **ikomosa > omosa**

i-omos-a > omos-a

REFL-warm maak-TERM > warm maak-TERM

'self warm maak' > 'warm maak'

Om die prefiks dus in hierdie geval te verwyder, moet **nk-** of **ik-** verwyder word as dit aan die begin van 'n ortografiese woord is en deur 'n klinker gevolg word: $[[\{n,i\},k]:[],\text{klinker},?^*]$. 'n Woord soos **nkitse** ('ken my') word dan verander na **itse** ('weet/ken'), nadat die reël toegepas is.

- Voor frikatiewe

Indien die *n-* of *i-* voor 'n stam wat met 'n **b-** begin, gebruik word, verander dit na **mp-** en **ip-** onderskeidelik, soos dit in voorbeelde (152) en (153) geïllustreer word.

(152) **mpona > bona**

n-bon-a > bon-a

AGR.OBJ-sien-TERM > sien-TERM

'sien my' > 'sien'

(153) **ipolela > bolela**

i-bolel-a > bolel-a

REFL-praat-TERM > praat-TERM

'praat' > 'bely'

Ten einde die prefiks te verwyder, moet **mp-** of **ip-** deur **b** vervang word as dit aan die begin van 'n ortografiese woord staan en nie deur 'n **h** gevolg word nie: $[[\{i,m\},p]:b,[? -h]]$. As hierdie reël geïmplementeer word, verander dit byvoorbeeld **mpetsa** ('my slaan') en **ipetsa** ('self slaan') na **betsa** ('slaan').

Waar die *n-* of *i-* voor 'n stam wat met 'n **f-** begin, gebruik word, verander dit na **mph-** en **iph-** onderskeidelik, soos dit in voorbeelde (154) en (155) na vore kom.

(154) **mpha > fa**

n-fa > fa

AGR.OBJ-gee-TERM > gee-TERM

'my gee' > 'gee'

(155) **iphitlha > fitlha**

i-fitlh-a > fitlha

REFL-versteek-TERM > versteek-TERM

'self versteek' > 'versteek'

Om die regte lemma te identifiseer, moet **mph-** of **iph-** deur **f** vervang word as dit aan die begin van 'n ortografiese woord staan: $[[\{i,m\},p,h]:f,[?*]]$. Hierdie reël bring mee dat **mpha**, **mphenya** ('my oorwin') en **iphitlha** onderskeidelik na **fa**, **fenya** ('oorwin') en **fitlha** verander word.

Die ander klankveranderings word kortliks met voorbeelde (156) - (163) toegelig (vergelyk Krüger, 2006a:178-181; Krüger & Snyman, 1986:121-135).

- *n-/i- + d- > nt- of it-*

(156) **ntirela > direla**

n-dir-el-a > dir-el-a

AGR.OBJ-werk-APPL-TERM > werk-APPL-TERM

'werk vir my' > 'werk vir'

- *n-/i- + g- > nkg- of ikg-*

(157) **ikgobatsa > gobatsa**

i-gobats-a > gobats-a

REFL-verwond-TERM > verwond-TERM

'self verwond' > 'verwond'

- *n-/i- + h- > nkh- of ikh-*

(158) **nkhumisa > humisa**

n-hum-is-a > hum-is-a

AGR.OBJ-ryk word-CAUS-TERM > ryk word-CAUS-TERM

'my ryk maak' > 'ryk maak'

- *n-/i- + j- > nj-/ntš- of ij/itš-*

(159) **itšela > jela**

i-j-el-a > j-el-a

REFL-eet-APPL-TERM > eet-APPL-TERM

'self eet vir' > 'eet vir'

- *n-/i- + l- > nt- of it-*

(160) **ntaela > laela**

n-lay-el-a > lay-el-a

AGR.OBJ-leer-APPL-TERM > leer-APPL-TERM

'my beveel' > 'beveel'

- *n-/i- + r- > nth- of ith-*

(161) **ithata > rata**

i-rat-a > rat-a

REFL-liefhê/hou van-TERM > liefhê/hou van-TERM

'hou van self/self liefhê' > 'liefhê/hou van'

- *n-/i- + s- > ntsh- of itsh-*

(162) **ntshega > sega**

n-seg-a > seg-a

AGR.OBJ-sny-TERM > sny-TERM

'my sny' > 'sny'

- *n-/i- + w- > nkw- of ikw-*

(163) **ikwela > wela**

i-w-el-a > w-el-a

REFL-val-APPL-TERM > val-APPL-TERM

'self val op/na' > 'val op/na'

3.4.2.1.2 Uitgangsmorfeem -ng

By die affigering van die uitgangsmorfeem *-ng* vind daar nie enige klankveranderings plaas nie (vergelyk Cole, 1955:285; Krüger, 2006a:257). Woorde met dié uitgangsmorfeem word in die relatiefkonstruksie gebruik, soos wat dit in voorbeeld (164) geïllustreer word. Die *-ng* word dus verwyder of deur die leë karakter vervang as dit aan die einde van 'n ortografiese woord staan: [?*, [n, g] : []]. In voorbeeld (165) word die *-ng* van **thusang** deur die leë karakter vervang.

(164) **Batho ba ba dirang, ba a humisa.**

Ba-tho ba ba-dir-a-ng, ba-a-hum-is-a.

CL1.PL-mens hulle (betrekkingswoord) AGR.SUBJ-werk-TERM-TERM, AGR.SUBJ-ryk word-CAUS-TERM.

Mense wat werk, word verrek.

(165) **thusang > thusa**

thus-a-ng > thus-a

help-TERM-TERM.REL > help-TERM

'wat help/' > 'help'

(**thusang** as bevel aan meer as een persoon, beteken 'help!')

3.4.2.1.3 Die passief in die perfektum

Gewoonlik word die perfektumsuffiks in morfologiese ontledings voor die passiewe suffiks uitgehaal (Krüger, 2006a:257). Vir die lemma-identifiseringsproses is dit sinvoller om eers die passiewe vorm wat in die perfektumuitgang voorkom uit te haal, want dan bly die perfektumvorm oor wat daarna deur die stel reëls vir die perfektum (sien 3.4.2.1.5) hanteer kan word.

Die passief van die perfektumvorm word gevorm deur *-w-* voor die perfektumuitgang te plaas (vergelyk Cole, 1955:228-229; Krüger, 2006a:249; Lombard *et al.*, 1985:116-117). In die meeste gevalle vind daar geen klankverandering plaas nie, soos in voorbeeld (166) die geval is, maar as die slotsillabe *-me* is, verander *-mwe* na *-ngwe* (Krüger & Snyman, 1986:137). Om dan die passiewe *-w-* uit te haal, moet *-we* deur *-e* vervang word as dit nie deur 'n *-ng* voorafgegaan word nie, as die werkwoordstam wel op *-ngwe* eindig moet dit deur *-me* vervang word: $\{[?*, [n, g, w, e] : [m, e]]\}$ en $\{[[\sim \text{ } \$[n, g]], [w, e] : e]\}$. Dié reëlmatige uitdrukking het tot gevolg dat **khubangwe** ('is gekniel/is geboorte gegee') in voorbeeld (167) na **khubame** ('het gekniel/het geboorte gegee') verander word.

(166) **badilwe > badile**

bal-il-w-e > bal-il-e

lees-PERF-PASS-TERM > lees-PERF-TERM

'is gelees' > 'het gelees'

(167) **khubangwe > khubame**

khubam-il-w-e > khubam-il-e

kniel/geboorte gee-PERF-PASS-TERM > kniel/geboorte gee -
PERF-TERM

'is gekniel/is geboorte gegee' > 'het gekniel/het geboorte gegee'

Die woorde **badile** en **khubame** is nog nie lemmas nie, maar sal in die volgende stap (3.4.2.1.5) na die regte lemma verander word.

3.4.2.1.4 Passief

In die vorming van passiewe werkwoorde vind daar verskeie klankveranderings plaas (vergelyk Cole, 1955:192-196; Krüger, 2006a:243-248; Krüger & Snyman, 1986:136-137; Lombard *et al.*, 1985:114-116).

- Werkwoordstamme wat eindig op **-ra**, **-na** en **-la** neem **-wa** in die plek van die uitgangsmorfeem **-a**. Twee- en drielettergrepige wortels neem by voorkeur **-wa** in plaas van **-iwa** (vergelyk voorbeelde (168) - (169)).

(168) **agwa > aga**

ag-w-a > ag-a

bou-PASS-TERM > bou-TERM

'word gebou' > 'bou'

(169) **romelwa > romela**

romel-w-a > romel-a

stuur-PASS-TERM > stuur-TERM

'word gestuur vir/na' > 'stuur vir/na'

- Werkwoordstamme eindigende op **s**, **ts**, **tsh** en **tsw** as die konsonant in die laaste sillabe neem **-iw-**, terwyl **-w-** ook af en toe gevind word (vergelyk voorbeeld (170)).

(170) **besiwa > besa**

bes-iw-a > bes-a

braai-PASS-TERM > braai-TERM

'word gebraai' > 'braai'

- Enkellettergrepige wortels neem weer die uitgang **-iwa**, soos in voorbeeld (171) geïllustreer.

(171) **fiwa > fa**

f-iw-a > f-a

gee-PASS-TERM > gee-TERM

'word gegee' > 'gee'

Wat die reëls vir die verwydering van dié suffikse vir lemma-identifisering betref, maak dit nie saak of die woorde twee- of drielettergrepig is nie. Daar moet net voorsiening gemaak word dat **-iwa** en **-wa** verwyder word. Die lemmas van die woorde in voorbeelde (168) - (171) kan met die volgende reëlmatige uitdrukkings bepaal word: Die **-iwa** word deur **a** vervang indien dit aan die einde van die ortografiese woord is en nie deur 'n **d**, **a** of **e** voorafgegaan word nie: $[?^*, ? - \{d, a, e\}, [i, w, a]: a]$ en **-wa** word deur 'n **a** vervang as dit nie deur **i**, **j**, **š**, **e**, **g**, **h** voorafgegaan word nie: $[?^*, ? - \{i, j, š, e, g, h, o\}, [w, a]: a]$.

- Indien **-ts-** in die laaste sillabe is, verander dit na **-d-** voordat **-iwa** die uitgangsmorfeem **-a** vervang (vergelyk voorbeeld (172)). Om dan in dié geval die passiewe suffiks te verwyder, moet **-diwa** vervang word deur **-tsa**: $[?^*, [d, i, w, a]: [t, s, a]]$.

(172) **bodiwa > botsa**

bots-iw-a > bots-a

vra-PASS-TERM > vra-TERM

'word gevra' > 'vra'

Die klankverandering wat in die laaste sillabes plaasvind, word in voorbeelde (173) - (180) geïllustreer (vergelyk Cole, 1955:192-196; Krüger, 2006a:243-248; Krüger & Snyman, 1986:136-137; Lombard *et al.*, 1985:114-116).

- **-b- + -w- > -jw-**

(173) **lejwa > leba**

leb-w-a > leb-a

kyk-PASS-TERM > kyk-TERM

'word gekyk' > 'kyk'

- **-p- + -w- > -tšw-**

(174) **botšwa > bopa**

bop-w-a > bop-a

skep-PASS-TERM > skep-TERM

'word geskep/gemaak/gevorm' > 'skep/maak/vorm'

- **-f- + -w- > -šw-**

(175) **lešwa > lefa**

lef-w-a > lef-a

betaal-PASS-TERM > betaal-TERM

'word betaal' > 'betaal'

- **-m- + -w- > -ngw-**

(176) **longwa > loma**

lom-w-a > lom-a

byt-PASS-TERM > byt-TERM

'word gebyt' > 'byt'

- **-ny- + -w- > -nngw-**

(177) **akanngwa > akanya**

akany-w-a > akany-a

dink/oorweeg-PASS-TERM > dink/oorweeg-TERM

'word gedink/oorweeg' > 'dink/oorweeg'

- **-ph- + -w- > -tshwa-/tšhwa-**

(178) **tlhotšhwa > tlhopha**

tlhop-w-a > tlhoph-a

kies-PASS-TERM > kies-TERM

'word gekies' > 'kies'

- As die werkwoord op **-aya** uitgaan, neem dit **-iwa** in plaas van **-wa** val die **-y-** weg. Soms vind assimilasië tussen **a** en **i** plaas en word dit 'n **e** (Krüger, 2006a:246).

(179) **tshwaiwa > tshwaya**

tshway-iw-a > tshway-a

merk-PASS-TERM > merk-TERM

'word gemerk' > 'merk'

(180) **tsewa > tsaya**

tsay-iw-a > tsay-a

neem-PASS-TERM > neem-TERM

'word geneem' > 'neem'

Die veranderings wat by enkellettergrepige wortels plaasvind, is so uiteenlopend dat dit as uitsonderings opgeneem is. Hieronder is enkele voorbeelde (181) - (183) van die enkellettergrepige wortels.

(181) **jewa > ja**

j-iw-a > j-a

eet-PASS-TERM > eet-TERM

'word geëet' > 'eet'

(182) **owa > wa**

w-iw-a > w-a

val-PASS-TERM > val-TERM

'word geval' > 'val'

(183) **fiwa > fa**

f-iw-a > f-a

gee-PASS-TERM > gee-TERM

'word gegee' > 'gee'

3.4.2.1.5 Perfektum

Wat die perfektum betref, is daar heelwat klankveranderings wat plaasvind, maar die algemene reël is dat die uitgangsmorfeem *-a* deur **-ile** vervang word (vergelyk Cole, 1955:224-229; Krüger, 2006a:249-255; Krüger & Snyman, 1986:131; Lombard *et al.*, 1985:116-120). Om dié suffiks dus te verwyder, moet **-ile** deur *-a* vervang word as dit aan die einde van die ortografiese woord voorkom: [?* , [i,l,e]:a], soos wat dit in voorbeeld (184) geïllustreer word.

(184) **robile > roba**

rob-il-e > rob-a

breek-PERF-TERM > breek-TERM

'gebreek' > 'breek'

Werkwoordstamme wat op **-sa**, **-tswa**, **-tsha** en **-tsa** eindig, neem **-itse** in plaas van die *-a*, soos in voorbeeld (185) - (186)) die geval is (Cole, 1955:226; Krüger, 2006a:249-250). Ten einde die lemma te kan identifiseer, moet **-itse** vervang word deur *-a* as dit aan die einde van die ortografiese

woord voorkom en nie deur 'n **d** voorafgegaan is nie: [?* , ? - d, [i, t, s, e]: a]. Daar is egter uitsonderings by sekere werkwoordstamme wat op **-tsa** eindig (Cole, 1955:226; Krüger, 2006a:249). In plaas daarvan dat die -a 'n **-itse** word, word **-tsa** deur **-ditse** vervang (vergelyk voorbeeld (187)). In hierdie geval word die perfektum verwyder deur **-ditse** deur **-tsa** te vervang as dit aan die einde van die ortografiese woord voorkom: [?* , [d, i, t, s, e]: [t, s, a]].

(185) **fisitse > fisa**

fis-il-e > fis-a

brand-PERF-TERM > brand-TERM

'gebrand' > 'brand'

(186) **tlhatswitse > tlhatswa**

tlhatsw-il-e > tlhatsw-a

was-PERF-TERM > was-TERM

'het gewas' > 'was'

(187) **biditse > bitsa**

bits-il-e > bits-a

roep-PERF-TERM > roep-TERM

'het geroep' > 'roep'

Werkwoordstamme wat op **-nya** eindig se **-ya** word deur **-tse** vervang, soos wat dit in voorbeeld (188) die geval is (Cole, 1955:226; Krüger, 2006a:250). Die perfektum kan in hierdie geval verwyder word deur **-ntse** deur **-nya** te vervang as dit aan die einde van die ortografiese woord is: [?* , [n, t, s, e]: [n, y, a]].

(188) **sentse > senya**

seny-il-e > seny-a

verniel-PERF-TERM > verniel-TERM

'het verniel' > 'verniel'

Die ander klankveranderings wat plaasvind as die suffiks *-il-* aan die werkwoordstam gevoeg word, word kortliks aan die hand van voorbeelde bespreek (vergelyk Cole, 1955:225-229; Krüger, 2006a:250-255).

- Tweelettergrepige werkwoordstamme wat op **-na** eindig, kry **-ne** in die plek van die uitgangsmorfeem *-a*, en meerlettergrepige werkwoordstamme kry **-e** of **-ê** in die plek van die *-a*, vergelyk voorbeelde (189) - (190) (Cole, 1955:227; Krüger, 2006a:250).

(189) **binne > bina**

bin-il-e > bin-a

dans-PERF-TERM > dans-TERM

'het gedans' > 'dans'

(190) **ratane > ratana**

rat-an-il-e > rat-an-a

liefhê-RECP-PERF-TERM > liefhê-RECP-TERM

'het mekaar liefgehad' > 'mekaar liefhê'

- Twee- en meerlettergrepige werkwoordstamme wat op **-ma** uitgaan en toestand aandui, word deur **-me** vervang (Cole, 1955:227; Krüger, 2006a:251). Vergelyk voorbeeld (191).

(191) **khubame > khubama**

khubam-il-e > khubam-a

kniel/geboorte gee-PERF-TERM > kniel/geboorte gee-TERM

'het gekniel/geboorte gegee' > 'kniel/geboorte gee'

- Werkwoordstamme wat op **-ara** eindig, kry **-ere** in die plek van **-ara**, mits dit 'n toestand aandui (Cole, 1955:227; Krüger, 2006a:251). Vergelyk voorbeeld (192).

(192) **apere > apara**

apar-il-e > apar-a

aantrek-PERF-TERM > aantrek-TERM

'het aangetrek' > 'aantrek'

- Werkwoordstamme wat op **-aya** eindig, neem **-ile** in die plek van **-aya** (Cole, 1955:225; Krüger, 2006a:252). Vergelyk voorbeeld (193).

(193) **tsamaile > tsamaya**

tsamay-il-e > tsamay-a

loop-PERF-TERM > loop-TERM

'het geloop' > 'loop'

- Twee- en meerlettergrepige stamme eindigende op **-ala** kry **-etse** in die plek van die **-ala**, soos wat dit in voorbeelde (194) - (195) geïllustreer word (Cole, 1955:225; Krüger, 2006a:253).

(194) **tletse > tlala**

tlal-il-e > tlal-a

vol word-PERF-TERM > vol word-TERM

'is vol' > 'word vol'

(195) **raletse > ralala**

ralal-il-e > ralal-a

deurkruis-PERF-TERM > deurkruis -TERM

'het deurkruis' > 'deurkruis'

- By alle meerlettergrepige werkwoordstamme waar die vokaal voor **-la** nie **-a-** is, nie en dit dui voltooidheid aan, verander **-la** na **-tse** (vergelyk Cole, 1955:225; Krüger, 2006a:253). Vergelyk voorbeeld (196).

(196) **tlametse > tlamela**

tlamel-il-e > tlamel-a

voorsien-PERF-TERM > voorsien-TERM

'het voorsien/versorg' > 'voorsien/versorg'

- Tweelettergrepige werkwoordstamme wat voltooidheid aandui en op **-la** eindig, neem **-dile** in die plek van **-la** (Cole, 1955:225; Krüger, 2006a:252-253). Vergelyk voorbeeld (197).

(197) **bedile > bela**

bel-il-e > bel-a

kook-PERF-TERM > kook-TERM

'het gekook' > 'kook'

Die perfektum van eenlettergrepige werkwoorde word op verskeie maniere gevorm. Daar is dié waar die uitgangsmorfeem **-a** normaal deur **-ile** vervang word (voorbeeld (198)), en dan die wortels waar daar klankveranderings plaasvind (voorbeelde (199) - (201)) (Cole, 1955:227-228; Krüger, 2006a:255). Die veranderings wat by enkellettergrepige wortels plaasvind, is so uiteenlopend dat dit as uitsonderings opgeneem is (die volledige lys woorde soos deur Cole (1955) en Krüger (2006a) bespreek, is in die bylae onder 2.12 opgeneem).

(198) **fa → file**

fa → f-il-e

gee-TERM → gee-PERF-TERM

'gee' → 'het gegee'

(199) **swa → sule/swile**

sw-a → sw-il-e

sterf-TERM → sterf-PERF-TERM

'sterf' → 'het gesterf'

(200) **lwa** → **lwele/lole**
lw-a → *lw-il-e*
veg-TERM → veg-PERF-TERM
'veg' → 'het geveg'

(201) **ja** → **jele**
ja → *j-il-e*
eet-TERM → eet-PERF-TERM
'eet' → 'het geëet'

Ander uitsonderings soos **budule** (voorbeeld (202)) en **amule** (voorbeeld (203)) is saam met die eenlettergrepige werkwoorde as uitsonderings op die lys opgeneem.

(202) **budule** > **butswa**
butsw-il-e > *butsw-a*
ryp/gaar word-PERF-TERM > ryp/gaar word-TERM
'is ryp/gaar' > 'ryp/gaar word'

(203) **amule** > **anya**
any-il-e > *any-a*
suip-PERF-TERM > suip-TERM
'het gesuip (kalf)' > 'suip (kalf)'

3.4.2.1.6 Uitgangsmorfeem -e

Werkwoorde se uitgangsmorfeem -a verander na -e in die ontkenning van die teenwoordige tyd (voorbeeld (204)) en in verskeie modi, soos die subjunktiewe modus (voorbeeld (205)) (vergelyk Cole, 1955:223-224; Krüger, 2006a:168,257). Nadat die perfektum (in 3.4.2.1.5) uitgehaal is, kan alle werkwoordstamme wat op -e uitgaan deur -a vervang word: [?*, e : a] (vergelyk voorbeeld (206)).

- (204) **ga ke kwale**
ga-ke-kwal-e
 NEG-AGR.SUBJ-skryf-TERM
 'ek skryf nie'
- (205) **...gore ke tsene.**
...gore ke-tsen-e.
 ...sodat AGR.SUBJ-ingaan-TERM
 '... sodat ek (kan) ingaan'
- (206) **dule > dula**
dul-e > dul-a
 sit/bly-TERM > sit/bly-TERM
 'sit/bly' > 'sit/bly'

3.4.2.1.7 Resiprokalis

Wat die vorming van resiprokale werkwoorde betref, word die uitgang **-ana** in die plek van die uitgangsmorfeem **-a** geplaas sonder dat daar enige klankveranderings plaasvind (Cole, 1955:209-211; Krüger, 2006a:239-243). Ten einde die lemma te identifiseer, moet die suffiks verwyder word deur **-ana** deur **-a** te vervang indien dit nie deur 'n **a** voorafgegaan word nie: [$?^*, ? - a, [a, n, a] : a$] (vergelyk voorbeeld (207)).

- (207) **thusana > thusa**
thus-an-a > thus-a
 help-RECP-TERM > help-TERM
 'help mekaar' > 'help'

Voorbeeld (208) is 'n uitsondering waar **-aana** deur **-aya** vervang word as dit aan die einde van die ortografiese woord voorkom: [$?^*, [a, a, n, a] : [a, y, a]$] (Cole, 1955:210; Lombard *et al.*, 1985:114).

- (208) **bolaana > bolaya**
bolay-an-a > bolay-a
 doodmaak-RECP-TERM > dood-TERM
 'maak mekaar dood' > 'doodmaak'

3.4.2.1.8 Applikatief

Vir die vorming van die applikatief is die algemene reël dat die uitgang **-ela** in die plek van **-a** aangevoeg word (Cole, 1955:199; Krüger, 2006a:236). Om die applikatief dus te verwyder, moet **-ela** deur **-a** vervang word: $[?*, [e, l, a] : a]$. Vergelyk voorbeeld (209).

- (209) **romela > roma**
rom-el-a > rom-a
 stuur-APPL-TERM > stuur-TERM
 'stuur vir/na' > 'stuur'

Daar vind ook by die suffigering van **-ela** klankveranderings plaas. Werkwoordstamme wat op **-sa**, **-tsha**, **-tswa**, **-nya** eindig se uitgangsmorfeem **-a** word deur **-etsa** vervang (Cole, 1955:200; Krüger, 2006a:236-237; Krüger & Snyman, 1986:132). Ten einde die applikatiewe suffiks te verwyder, moet **-etsa** deur **-a** vervang word indien dit nie deur **b**, **i**, **l** of **f** voorafgegaan word nie: $[?*, ? - \{b, i, l, f\}, [e, t, s, a] : a]$ (vergelyk voorbeelde (210)-(212)).

- (210) **disetsa > disa**
dis-el-a > dis-a
 oppas-APPL-TERM > oppas-TERM
 'oppas vir' > 'oppas'

- (211) **ntshetsa > ntsha**
ntsh-el-a > ntsh-a
 uithaal-APPL-TERM > uithaal-TERM
 'uithaal vir' > 'uithaal'

(212) **rulaganyetsa > rulaganya**

rulagan-y-el-a > rulagan-y-a

georden raak-CAUS-APPL-TERM > georden raak-CAUS-TERM

'rangskik vir' > 'rangskik'

Werkwoordstamme wat op **-tsa** eindig se **-ts-** verander na **-lets-** (Cole, 1955:200; Krüger, 2006a:237; Krüger & Snyman, 1986:132; Lombard *et al.*, 1985:109-110). Die applikatiewe suffiks word dan verwyder tydens lemma-identifisering deur **-letsa** deur **-tsa** te vervang: $[?+, [l, e, t, s, a] : [t, s, a]]$, (vergelyk voorbeeld (213))

(213) **biletsa > bitsa**

bits-el-a > bits-a

roep-APPL-TERM > roep-TERM

'uitnoot' > 'roep'

In die vorming van die applikatief word **-ya** van werkwoordstamme wat eindig op

-aya na **-ela** verander (vergelyk voorbeeld (214)) (Cole, 1955:200). Om dan die applikatiewe suffiks te verwyder, moet **-aela** of **-eela** vervang word deur **-ya**: $[?*, \{ \{ [a, e], e, l, a \} : [a, y, a] \}]$. Daar moet vir **-eela** voorsiening gemaak word, omdat daar soms klankverandering plaasvind en die **a** 'n **e** word (vergelyk voorbeeld (215)) (Cole, 1955:200).

(214) **bolaela > bolaya**

bolay-el-a > bolay-a

doodmaak-APPL-TERM > doodmaak-TERM

'doodmaak vir' > 'doodmaak'

- (215) **neela** > **naya**
nay-el-a > *nay-a*
gee-APPL-TERM > gee-TERM
'gee vir' > 'gee'

Daar is egter gevalle van verstarrings, waar die 'eenvoudiger' vorms, ***rapa** van **rapela** (voorbeeld (216)), ***bola** van **bolela** (voorbeeld (217)), en **duma** ('dreun') van **dumela** (voorbeeld (218)) nie bestaan of meer gebruik word nie. (Die verstarde applikatiewe werkwoorde van Krüger (2006a:238-239) is as lemmas in die lemmalys opgeneem.)

- (216) **rapela**
rapel-a
bid-TERM
'bid'

- (217) **bolela**
bolel-a
vertel-TERM
'vertel'

- (218) **dumel-a**
dumel-a
groet-TERM
'groet'

3.4.2.1.9 Kousatief

Wat die kousatief betref, is daar heelwat klankveranderings wat plaasvind, maar die algemene reël is dat die uitgangsmorfeem -a deur die kousatiewe uitgang **-isa** vervang word (Cole, 1955:203-209; Krüger, 2006a:217-232; Krüger & Snyman, 1986:124,130; Lombard *et al.*, 1985:111-113). Om dié suffiks dus te verwyder, moet **-isa** deur -a vervang word as dit aan die einde

van die ortografiese woord voorkom en nie deur 'n klinker of **d** voorafgegaan word nie: $[?^*, ? -\{d, \text{klinker}\}, [i, s, a]: a]$ (vergelyk voorbeeld (219))

(219) **lapisa > lapa**

lap-is-a > lap-a

moeg raak-CAUS-TERM > moeg raak-TERM

'moeg maak' > 'moeg raak'

Indien die laaste lettergreep **-la** is, verander dit na **-disa**, of **-tsa** (Cole, 1955:204; Krüger, 2006a:219-220,226). Ten einde die kousatiewe suffiks te verwyder, moet **-disa** of **-tsa** na **-la** verander word: $[?^*, \{[d, i, s, a], [t, s, a]\}: la]$, (vergelyk voorbeelde (220) - (221)).

(220) **badisa > bala**

bal-is-a > bal-a

lees-CAUS-TERM > lees-TERM

'laat lees' > 'lees'

(221) **tlatsa > tlala**

tlal-y-a > tlal-a

vol word-CAUS-TERM > vol word-TERM

'vul, volmaak' > 'vol word'

Werkwoordstamme wat op **-ga** eindig se uitgang **-ga** word deur **-sa** vervang (Cole, 1955:204; Krüger, 2006a:227). Ten einde die suffikse te verwyder, moet **-sa** deur **-ga** vervang word indien dit nie deur **i, t, a, u** of **e** voorafgegaan word nie: $[?^*, ? -\{i, t, a, u, e\}, [s, a]: [g, a]]$ (vergelyk voorbeeld (222)):

(222) **lemosa > lemoga**

lemog-y-a > lemog-a

verstaan-CAUS-TERM > verstaan-TERM

'laat verstaan' > 'verstaan, begryp'

Ander klankverandering wat plaasvind as *-is-* gesuffigeer word, word kortliks aan die hand van voorbeelde bespreek (vergelyk Cole, 1955:204-206; Krüger, 2006a:221-223,228,230).

- Twee en meerlettergrepige werkwoordstamme wat op *-na* eindig, neem die kousatiewe suffiks *-y-* of *-is-* (vergelyk voorbeelde (223) - (224)). In die geval van *-is-* (voorbeeld (224)) vind daar klankverandering plaas as dit gesuffigeer word (Cole, 1955:205; Krüger, 2006a:223,228).

(223) **tsenya > tsena**

tsen-y-a > tsen-a

ingaan-CAUS-TERM > ingaan-TERM

'insit' > 'ingaan'

(224) **gantsha > gana**

gan-is-a > gan-a

weier-CAUS-TERM > weier-TERM

'laat weier' > 'weier'

- Die uitgang *-aya* verander gewoonlik na *-isa*, maar daar is ook uitsonderings, soos dit in voorbeeld (225) - (226) getoon word (Cole, 1955:204; Krüger, 2006a:222).

(225) **apeisa > apaya**

apay-is-a > apay-a

kook-CAUS-TERM > kook-TERM

'laat kook' > 'kook'

(226) **beisa > baya**

bay-is-a > bay-a

neersit-CAUS-TERM > neersit-TERM

'laat neersit' > 'neersit'

- Eenlettergrepige werkwoordstamme neem **-isa** in die plek van die uitgangsmorfeem **-a** (Cole, 1955:206; Krüger, 2006a:230). Die **a** en **i** assimileer tot **e** in 'n paar gevalle (vergelyk voorbeeld (227)).

(227) **jesa > ja**

j-is-a > ja

eet-CAUS-TERM > eet.TERM

'laat eet/voer' > 'eet'

Daar is, soos by die ander suffikse, ook uitsonderings (vergelyk Cole, 1955:206-207; Krüger, 2006a:231-232) wat as reëlmatige uitdrukking opgeneem is. Hier is 'n paar voorbeelde van die uitsonderings:

{ [b,u,s,a]:[b,o,a], [a,p,e,s,a]:[a,p,a,r,a], [u,s,a]:[w,a],
[l,o,s,a]:[l,w,a] } (vergelyk voorbeelde (228) - (231)).

(228) **busa > boa**

bo-is-a > bo-a

terugkeer-CAUS-TERM > terugkeer-TERM

'terugbring' > 'terugkeer'

(229) **apesa > apara**

apar-is-a > apara

aantrek-CAUS-TERM > aantrek-TERM

'laat aantrek' > 'aantrek'

(230) **osa > wa**

w-is-a > w-a

val-CAUS-TERM > val-TERM

'laat val' > 'val'

(231) **losa > lwa**

lw-is-a > lw-a

veg-CAUS-TERM > veg-TERM

'laat veg' > 'veg'

3.4.2.1.10 Iteratief

Die iteratiewe werkwoord word gevorm deur die uitgangsmorfeem *-a* deur die uitgang **-aka** te vervang (Cole, 1955:209; Krüger, 2006a:215; Lombard *et al.*, 1985:125). Om die iteratiewe suffiks dus weer uit te haal, moet **-aka** deur *-a* vervang word: [$?^*$, [a, k, a]:a]. As die reël toegepas word, word **gataka** na **gata** verander (voorbeeld (232)):

(232) **gataka > gata**

gat-ak-a > gat-a

trap-ITER-TERM > trap-TERM

'vertrap' > 'trap'

Daar is geen klankveranderings of uitsonderings as die iteratiewe suffiks *-ak-* gesuffigeer word (Cole, 1955:209; Krüger, 2006a:215; Lombard *et al.*, 1985:125).

3.4.2.1.11 Neutro-passief

Die neutro-passiewe uitgange **-ega** en **-esega** word in die plek van die uitgangsmorfeem *-a* gebruik (Cole, 1955:196-199; Krüger, 2006a:208-211; Lombard *et al.*, 1985:123). Daar is dus twee reëls waarmee die suffikse verwyder kan word: **-esega** word deur **-a** vervang: [$?^*$, [e, s, e, g, a]:a]; en **-ega** word deur *-a* vervang indien dit nie deur **-es-** voorafgegaan word nie: [$?^*$, ? -e, ? -s, [e, g, a]:a] (vergelyk voorbeeld (233) - (234)).

(233) **alafesega > alafa**

alaf-eseg-a > alaf-a

genees/gesond raak-N.PASS-TERM > genees/gesond raak-TERM

'geneesbaar' > 'genees/gesond raak'

(234) **lemega > lema**

lem-eg-a > lem-a

ploeg-N.PASS-TERM > ploeg-TERM

'ploegbaar' > 'ploeg'

'n Uitsondering wat nie in die twee vorige reëls vervat word nie, is **itsege** (voorbeeld (235)) en kan só as 'n reëlmatige uitdrukking hanteer word: [i,t,s,e,g,e]:[i,t,s,e].

(235) **itsege > itse**

its-eg-e > its-e

weet-N.PASS-TERM > weet-TERM

'bekend raak/wees' > 'ken, weet'

3.4.2.2 Die hulpwerkwoord

Die egte hulpwerkwoorde is 'n geslote groep woorde en word op grond van hulle betekeniswaardes geklassifiseer (vergelyk Krüger, 2006a:261-275; Pretorius, 1997).

- Voorbeelde van hulpwerkwoorde met 'n negatiewe betekenis, is **ise** ('nog nie'), **(se) kitla** ('sal nie'), **ketla** ('sal nie') en **nke** ('sal nie/gaan nie'). Hoe dit in sinsverband voorkom, word in voorbeeld (236) geïllustreer.

(236) **Ga re kitla re ja dijo.**

Ga-re-kitla re-j-a di-jo.

NEG-AGR.SUBJ-sal nie AGR.SUBJ-eet-TERM CL4.PL-
kos.

‘Ons sal nie kos eet nie’

- Enkele voorbeelde van hulpwerkwoorde wat logiese tyd aandui, is **bolo** (‘lank gelede/redelik lank terug/lank nadat/nie onlangs nie’), **sena** (‘nadat’), **jafile** (‘op die regte oomblik’), **kile** (‘eens/eenmaal lank gelede’) en **setse** (‘reeds’) (vergelyk voorbeeld (237)).

(237) **Monna o jafile a tla go nthusa.**

Mo-nna o-jafile a-tl-a go-n-thus-a.

CL1.SG-man AGR.SUBJ-op die regte oomblik
AGR.SUBJ-kom-TERM INF-AGR.OBJ-help-TERM.

‘Die man het op die regte tyd gekom om my te help.’

- Van die hulpwerkwoorde wat konsekwatheid aandui, is **ba** (‘en/en toe/en ook/verder’), **bo** (‘en/en toe/en ook/verder’), **be** (‘en/en toe/en ook/verder’), **kabo** (‘en/en toe/en ook/verder’) en **bile** (‘en/en toe/en ook/verder’) (vergelyk voorbeeld (238)).

(238) **Ngaka ya nkalafa, ke be ke fola.**

Ngaka e-a-n-alaf-a, ke-be ke-fol-a.

CL5.SG.dokter AGR.SUBJ-TEMP-AGR.OBJ-
genees/gesond raak-TERM, AGR.SUBJ-en/en toe/en
ook/verder AGR-OBJ-herstel-TERM.

‘Die dokter het my behandel en toe herstel ek.’

Daar is relatief min egte hulpwerkwoorde en hulle word net soos wat hulle voorkom as lemmas gebruik. Die woorde sal soos wat hulle deur Krüger (2006a:261-275) en Pretorius (1997) opgeteken is, in Bylaag 2.4 as reëlmatige uitdrukkings opgeneem word.

3.4.2.3 Die kopulatiewe werkwoord

Daar is drie soorte kopulatiewe, naamlik die identifiserende, definiërende en assosiatiewe kopulatiewe werkwoorde. Die kopulatiewe werkwoord word modaal verbuig en kom in verskillende tye, aspekte, positief en negatief voor, maar kopulatiewe werkwoorde se lemma is die basiese infinitiewe kopulatiewe werkwoord **go nna** ('om te wees') (vergelyk Cole, 1955:336-339; Krüger, 2006a:275-278). Hieronder volg voorbeelde van enkele vorms van die identifiserende, definiërende en assosiatiewe werkwoorde wat met behulp van reëlmatige uitdrukkings na **nna** (sonder *go*, in navolging van die egte werkwoord) herlei word.

- Die identifiserende kopulatiewe werkwoord in die indikatiefmodus se teenwoordige tyd positief (**ke** in voorbeeld (239)) en negatief (**ga se** in voorbeeld (240)) verander na **nna** nadat die reëlmatige uitdrukking, $\{ \{ [k, e], [[g, a], [' '], [s, e]] \} : [n, n, a] \}$, toegepas is. Die $[' ']$ maak voorsiening vir kopulatiewe werkwoord wat disjunktief geskryf word.

(239) **Mosadi ke morutabana. > nna**

Mo-sadi ke mo-ruta-ba-ngwana. > nn-a

CL1SG-vrou dit is CL1SG-onderrig-CL1PL-kind. > wees-
TERM

'Die vrou is 'n onderwyser.' > 'wees'

(240) **Monna ga se moruti. > nna**

Mo-nna ga-se mo-rut-i. > nn-a

CL1SG-man NEG-is CL1SG-onderrig-TERM. > wees-
TERM

'Die man is nie 'n predikant nie.' > 'wees'

- Die definiërende kopulatiewe werkwoord in die indiktatiefmodus teenwoordige tyd positief (**o** in voorbeeld (241)) en in die negatief (**ga a** in voorbeeld (242)) verander na **nna** nadat die reëlmatige uitdrukking,

{[[[g,a],['
 ']]^,{[b^,o],[b^,a],[{l,s,r,k,[]},e],[d,i],[{g,l},o]
 }]:[n,n,a]}, geïmplementeer is.

(241) **Molemi o botlhale. > nna**

Mo-lem-i o bo-tlhale.

CL1.SG-ploeg-TERM is CL7.SG-kwaaiheid.

'Die saaiboer is slim.'

(242) **Monna ga a bogale. > nna**

Mo-nna ga-a bo-gale. > nn-a

CL1.SG-man NEG-is CL7.SG-kwaaiheid. > wees-TERM

'Die man is nie kwaai nie.' > 'wees'

- Die assosiatiewe kopulatiewe werkwoord in die indiktatiefmodus se teenwoordige en toekomstige tyd verander na **nna** as hierdie reël, $\{[[[b^,o],[b^,a],[{l,s,r,k,[]},e],[d,i],[{g,l},o]], [[[' '],[t,l,a]]^,[' '],[n,a]]:[n,n,a]\}$, geïmplementeer word. Die kopulatiewe werkwoord **e na** in voorbeeld (243) en **o tla na** in voorbeeld (244) verander na **nna**. Hierdie deel van die reëlmatige uitdrukking,

$[b^,o],[b^,a],[{l,s,r,k,[]},e],[d,i],[{g,l},o]$, is die die onderwersmorfeme.

(243) **Ntšwa e na le dikgofa. > nna**

Ntšwa e-na le di-kgofa. > nn-a

CL5.SG.hond AGR.SUBJ-is saam met CL5.PL-bosluis. > wees-TERM

'Die hond het bosluise.' > 'wees'

(244) **Monna o tla nna le koloi. > nna**

Mo-nna o-tla-nn-a le koloi >nn-a

CL1.SG-man AGR.SUBJ-FUT-is-TERM saam met
CL5.SG.motor. > wees-TERM

'Die man sal 'n motor hê.' > 'wees'

(Die volledige reëls vir die lemma-identifisering van die kopulatiewe werkwoord is in Bylaag 2.5 opgeneem.)

3.4.3 Ander woordsoorte

Soos reeds in Hoofstuk 2 onder 2.4.3 - 2.4.7 bespreek, is die vyf woordsoorte, naamlik voornaamwoord, betrekkingwoord, bywoord, interjeksie en ideofoon, die meer geslote woordsoorte (vergelyk Cole, 1955:127-170,370-400; Krüger, 2006a:127-166; Lombard *et al.*, 1985:166-184; Ntsime & Krüger, 1995:177-182; Ras, 1991; Zerwick, 1997:76-88). Dié woordsoorte kan op een van twee maniere in 'n lemma-identifiseerder geïmplementeer word: deur al die woorde in 'n teksleër op te neem, of om dié woorde as outomate te implementeer. Die voordeel van laasgenoemde metode is dat dit minder stoorplek vereis (Jurafsky & Martin, 2000:71; Kaplan, 1996; Ritchie *et al.*, 1987:253). Die vraende, eksklusiewe en seperatiewe voornaamwoord kan byvoorbeeld soos 'n enkele outomaton geïmplementeer word: $\{[[[o, [b, \{a, o\}], e, [l, e], a, [s, e], [d, i], [l, o], [g, o]]], \{[\{p, f\}, e], [s, e, l, e]]\}]$. Hierdie outomaton verteenwoordig die vraende (**ofe, bafe, efe, lefe, afe, sefe, dife, lofe, bofe** en **gofe** - 'watter'), eksklusiewe (**ope, bape, epe, lepe, ape, sepe, dipe, lope, bope** en **gope** - 'geeneen') en seperatiewe voornaamwoorde (**osele, basele, esele, lesele, asele, sesele, disele, losele, bosele** en **gosele** - 'ander/vreemde').

(Die outomate van die geslote woordsoorte is in Bylaag 2.6 - 2.10 opgeneem.)

3.5 Gevolgtrekking

In die ontwikkeling van 'n lemma-identifiseerder handel dit nie net oor die bepaling van wat 'n lemma is nie, maar ook oor hoe die lemma geïdentifiseer moet word. Die doel van die hoofstuk was daarom om die bestaande

linguistiese reëls in Setswana te formaliseer ten einde die lemma outomaties te kan identifiseer. Daar is reeds in Hoofstuk 1 besluit om die linguistiese benadering in die ontwikkeling van die lemma-identifiseerder te volg omdat die morfologie van die naamwoorde en werkwoorde in Setswana redelik reëlmatig is en die ander woordsoorte morfologies onproduktief is.

In die linguistiese benadering is daar onder andere van eindigestaatsageware en reëlmatige uitdrukkings wat via die eindigestaatsoutomaat en die eindigestaatoorvormer geïmplementeer word (Jurafsky & Martin, 2000:21-22). Eindigestaatoorvormers 'n integrale komponent in die ontwikkeling van 'n lemma-identifiseerder. Die eindigestaatsageware *FSA 6* kan hierdie eindigestaatoorvormers hanteer en is daarom gebruik om die bestaande linguistiese reëls in Setswana mee te formaliseer. Die reëlmatige uitdrukkingsbewerkingstekens is daarom vervolgens soos wat dit in *FSA 6* gebruik word, bespreek.

Bestaande grammatikas van Cole (1955) en Krüger (2006a) is daarna met behulp van reëlmatige uitdrukkings in die notasievorm van *FSA 6* geformaliseer. Die morfologiese reëls en morfo-fonologiese veranderings is bespreek soos wat dit in die vorming van die naamwoord en werkwoord voorkom, en aangepas om die lemmas te identifiseer. Die ander woordsoorte is as outomate geskryf om in die lemma-identifiseerder geïmplementeer te kan word.

In Hoofstuk 1 is genoem dat die ontwikkeling van 'n reëlgebaseerde lemma-identifiseerder 'n geleentheid bied om te bepaal hoe goed die bestaande grammatikas sal vaar indien dit geformaliseer sou word. In Hoofstuk 3 is die bestaande grammatikas wel met behulp van reëlmatige uitdrukkings in *FSA 6* se notasievorm geformaliseer, maar in Hoofstuk 4 sal al die *FSA*'s en *FST*'s bymekaarkom in 'n naamwoordmodule, 'n egtewerkwoordmodule en 'n lemma-identifiseerder. Die reëls wat in Hoofstuk 3 geformaliseer is, sal dan geëvalueer word, en daarmee sal bepaal word of bestaande grammatikas effektief geformaliseer kan word om die lemma in Setswana te identifiseer.

Hoofstuk 4

'n Lemma-identifiseerder vir Setswana: Implementering en Evaluering

4.1 Inleiding

In Hoofstuk 3 is die reëls van die Setswanagrammatikas geformaliseer. Hierdie reëls word in dié hoofstuk geïmplementeer en geëvalueer. Eerstens word die algemene argitektuur van die lemma-identifiseerder bespreek. Die lemma-identifiseerder bestaan onder andere uit die egtewerkwoordmodule en naamwoordmodule wat afsonderlik bespreek word. Die lemma-identifiseerder word in die laaste afdeling van die hoofstuk geëvalueer. In dié afdeling word die meetinstrumente onder die loep geneem en die resultate aan die hand van voorbeelde bespreek.

4.2 Implementering

4.2.1 Argitektuur van lemma-identifiseerder

Die lemma-identifiseerder bestaan uit gewone outomate en eindigestaatoorvormers. Die outomaat lewer een van twee “antwoorde” op: “ja” (waar) of “nee” (vals). As die antwoord “ja” is (dus, as die woord deur die outomaat herken word), is dit 'n lemma. Die eindigestaatoorvormers verander/verwyder die affikse ten einde die lemma te identifiseer.

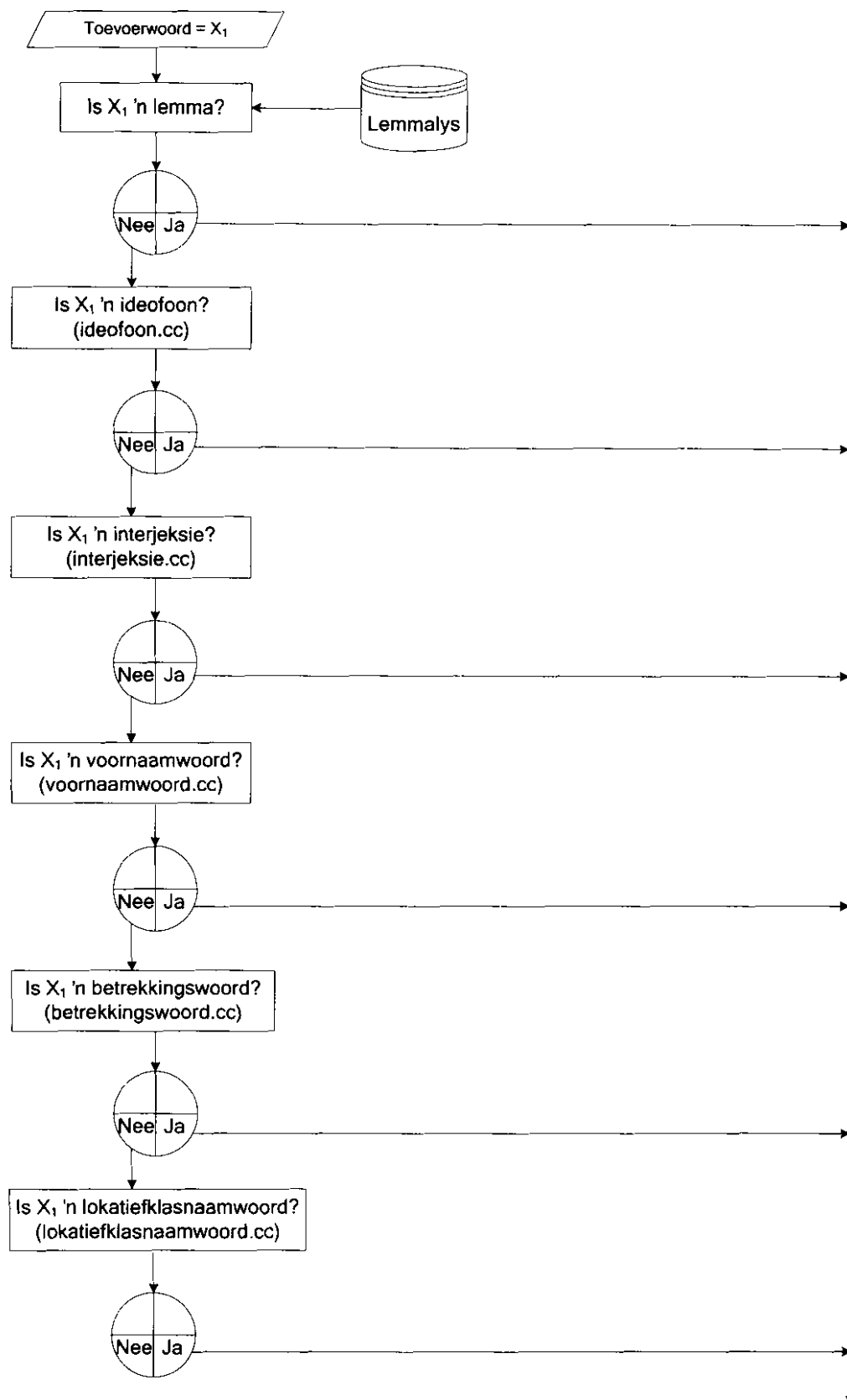
Daar is in *FSA 6* die opsie om die reëlmatige uitdrukkings in C++-formaat te stoor (Van Noord, 2002b:51). Hierdie kode kan dan in 'n alleenstaande program gebruik word. Die sisteem (lemma-identifiseerder) is in C++ geprogrammeer, wat 'n toevoerlêer neem (met een woord per lyn) en die woorde een vir een deur al die stappe, soos in Figuur 4.1 uiteengesit, neem. Die woorde (die toevoerwoord en ooreenstemmende lemma) word in 'n afvoerlêer geskryf. Die gebruiker kan, nadat die proses voltooi is, die resultate in die afvoerlêer sien.

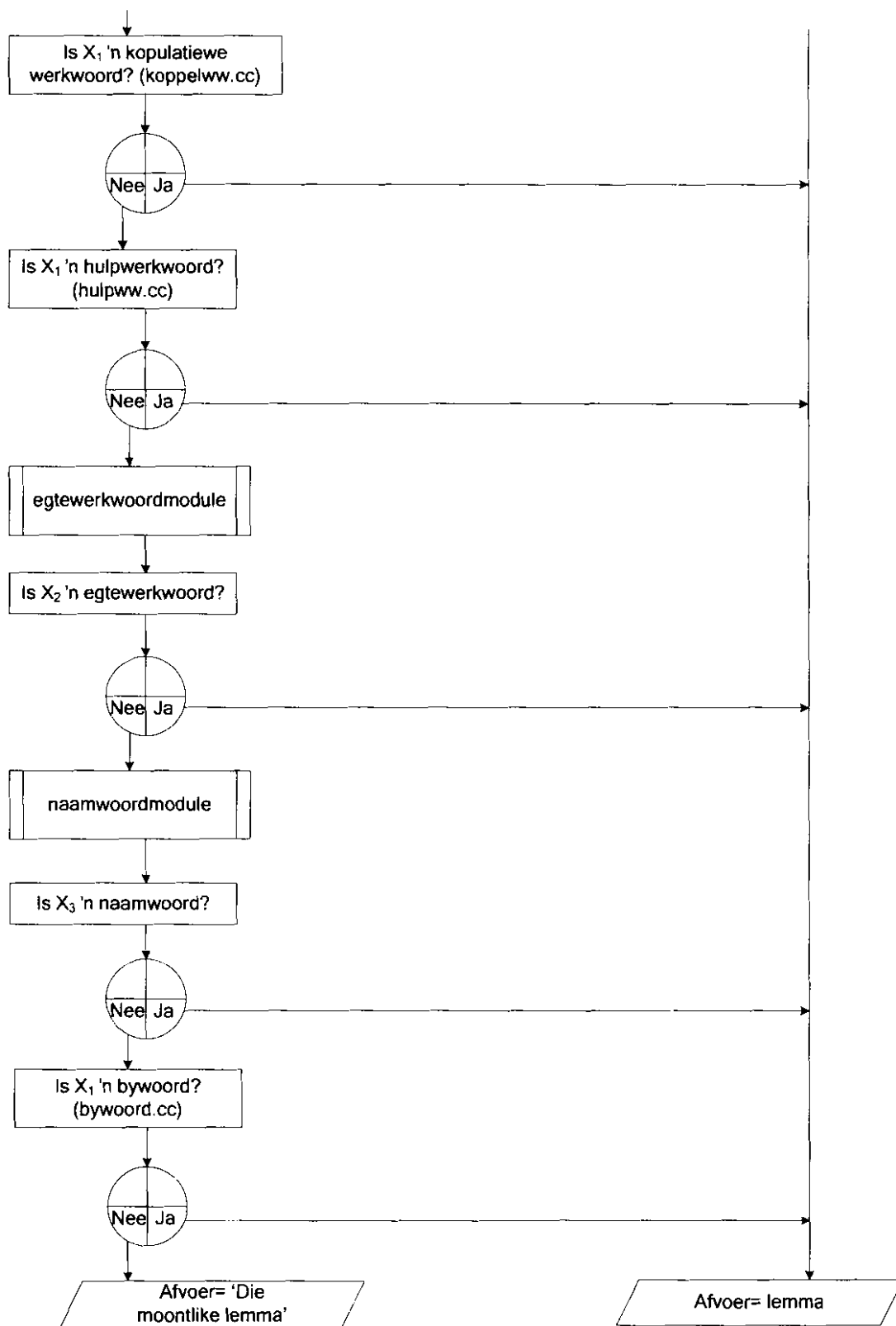
Die reëlmatige uitdrukkings is volgens die verskillende suffikse en prefikse, soos dit onder 3.4 bespreek is, ingedeel, en verskillende C++-lêers is vir groepe suffikse en prefikse geskep. In sommige gevalle is die groepe verder verdeel,

omdat sommige reëlmatige uitdrukkings meer as een oplossing in *FSA 6* kan genereer. So byvoorbeeld lewer die reëlmatige uitdrukking $\{[?*, [d, i, l, e]: [l, a]], [?*, [i, l, e]: a]\}$ twee resultate in *FSA 6* as die woord **badile** ('het gelees') ingevoer word, te wete **bala** ('lees') en ***bada**. Soms lewer 'n reëlmatige uitdrukking soos $\{[\{i, n\}: [], ?*], [\{i, n\}, t, h]: r, ?*]\}$ twee resultate wat albei geldige woorde is, soos byvoorbeeld **thusa** ('help') en **rusa** ('roes') van die toevoerwoord **nthusa** ('my help'). Die vereiste is egter dat elke reëlmatige uitdrukking net een oplossing moet genereer. Daarom moet die reëlmatige uitdrukkings geskei word en moet $\{[?*, [d, i, l, e]: [l, a]]\}$ byvoorbeeld voor $\{[?*, [i, l, e]: a]\}$ uitgevoer word. Die rede daarvoor sal uit die bespreking van die volgorde duidelik word.

Die volgorde van die sisteem en in besonder die volgorde van die stappe in die egtewerkwoord- en naamwoordmodule is baie belangrik omdat die suffikse in elke stap reg verwyder moet word sodat die oorblywende suffikse in die volgende stappe reg herken kan word. In die eerste stap word die toevoerwoord aan 'n lemmalys getoets om te bepaal of dit nie alreeds 'n lemma is nie. Hierdie stap identifiseer onder andere ook al die woorde wat in die geslote klasse is. Indien dit wel 'n lemma is, word dit na die afvoerlêer as die lemma geskryf. Indien die woord nog nie 'n lemma is nie, gaan die woord deur die eerste reël. Die reël mag net een moontlike antwoord gee, want die sisteem soek net een woord op 'n slag op. Nadat 'n reël toegepas is, word die woord weer aan die lemmalys getoets. So gaan die proses aan totdat die lemma in die lys gevind word, of totdat al die reëls uitgevoer is.

Daar is een lemma-identifiseerder gemaak wat uit verskeie prosedures bestaan, waaronder ook 'n egtewerkwoordmodule en 'n naamwoordmodule. Hierdie twee modules kan ook as losstaande komponente gebruik en geëvalueer word. Die volgorde van die sisteem word in Figuur 4.1 hieronder geïllustreer (die X_1 stel die woord voor en die syfer dui die inkrementele verandering aan).





Figuur 4.1 'n Diagrammatiese voorstel van die sisteem

4.2.2 Die naamwoordmodule

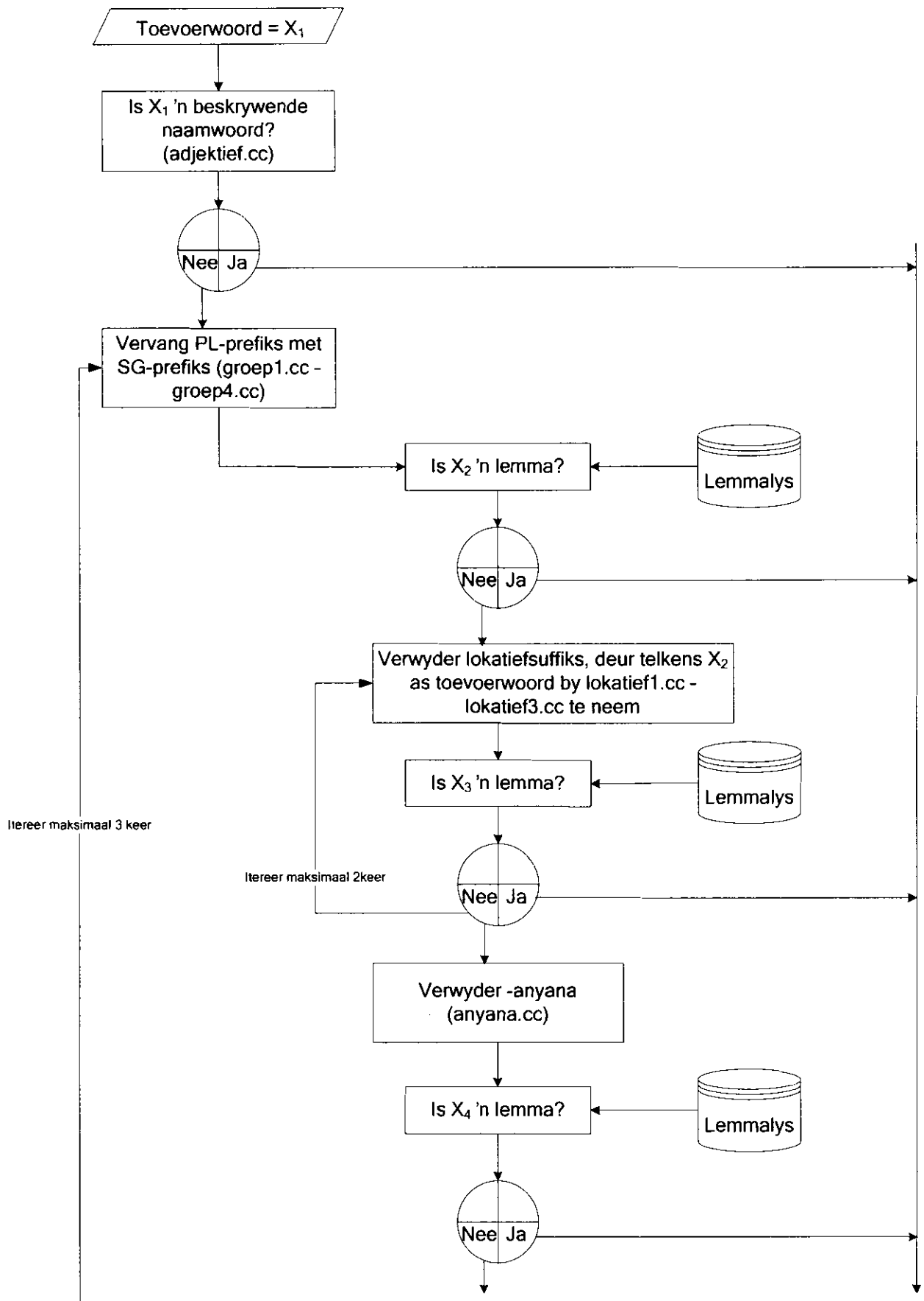
Die ordening van die reëls in die naamwoordmodule is in 'n ander (omgekeerde) volgorde as die reëls wat vir woordvorming in die grammatika gevolg word. Dit volg dieselfde hiërargiese orde as wat in morfologiese analises gebruik word. Dit beteken dat die lokatiewe suffiks eerste verwyder moet word, gevolg deur die diminutiewe suffiks en laastens feminitiewe/augmentatiewe suffikse. Soos in Hoofstuk 3 bespreek, moet die meervoudsprefiks ten slotte deur die enkelvoudsprefiks vervang word.

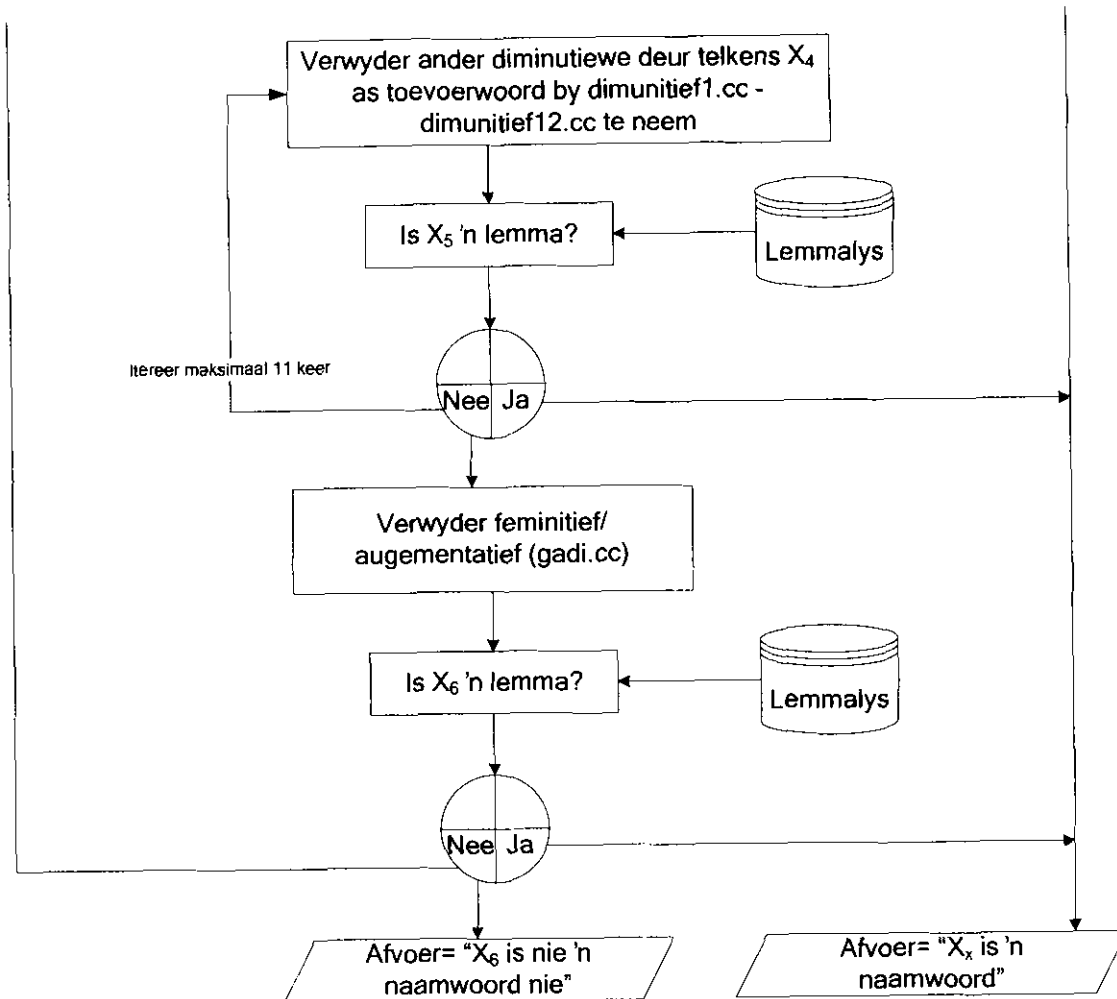
Vir die lemma-identifisering van naamwoorde is die volgorde soos volg (vergelyk Figuur 4.2 vir 'n diagrammatiese voorstelling van hierdie module – die X stel die woord voor en die syfer onderaan dui die inkrementele verandering aan):¹⁰

- Die beskrywende naamwoorde word eerste geïdentifiseer.
- In die tweede stap moet die meervoudsprefikse deur die enkelvoudsprefikse vervang word. Die reëls word nie in een lêer gestoor nie, maar wel in vier verskillendes (groep1.cc, groep2.cc, groep3.cc, groep4.cc), sodat dit nie meervoudige resultate lewer nie. Dit verhoed dan byvoorbeeld dat die *di*-prefiks van *dinku* ('skape') terselfdertyd na *se*- (klas 3) en *lo*- (klas 6) verander en in die geval van klas 5 deur die *leë* karakter vervang word, wat onderskeidelik **senku*, **lonku* en *nku* (skaap) as resultaat lewer. Elke groep reëls lewer daarom net een resultaat op 'n slag op en die resultaat word telkens aan die lemmalys getoets totdat die lemma, in dié geval *nku*, geïdentifiseer is. Die woord word in hierdie tweede stap net aan groep1.cc se reëls onderwerp.
- Gestel die woord het verander, maar dit is nie in die lemmalys nie, dan word die veranderde toevoerwoord na *lokatief1.cc* - *lokatief3.cc* gestuur, wat dan die lokatiewe *-ng* verwyder, indien dit daar sou wees.

¹⁰ Indien die module alleenstaande gebruik word, word die woord eers teen die lemmalys getoets om te bepaal of dit nie alreeds 'n lemma is nie.

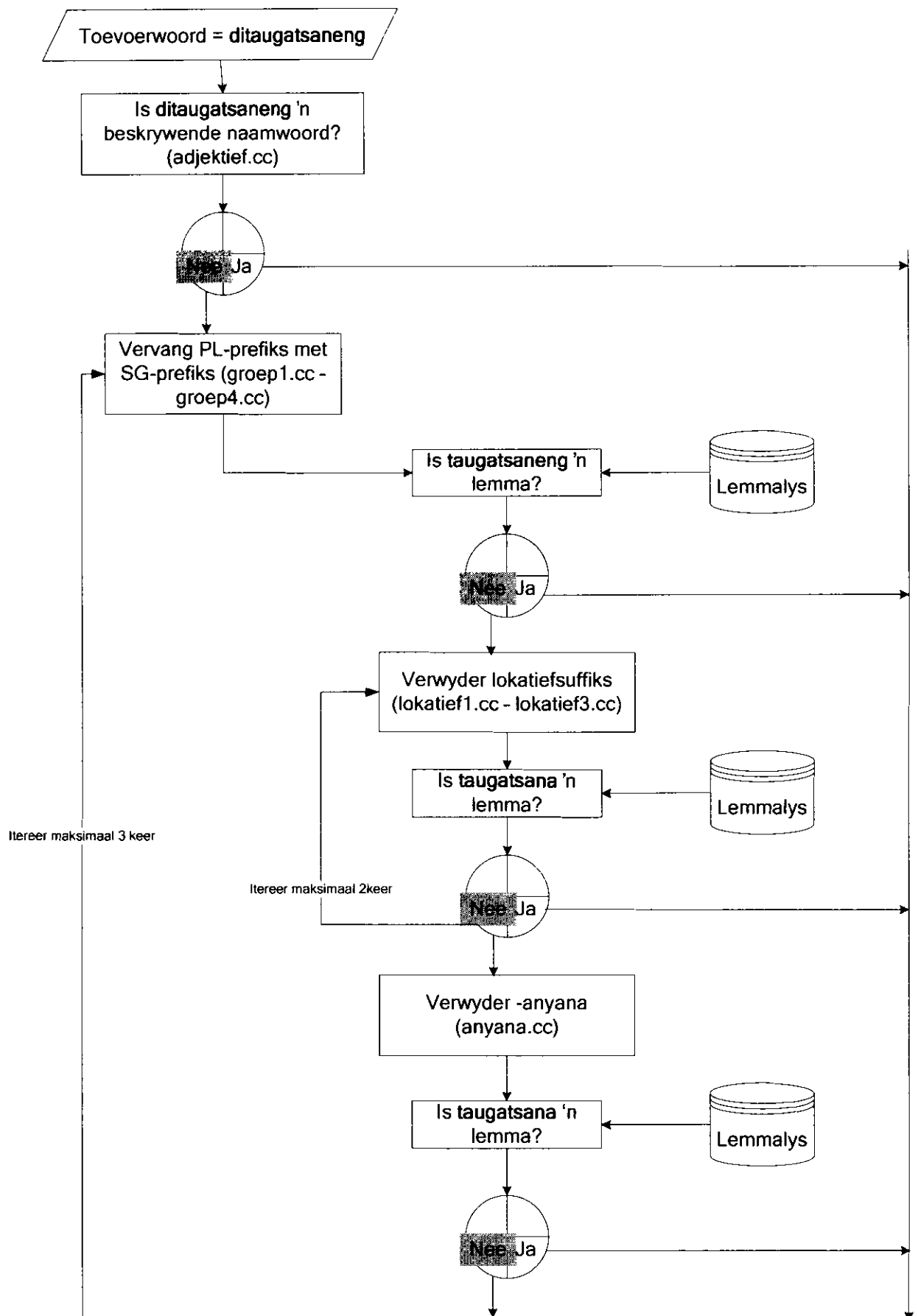
- Indien die woord nog steeds nie in die lemmalys is nie, gaan die woord na `anyana.cc`. Die suffiks *-anyana* (wat verdere verkleining aandui) moet eerste verwyder word, voor die gewone diminutiefsuffiks *-ana* afgehaal kan word (vergelyk die bespreking onder 3.4.1.2).
- Daarna word die veranderde woord na `diminutief1.cc` tot `diminutief12.cc` gestuur, maar hier moet die veranderde woord soos dit voor `diminutief1.cc` gelyk het na `diminutief2.cc`, `diminutief3.cc`, ens. gestuur word, totdat die woord 'n lemma is, of `diminutief12.cc` bereik is.
- Die veranderde woord word daarna, indien dit nog nie 'n lemma is nie, na `gadi.cc` gestuur, waar die feminitiewe/augmentatiewe suffiks verwyder word.
- Indien die woord nog nie in die lemmalys is nie, word die hele proses met `groep2.cc` as die eerste stap herhaal. Die proses kan herhaal word tot en met `groep4.cc` as die eerste stap.
- Die afvoer (die lemma of moontlik lemma) word na 'n lêer met twee kolomme vir die toevoer- en afvoerwoord geskryf.

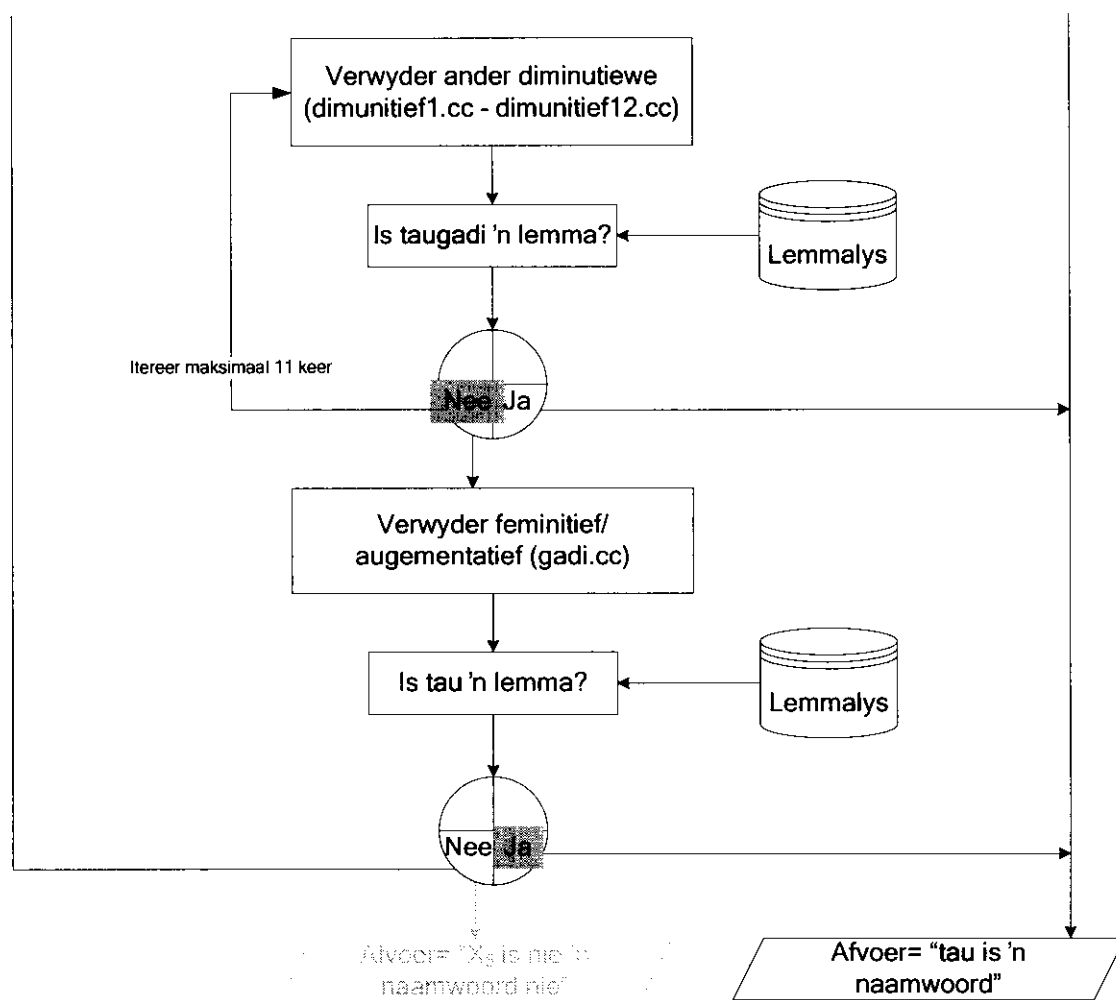




Figuur 4.2 'n Diagrammatiese voorstelling van die naamwoordmodule

Die proses hierbo kan aan die hand van 'n voorbeeld geïllustreer word. Indien **ditaugatsaneng** ('by die leeuwyfietjies') deur die naamwoordmodule geanaliseer word, sal die resultaat nadat al die affikse verwyder/verander is, **tau** ('leu') wees, soos in Figuur 4.3 geïllustreer word.





*Figuur 4.3 'n Diagrammatiese voorstel van die ontleding van **ditaugatsaneng** deur die naamwoordmodule*

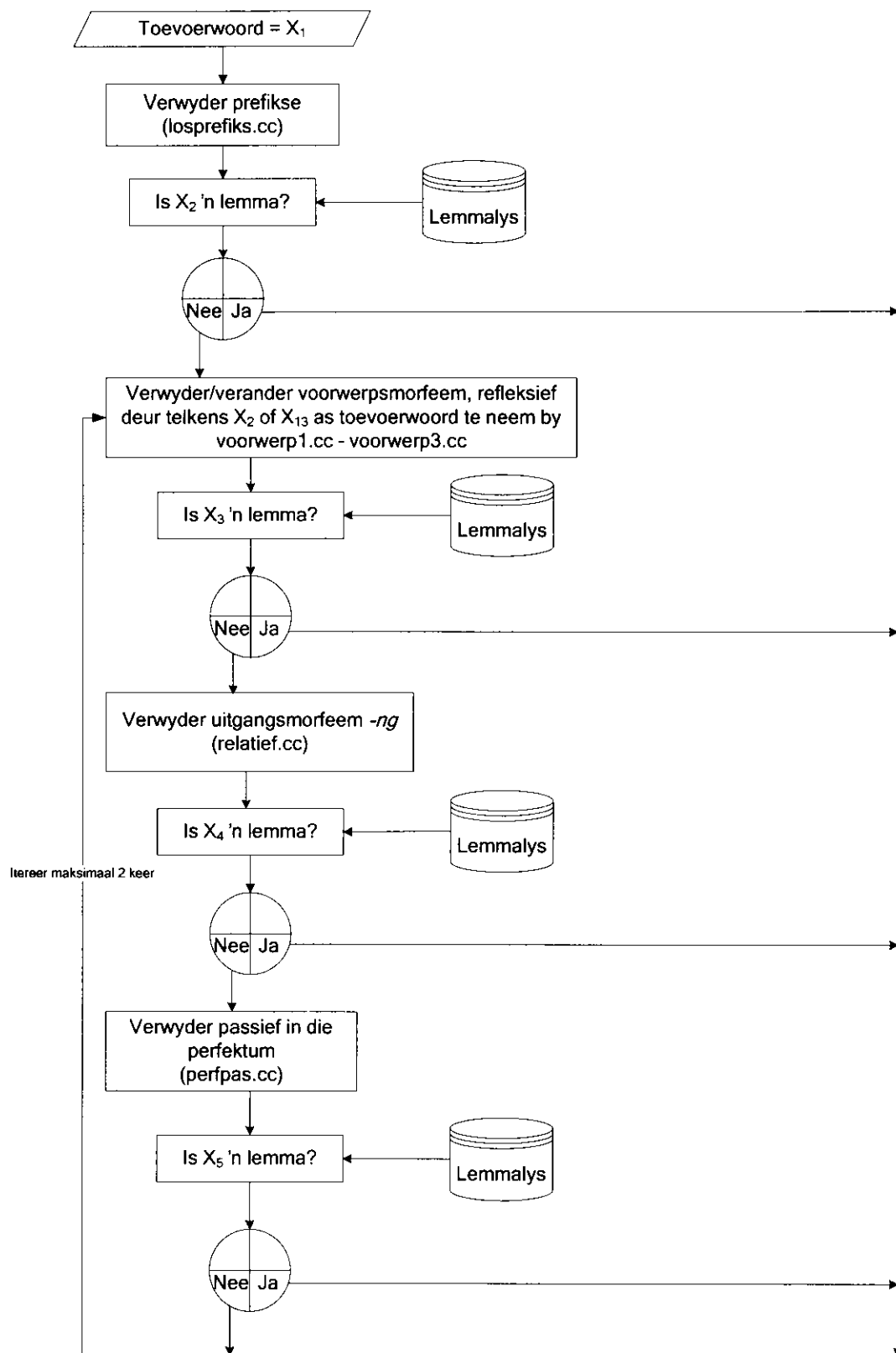
4.2.3 Die egtewerkwoordmodule

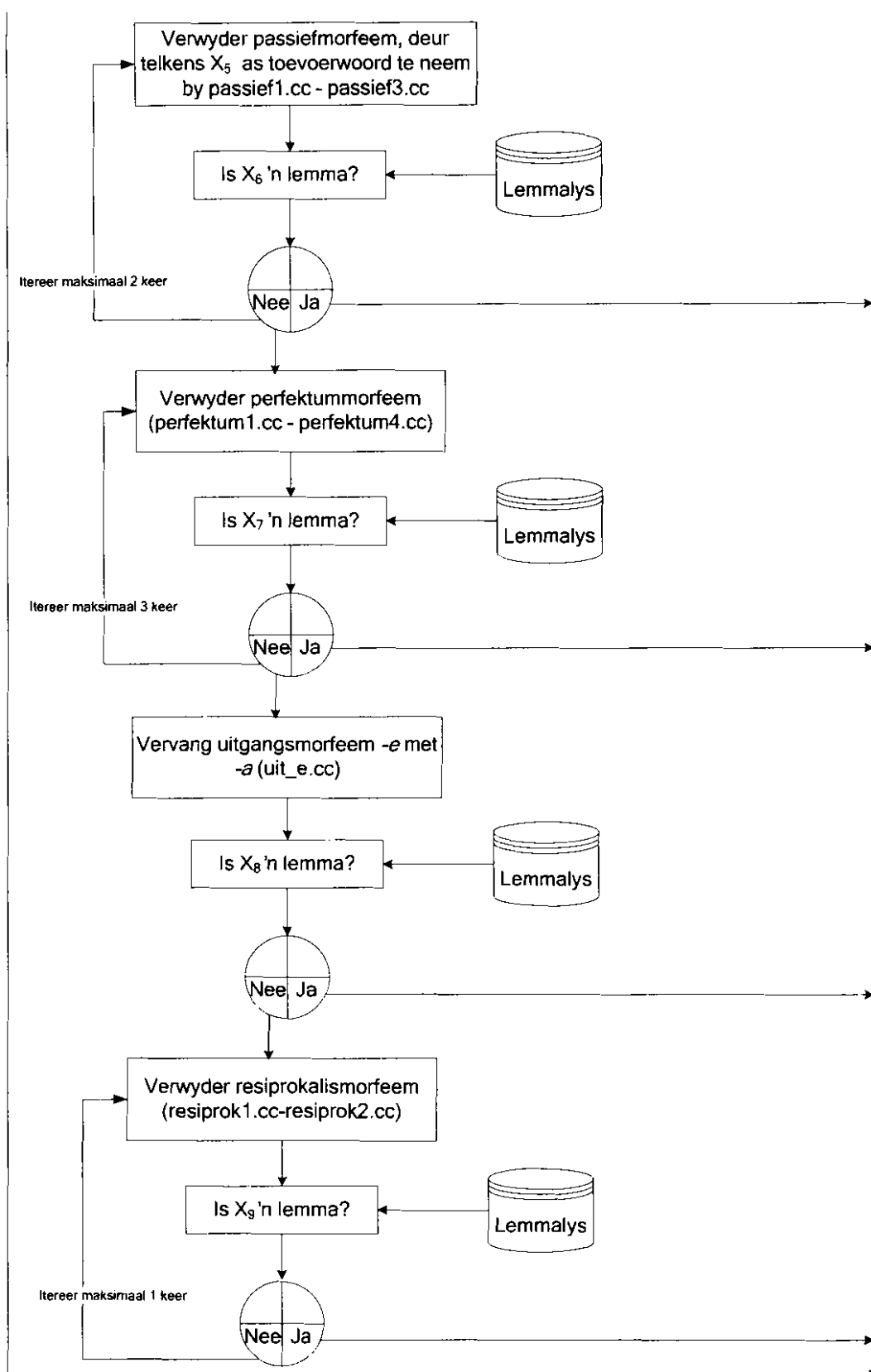
Die ordening van die reëls in die egtewerkwoordmodule is 'n ander (omgekeerde) volgorde as die reëls wat vir woordvorming in die grammatika gevolg word. Dit is dieselfde volgorde wat in morfologiese analyses gebruik word, en in hierdie geval dieselfde volgorde as wat Krüger (2006a:257) gebruik. Die eerste suffiks wat verwyder moet word, is die uitgangsmorfeem *-ng*, gevolg deur die passief *-iw-*, perfektum *-il-*, die uitgangsmorfeem *-e*, die resiprokalis *-an-*, die applikatief *-e/-*, die kousatief *-is-*, die iteratief *-ak-*, en laastens die neutro-passief *-eg-*.

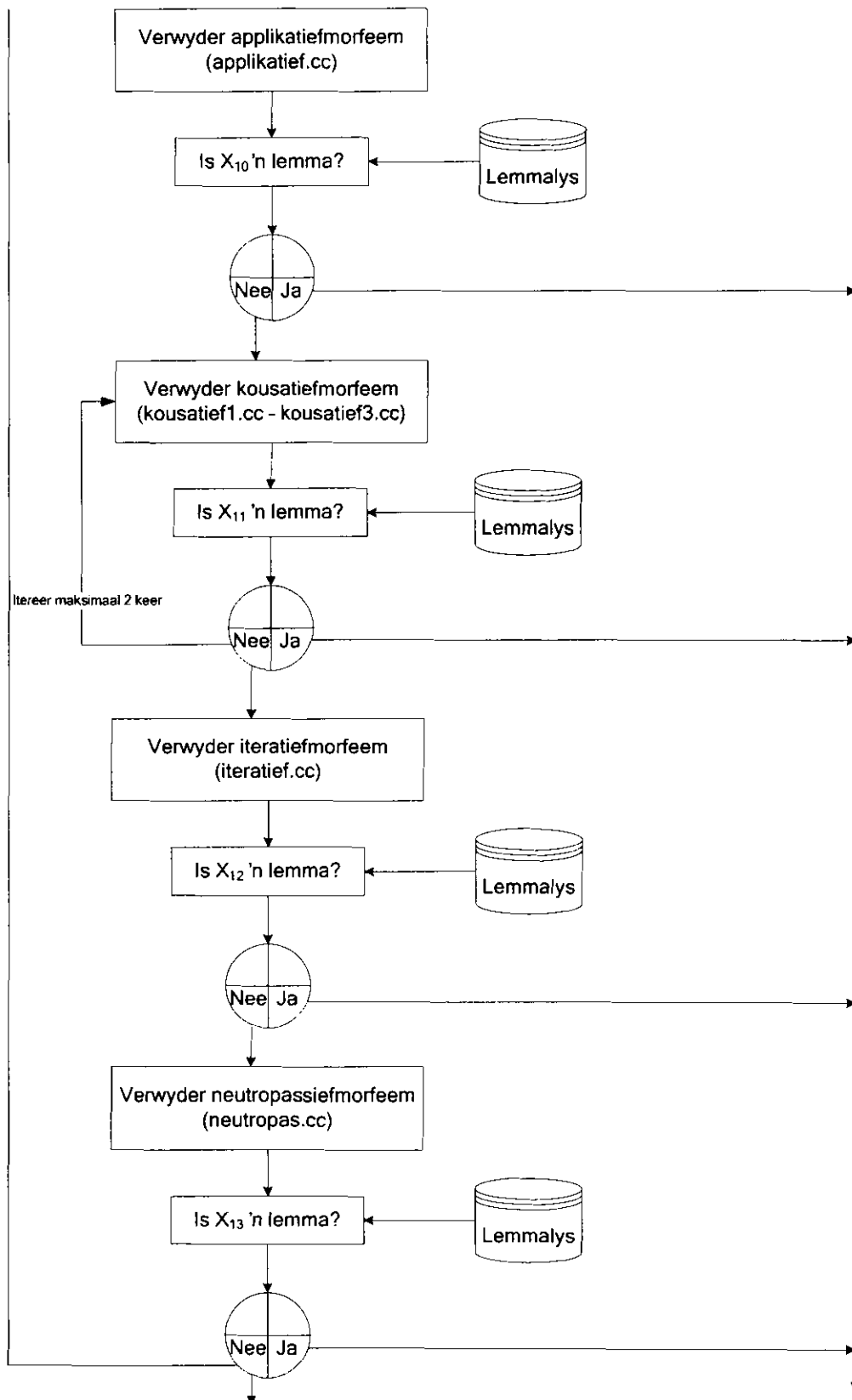
Die proses verloop daarom soos volg by die werkwoord (vergelyk Figuur 4.4 vir 'n diagrammatiese voorstelling van die proses – die X stel die woord voor en die syfer onderaan dui die inkrementele verandering aan):

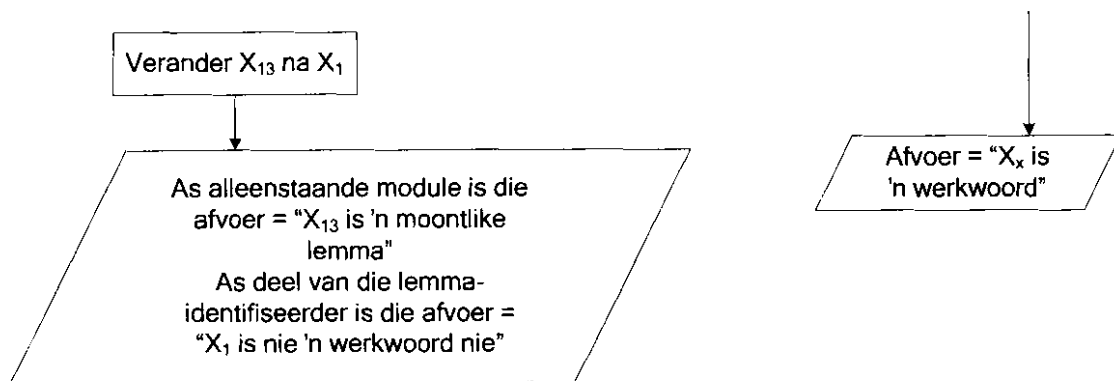
- Indien die egtewerkwoordmodule alleenstaande is, word die woord eers teen die lemmalys getoets om te bepaal of dit nie al 'n lemma is nie. Andersins word die prefikse wat los van die werkwoord geskryf is, verwyder.
- In die tweede stap moet die voorwerpsmorfeme en die refleksief verwyder, en in sekere gevalle verander, word. Dit word met behulp van die lêers `voorwerp1.cc`, `voorwerp2.cc` en `voorwerp3.cc` gedoen. Al die reëls is nie in een lêer gestoor nie, maar wel in drie verskillendes, sodat dit nie meervoudige resultate lewer nie. Die *i*- (refleksief) in **ithuta** ('self leer') kan byvoorbeeld na **ruta** ('onderrig') verander word of verwyder word met ***thuta** as resultaat. Die woord word aan die reëls in `voorwerp1.cc`, `voorwerp2.cc` en `voorwerp3.cc` onderwerp. Die resultaat word telkens aan die lemmalys getoets of dit 'n lemma is.
- Indien die resultaat in die vorige stap nie 'n lemma is nie, word dit aan die reëls in `relatief.cc`, waarin die uitgangsmorfeem *-ng* verwyder word, onderwerp. Daarna word die woord weer aan die lemmalys getoets.
- Die woord word vervolgens aan die reëls in die lêer genaamd `perfpas.cc` onderwerp. In hierdie reëls word die *-w-* wat in die perfektumuitgange voorkom, verwyder. Die resultaat in hierdie stap is 'n werkwoord met 'n perfektumuitgang, wat twee stappe verder verwyder sal word.
- Daarna gaan die woord na `passief1.cc`, `passief2.cc` en `passief3.cc` waar die gewone passiewe suffikse verwyder word. Die woord word na die proses in die lemmalys opgesoek om te bepaal of dit 'n lemma is.

- Van hier gaan die woord na `perfektum1.cc` - `perfektum4.cc` sodat die perfektumsuffikse verwyder kan word. Die resultaat word weer aan die lemmalys getoets.
- In die volgende stap word die uitgangsmorfeem `-e` met behulp van `uit_e.cc` verwyder. Die resultaat word na hierdie stap aan die lemmalys getoets of dit 'n lemma is.
- Die werkwoord gaan daarna verder na `resiprok1.cc` en `resiprok2.cc`, waar die resiproke suffiks verwyder word. Soos tevore word getoets of die resultaat 'n lemma is.
- Vandaar gaan dit na `applikatief.cc`, waar die applikatiewe suffiks verwyder word. Die resultaat word weer aan die lemmalys getoets.
- Daarna word die kousatiewe suffiks met behulp van `kousatief1.cc`, `kousatief2.cc` en `kousatief3.cc` verwyder. Die resultaat word na hierdie stap aan die lemmalys getoets om te bepaal of dit 'n lemma is.
- Indien die werkwoord nog nie 'n lemma is nie, gaan die resultaat van die vorige stap na die lêer `iteratief.cc` vir die verwydering van die iteratiewe suffiks. Ná hierdie stap word die resultaat aan die lemmalys getoets.
- In die volgende stap word die neutro-passiewe suffiks met behulp van `neutropas.cc` verwyder, waarna dit weer aan die lemmalys getoets word.
- Die proses word met 'n herhaling van `voorwerp1.cc`, `voorwerp2.cc` en `voorwerp3.cc` afgesluit, waarna daar telkens gekyk word of die resultaat 'n lemma is.
- Die afvoer (die lemma of moontlike lemma) word na 'n lêer met twee kolomme vir die toevoer- en afvoerwoord geskryf. Indien dit as module in die sisteem gebruik word, word die toevoerwoord, wat nie as 'n werkwoordlemma herken is nie, in die vorm soos wat dit voor die module gelyk het, na die naamwoordmodule gestuur.





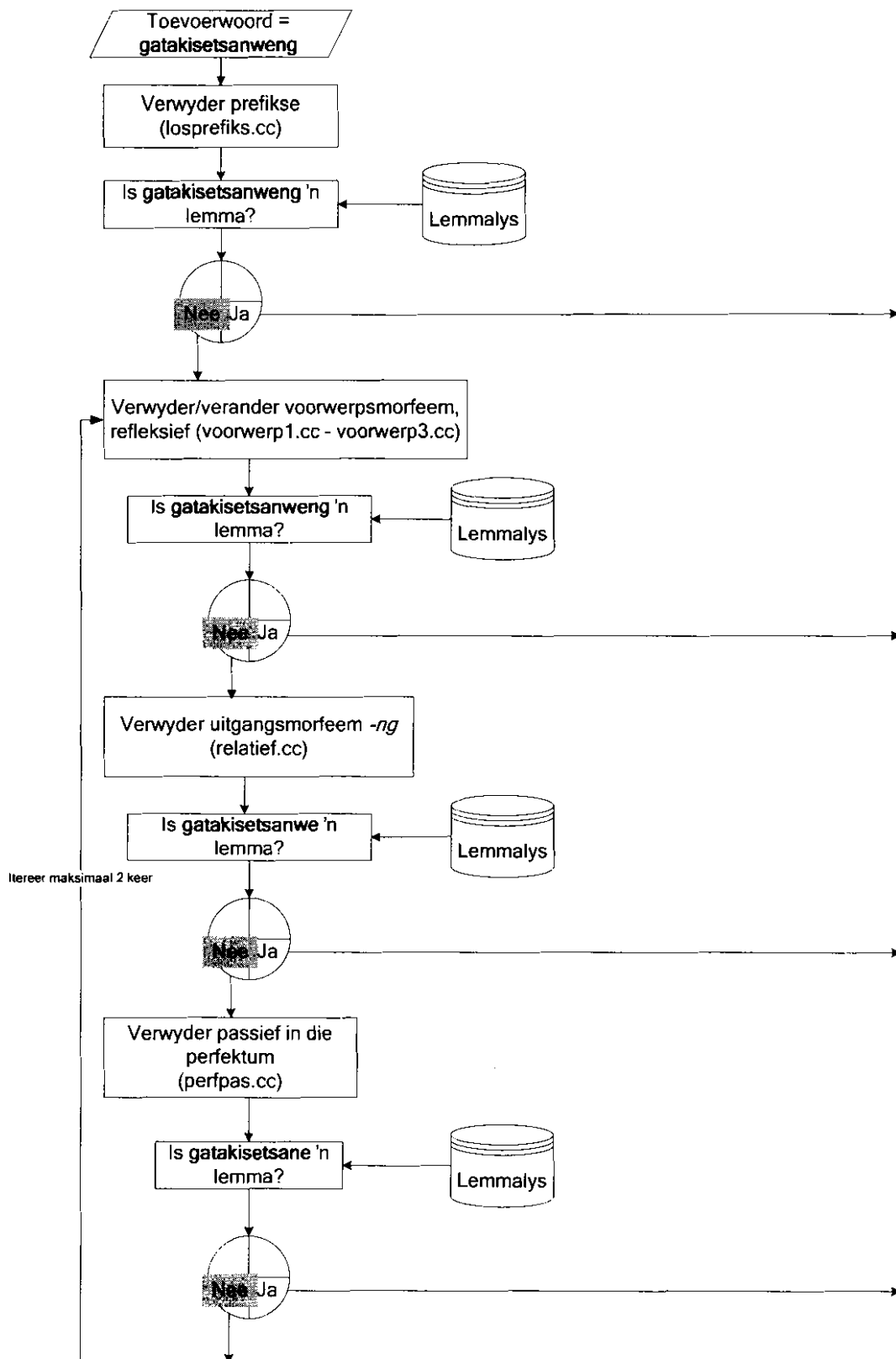


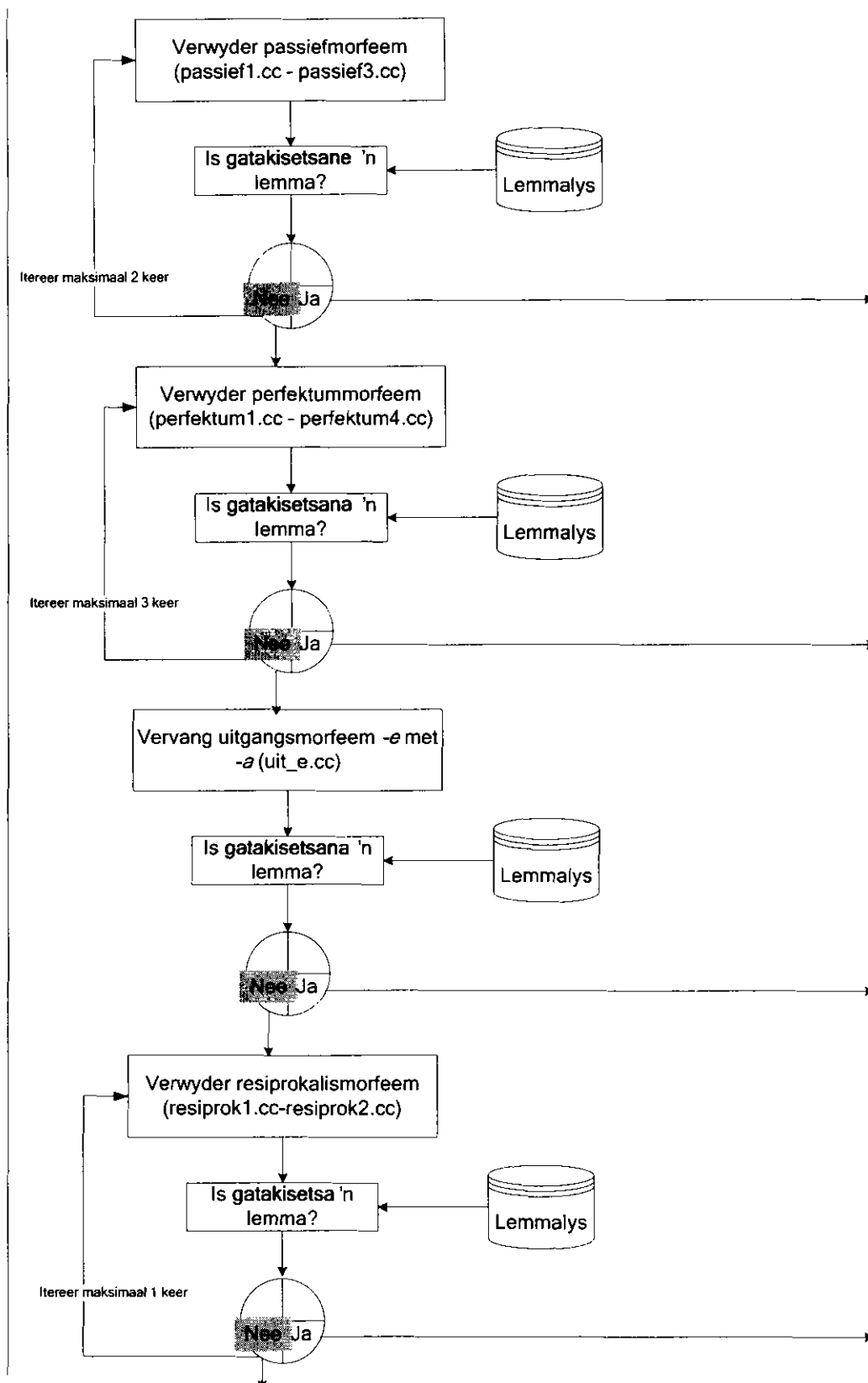


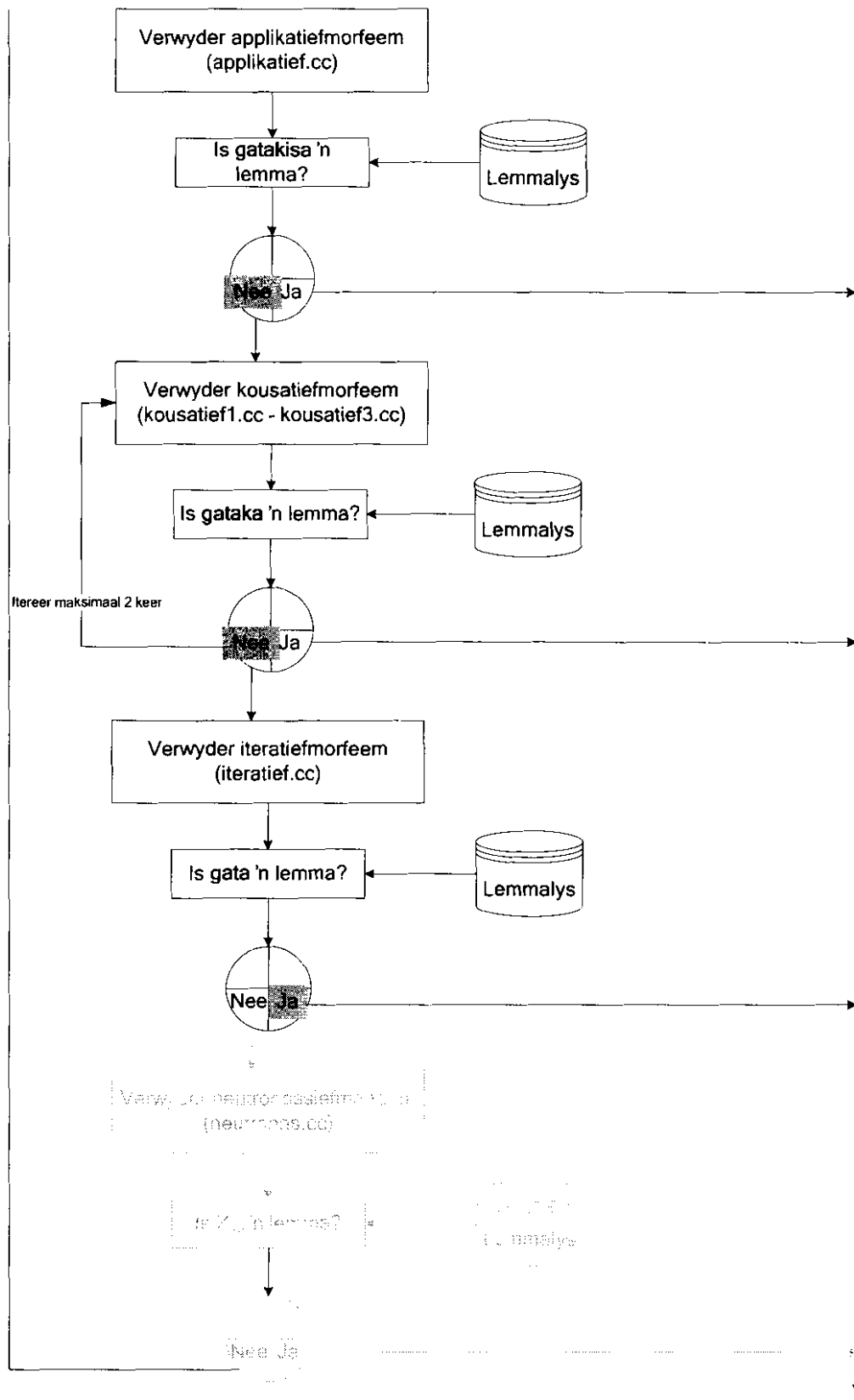
Figuur 4.4 Diagrammatiese voorstelling van die egtewerkwoordmodule

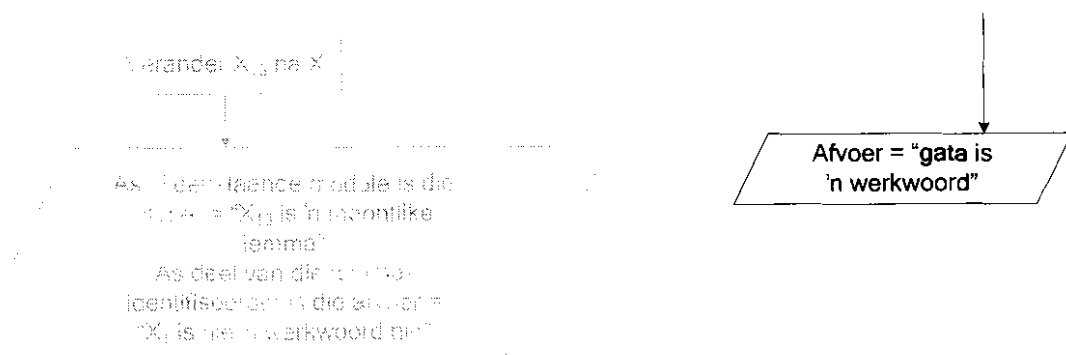
Indien die woord **gatakisetanweng**¹¹ ('vir mekaar laat stukkend trap') deur die egtewerkwoordmodule gestuur word, sal die resultaat, nadat al die affikse verwyder/verander is, **gata** ('trap') wees, soos dit in Figuur 4.5 geïllustreer word.

¹¹ Die woord klink ietwat geforseerd met al die suffikse, maar dit is om te illustreer hoe die module werk.









Figuur 4.5 'n Diagrammatiese voorstel van die ontleding van gatakisetsanweng deur die egtewerkwoordmodule

4.2.4 Samevatting

In die voorafgaande afdeling is die samestelling van die lemma-identifiseringsstelsel en die ordening van die reëls in die naamwoord- en egtewerkwoordmodule bespreek. Die stelsel bestaan onder andere uit FSA's waarmee woorde in die geslote woordklasse (ideofone, interjeksies, voornaamwoorde, betrekkingswoorde, bywoorde) en die lokatiefklasnaamwoorde, kopulatiewe werkwoorde en hulpwerkwoorde mee geïdentifiseer word. Voorts bestaan dit ook uit FST's in die naamwoord- en werkwoordmodules. Die ordening van die reëls in die naamwoord- en egtewerkwoordmodule is in 'n teenoorgestelde volgorde as die reëls vir woordvorming in die grammatika (vergelyk Cole (1955) en Krüger (2006a)). Dit is dieselfde volgorde wat in morfologiese analyses gebruik word, en in hierdie geval dieselfde volgorde as wat Krüger (1994) gebruik. Vir die naamwoord beteken dit dat die lokatiewe suffiks *-ng* eerste verwyder moet word, gevolg deur die diminutiewe suffiks *-ana* en laastens die feminitiewe/augmentatiewe suffiks *-gadi*, waarna die meervoudprefiks deur die enkelvoudsprefiks vervang word. Wat die werkwoorde betref, is die eerste suffiks wat verwyder moet word die uitgangsmorfeem *-ng*, gevolg deur die passief *-iw-*, perfektum *-il-*, die uitgangsmorfeem *-e*, die resiprokalis *-an-*, die applikatief *-el-*, die kousatief *-is-*, die iteratief *-ak-* en laastens die neutro-passief *-eg-*. In die volgende afdeling kom die evaluering van die lemma-identifiseringsstelsel aan die beurt.

4.3 Evaluering

In Hoofstuk 3 en afdeling 4.2 van hierdie hoofstuk is eindigestaatoutomate en reëls vir die bepaling van die lemma bespreek. Dit is egter van geen waarde as die reëls teoreties werk, maar nooit op werklike data getoets word nie. In hierdie afdeling word aangedui hoe die lemma-identifiseerder se akkuraatheid getoets word. Vervolgens word die samestelling van die toetsdata en die meetinstrumente wat gebruik word om die lemma-identifiseerder mee te evalueer, bespreek. Daarna word die resultate van die naamwoordmodule, egtewerkwoordmodule en die volledige sisteem weergegee, en die foute wat die resultate beïnvloed het, word bespreek.

4.3.1 Samestelling van toetsdata

Die twee morfologies komplekse woordsoorte is, soos in Hoofstuk 2 aangedui, die naamwoord en die werkwoord. Derhalwe was die ontwikkeling van eindigestaatoorvormers (Hoofstuk 3) op die naamwoord en werkwoord toegespits, en moet die naamwoord en werkwoord spesifiek aan evaluering onderwerp word. Twee stelde woorde is aan toetse onderwerp: ten eerste is die naamwoord- en werkwoordreëls getoets op 'n ewekansige steekproef van 500 komplekse naamwoorde, en 500 komplekse en simplekse naamwoorde, asook 500 komplekse werkwoorde en 500 komplekse en simplekse werkwoorde, en ten tweede is die volledige stel reëls met 'n ewekansige steekproef van 1,000 woorde uit alle woordsoortkategorieë getoets.

Die steekproef is met onderstaande formule bepaal, waar n die minimumgrootte van die steekproef is, N die populasiegrootte is, en d die persentasie sekerheid ("confidence") uitdruk.

$$n = \frac{N}{1 + \frac{N \times d^2}{10000}}$$

Deur middel van 'n lukrake, geoutomatiseerde steekproef word die datastelle uit die woordelys van 'n Setswanaspeltoetser (NWU, 2005) met 'n grootte van 73

119 woorde onttrek. Indien daar met 'n steekproefgrootte van 'n 1,000 woorde gewerk word, is die sekerheid ("confidence") 97%; 'n sekerheid van 95% word allerweë as aanvaarbaar beskou in die bepaling van 'n steekproefgrootte (Lohr, 1999:40).

Daar was tydens die ontwikkeling van die lemma-identifiseerder nie 'n amptelike volledige lemmalys wat aan die vereistes voldoen – soos in Hoofstuk 2 bespreek – beskikbaar nie. 'n Lemmalys is saamgestel met behulp van die *South African multi-language dictionary and phrase book* (Reynierse, 1991) en 'n frekwensielys (Prinsloo, datum onbekend). Die woorde in die woordeboek en in Prinsloo (datum onbekend) se frekwensielys is nie almal lemmas, soos wat dit in Hoofstuk 2 gedefinieer is nie. Die lysie moes dus handmatig gekontroleer word, en die woorde wat nog nie lemmas was nie, se lemmas moes aan die hand van die kriteria, soos dit in Hoofstuk 2 uiteengesit is, geïdentifiseer word. Hierdie word die standaardlemmalys genoem en is weer by 4.3.3 ter sprake.

'n Tekseenheididentifiseerder wat voorsiening maak vir die disjunktiewe ortografie van Setswana was ook nie tydens die ontwikkeling van die lemma-identifiseerder beskikbaar nie. Die rede waarom dit by lemma-identifisering van belang is, is dat lopende teks se lemmas dan ook geïdentifiseer kan word. Daar is egter in die lemma-identifiseerder voorsiening gemaak vir spasies in die werkwoorde (egte, hulp- en koppelwerkwoorde). Indien 'n tekseenheididentifiseerder later van tyd ontwikkel sou word, sal die tekseenheididentifiseerder saam met die lemma-identifiseerder op lopende teks gebruik kan word.

4.3.2 Meetinstrumente

Volgens Paice (1994:42) is daar wat stamidentifisering betref twee soorte foute, wat 'n stamidentifiseerder kan maak, naamlik "overstemming" en "understemming". Die stamidentifiseerder lewer dus stamme waarvan te veel ("overstemming") of te min affikse ("understemming") verwyder is. Dit is ook die

geval by lemma-identifisering, waar daar te veel of te min suffikse verwyder word en 'n resultaat gelewer word wat nie 'n lemma is nie.

Op grond van die verband tussen stam- en lemma-identifiseerders, kan die kriteria wat gewoonlik in die evaluering van stamidentifiseerders geld, gebruik word (vergelyk Daelemans & Strik, 2002:37). Die kriteria wat in die evaluering veral ten opsigte van inligtingonttrekking geld, is akkuraatheid ("precision") en herkenning ("recall"). As akkuraatheid nie van groot belang is nie (in inligtingonttrekking is dit irrelevante inligting saam met die bruikbare inligting), is 'n hoë herkenning aan die orde van die dag, maar as 'n mens relevante en juiste inligting wil hê, geld hoë akkuraatheid.

Herkenning in inligtingonttrekking handel oor die hoeveelheid relevante dokumente wat onttrek is deur 'n sisteem, uit die totale hoeveelheid dokumente beskikbaar in die databasis. Aan hierdie dokumente word punte toegeken op grond van die kwaliteit van die dokumente – oftewel die akkuraatheid (vergelyk Kraaijman & Pohlmann, 1996). Vir lemma-identifisering sal 'n hoë herkenning sonder akkuraatheid nie soveel waarde hê nie. Die doel van lemma-identifisering is nie om inligting te onttrek nie, maar om die regte lemma van woorde te identifiseer. Wat die lemma-identifiseerder betref, is akkuraatheid die belangrikste, veral omdat dit die linguistiese variant van die stamidentifiseerder is (vergelyk 1.2 se bespreking oor die verskille tussen die lemma- en stamidentifiseerder).

Dit is dan in die geval van 'n lemma-identifiseerder belangriker om die akkuraatheid te bepaal. Alvorens dit gedoen kan word, moet 'n paar begrippe wat in die formule om die akkuraatheid te bepaal voorkom, bespreek word. Daar is byvoorbeeld woorde wat reeds lemmas is; hulle word ter wille van die bespreking van die formule die "simplekse" genoem. Dan is daar woorde wat nog omvorm moet word; dit is die "komplekse". Die veranderlike

AsSimplekseHanteer is al die woorde wat soos simplekse hanteer is en die **AsKomplekseHanteer** is al die woorde wat soos komplekse hanteer is.

Al die komplekse wat korrek geanaliseer is, kan **KomplekseKorrekHanteer** genoem word. Alle simplekse wat korrek hanteer is, kry die veranderlikenaam **SimplekseKorrekHanteer**. Ten einde dan te bepaal hoe akkuraat die lemma-identifiseerder is, moet dit soos volg bereken word:

- **Akkuraatheid op komplekse naamwoorde en werkwoorde afsonderlik:**

$$\text{KompleksAkkuraatheid} = \frac{\text{KomplekseKorrekHanteer}}{\text{AsKomplekseHanteer}}$$

Hiermee word dan bepaal hoe goed die naamwoord- of egtewerkwoordmodule die lemmas van komplekse woorde identifiseer.

- **Akkuraatheid op komplekse en simplekse naamwoorde en werkwoorde afsonderlik:**

$$\text{VoorspeldeAkkuraatheid} = \frac{\text{KomplekseKorrekHanteer} + \text{SimplekseKorrekHanteer}}{\text{AsKomplekseHanteer} + \text{AsSimplekseHanteer}}$$

Op hierdie manier word bepaal hoe goed die lemma-identifiseerder op die kategorieë naamwoord en werkwoord afsonderlik is.

- **Akkuraatheid op alle woorde (lopende teks)**

$$\text{VoorspeldeAkkuraatheid} = \frac{\text{KomplekseKorrekHanteer} + \text{SimplekseKorrekHanteer}}{\text{AsKomplekseHanteer} + \text{AsSimplekseHanteer}}$$

Hiermee kan bepaal word hoe goed die lemma-identifiseerder op 'n versameling woorde wat ander woordsoorte ook insluit, is.

4.3.3 Resultate

In Tabel 4.1 is die resultate van die evaluering van die lemma-identifiseerder saamgevat. Die akkuraatheid van die lemma-identifiseerder met die standaardlemmalys is 68,97% op komplekse naamwoorde, 63,5% op komplekse werkwoorde, 76,29% op komplekse en simplekse naamwoorde, 73,95% op komplekse en simplekse werkwoorde, en 68,54% op alle woorde. Die uitgebreide lys (met die toegevoegde lemmas wat uit die toetsdata kom) verbeter die resultate aansienlik (vergelyk die bespreking van die lemmalys onder 4.4.1, 4.4.2 en 4.4.3).

	Standaard- lemmalys	Uitgebreide lemmalys
Akkuraatheid op komplekse naamwoorde	68,97%	70,96%
Akkuraatheid op komplekse werkwoorde	63,50%	70,52%
Akkuraatheid op alle naamwoorde (kompleks en simpleks)	76,29%	78,45%
Akkuraatheid op alle werkwoorde (kompleks en simpleks)	73,95%	79,59%
Akkuraatheid op alle woordsoorte	68,54%	75,92%

Tabel 4.1 Resultate van die evaluering

Die resultate by die naamwoorde is onder andere beïnvloed deur 'n ontoereikende lemmalys, onvoorsiene klankveranderings, leenwoorde, samestellings, naamwoordklasverskuiwings, dubbele klasprefikse en homograwe. Die omvang van die foute is persentasiegewys in Tabel 4.2 uitgedruk. Die persentasie is die aantal foute (in 'n kategorie soos byvoorbeeld die ontoereikende lemmalys) in die toetsdata.

Naamwoordmodule	
Ontoereikende lemmalys	11,2%
Onvoorsiene klankveranderings	2,3%
Leenwoorde	0,6%
Samestellings	0,7%
Naamwoordklasveranderings	0,7%
Dubbele prefikse	0,2%
Homograwe	0,2%

Tabel 4.2 Omvang van foute by die naamwoordmodule

Uit Tabel 4.2 blyk dit dat die foute hoofsaaklik aan 'n ontoereikende lemmalys, en tot 'n mate ook aan onvoorsiene klankveranderings, toegeskryf kan word. Die dubbele prefikse en homograwe het die kleinste invloed op die resultate gehad.

Die foute by werkwoorde kan, soos by die naamwoorde, aan 'n ontoereikende lemmalys, onvoorsiene klankveranderings, die komplekse aard van die verdubbelings en samestellings, dubbele suffikse, homograwe asook meerduidige ("ambiguous") uitgange, verandering in die reëlvolgorde, en uitsonderings toegeskryf word. Die invloed van die foute word persentasiegewys in Tabel 4.3 aangedui. Die persentasie is die aantal foute (in 'n kategorie soos byvoorbeeld die ontoereikende lemmalys) in die toetsdata.

Werkwoordmodule	
Ontoereikende lemmalys	6,6%
Onvoorsiene klankveranderings	0,2%
Verdubbelings en samestellings	0,5%
Dubbele suffikse	0,9%
Homograwe	0,8%
Meerduidige uitgange	7,9%
Verandering in die reëlvolgorde	2,0%
Uitsonderings	0,2%

Tabel 4.3 Omvang van foute by die werkwoordmodule

Anders as by die naamwoorde is die grootse oorsaak van foute nie die ontoereikende lemmalys nie, maar die meerduidige uitgange. Onvoorsiene klankveranderings en uitsonderings het die minste tot die foute bygedra.

Die kwantitatiewe prestasie van die naamwoordmodule, egtewerkwoordmodule en volledige lemma-identifiseerder is egter nie die fokuspunt van hierdie studie nie, aangesien dit oor die evaluering van die grammatikas gaan. Vervolgens word die kategorieë van foute in Tabel 4.2 en Tabel 4.3 dan kwalitatief bespreek. (Sien Bylaag 4.1 - 4.3 vir die toetsdata)

4.4 Bespreking van resultate

4.4.1 Naamwoorde

Na bestudering van die resultate kon die oorsake van die afwykings bepaal word, naamlik 'n ontoereikende lemmalys, onvoorsiene klankveranderings, leenwoorde, samestellings, naamwoordklasverskuiwings, dubbele klasprefikse en homograwe. Hierdie oorsake word vervolgens aan die hand van die voorbeelde in Tabel 4.4 bespreek.

Nommer	Toevoer	Verwagte afvoer	Afvoer deur module
245	sejabana <i>se-jabana</i> CL4.SG-elmboog 'elmboog'	sejabana	*sejaba¹²
246	mokgomong <i>mo-kgomo-ng</i> CL1.SG-die totem kgomo- LOC 'persoon wat tot die totem kgomong behoort'	mokgomong	*mokgomo
247	mokgadi <i>mo-kgal-a-i</i> CL1.SG-teregwys-TERM- TERM 'persoon wat teregwys'	mokgadi	*mok
248	maikgaoganyo <i>ma-ikgaoganya-o</i> CL3.PL-verdeel-DEV.TERM 'selfverdelings'	maikgaoganyo	*leikgaoganyo
249	ditlhabelo <i>di-tlhabelo</i> CL5.PL-offer 'offers'	tlhabelo <i>tlhabelo</i> CL5.SG.offer 'offer'	*lollhabelo

¹² Hoewel die afvoer deur die module telkens aan die lemmalys getoets word, word dit nie so rigied toegepas dat vreemde woorde nie geanaliseer kan word nie. Dit lei egter in sommige gevalle daartoe dat die afvoer non-woorde is.

250	dikgala <i>di-kgala</i> CL4.PL-afstand 'afstande'	sekgala <i>se-kgala</i> CL4.SG-afstand 'afstande'	kgala <i>kgal-a</i> kritiseer/verwy- TERM 'kritiseer/verwy'
251	letlhalwane <i>le-tlhale-ana</i> CL3.SG-genie-DIM 'klein genie/iemand wat dink dat hy/sy slim is'	letlhale <i>le-tlhale</i> CL3.SG-genie 'genie'	letlhalwane
252	bahutsanyana <i>ba-huts-a-i-ana-nyana</i> CL1.PL-vervloek-TERM- TERM-DIM-DIM 'arm mense/verloekte mensies'	mohutsi <i>mo-huts-a-i</i> CL1.SG-vervloek- TERM-TERM 'verloekte mens'	*mohutsa
253	kgetsanyana <i>kgetsi-ana-nyana</i> CL5.SG.sak-DIM-DIM 'n (ou) sakkie soos 'n koffie- of teesakkie'	kgetsi <i>kgetsi</i> CL5.SG.sak 'sak'	*kgetsa
254	tlhotshwana <i>tlhoph-a-o-ana</i> CL5.SG.kies-TERM-TERM- DIM 'groepie'	tlhopho <i>tlhoph-a-o</i> CL5.SG.kies-TERM- TERM 'groep/stemming'	tlhotshwana
255	setlhotšhwana <i>se-tlhoph-a-ana</i> CL4.SG.kies-TERM-DIM 'groepie/spannetjie'	setlhopha <i>se-tlhoph-a</i> CL4.SG.kies-TERM 'groep'	*setlhofi

256	materiki <i>materiki</i> CL5.SG.matriek 'matriek'	materiki	*leteriki
257	bakinyane <i>baki-nyane</i> CL5.SG.baadjie-DIM 'baadjietjie'	baki <i>baki</i> CL5.SG.baadjie 'baadjie'	*mokinyane
258	bathobatsho <i>ba-tho-ba-tsho</i> CL1.PL-mens-CL1.PL-swart 'swartmense'	mothomotsho <i>mo-tho-mo-tsho</i> CL1.SG-mens- CL1.SG-swart 'swartmens'	*mothobatsho
259	mafatlhamabedi <i>ma-fatlha-ma-bedi</i> CL3.PL-long-CL3.PL-twee 'longe/tweeling'	mafatlhamabedi	*lefatlhamabedi
260	dikgorokgorwana <i>di-kgoro-kgoro-ana</i> CL6.PL-opening-opening- DIM 'openinkies/stukkende goed'	legorogorwana <i>le-goro-goro-ana</i> CL3.SG- opening- opening-DIM 'openinkie/plaag'	*kgorokgorwana
261	matlo <i>ma-tlo</i> CL3.PL-huis 'huise'	ntlo <i>ntlo</i> CL5.SG.huis 'huis'	*botlo
262	bomaitsheka <i>bo-ma-i-sek-a</i> CL7.SG-CL7.PL-REFL- ondersoek-TERM 'selfondersoekinge'	boitsheka <i>bo-i-sek-a</i> CL7.SG-REFL- ondersoek-TERM 'selfondersoeking'	bomaitsheka

263	mokong <i>mo-ko-ng</i> CL2.SG-murg-LOC 'in die murg'	moko <i>mo-ko</i> CL2.SG-murg 'murg'	moko
264	mokong <i>mo-kong</i> CL2.SG-steil opdraende 'steil opdraende/steilte'	mokong	moko

Tabel 4.4 Ontledings van die naamwoordmodule se resultate

- 'n Ontoereikende lemmalys

Indien die lemmalys nie volledig genoeg is nie, beteken dit dat 'dele' wat soos suffikse lyk (maar eintlik deel van die woord is), verwyder word (voorbeelde (245) – (247)).

Soos in die vorige afdeling bespreek, is die sisteem afhanklik van 'n volledige lemmalys (die lys waaraan daar telkens na elke stap in die sisteem getoets word of dit die regte woord is). Daarsonder kan daar nie bepaal word of 'n woord met 'n klasprefiks *bo-* reeds in die enkelvoud (dus klas 7) of in die meervoud (dus klas 1) is nie. Dit veroorsaak ook probleme met klasse 3 en 7 se klasprefiks *ma-* (vergelyk voorbeeld (248)). So ook kan daar nie bepaal word of die klasprefiks *di-* (klas 4, 5 en 6) se enkelvoudsprefiks *se-* of *lo-* is, of net verwyder moet word vir klas 5 se enkelvoud nie. In voorbeeld (249) word **ditlhabelo** ('offers') byvoorbeeld verkeerdlik na ***lollhabelo** verander.

Heelwat van hierdie probleme kan uitgeskakel word indien die lemmalys volledig genoeg is om hierdie probleme te ondervang. Die akkuraatheid op komplekse naamwoorde het daarom met die uitgebreide lemmalys van 68,97% tot 70,96% verbeter, en die akkuraatheid op alle werkwoorde (kompleks en simpleks) het van 76,29% tot 78,45% verbeter (vergelyk Tabel 4.1).

’n Volledige lemmalys is ook nie in alle gevalle ’n oplossing nie, want die lemmas in ’n lemmalys kan self ook enkele foute veroorsaak. In voorbeeld (250) is die lemma **sekgala** (‘afstand’), maar **kgala** (‘teregwys/opdroog’) word na die toepassing van die reël, $[[d, i]: [], ?^*]$, aanvaar omdat dit in die lemmalys voorkom. Hierdie meerduidigheidsfoute is egter “aanvaarbaar”, aangesien dit slegs opgelos kan word in ’n sisteem wat kontekssensitief is (maar wat normaalweg buite die bestek van lemma-identifisering val). Dit is dus uiteindelik die werk van ’n menslike kontroleerder om te sorg dat sodanige foute nie as korrek aanvaar word nie (byvoorbeeld in die konteks van lemma-identifisering vir leksikografiese doeleindes).

- **Onvoorsiene klankveranderings**

Volgens die reëls (vergelyk Hoofstuk 3) is dit net as die naamwoord op ’n **u** of ’n **o** eindig dat dit met die suffigering van die diminutiewe **-ana** verander na die semivokaal **w**, behalwe as die voorafgaande konsonant **b**, **p**, **ph**, **f** of **d** is. Die uitsondering op die reël blyk ’n woord soos **letlhale** (‘genie’) in voorbeeld (251)) te wees wat op ’n sillabe uitgaan wat met ’n **l** begin. Krüger (2006b) sê in sy kommentaar op die ontleding in voorbeeld (250) dat diminutiewe van woorde wat op **l** + voorvokaal eindig gewoonlik **-jana** word, maar dat **l** + agtervokaal na **lw** verander. Na konsultasie met ’n moedertaalspreker is bevind dat die vorm **letlhajana**, wat die lemma-identifiseerder dan ook reg na **letlhale** herlei, verkies word. Ondersoek sal op ’n verteenwoordigende korpus gedoen moet word om te bepaal of daar meer sulke gevalle in die taal is, en of hierdie net ’n uitsondering op die reël is.

Die lemma van **bahutsanyana** (in voorbeeld (252)) is **mohutsi** (‘’n vervloekte mens/arm mens’), wat nie met behulp van die bestaande reël $([?^*, [a, n, y, a, n, a]: [a]])$ bepaal kan word nie. Die naamwoordmodule lewer die resultaat ***mohutsa**. ’n Soortgelyke voorbeeld is die van **kgetsanyana** (lemma is **kgetsi**) in voorbeeld (253). Volgens Krüger

(2006b) is **-anyana** 'n verkorting van **-ananyana**, wat 'n kombinasie van **-ana** en **-nyana** is. Daar is nie eksplisiete reëls daarvoor in Krüger (2006a) nie, maar vir die toekoms kan die reëls $([?^*, [a, n, y, a, n, a]: [i]])$ en $([?^*, [a, n, y, a, n, a]: [e]])$ ook by die lemma-identifiseerder se reëls gevoeg word.

Die lemma van **tlhotswana** (in voorbeeld (254) is **tlhopho** ('groep/keuse'), maar die reëls maak net voorsiening vir die klankverandering by voorvokale as **-ana** by die naamwoord gevoeg word waar die laaste sillabe op **ph** uitgaan $([?^*, [t, š, h, w, a, n, a]: [p, h, i]])$ (vergelyk Krüger, 2006a:86). Dieselfde is by voorbeeld (255) die geval, waar daar nie voorsiening gemaak is dat **-tšhwana** weer **-pha** kan word nie, en so kan **setlhotšhwana** ('groepie/spannetjie') se lemma **setlhopha** ('groep/span') ook nie geïdentifiseer word nie. In die toekoms kan die reëls $[?^*, [t, š, h, w, a, n, a]: [p, h, o]]$ en $[?^*, [t, š, h, w, a, n, a]: [p, h, a]]$ ook in die lemma-identifiseerder geïmplementeer word.

Indien die reëls wat vir hierdie onvoorsiene klankveranderings voorsiening maak, bygewerk word, sal die persentasie van 78,45% (met die uitgebreide lemmalys) waarskynlik na 78,90% kan opskuif.

- **Leenwoorde**

Sommige leenwoorde veroorsaak ook foute, soos dit duidelik uit die voorbeelde (255-256) blyk. As die leenwoorde ook by die lemmalys bygewerk kan word, sal dit ook dié foute uit die weg kan ruim. **Materiki** (voorbeeld (257)) sal dus nie meer as 'n meervoud in klas 3 gesien word nie, maar as 'n leenwoord in klas 5, wat reeds 'n lemma is. Die persentasie kan in dié geval moontlik van 78,90% na 79,02% verbeter.

- **Samestellings**

Samestellings wat uit 'n naamwoord en adjektief bestaan, is dikwels verkeerd geanaliseer – veral dié wat na enkelvoud verander moes word (voorbeeld (258)). Dit is te wyte aan die veranderings wat binne die woord moes plaasvind en waarvoor daar op hierdie stadium nie morfologiese reëls in die sisteem ingewerk is nie. Al was daar nie so baie van dié soort samestellings in die steekproef vir die evaluering nie (in totaal 7 uit 1000 naamwoorde), sou dit goed wees om 'n ondersoek op 'n verteenwoordigende korpus te doen om te bepaal hoe gereeld dié soort samestelling in die taal voorkom. Indien die voorkoms hoog is, sou dit in hierdie opsig handig wees om 'n kompositumanaliseerder vir Setswana te ontwikkel. Indien 'n kompositumanaliseerder toegepas sou word, sou die persentasie moontlik van 79,02% na 79,14% verbeter het.

Sommige samestellings kan weer nie in die enkelvoud gebruik word nie omdat dit nie semanties moontlik is nie, soos dit in voorbeeld (259) die geval is. Hierdie soort woorde moet daarom ook in die lemmalys opgeneem word (In die toetsdata was daar net een geval van 'n samestelling wat nie in die enkelvoud gebruik word nie.)

- **Naamwoordklasverskuiwing (“nominal shift”)**

Volgens Krüger (2006a:101) is daar twee soorte verskuiwings by die naamwoord: in die eerste plek is daar die verskuiwings tussen die nie-infinitiewe naamwoordklasse self en tweedens tussen die infinitiewe naamwoordklas (*go*-klas) en nie-infinitiewe naamwoordklasse (dit staan ook as die vorming van deverbatiwe bekend). Die eerste soort verskuiwing is egter hier van belang. 'n Bekende verskuiwing is dié tussen klas 3 (*le-/ma-*) en klas 6 (*lo-/di-*), wat dan ook in voorbeeld (260) die geval is (Krüger, 2006a:104). Normaalweg sou die klasprefiks *di-* na *lo-* moes verander, maar hier verskuif die woord na klas 3 toe (vergelyk Cole, 1955:91-93). In voorbeeld (261) is klas 3 se meervoudprefiks *ma-* 'n variante prefiks vir klas 5

se meervoudprefiks *di-* (Krüger, 2006a:108). Daar is nie in die reëls vir naamwoordklasverskuiwings voorsiening gemaak nie. Indien die reëls, wat voorsiening maak vir naamwoordklasverskuiwings, bygewerk word, sal die persentasie waarskynlik van 79,14% na 79,26% opskuif.

- **Dubbele klasprefikse**

Daar is ook vreemde verskynsels soos hierdie, waar daar dubbele klasprefikse is. Dit blyk hier uit voorbeeld (262) dat die *bo*-klasprefiks verwyder moet word en die meervoudprefiks *ma-* na *bo-* toe verander moet word om die regte lemma te identifiseer. Die naamwoordmodule het dit egter as 'n woord in klas 7 beskou, wat nie hoef te verander nie. Dubbele prefikse kan volgens Cole (1955:104) aangetref word by woorde soos **bomodimo** (voorbeeld (265)) en **bomorafe** (voorbeeld (266)). Dit is egter nie iets wat dikwels voorkom nie, en kan daarom as uitsonderings in die lemmalys hanteer word. Die resultate sou dan van 79,26% na 79,30% verbeter.

(265) **bomodimo**

bo-mo-dimo

CL7.SG-CL2.SG-god

'godheid'

(266) **bomorafe**

bo-mo-rafe

CL7.SG-CL2.SG-nasie

'nasieskap'

- **Homograwe**

Soos hierbo genoem, is dit soms moeilik om die lemmas sonder die breër konteks te bepaal. Die woord **mokong** kan 'in die murg' of 'steil opdraende' beteken; in voorbeeld (263) is **mokong** se lemma dan **moko** ('murg'), maar in voorbeeld (264) is dit **mokong** 'steil opdraende/steilte'. Dit kan opgelos word

deur albei woorde as lemma te gee en die keuse aan die gebruiker oor te laat, of om 'n kontekssensitiewe lemma-identifiseerder te ontwikkel. Indien hierdie probleme uit die weg geruim is, sou die lemma-identifiseerder se resultate van 79,30% tot 79,34% kon verbeter.

Indien die probleme soos hierbo bespreek sinvol opgelos kan word, sal dit 'n moontlike verbetering in die akkuraatheid van 76,29% na 79,35% beteken. Vervolgens word die resultate van die werkwoordmodule bespreek.

4.4.2 Werkwoorde

Net soos by die naamwoorde kan sommige van die foute by werkwoorde toegeskryf word aan 'n ontoereikende lemmalys, onvoorsiene klankveranderinge, die komplekse aard van die verdubbelings en samestellings, dubbele suffikse, homograwe, meerduidige uitgange, verandering in die reëlvolgorde en uitsonderings (vergelyk Tabel 4.3). Hierdie verskynsels word vervolgens aan die hand van die voorbeelde in Tabel 4.5 bespreek.

Nommer	Toevoer	Verwagte afvoer	Afvoer deur module
267	ntlhologelela <i>n-tlhologelel-a</i> AGR.OBJ-verlang-TERM 'verlang na my'	tlhologelela	*dlhologela
268	thanya <i>thany-a</i> wakker word-TERM 'wakker word'	thanya	ta ('n ideofoon wat algehele vernietiging aandui)

269	thamela <i>thamel-a</i> vassit-TERM 'vassit'	thamela	thama <i>tham-a</i> weer moet te skep- TERM 'weer moet te skep'
270	gwaega <i>gwaeg-a</i> gekrap word-TERM 'gekrap word'	gwaega	gwaa (n ideofoon wat 'n harde bui reën aandui)
271	nthole <i>n-rol-e</i> OBJ-uittrek (van skoene/handskoene/hoed)- TERM 'my skoene/handskoene/ hoed uittrek'	rola	thole CL5.SG.stof 'stof'
272	itetlile <i>i-letl-il-e</i> REFL-toelaat-PERF-TERM 'het self toegelaat'	letla	tetla CL5.SG.toestemming 'toestemming'
273	lwetseng <i>lwal-il-e-ng</i> siek wees-PERF-TERM-TERM 'wat siek was' (kan ook 'het baklei vir' beteken)	lwala	Lwetse 'September'
274	ntebane <i>n-leb-an-e</i> AGR.OBJ-draai na-RECP-TERM 'kom my toe/behoort aan my'	leba <i>leb-a</i> <i>draai na-TERM</i> 'draai na/kyk'	teba <i>teb-a</i> gelykmatig raak- TERM 'gelykmatig raak'

275	tlhorontshitseng (> tlhoro- ntshitse > tlhorontsha) <i>tlhor-is-il-e-ng</i> kwel-CAUS-PERF-TERM-TERM 'wat besorg/gekwel/getreiter/ ongelukkig gemaak het'	tlhora <i>tlhor-a</i> kwel- TERM 'kwel'	*tlhorona
276	utlwisisa <i>utlw-is-is-a</i> hoor-CAUS-CAUS-TERM 'goed verstaan of hoor'	utlwa <i>utlw-a</i> hoor-TERM 'hoor'	*utlwisa
277	tlhabeletse <i>tlhab-el-el-a</i> steek-APPL-APPL-TERM 'n lied insit'	tlhaba <i>tlhab-a</i> steek-TERM 'steek'	*tlhabelela
278	nteseletsa <i>n-les-el-el-a</i> AGR.OBJ-alleen laat-APPL- APPL-TERM 'my heeltemal uitlos'	lesa <i>les-a</i> alleen laat-TERM 'alleen laat'	*desela
279	ntshireletsang <i>n-sir-el-el-y-a-ng</i> AGR.OBJ-uitsig belemmer- APPL-APPL-CAUS-TERM- TERM 'wat my beskerm'	sira <i>sir-a</i> uitsig belemmer- TERM 'uitsig belemmer'	sirela <i>sir-el-a</i> uitsig belemmer- APPL-TERM 'uitsig belemmer vir'
280	omeleditswe <i>om-el-el-y-il-w-e</i> droog word-APPL-APPL-CAUS- PERF-PASS-TERM 'is heeltemal droog gemaak'	oma <i>om-a</i> droog word-TERM 'droog word'	omela <i>om-el-a</i> droog word-APPL- TERM 'droog word vir'

281	kabetsa ¹³ <i>kabal-y-a</i> toegeprop-CAUS-TERM 'laat toestop/verseël/toemaak vir'	kabala <i>kabal-a</i> toegeprop-CAUS-TERM 'toegeprop raak'	kabela <i>kabel-a</i> plant-TERM 'plant met behulp van 'n skoffelpik'
282	kabetsa <i>kabets-a</i> 'n uier wat min melk het, leegmelk-TERM 'n uier wat min melk het, leegmelk	kabetsa	kabela
283	arametsa <i>aramel-y-a</i> in die son koester/stoom onder 'n kombers-CAUS-TERM 'in die son laat koester/laat stoom onder 'n kombers'	aramela <i>aramel-a</i> in die son koester/stoom onder 'n kombers- TERM 'in die son koester/stoom onder 'n kombers'	*arama
284	thobuetsa <i>thobuel-y-a</i> in vloeistof inval-CAUS-TERM 'laat inplons/in vloeistof dompel'	thobuela <i>thobuel-a</i> in vloeistof inval-TERM	*thobua

¹³ **Kabetsa** kan ook só ontleed word: *kab-el-y-a*, toeprop-APPL-CAUS-TERM, 'laat toestop/verseël/toemaak vir'. In dié geval is daar 'n verandering in die reëlvolgorde.

285	kaleditsweng (> kaleditswe > kaleditse > kaletsa) <i>kalel-il-y-w-e-ng</i> hang-CAUS-PERF-PASS-TERM-TERM 'wat opgehang is'	kalela <i>kalel-a</i> hang-TERM 'hang'	*kala
286	belaela <i>belael-y-a</i> twyfel-CAUS-TERM 'verdink/ongemaklik laat voel'	belaela <i>belael-a</i> twyfel-TERM 'verdink/ongemaklik voel'	*belaa
287	tlhatsweng <i>tlhatsw-e-ng</i> was-TERM-TERM 'wat iets nie-lewends was'	tlhatswa <i>tlhatsw-a</i> was-TERM 'iets nie-lewends was'	tlhala <i>tlhal-a</i> skei-TERM 'skei'
288	utlwe <i>utlw-e</i> hoor/voel/proe/ruik-TERM 'hoor/voel/proe/ruik'	utlwa <i>utlw-a</i> hoor/voel/proe/ruik-TERM 'hoor/voel/proe/ruik'	utla <i>utl-a</i> gryp-TERM 'gryp'
289	tshwantshanngwe <i>tshwan-is-an-y-w-a</i> lyk na/eenders wees-CAUS-RECP-CAUS-PASS-TERM 'met mekaar vergelyk word'	tshwana <i>tshwan- a</i> lyk na/eenders wees- TERM 'lyk na/eenders wees'	*tshwantshanma

290	menolwe <i>menol-iw-e</i> omverwerp-PASS-TERM 'word omgedop'	menola <i>menol-a</i> omverwerp-TERM 'omverwerp'	*menwa
291	buisiwe <i>bu-is-iw-e</i> praat-CAUS-PASS-TERM 'gelees word'	bua <i>bu-a</i> praat-TERM 'praat'	*buisia
292	utlwiwe <i>utlw-iw-e</i> hoor/voel/proe/ruik-PASS-TERM 'word gehoor/gevoel/geproe/geruik'	utlwa <i>utlw-a</i> hoor/voel/proe/ruik-TERM 'hoor/voel/proe/ruik'	*utlwia
293	fetsanye <i>fel-y-an-y-a</i> klaarmaak-CAUS-RECP-CAUS-TERM 'mekaar klaar laat maak/mekaar dood maak'	fela <i>fel-a</i> klaarmaak-TERM 'klaarmaak'	fetsana <i>fel-y-an-a</i> klaarmaak-CAUS-RECP-TERM 'mekaar klaar maak'
294	matologanya <i>matolog-an-y-a</i> losraak-RECP-CAUS-TERM 'mekaar laat losraak'	matologa <i>matolog-a</i> losraak-TERM 'losraak'	matologana <i>matolog-an-a</i> losraak-RECP-TERM 'van mekaar losraak'
295	ntshwaraganya <i>n-tshwar-eg-an-y-a</i> AGR.OBJ-vang-N.PASS-RECP-CAUS-TERM 'my verenig/saamvoeg'	tshwara <i>tshwar-a</i> vang-TERM 'vang'	*swaragana

296	belаетsegang <i>belael-y-eg-a-ng</i> twyfel-CAUS-N.PASS- TERM-TERM 'wat laat twyfel/betwyfel/ verdink'	belaela <i>belael-a</i> twyfel-TERM 'wat laat twyfel/ betwyfel/verdink'	belаетsa <i>belael-y-a</i> twyfel-CAUS-TERM 'laat twyfel'
297	tsibiketsega <i>tsibikel-y-eg-a</i> duiselig raak-CAUS- N.PASS-TERM 'laat duiselig raak/naar raak'	tsibikela <i>tsibikel-a</i> duiselig raak-TERM 'duiselig raak/naar raak'	tsibiketsa <i>tsibikel-y-a</i> duiselig raak-CAUS- TERM 'laat duiselig raak/naar raak'
298	dirisege <i>dir-is-eg-e</i> werk-CAUS-N.PASS-TERM 'bruikbaar wees/gebruik kan word'	dira <i>dir-a</i> werk-TERM 'werk'	dirisa <i>dir-is-a</i> werk-CAUS-TERM 'gebruik'
299	porometsega <i>porom-el-y-eg-a</i> bulder/bas sing-APPL- CAUS-N.PASS-TERM 'bulderbaar vir/bas singbaar vir'	poroma <i>porom-a</i> bulder/bas sing- TERM 'bulder/bas sing'	porometsa <i>porom-el-y-a</i> bulder/bas sing- APPL-CAUS-TERM 'laat bulder/bas sing vir'
300	goelega <i>go-el-eg-a</i> hard skreeu-APPL-N.PASS- TERM 'hard skreeubaar vir'	goa <i>go-a</i> hard skreeu-TERM 'hard skreeu'	goela <i>go-el-a</i> hard skreeu-APPL- TERM 'hard skreeu vir'

301	nosa <i>nw-is-a</i> drink-CAUS-TERM 'laat drink/besproei'	nwa <i>nw-a</i> drink-TERM 'drink'	noga CL5.SG.slang 'slang'
-----	---	--	--

Tabel 4.5 Ontledings van die egtewerkwoordmodule se resultate

- 'n Ontoereikende lemmalys

Die sisteem is hier ook, soos dit by die naamwoordmodule bespreek is, afhanklik van die volledigste lemmalys moontlik. Dit is veral van belang by die klankveranderinge van die refleksief (*i-*) en die eerste persoon enkelvoud se voorwerpsmorfeem (*n-*), wat baie moontlikhede kan oplewer. Die kombinasie **nt** kan byvoorbeeld **t**, **l** of **d** word, soos in die geval van voorbeeld (267) (vergelyk die bespreking onder 3.4.2.1.1). Soms beteken dit dan dat dele verwyder is wat soos suffikse lyk, maar deel is van die woord, soos dit in voorbeeld (268) - (270) geïllustreer word. 'n Uitgebreide lemmalys het die akkuraatheid op komplekse werkwoorde van 64,50% tot 70,52% verhoog en die akkuraatheid op alle werkwoorde (kompleks en simpleks) van 73,95% tot 79,59% verhoog (vergelyk Tabel 4.1).

Soms gebeur dit ook dat een van die stappe se resultate in die lemmalys voorkom, hoewel dit nie die lemma van die betrokke woord is nie. Die woord **nthole** in voorbeeld (271) se lemma is **rola**, maar met die uitgebreide lemmalys stop dit by **thole** ('stof') wat in die uitgebreide lemmalys opgeneem is. Dit sou reg geïdentifiseer gewees het as dit só ontleed is: **nthole** > **nthola** > **rola**. Ander soortgelyke gevalle word in voorbeelde (272) - (274) geïllustreer. Meerdere afvoermoontlikhede waaruit 'n gebruiker kan kies, sal in hierdie geval 'n oplossing wees.

- **Onvoorsiene klankveranderings**

Die kousatiefreëls (vergelyk Krüger (2006a:223,228) maak voorsiening dat die uitgang **-ntsha** verander kan word na **-na**

([?*, {[n, y, a], [n, t, s, h, a]}:na]), maar daar is ook 'n uitsondering teëgekom waar die **-ontsha** na **-a** moes verander (voorbeeld (275)). Hierdie probleem kan opgelos word deur [?*, {[o, n, t, s, h, a]}:a] as 'n reël te implementeer.

Daar is nie voorsiening gemaak vir die imperatiewe uitgänge van die eenlettergrepige werkwoorde nie, omdat dit nie in Setswana-morfologie formeel as suffigse beskou word nie (Cole (1955:240) verwys daarna as suffigale elemente). Die suffigale element **-a** van **kgwaa** ('spoeg!/braak!') word daarom nie verwyder om die lemma **kgwa** ('spoeg/braak') te identifiseer nie. Die eenlettergrepige werkwoorde ('n beperkte lys) met imperatiewe uitgänge kan in die toekoms gelys word en as uitsonderings deur die sisteem hanteer word. Die suffigale elemente **-a** is 'n saak wat verder ondersoek kan word; toetse vir morfeemstatus kan byvoorbeeld aangelê word om te wys dat dit wel 'n morfeem is.

Indien bogenoemde reël en lys bygewerk word, kan die persentasie van 78,45% na 78,46% verbeter. Omdat dit onvoorsiene klankveranderings is wat net in uitsonderlike gevalle voorkom, was daar net twee voorbeelde in die toetsdata.

- **Verdubbelings en samestellings**

Daar is nog nie klarigheid oor hoe die verdubbelings en samestellings by werkwoorde hanteer moet word nie. Moet 'n woord soos **tsamaisatsamaisa** ('administreer/vinniger loop') na ***tsamaisatsamaya** of **tsamaya** ('loop') verander word? Uit die bespreking in Hoofstuk 2 volg dat die verdubbelings, soos hierdie, as lemma aanvaar sal word. Dit is dan uiteraard nodig dat die verdubbelings en samestellings ook in die lemmalys opgeneem word, of

alternatiewelik moet 'n reduplikasie-analiseerder ontwikkel word. Soos dit by die naamwoorde die geval was, is daar nie baie van dié soort verdubbelings in die steekproef nie (5 uit 'n totaal van 1000 naamwoorde), maar dit sou goed wees om 'n ondersoek op 'n verteenwoordigende korpus te doen om te bepaal hoe produktief die verdubbelingsprosesse is. Indien dit produktief voorkom, sou dit goed wees om 'n kompositum- en reduplikasie-analiseerder vir Setswana te ontwikkel. Die resultate kan, as die probleme ondervang is, van 78,46% tot 78,50% verbeter.

- **Dubbele suffikse**

Dubbele kousatiewe kom selde voor (in die toetsdata was daar slegs die een geval), maar **utlwisisa** ('om goed te verstaan of te hoor'), wat 'n dubbele kousatief het, se lemma is **utlwa** (vergelyk voorbeeld (276). (Daar is egter nie in die reëls voorsiening gemaak daarvoor nie, omdat dit volgens die grammatikas nie dikwels voorkom nie – vergelyk Krüger (2006a:232) en Cole (1955:209).)

Daar is ook 'n voorkoms van die dubbele applikatief in voorbeelde (277) - (280) (wat nie met die reduplikasievorm *-e/el-* van Cole (1955:203) verwar moet word nie: vergelyk Krüger (2006a:239) se opmerkings daaroor). Volgens Krüger (2006b) het die verdubbeling van die applikatief die semantiese waarde van intensifisering.

Die bestaande reëls vir kousatiewe en applikatiewe suffikse kan twee keer uitgevoer word om die gevalle te hanteer. Indien dit toegepas word, kan die persentasie moontlik van 78,50% na 79,27% verbeter.

- **Homograwe**

Soos wat dit by die naamwoorde die geval is, is dit by die werkwoorde ook soms moeilik om die lemmas sonder die breër konteks te bepaal. Die woord **kabetsa** ('n uier wat min melk het, leegmelk') is 'n lemma, maar dit lyk ook

soos die kousatiewe vorm van **kabala** ('toegeprop raak') (voorbeelde (281) – (282)). Die probleem kan opgelos word deur meerdere afvoermoontlikhede en deur die keuse aan die gebruiker oor te laat, of om 'n kontekstsensitiewe lemma-identifiseerder te ontwikkel. Indien hierdie probleme uit die weg geruim is, sou die werkwoordmodule se resultate van 79,27% tot 80,04% kon verbeter.

- **Meerduidige uitgange**

Die uitgang **-etsa** kom by die applikatief en kousatief voor. Soms gebeur dit dat **-etsa** as kousatiewe uitgang verwyder moet word, maar omdat die applikatiefreël $[?*, [e, t, s, a] : a]$ eerste uitgevoer word, lei dit daartoe dat die lemma verkeerd geïdentifiseer word. Dus: in plaas daarvan dat die reël $[?*, \{[d, i, s, a], [t, s, a]\} : [1, a]]$ uitgevoer word, verander **-etsa** na **-a** soos dit in die voorbeelde (283) - (284) geïllustreer word (vergelyk die bespreking van die applikatief (3.4.2.1.8) en kousatief (3.4.2.1.9)). Die probleem kan opgelos word deur die werkwoord by die kousatiefreël te neem soos wat dit voor die toepassing van die applikatiefreël gelyk het (as die werkwoord nie 'n lemma is na die uitvoer van die applikatiefreël nie).

In voorbeelde (285) - (292) word die uitgang **-we** dikwels verkeerdelik as 'n passiewe uitgang in die perfektum gesien, en dit lei daartoe dat die lemma verkeerd geïdentifiseer word (vergelyk die bespreking van die passief in die perfektum onder 3.4.2.1.3). In voorbeeld (286) word die volgende roete gevolg: **tlhatsweng** > **tlhatswe** > **tlhatse** > **tlhala**. Die **-tse** word as 'n perfektumuitgang gesien en word daarom na **tlhala** verander na aanleiding van die reël $[?*, ? \sim \{i, n, e\}, [t, s, e] : [1, a]]$. In voorbeeld (287) word die **-we** in **utlwe** as 'n passiewe uitgang beskou, en is die resultaat daarom **utla** in plaas van **utlwa**.

Die reël $[?*, [n, g, w, e] : [m, e]]$ word byvoorbeeld (288) toegepas en dit lei daartoe dat **tshwantshannngwe** verander word na ***tshwantshanme** en

dat die 'uitgangsmorfeem' -e na 'n -a verander word en die lemma as ***tshwantshanma** geïdentifiseer word. Die regte stappe sou gewees het dat die -a eers 'n -e word en dat die passief daarna uitgehaal word: **tshwantshanngwe > tshwantshanngwa > tshwantshanya > tshwana**. In voorbeeld (289) word die passiewe -w- verwyder uit wat soos die 'perfektumuitgang' -ole lyk, wat na aanleiding van die reël [?* , [o, ɪ, e] : [w, a]] weer -wa word en ***menwa** as resultaat lewer. Ook in voorbeelde (290) - (291) lyk die uitgange soos 'n passiewe suffiks in die perfektum en word die passiewe suffiks uit die 'perfektumuitgang' verwyder, in plaas daarvan dat die uitgangsmorfeem -e eers 'n -a word en die passiewe suffiks daarna verwyder word. Die foute kan in al bogenoemde gevalle aan die sisteem, en nie aan die grammatikas nie, toegeskryf word. Dit kan opgelos word deur die werkwoord by die volgende reël te neem soos wat dit voor die toepassing van die passiefreël gelyk het (as die werkwoord nie 'n lemma is na die uitvoer van die passiefreël nie). Die resultaat kan moontlik na toepassing van die maatreëls van 80,04% tot 86,79% verbeter.

- **Verandering in die reëlvolgorde**

Menoganya¹⁴ (menog-an-y-a; omdraai-RECP-CAUS-TERM; 'omrol/laat kreun van pyn') se lemma is **menoga** ('omdraai'), maar die lemma-identifiseerder stop by **menogana** ('n paar maal omslaan'). Daar blyk 'n verandering in die gewone volgorde van Krüger (2006a:257) te wees, waar die resiprokale suffiks voor die kousatiewe suffiks verwyder word. In hierdie geval moet eers die kousatiewe en dan die resiprokale suffiks verwyder word (vergelyk ook soortgelyke voorbeelde (293) - (295)). Behalwe dat die volgorde in voorbeeld (292) omgeruil is, is daar ook 'n ekstra kousatiewe suffiks by. Volgens Krüger (2006b) kom dié omgekeerde volgorde (resiprokalis + kousatief) "uiters selde (indien ooit)" voor; tog is dit 'n voorval wat verder ondersoek kan word.

¹⁴ **Menoga** sou in 'n morfologiese analise verder tot **mena** ('vou') ontleed kon word, indien die rewersiewe suffiks -og- verwyder word. Dit is egter nie van toepassing op die lemma-identifiseerder nie (vergelyk die bespreking onder 2.4.2.1).

In die gewone volgorde van Krüger (2006a:257) moet die kousatiewe suffiks voor die neutro-passiewe suffiks afgehaal word. Dit was egter nie die geval in voorbeelde (296) - (298) waar die volgorde omgeruil het nie. Die volgorde moet in dié gevalle aangepas word om die lemma te identifiseer (d.i. die neutro-passiewe moet voor die kousatiewe suffiks afgehaal word).

Dieselfde verandering in volgorde het ook by die kombinasie van die applikatiewe en neutro-passiewe suffikse plaasgevind. In die gewone volgorde van Krüger (2006a:257) moet die kousatiewe en applikatiewe suffiks voor die neutro-passiewe suffiks afgehaal word. In voorbeelde (299) - (300) moet die volgorde aangepas word om die lemma te identifiseer (d.i. die neutro-passiewe moet voor die kousatiewe/applikatiewe suffiks afgehaal word).

'n Soektog kan op 'n verteenwoordigende korpus gedoen word om te bepaal of daar meer sulke gevalle in die taal is en of dit 'n nuwe tendens in Setswana is. Indien daar aanduibare veranderings in die reëlvolgorde is, sal die volgorde in die lemma-identifiseerder aangepas moet word. Indien dit egter net die twintig gevalle is wat in die steekproef van die werkwoorde voorgekom het, sal dit as uitsonderings hanteer kan word. Op welke manier die probleem ook al opgelos word, sal dit 'n moontlike verbetering van 86,79% tot 88,50% beteken.

- **Uitsonderings**

Hoewel daar vooraf vir uitsonderings, soos deur Krüger (2006a:230-231) genoem, in die kousatiefreëls voorsiening gemaak is, moet voorbeeld (301) nog by die uitsonderings gevoeg word (vergelyk die bespreking oor die uitsonderings by die kousatief onder 3.4.2.1.9). Ander moontlike uitsonderings sou op dieselfde manier hanteer kon word, wat 'n verbetering van 88,50% tot 88,67% tot gevolg sal hê.

Deur 'n volledige lemmalys saam te stel en probleme met onvoorsiene klankveranderings, verdubbelings en samestellings, dubbele suffikse, homograwe, meerduidige uitgange en die verandering in die reëlvolgorde en uitsonderings op te los, word die akkuraatheid van 73,95% na 'n moontlike 88,67% opgeskuif.

4.4.3 Verskillende woordsoorte

Die volledige lemma-identifiseerder is met 'n ander datastel as by die naamwoord- en egtewerkwoordmodule geëvalueer (vergelyk die bespreking onder 4.3.1); waar dieselfde patrone waargeneem is as in die naamwoord- en egtewerkwoordmodule (dit het ook dieselfde persentasietendense getoon). Die fokus in die bespreking wat volg is daarom op enkele nuwe voorbeelde en die resultate as gevolg van die volgorde van die modules. Daar is egter geen probleme met die ander woordsoorte ondervind nie. Daar is dus 'n 100% akkuraatheid op die ander woordsoorte.

Dit was veral die werkwoorde wat die grootste aantal foute veroorsaak het, en hier kan die oorsaak weer aan die lemmalys, maar ook aan die sisteem toegeskryf word. As die toevoerwoord se lemma nie in die egtewerkwoordmodule van die sisteem bepaal kan word nie – dus nie in die lemmalys opgespoor word nie – word daar aangeneem dat dit 'n naamwoord is. Al sou dit dan ook na die regte lemma verander wees, maar dit was nie in die lemmalys nie, word die toevoer, soos wat dit aanvanklik aan die sisteem gegee is, aan die naamwoordgedeelte van die sisteem gegee. Dit is om te verseker dat veranderings wat deur die egtewerkwoordmodule aan 'n naamwoord in die sisteem aangebring is, nie na die naamwoordmodule oorgedra word nie. Dit het egter tot gevolg dat, as die werkwoordlemma nie in die lemmalys voorkom nie, al die veranderings aan die woord tot niet gemaak word omdat die sisteem dit dan as 'n naamwoord beskou. Die werkwoorde bly as gevolg van hierdie werkwyse meestal onveranderd. Dit is dus hier van kritiese belang dat die lemmalys so volledig moontlik moet wees. Die akkuraatheid op al die woordsoorte het

daarom met die uitgebreide lemmalys van 68,54% tot 75,92% verhoog (vergelyk Tabel 4.1). Die res van die foute kan, soos by die naamwoord- en egtewerkwoordmodule, onder andere toegeskryf word aan onvoorsiene klankveranderings, verdubbelings en samestellings, meerduidige uitgange en wisselvorme. Die oorsake van die foute word aan die hand van die voorbeelde in Tabel 4.6 bespreek.

Dit blyk uit die resultate (vergelyk Tabel 4.1) dat die linguistiese benadering nie noodwendig die juiste benadering is om lemma-identifisering te hanteer nie. Hoewel die morfologiese patrone van die naamwoorde en werkwoorde redelik reëlmatig is en die Setswanagrammatikas goed gedokumenteerd is, is daar te veel veranderlikes wat die resultate beïnvloed het. Masjienleertegnieke sal in die opsig 'n alternatief wees om te ondersoek. Die ontwikkelingswerk wat met die lemma-identifiseerder gedoen is, is egter nie verniet nie. Hierdie lemma-identifiseerder kan in skoelussteekeproefneming ("bootstrapping") gebruik word. Dit beteken dat afrigtingsdata deur middel van die proses saamgestel kan word vir 'n lemma-identifiseerder wat met masjienleertegnieke ontwikkel word. Dus: in plaas daarvan om woorde se lemmas met die hand te identifiseer om afrigtingsdata saam te stel, kan meer afrigtingsdata (wat net gekontroleer moet word) met behulp van die linguistiese lemma-identifiseerder saamgestel word. Op hierdie manier kan 'n lemma-identifiseerder in die masjienleergebaseerde benadering inkrementeel verbeter word.

Nommer	Toevoer	Verwagte afvoer	Afvoer deur module
302	bolaesegile <i>bolay-eseg-il-e</i> doodmaak-N.PASS-PERF-TERM 'was doodmaakbaar'	bolaya <i>bolay-a</i> doodmaak-TERM 'doodmaak'	*laesegile

303	montsintsi <i>mo-ntsi-ntsi</i> CL1.SG-baie-baie 'baie-baie'	<i>-ntsintsi</i> 'baie-baie'	montsintsi
304	bomaratahelele <i>bo-ma-rata-helele</i> CL1-CL7-hou van- klanknaboetsing 'nuuskierige agie/persoon wat afluister as mense praat'	bomaratahelele	*maratahelele
305	tsetsepetsa <i>tsetsepel-is-a</i> vasstaan-CAUS-TERM 'laat vasstaan'	tsetsepela	tsetsepetsa
306	tlhanole <i>tlhanol-e</i> omdop-TERM 'omdop'	tlhanola <i>tlhanol-a</i> omdop-TERM 'omdop'	tlhanole
307	tsholeng <i>tshol-e-ng</i> dra/inskep/hanteer-TERM- TERM 'wat dra/inskep/hanteer'	tshola <i>tshol-a</i> dra/inskep/hanteer- TERM 'dra/inskep/hanteer'	tshole <i>tshol-e</i> dra/inskep/hanteer- TERM 'dra/inskep/hanteer'
308	tsiseditse <i>tl-is-el-il-e</i> kom-CAUS-APPL-PERF- TERM 'het laat kom vir'	tla <i>tl-a</i> kom-TERM 'kom'	tsa 'n possessiewe betrekkingswoord

Tabel 4.6 Ontledings van die volledige lemma-identifiseerder se resultate

- **Onvoorsiene klankveranderings**

Daar is geen voorsiening deur Krüger (2006a:208-214) gemaak dat die uitgange **-esega** na **-ya** verander kan word nie. Dit veroorsaak dat die woord **bolaesegile** in voorbeeld (302) se lemma nie geïdentifiseer kan word nie.

Die lemma-identifiseerder kan verbeter word deur die reël

$[?*, [e, s, e, g, a] : [y, a]]$ te implementeer. Die lemma-identifiseerder beskou dit egter in dié geval as 'n naamwoord, en verwyder die **bo-** wat soos 'n klasprefiks lyk.

- **Verdubbelings en samestellings**

Daar is in Hoofstuk 3 (3.4.1.2) besluit om die beskrywende naamwoordwortels (wat betreklik min is) in 'n lys te hanteer. Daar is egter nie voorsiening gemaak dat die wortel ook kan verdubbel nie, soos in (303). 'n Samestelling soos in voorbeeld (304) kan net so in die lemmalys opgeneem word, of alternatiewelik kan 'n kompositum- en reduplikasie-analiseerder ontwikkel word (vergelyk die bespreking onder 4.4.1 en 4.4.2).

- **Meerduidige uitgange**

Soos dit by die resultate van die egtewerkwoordmodule ook bespreek is, kom die uitgang **-etsa** by die applikatief en kousatief voor. Só gebeur dit dan dat **-etsa** as kousatiewe uitgang verwyder moet word, maar omdat die applikatiefreël eerste uitgevoer word, lei dit daartoe dat die regte lemma nie geïdentifiseer word nie. Omdat dit nie 'n lemma is nie, word dit weer na die toevoerwoord verander (dus net soos dit voor die proses gelyk het; vergelyk voorbeeld (305)).

In voorbeeld (306) word die gedeelte wat soos die 'perfektumuitgang' **-ole** lyk, na aanleiding van die reël $[?*, [o, l, e] : [w, a]]$ verkeerdelik verander na **-wa**, en word die lemma as ***tlhanwa** geïdentifiseer. Omdat dit nie 'n lemma is nie, word dit weer na **tlhanole** verander (dus net soos dit voor die proses gelyk het). Dieselfde gebeur in voorbeeld (307), behalwe dat die **-ng** by die

naamwoordmodule verwyder word. Die woord word derhalwe verkeerderlik as naamwoord geanaliseer.

- **Wisselvorme**

Die wisselvorm van **tliseditse** in voorbeeld (308) is **tsiseditse**, en omdat daar nie vir die wisselvorm voorsiening gemaak is nie, word die lemma verkeerd geïdentifiseer. Dit was egter die enigste wisselvorm wat in die toetsdata aangetref is. Foute soos hierdie kan voorkom word deur 'n lys van sodanige wisselvorme te maak en as uitsonderings te hanteer. Indien dit in die konteks van lemma-identifisering vir leksikografiese doeleindes sou plaasvind, is dit die werk van 'n menslike kontroleerder om te sorg dat hierdie foute nie as korrek aanvaar word nie.

4.4 Gevolgtrekking

Daar is in Hoofstuk 1 van die standpunt uitgegaan dat die bestaande Setswanagrammatikas effektief geformaliseer kan word om 'n outomatiese lemma-identifiseerder vir Setswana te ontwikkel. Tog het die naamwoord- en werkwoordreëls swakker resultate gelewer as wat verwag is. Hoewel die morfologiese struktuur van Setswana goed gedokumenteer is, is dit die eerste keer dat bestaande grammatikas op só 'n manier getoets is. Die evaluering in die hoofstuk dui daarop dat die akkuraatheid van die lemma-identifiseerder met die basiese lemmalys 68,97% op komplekse naamwoorde, 63,5% op komplekse werkwoorde, 76,29% op komplekse en simplekse naamwoorde, 73,95% op komplekse en simplekse werkwoorde, en 68,54% op alle woorde is. Dit blyk egter dat 'n goeie omvattende lemmalys die resultate aansienlik sal kan verbeter, want die akkuraatheid van die lemma-identifiseerder is met die uitgebreide lemmalys 70,96% op komplekse naamwoorde, 70,52% op komplekse werkwoorde, 78,45% op komplekse en simplekse naamwoorde, 79,59% op komplekse en simplekse werkwoorde, en 75,92% op alle woorde.

Die foute van die lemma-identifiseerder lê nie by die grammatika as sodanig nie, maar ook in die sisteem, soos met die voorbeelde aangedui is. Die ontwikkeling van 'n lemma-identifiseerder is daarom nie noodwendig die beste metode om grammatikas kwantitatief mee te evalueer nie, maar dit gee wel 'n aanduiding van die relatiewe akkuraatheid van die grammatikas (waar die beperkinge van die sisteem in gedagte gehou moet word, soos byvoorbeeld die beperkte lemmalys). Dit bied ook geanaliseerde data waarmee die grammatikas kwalitatief geëvalueer kan word (soos dit dan ook in die hoofstuk gedoen is).

Indien 'n volledige Setswanalemmalys (leenwoorde ingesluit) saamgestel kan word en onvoorsiene klankveranderings, naamwoordklasverskuiwings, dubbele klasprefikse, dubbele suffikse, meerduidige uitgange en die verandering in die reëlvolgorde ondervang kan word (soos aangedui in hierdie hoofstuk), kan die lemma-identifiseerder vir Setswana aansienlik verbeter word. Die verdubbelings en samestellings in Setswana is egter nog nie bevredigend hanteer nie, en daar is nog ruimte vir verdere navorsing oor dié aspek van die lemma-identifiseerder.

Die ontwikkeling van 'n lemma-identifiseerder met behulp van masjienleertegnieke kan egter ook ondersoek word. In so 'n geval sal die lemma-identifiseerder wat in hierdie studie ontwikkel is, gebruik kan word om deur middel van skoenlussteekproefneming 'n lemmalys vir afrigtingsdata te ontwikkel en aan te pas.

Hoofstuk 5

Slot

5.1 Inleiding

In Hoofstuk 1 van hierdie verhandeling is die konsep “lemma-identifisering” onder die loep geneem en is bepaalde probleme, navorsingsvrae en doelstellings geïdentifiseer. Ten einde ’n lemma-identifiseerder vir Setswana te ontwikkel, is daar in Hoofstuk 2, ná ’n deeglike bespreking van die woordmorfologie, bepaal wat as ’n lemma in Setswana beskou kan word. Met dié definisies as riglyn is die bestaande morfologiese analisereëls, wat die teenoorgestelde volgorde as woordvormingsreëls volg, in Hoofstuk 3 geformaliseer. Vervolgens is die reëlmatige uitdrukkings wat in die eindigestaatsprogrammatuur (FSA 6) geïmplementeer is, in ’n program geïnkorporeer om die lemmas outomaties te identifiseer. In Hoofstuk 4 is die lemma-identifiseringsstelsel bespreek en geëvalueer. Die oorsake van foute is aan die hand van voorbeelde bespreek.

Die belangrikste bevindinge van die studie is in hierdie hoofstuk saamgevat, waarna algemene gevolgtrekkings gemaak is (in antwoord op die navorsingsvrae van Hoofstuk 1). Ten slotte is vooruitskouings vir verdere navorsing bespreek.

5.2 Samevatting

Die volgende werksdefinisie vir lemma-identifisering is in die eerste hoofstuk aangebied: Lemma-identifisering is ’n natuurliketaalprosesseringsprosedure wat die korrekte basisvorm/lemma van ’n toevoerwoord bepaal deur die verwydering van fleksie-affikse. As sodanig word dit as ’n eenvoudige proses van morfologiese analise beskou wat vele toepassingsmoontlikhede in korpusgebaseerde navorsings- en ontwikkelingswerk het.

Uit die werksdefinisie volg dit dat lemma-identifisering met die konsep “lemma” verband hou. Wat as lemma beskou word, moes dus ook duidelik gedefinieer word. ’n Lemma word oor die algemeen gedefinieer as die

betekenisvolle basisvorm, gestroop van fleksiemorfeme (Choueka *et al.*, 2000:74; Mitkov, 2003:728; Trost, 2003:38). Die bepaling van die konsep "lemma" is egter, met inagneming van algemene teorieë daaroor, taalspesifiek, en voordat 'n lemma-identifiseerder vir Setswana ontwikkel kon word, moes deeglik besin word oor wat 'n lemma in Setswana is. Daar is egter nie klarigheid oor wat die lemma in Setswana is nie, en die definisie van wat die lemma is, moes aan die hand van die Setswanamorfologie bepaal word. Die terme "wortel" en "stam" is derhalwe onder die loep geneem om te bepaal wat die lemma in Setswana kan wees.

Beide die stam en die wortel is leksikale morfeme: die wortel is die semantiese kern van die woord, maar realiseer onder die woordvlak, terwyl die stam op die woordvlak realiseer. Daar is in Hoofstuk 2 tot die gevolgtrekking gekom dat die lemma in Setswana die stam in sy eenvoudigste vorm is, voordat dit as 'n wortel beskou kan word. Prakties beteken dit dat die naamwoordlemma in die enkelvoud sonder enige suffikse (behalwe die deverbatiwewoord-suffiks) is. Die egte werkwoorde se eenvoudigste stam is in die infinitief, sonder enige suffikse (behalwe die onproduktiewe suffikse). Die doel van outomatiese lemma-identifisering vir Setswana is om naamwoorde en egte werkwoorde na hierdie eenvoudigste vorms te verander.

In natuurliketaalprosessering is daar twee benaderings wat gevolg kan word in die ontwikkeling van 'n lemma-identifiseerder, naamlik die linguistiese en die datagedrewe benadering. In hierdie studie is die linguistiese benadering gevolg in die ontwikkeling van 'n reëlgebaseerde lemma-identifiseerder, waarmee ook bepaal kan word hoe effektief die bestaande Setswanagrammatikas geformaliseer kan word.

In die linguistiese benadering word daar onder andere van eindigestaatsagteware gebruik gemaak. Die eindigestaatsagteware maak van reëlmatige uitdrukkings gebruik, wat die standaardnotasies vir die beskrywing van teksstringe is. Die bestaande grammatikas is daarom in die reëlmatige uitdrukkings herskryf (of anders gestel: die bestaande

grammatikas is geformaliseer). Reëlmatige uitdrukkings word via die eindigestaatoutomaat (FSA) en variasies daarvan, soos die eindigestaatoorvormer (FST), geïmplementeer (Jurafsky & Martin, 2000:21-22). Die morfologiese reëls en morfo-fonologiese veranderings soos dit in morfologiese analyses gebruik word, is aangepas om reëlmatige uitdrukkings te skryf in die ontwikkeling van 'n lemma-identifiseerder.

Die reëlmatige uitdrukkings is daarna in 'n lemma-identifiseringsstelsiem geïnkorporeer. Die hiërargie wat in morfologiese analyses gebruik word, is ook in die ontwikkeling van die lemma-identifiseringsstelsiem gebruik. Die stelsiem bestaan daarom uit FSA's waarmee woorde in die geslote woordklasse (ideofone, interjeksies, voornaamwoorde, betrekkingwoorde, bywoorde) en die lokatiefklasnaamwoorde, kopulatiewe werkwoorde en hulpwerkwoorde geïdentifiseer word. Verder bestaan dit ook uit FST's waarmee die naamwoord- en werkwoordlemma geïdentifiseer word.

Ten slotte is die lemma-identifiseerder geëvalueer. Die naamwoord- en werkwoordreëls, asook die reëls wat betrekking het op die ander woordsoortkategorieë, is getoets met twee stelling woorde. Die kriteria wat in die evaluering gegeld het, was akkuraatheid ("precision").

5.3 Gevolgtrekking

Die uitgangspunt en vooronderstelling van hierdie studie was dat die linguistiese benadering 'n gepaste benadering tot lemma-identifisering vir Setswana kan wees. Dié vooronderstelling berus op die feit dat die morfologiese patrone van die naamwoorde en werkwoorde in Setswana redelik reëlmatig is, terwyl die ander woordsoorte morfologies onproduktief is. Voorts bestaan daar 'n goeie kennisbasis vir die ontwikkeling van 'n reëlgebaseerde lemma-identifiseerder. Dit bied dan 'n geleentheid om te bepaal hoe goed die bestaande grammatikas sal vaar indien dit geformaliseer word (d.i. aangepas word om rekenaarmatig aangewend te word).

Hieruit volg dan dat die doel van die studie was om die lemma in Setswana te definieer en te identifiseer, en om te bepaal of bestaande

Setswanagrammatikas effektief geformaliseer kan word om die lemmas te identifiseer. Vervolgens word die navorsingsvrae, soos wat dit in Hoofstuk 1 gestel is, bespreek.

(i) Hoe moet die lemma in Setswana gedefinieer en geïdentifiseer word?

Soos dit in Hoofstuk 1 bespreek is, word die eenvoudigste stamme (wat as woorde kan optree) as die lemma aanvaar. Die term “lemma” kan vanuit verskillende perspektiewe gedefinieer word, maar lemma-identifisering word in hierdie studie gesien as ’n vorm van vereenvoudigde morfologiese analise; daarom geld morfologiese perspektiewe op wat ’n lemma is sterker binne dié konteks. As sodanig word ’n lemma in Setswana gesien as die eenvoudigste stam in die kanonieke vorm.

Dit beteken vir die naamwoord dat die lemma in die enkelvoud sonder enige suffikse (behalwe die deverbatieme suffikse waar dit voorkom) is. Die suffikse wat nie in die lemma ingesluit word nie, is die lokatiewe, diminutiewe, feminitiewe en augmentatiewe suffikse. Wat die beskrywende naamwoord betref, is daar van die algemene riglyn van die eenvoudigste stam afgewyk en word die wortel as lemma geneem. Die rede daarvoor is dat die beskrywende naamwoorde baie verskillende vorms aanneem omdat hulle vormlik afhanklik is van die naamwoord wat hulle beskryf. Voorts is die beskrywende naamwoorde ’n beperkte lys. Plek- en eiename word onveranderd gelos.

Die egtewerkwoordlemma is die eenvoudigste stam in die infinitief, sonder enige prefikse. Dit is dan die stam van die werkwoord sonder die volgende affikse: die negatiewe, kongruensie-, aspek- of temporale morfeme, neutropassiewe, iteratiewe, kousatiewe, applikatiewe, resiproke, perfektum, passiewe suffikse en die uitgangsmorfeme **-e** en **-ng**. Die hulpwerkwoord word in egte en onegte hulpwerkwoorde onderverdeel. Egte hulpwerkwoorde word net so as lemmas geneem. Waar die onegte hulpwerkwoord vormlik met die egte werkwoord ooreenstem, is dieselfde beginsels wat by die bepaling van die egtewerkwoordlemma geld ook by die onegte hulpwerkwoorde van toepassing. Hoewel die kopulatiewe werkwoord,

afhangende van die tyd en modus waarin dit verskyn, verskillende vorms kan aanneem, kan die basiese kopulatiewe werkwoord **nna** as die lemma beskou word.

Die lemmas van die verskillende voornaamwoorde, betrekkingwoorde, bywoorde, interjeksies en ideofone is net soos wat hulle in die taal voorkom, omdat dié woordsoorte geslote is en dus nie soos in die geval van die naamwoord en werkwoord morfologies uitgebrei word nie.

(ii) Kan bestaande Setswanagrammatikas effektief geformaliseer word om die lemmas te kan identifiseer?

Daar is in Hoofstuk 1 van die standpunt uitgegaan dat die bestaande Setswanagrammatikas effektief geformaliseer kan word om 'n outomatiese lemma-identifiseerder vir Setswana te ontwikkel. Die morfologiese struktuur van die Sothotale, spesifiek Setswana, is goed gedokumenteerd in handboeke en artikels van verskeie navorsers, en het as basis vir kennisgebaseerde reëls in die volg van 'n linguistiese benadering tot die ontwikkeling van 'n lemma-identifiseerder vir Setswana gedien. Dit is die eerste keer dat bestaande grammatikas op só 'n manier getoets is. In syfers uitgedruk, is die akkuraatheid van die lemma-identifiseerder met die uitgebreide lemmalys 70,96% op komplekse naamwoorde, 70,52% op komplekse werkwoorde, 78,45% op komplekse en simplekse naamwoorde, 79,59% op komplekse en simplekse werkwoorde, en 75,92% op alle woorde. Die ander woordsoorte het 'n akkuraatheid van 100%.

Die *kwantitatiewe* prestasie van die lemma-identifiseerder is egter nie die fokuspunt van die studie nie, aangesien dit oor die evaluering van die grammatikas gaan. Die foute van die lemma-identifiseerder lê nie net by die grammatika nie, maar ook in die sisteem. Dit is daarom nie die beste metode om grammatikas kwantitatief mee te evalueer nie. Die kwantitatiewe prestasie gee wel 'n aanduiding van die relatiewe akkuraatheid van die grammatikas, maar dit bied veral ook geanaliseerde data waarmee die grammatikas *kwalitatief* geëvalueer kon word (soos dit in 4.4 aan die hand van voorbeelde gedoen is).

Die Setswanagrammatikas kan tot 'n redelike mate geformaliseer word, maar daar sal aan sake soos die uitbreiding van 'n lemmalys, reëlvolgorde, naamwoordklasverskuiwings, dubbele klasprefikse, dubbele suffikse, meerduidige uitgange, verdubbelings en samestellings aandag gegee moet word.

5.4 Vooruitskouing

Indien 'n volledige Setswanalemmalys (leenwoorde ingesluit) saamgestel kan word en onvoorsiene klankveranderings, naamwoordklasverskuiwings, dubbele klasprefikse, dubbele suffikse, meerduidige uitgange en die verandering in die reëlvolgorde ondervang kan word, kan die lemma-identifiseerder vir Setswana met groot vrug in die ontwikkeling van 'n tweedegenerasiespeltoetser vir Setswana aangewend word. Dit kan ook suksesvol gebruik word in die leksikografiese veld – ten opsigte van die samestelling van woordeboeke. Hierdie eerste lemma-identifiseerder vir Setswana kan ook in die navorsingsveld nuttig gebruik word in die verdere ontwikkeling van 'n morfologiese analiseerder vir Setswana, en kan selfs as 'n model gebruik word vir soortgelyke ontwikkelings in ander Afrikatale, spesifiek die Sothotale. Indien 'n lemma-identifiseerder met behulp van masjienleertegnieke in die toekoms ontwikkel sou word, sal die linguistiese lemma-identifiseerder gebruik kan word om deur middel van skoenlussteekproefneming die lemmalys makliker saam te stel en uit te brei.

Vir die hantering van verdubbelings en samestellings in Setswana moet die ontwikkeling van kompositum- en reduplikasie-analiseerders ondersoek word. In hierdie opsig kan dié analiseerders wat onder andere reeds vir Afrikaans (Van Huyssteen & Van Zaanen, 2004) ontwikkel is, as prototipe dien vir die ontwikkeling van soortgelyke analiseerders vir Setswana. Wat die eenlettergrepige werkwoorde met imperatiewe uitgange betref, kan dit in die toekoms gelys word en as uitsonderings deur die sisteem hanteer word. Die imperatiewe uitgang of suffigale element **-a** is voorts 'n saak wat verder ondersoek behoort te word; toetse vir morfeemstatus kan byvoorbeeld

aangelê word om te wys dat dit wel 'n morfeem is. Die geval van die verandering in die reëlvolgorde moet ook verder ondersoek word.

'n Tekseenheididentifiseerder wat voorsiening maak vir die disjunktiewe ortografie van Setswana was nie tydens die ontwikkeling van die lemma-identifiseerder beskikbaar nie. Indien so 'n tekseenheididentifiseerder ontwikkel sou word, sal die lemma-identifiseerder die spasies in die werkwoorde (egte, hulp- en kopulatiewe werkwoorde) kan hanteer (sien Bylaag 2.11). Dit word in elk geval, om die lemma-identifiseerder nog wyer te kan gebruik, aanbeveel dat 'n tekseenheididentifiseerder ontwikkel word om saam met die lemma-identifiseerder gebruik te word.

Bronnelys

BEESLEY, K.R. & KARTTUNEN, L. 2003. Finite state morphology. Stanford : CSLI Publications. 510 p.

BEYER, G. 1920. A handbook of the Pedi-Transvaal Suto language: practical grammar with exercises, phrases, dialogues and vocabularies. Morija : Sesuto Book Depot. 99 p.

BOUMA, G. 2004?. FSA tutorial. [Web:] <http://odur.let.rug.nl/~gosse/tt/fsa.html> [Datum van gebruik: 22 Mrt. 2005].

BROWN, J.T. 1988. Setswana-English dictionary. Johannesburg : Pula Press. 593 p.

BUSSMAN, H. 1996. Routledge dictionary of language and linguistics. London : Routledge. 530 p.

CHOUKEA, Y., CONLEY, E.S. & DAGAN, I. 2000. A comprehensive bilingual word alignment system. (*In* Véronis, J., ed. Parallel text processing. Alignment and use of translation corpora. Dordrecht : Kluwer Academic Publishers. p. 69-96.)

COHEN-SYGAL, Y. & WINTNER, S. 2006. Finite-state registered automata for non-concatenative morphology. *Computational linguistics*, 32(1):49-82, Mar.

COLE, D.T. 1955. An introduction to Tswana grammar. Cape Town : Longmans & Green. 473 p.

CREISSELS, D. 1996. Conjunctive and disjunctive verb forms in Setswana. *South African journal of African languages*, 16(4):109-115, Jan.

CROFT, W. 2003. Typology and universals. 2nd ed. New York : Cambridge University Press. 341 p.

CRYSTAL, D. 1992. An encyclopedic dictionary of language and languages. Oxford : Blackwell. 428 p.

CRYSTAL, D. 2003. A dictionary of linguistics and phonetics. 5th ed. Oxford : Blackwell. 508 p.

DAELEMANS, W. & STRIK, H., eds. 2002. Het Nederlands in de taal- en spraaktechnologie: prioriteiten voor basisvoorzieningen. [Dutch in language and speech technology: priorities for basic provisions]. Final report (01/07/2002). Dutch Language Union. [Web:] taalunieversum.nl/taal/technologie/docs/daelemans-strik.pdf [Datum van gebruik: 4 Des. 2005].

DE SAUSSURE, F. 1916. Cours de linguistique générale publié. Paris : Payot. 331 p.

DE SCHRYVER, G-M & PRINSLOO, D.J. 2001. Towards a sound lemmatisation strategy for the Bantu verb through the use of frequency-based tail slots – with special reference to Cilubà, Sepedi and Kiswahili (*In* Mdee, J.S. & Mwansoko, H.J.M., eds. 2001. Makala ya kongamano la kimataifa Kiswahili 2000. Dar es Salaam : TUKI, Chuo Kikuu cha Dar es Salaam. p. 216-242.)

DENT, G.R. 1994. Kompakte Setswana woordeboek. Pietermaritzburg : Shuter & Shooter. 207 p.

DOKE, C.M. 1931. Text-book of Zulu grammar. London : Longmans. 351 p.

ENDEMANN, K. 1976. Versuch einer Grammatik des Sotho. Berlin : Hertz. 201 p.

ERJAVEC, T. & DŽEROSKI, S. 2004. Machine learning of morphosyntactic structure: lemmatising unknown Slovene words. *Applied artificial intelligence*, 18(1):17-40.

GEARAILT, D.N. 2005. Dictionary characteristics in cross-language information retrieval. Cambridge: University of Cambridge Computer Laboratory. (Technical report UCAM-CL-TR-616.) [Web:] <http://www.cl.cam.ac.uk/TechReports/UCAM-CL-TR-616.pdf> [Datum van gebruik: 4 Des. 2005].

GOLDSMITH, J.A., HIGGINS, D. & SOGLASNOVA, S. 2000. Automatic language-specific stemming in information retrieval (*In* Peters, C., ed. Cross-language information retrieval and evaluation: revised papers from the workshop of cross-language evaluation forum (CLEF), Lisbon, Portugal, September 21-22. London : Springer. p. 273-284.)

GOUWS, R.H. 1989. Leksikografie. Pretoria : Akademica. 287 p.

GOUWS, R.H. & PRINSLOO, D.J. 2005. Principles and practice of South African lexicography. Stellenbosch : SUN PReSS. 210 p.

HANDKE, J. 1995. The structure of the lexicon: human versus machine. Berlin : Mouton de Gruyter. 388 p.

HARTMANN, R.K.K. 2001. Teaching and researching lexicography. Harlow : Longman. 211 p.

HARTMANN, R.K.K. & JAMES, G. 1998. Dictionary of lexicography. London : Routledge. 176 p.

- HASPELMATH, M. 2002. Understanding morphology. New York : Oxford University Press. 290 p.
- HAUSSER, R. 1999. Foundation of computational linguistics: man-machine communication in natural language. Berlin : Springer. 516 p.
- HOARD, J.E. 1998. Language understanding and the emerging alignment of linguistics and natural language processing (*In* Lawler, J.M. & Dry, H.A., eds. Using computers in linguistics. London : Routledge. p. 197-230.)
- JACKSON, H. 1988. Words and their meanings. New York : Longman. 279 p.
- JURAFSKY, D. & MARTIN, J.H. 2000. Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition. Upper Saddle River, N.J. : Prentice Hall. 934 p.
- KAPLAN, R.M. & KAY, M. 1981. Phonological rules and finite-state transducers. (Paper presented to the Winter Meeting of the Linguistic Society of America.) New York.
- KAPLAN, R.M. 1996. Finite state technology. [Web:] <http://cslu.cse.ogi.edu/HLTsurvey/ch11node8.html#SECTION116> [Datum van gebruik: 28 Des. 2004].
- KATAMBA, F. 2005. English words: structure, history, usage. 2nd ed. New York : Routledge. 322 p.

KHOJA, S. 2002. Arabic computational linguistics: stemming [Web:] <http://www.comp.lancs.ac.uk/computing/users/khoja/research.htm> [Datum van gebruik: 27 Apr. 2004]

KLEENE, S.C., 1951. Representation of events in nerve nets and finite automata. The Rand Corporation. Rand Research Memorandum. (Technical Report RM-704.)

KLEENE, S.C. 1956. Representation of events in nerve nets and finite automata. (In Shannon, C. & McCarthy, J., eds. Automata studies. Princeton, N.J. : Princeton University Press. p. 3-41.)

KNOWLES, G. & DON, Z.M. 2004. The notion of a “lemma”: headwords, roots and lexical sets. *International journal of corpus linguistics*, 9(1):69-81.

KRAAIJ, W. & POHLMANN, R. 1996. Using linguistic knowledge in information retrieval. [Web:] <http://www.uilots.let.uu.nl/%7Euplift/workpap.ps.gz> [Datum van gebruik: 16 Aug. 2005].

KRIEL, T.J. 1958. The new Sesotho-English dictionary. Johannesburg : Afrikaanse Pers-Boekhandel. 452 p.

KRIEL, T.J., VAN WYK, E.B. & MAKOPO, S.A. 1989. Pukuntšu. 4de uitgawe. Pretoria : Van Schaik. 517 p.

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

KRÜGER, C.J.H. 1994. Word-based versus root-based morphology in the African languages. *South African journal of African languages*, 14(1):15-23.

KRÜGER, C.J.H. 2006a. Introduction to the morphology of Setswana. Manuscript to be published in München (Lincom).

KRÜGER, C.J.H. 2006b. 'n Faks met kommentaar op ontledings deur die outeur. Pretoria. (Kopie in besit van die outeur.)

KRÜGER, C.J.H. & SNYMAN, J.W. 1986. The sound system of Setwana. Goodwood : Via Afrika. 185 p.

LAAS, J.A.M. 1974. Woordbou en woordanalise in Suid-Sotho. Potchefstroom : PU vir CHO. (Verhandeling – MA).

LANCASTER UNIVERSITY COMPUTING DEPARTMENT. 2000. What is stemming? [Web:]
<http://www.comp.lancs.ac.uk/computing/research/stemming/general/> [Datum van gebruik: 4 Apr. 2004]

LENNON, M., PIERCE, D.C. & WILLETT, P. 1981. An evaluation of some conflation algorithms. *Journal of information science*, 3: 177-183.

LOHR, S.L. 1999. Sampling: design and analysis. Pacific Grove, Calif. : Duxburg Press. 494 p.

LOMBARD, D.P., VAN WYK, E.B. & MOKGOKONG, P.C. 1985. Inleiding tot die grammatika van Noord-Sotho. Pretoria : Van Schaik. 205 p.

LOUWRENS, L.J. 1991. Aspects of Northern Sotho grammar. Pretoria : Via Afrika. 176 p.

LOUWRENS, L.J. 1994. Dictionary of Northern Sotho grammatical terms. Pretoria : Via Afrika. 257 p.

MANNING, C.D. & SCHÜTZE, H. 1999. Foundations of statistical natural language processing. Cambridge : The MIT Press. 680 p.

MATTHEWS, P.H. 1997. Oxford concise dictionary of linguistics. New York : Oxford. 410 p.

MCENERY, T. & WILSON, A. 2001. Corpus linguistics: an introduction. Edinburgh : Edinburgh University Press. 235 p.

MITKOV, R. 2003. The Oxford handbook of computational linguistics. New York : Oxford University Press. 784 p.

NASIONALE DEPARTEMENT van Kuns en Kultuur **kyk** SUID-AFRIKA.
Nasionale Departement van Kuns en Kultuur.

NTSIME, J.M & KRÜGER, C.J.H. 1995. E antswa letseleng: Setlhophsa sa 9. Kgatiso ya 4. Pretoria : Via Afrika. 394 p.

NWU. 2003. *Setswana spelling checker 1.0 and hyphenator*. Potchefstroom : North-West University, Potchefstroom Campus.

OOI, V.B.Y. 1998. Computer corpus lexicography. Edinburgh : Edinburgh University Press. 243 p.

PAICE, C.D. 1994. An evaluation method for stemming algorithms. (*In* Croft, W.B & Van Rijsbergen, C.J., eds. ACM-SIGIR 1994 Ireland, Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. Paper read at the 17th international ACM-SIGIR conference. Dublin. p. 42-50.)

PISKORSI, J. & NEUMANN, G. 1999. An intelligent text extraction and navigation system. [Web:] <http://www.coli.uni-sb.de/publikationen/softcopies/Piskorski:2000:ITE.pdf> [Datum van gebruik: 11 Jan. 05].

PLISSON, J., LAVRAC, N. & MLADENIC, D. 2004. A rule-based approach to word lemmatization. (*In Proceedings C of the 7th International Multi-Conference Information Society IS 2004, October 2004, Ljubljana, p. 83-86*)

POSTHUMUS, L.C. 1994. Notes on morphology with special reference to Tswana. *South African journal of African languages*, 14(1):28-36, Feb.

POULOS, G. & LOUWRENS, L.J. 1994. A linguistic analysis of Northern Sotho. Pretoria : Via Afrika. 524 p.

PRETORIUS, L. & BOSCH S.E. 2002. The significance of computational morphological analysis for Zulu lexicography. *South African journal of African languages*, 1:11-20, Mar.

PRETORIUS, L. & BOSCH S.E. 2003. Computational aids for Zulu natural language processing. *Southern African linguistics and applied language studies*, 21(4):276-282, Nov.

PRETORIUS, R.S. 1997. Auxiliary verbs as a subcategory of the verb in Tswana. Potchefstroom : PU vir CHO. (Thesis – PhD).

PRETORIUS, W.J. 2000. Die identifisering en beskrywing van die begrippe stam en wortel in die Afrikatale, met besondere verwysing na die Sothotale. *Journal for language teaching*, 34(1):51-62.

PRINSLOO, D.J. Datum onbekend. Frekwensielys van Setswana woorde. Pretoria : Universiteit van Pretoria.

RAS, W.A.P. 1991. Die ideofoon in Tswana. Potchefstroom : PU vir CHO. (Verhandeling – MA).

REYNIERSE, C. 1991 South African multi-language dictionary and phrase book: English, Afrikaans, Northern Sotho, Sesotho, Tswana, Xhosa, Zulu. Cape Town : Reader's Digest. 495 p.

RITCHIE, G.D., PULMAN, S.G., BLACK, A.W.G. & RUSSELL, J. 1987. A computational framework for lexical description. *Computational linguistics*, 13(3-4): 290-307.

SIEMENS, R.G. 1996. Lemmatization and parsing with TACT preprocessing programs. *CHWP A.1.*, Feb. 1996. [Web:] <http://www.chass.utoronto.ca/epc/chwp/siemens2/siemintr.html> [Datum van gebruik: 4 Apr. 2004].

SNOXALL, R.A. 1965. Some problems and principles of lexicography in Luganda. *African language studies*, 6: 27-31.

SNYMAN, J.W., SHOLE, J.S. & LE ROUX, J.C. 1990. Setswana-Engels-Afrikaanse woordeboek. Pretoria : Via Afrika. 527 p.

SUID-AFRIKA. Nasionale Departement van Kuns en Kultuur. 2000. Language policy and plan for South Africa. [Web:] http://www.dac.gov.za/about_us/cd_nat_language/language_policy/Language%20Policy%20and%20Plan%20for%20South%20Africa.htm#lang_tech [Datum van gebruik: 1 Feb. 2005].

TRASK, R.L. 1996. A dictionary of phonetics and phonology. London : Routledge. 424 p.

TROST, H. 2003. Morphology. (*In* Mitkov, R., ed. The Oxford handbook of computational linguistics. New York : Oxford University Press. p. 25-47.)

TURING, A.M. 1936. On computable numbers, with an application to the Entscheidungsproblem. (*In* Proceedings of the London Mathematical Society, 42:230-265. Read to the Society in 1936, but published in 1937. Correction in 34:544-537.)

UNIVERSITY OF WESTERN ONTARIO DEPARTMENT OF COMPUTER SCIENCE. 2002. Finite-state machine software, products, and projects. [Web:] <http://www.csd.uwo.ca/research/grail/links.html> [Datum van gebruik: 12 Okt. 2004].

VAN EENDEN, B.I.C. 1956. Zoeloe-grammatika. Stellenbosch : U.U.B. 777 p.

VAN HUYSSTEEN, G.B. & VAN ZAAZEN, M.M. 2004. Learning compound boundaries for Afrikaans spelling checking. (*In* Proceedings of the First Workshop on International Proofing Tools and Language Technologies. Patras: University of Patras. p. 101-108.)

VAN NOORD, G. 2002. Finite state utilities. [Web:] <http://odur.let.rug.nl/~vannoord/Fsa/fsa.html> [Datum van gebruik: 27 Sep. 2004].

VAN NOORD, G. 2002b. FSA reference manual. [Web:] <http://odur.let.rug.nl/~vannoord/Fsa/fsa.html> [Datum van gebruik: 27 Sep. 2004].

VAN NOORD, G. (vannoord@let.rug.nl) 2004. Bespreking van Gertjan van Noord se "FSA Utilities toolbox". [E-pos aan:] Brits, J.H. (bkkjhb@puk.ac.za) Okt. 27.

VAN ROOY, A.J. & PRETORIUS, R.S. 2003. A word-class tagset for Setswana. *Southern African linguistics and applied language studies*, 21(4):203–222.

VAN STERKENBURG, P. 2003. A practical guide to lexicography. Amsterdam : Benjamins. 459 p.

VAN WYK, E.B. 1961. Ongepubliseerde klasaantekeninge. Universiteit van Pretoria.

VAN WYK, E.B. 1966. The word-classes in Northern Sotho. Die woord as linguistiese eenheid. *Lingua*, 17(2):230-261.

VAN WYK, E.B. 1968. Die woord as linguistiese eenheid. *Student en dosent: klasgids*. November.

VAN WYK, E.B. 1969. Kritiese ontleding van die morfeemopvatting. *Klasgids*, 4(3):35-43.

VERMEULEN, G.J.G. & PRETORIUS, R.S. 1999. Tswana diktaat. Potchefstroom : PU vir CHO. 312 p.

VOUTILAINEN, A. 1999. A short history of tagging (*In* Van Halteren, H., ed. Syntactic wordclass tagging. Dordrecht : Kluwer Academic Publishers. p. 9-19.)

XEROX. 2004. Finite state technology. [Web:]
<http://www.xrce.xerox.com/competencies/content-analysis/fst/home.en.html>
[Datum van gebruik: 27 Sep. 2004].

ZERWICK, M. 1997. Die voornaamwoord in Tswana: identifikasie en klassifikasie. Potchefstroom : PU vir CHO. (Verhandeling – MA).

ZGUSTA, L. 1971. Manual of lexicography. Paris : Mouton. 360 p.

ZIERVOGEL, D. & MOKGOKONG, P.C. 1985. Groot Noord-Sotho woordeboek. 2de uitgawe. Pretoria : Van Schaik. 1536 p.