

*REËLGEBASEERDE
KLEMTOONTOEKENNING
IN 'N GRAFEEM-NA-FONEEMSTELSEL
VIR*

AFRIKAANS

E.W. MOUTON

12775118

*VERHANDELING VOORGELÊ VIR DIE GRAAD
MAGISTER ARTIUM IN ALGEMENE TAAL EN LITERATUURWETENSKAP
AAN DIE POTCHEFSTROOM-KAMPUS VAN DIE NOORDWES-UNIVERSITEIT*

STUDIELEIER: PROF. D.P. WISSING

MEI 2010

VOORWOORD

Graag word die volgende persone en instansies bedank:

- Prof Daan Wissing, vir al die kennis wat Prof na my toe oorgedra het en omdat dit altyd lekker is om met Prof oor taalreëls te redeneer. Dankie dat ek, selfs toe ek nog 'n voorgraadse student was, so baie by Prof kon leer.
- Martin, omdat jy maar altyd bereid was om nog 'n woordelysie vir my te "CKarma en Calomo" en ook vir die hulp met die program en met Perl.
- CTeXT, omdat julle my alles geleer het wat ek vandag van Taaltegnologie af weet.
- Paul, omdat jy altyd maar net daar was as ek 'n skouer nodig gehad het en maar gewillig my werk oorgeneem het as dit bietjie druk gegaan het met my studies.
- My pa en ma, vir al die moed wat julle my ingepraat het en dat julle deurentyd agter my gestaan het tydens my studies.
- Elrie, vir al die tyd wat jy opgeoffer het om vir my boeke by die biblioteek te gaan uitneem en aan te stuur en al die ander gunsies wat jy vir my moes doen, sodat ek my studies kon voltooi.
- My ander pa en ma, dat julle maar altyd verstaan het as my werk in die boerdery moes agterbly sodat ek tyd kom maak vir my studies.
- Ouma Ella, Oupa Fires, Ouma Engela, Abrie en die res van my familie vir al die ondersteuning en telefoonoproepe.
- Die NRF, vir die befondsing wat ek ontvang het vir die studie.
- Almal wat nie hierbo genoem is nie, maar wat my ondersteun het en maar altyd bereid was om te luister.

ABSTRACT

RULE-BASED STRESS ASSIGNMENT IN A GRAPHEME-TO-PHONEME SYSTEM FOR AFRIKAANS

by Elsamien Mouton

Text-to-speech systems currently are of great importance in the community. One core technology in this human language technology resource is stress assignment which plays an important role in any text-to-speech system. At present no automatic stress assigner for Afrikaans exists. For these reasons, the two most important aims of this project will be: a) to develop a complete and accurate set of stress rules for Afrikaans that can be implemented in an automatic stress assigner, and b) to develop an effective and highly accurate stress assigner in order to assign Afrikaans stress to words quickly and effectively.

A set of stress rules for Afrikaans was developed in order to reach the first goal. It consists of 18 rules that are divided into groups for words that contain a schwa, derivations, and disyllabic, trisyllabic and polysyllabic simplex words.

Next, different approaches that can be used to develop a stress assigner were examined, and the rule-based approach was used to implement the developed stress rules within the stress assigner. The programming language, Perl, was chosen for the implementation of the rules.

The chosen algorithm was used to generate a stress assigner for Afrikaans by implementing the stress rules developed. The hyphenator, Calomo and the compound analyser, CKarma was used to hyphenate all the test data and detect word boundaries within compounds. A data set of 10 000 correctly annotated tokens was developed during the testing process.

The evaluation of the stress assigner consists of four phases. During the first phase, the stress assigner was evaluated with the 10 000 tokens and achieved an accuracy of 92.09%. The grapheme-to-phoneme converter was evaluated with the same data and scored 91.9%. The influence of various factors on stress assignment was determined, and it was established

that stress assignment is an essential component of rule-based grapheme-to-phoneme conversion.

In conclusion, it can be said that the stress assigner achieved satisfactory results, and that the stress assigner can be successfully utilized in future projects to develop training data for further experiments with stress assignment and grapheme-to-phoneme conversion for Afrikaans. Experiments can be conducted in future with data-driven approaches that possibly may lead to better results in Afrikaans stress assignment and grapheme-to-phoneme conversion.

KEY WORDS: AFRIKAANS, STRESS, ASSIGNMENT, GRAPHEME-TO-PHONEME CONVERSION, G2P, STRESS RULES, WORD STRESS, NATURAL LANGUAGE PROCESSING, COMPUTER LINGUISTICS, RULE-BASED.

OPSOMMING

REËLGEBASEERDE KLEMTOONTOEKENNING IN 'n GRAFEEM-NA- FONEEMSTELSEL VIR AFRIKAANS

deur Elsamien Mouton

Teks-na-spraakstelsels is tans van groot belang vir die samelewing. Een kerntegnologie binne dié mensetaaltegnologietoepassing is klemtoontoekenning, wat 'n baie groot rol speel binne 'n teks-na-spraakstelsel. Daar bestaan tans geen klemtoontoekenner vir Afrikaans nie en daarom is die twee belangrikste doelstellings binne dié projek a) om 'n volledige en akkurate stel klemtoonreëls vir Afrikaans te ontwikkel wat in 'n outomatiese klemtoontoekenner geïmplementeer kan word en b) om 'n effektiewe en hoogs akkurate klemtoontoekenner te ontwikkel, sodat Afrikaanse klemtoon vinnig en doeltreffend toegeken kan word.

Om die eerste doelstelling te bereik is daar 'n stel klemtoonreëls vir Afrikaans ontwikkel. Dit bestaan uit 18 reëls wat onderverdeel is in groepe woorde wat 'n sjwa bevat, afleidings, en tweelettergrepe, drielettergrepe, en vier- en meerlettergrepe simplekse.

Daarna is die verskillende benaderings ondersoek wat gevolg kan word om 'n klemtoontoekenner te ontwikkel, en die reëlgebaseerde benadering is gebruik om die klemtoonreëls in die klemtoontoekenner te implementeer. Die programmeertaal, Perl, is vir die implementering van die reëls gekies.

Vervolgens word die gekose algoritme ingespan om 'n klemtoontoekenner vir Afrikaans te genereer met behulp van dié klemtoonreëls. Die lettergreepverdeler, Calomo en die kompositumanaliseerder, CKarma word vooraf gebruik om die toetsdata in lettergrepe te verdeel en alle samestellings uit te wys. Tydens die toetsproses word 'n datastel van 10 000 korrek geannoteerde tekseenhede ontwikkel.

Die evaluering van die klemtoontoekenner bestaan uit vier fases. Tydens die eerste fase word die klemtoontoekenner geëvalueer met die 10 000 tekseenhede en behaal dit 'n akkuraatheid van 92.09%. Die grafeem-na-foneem-omsetter waarin die klemtoontoekenner

ingebou is, word ook geëvalueer met dieselfde data en is 91.9% akkuraat. Die invloed van verskeie faktore op klemtoontoekenning word bepaal, en daar is vasgestel dat klemtoontoekenning 'n noodsaaklike komponent vir reëlgebaseerde grafeem-na-foneem-omsetting is.

Die gevolgtrekking van die studie is dat die klemtoontoekenner bevredigende resultate behaal het en dat die klemtoontoekenner in toekomstige projekte suksesvol gebruik kan word om afrigtingsdata te genereer vir verdere eksperimente met klemtoontoekenning en grafeem-na-foneem-omsetting vir Afrikaans. Daar kan ook in die toekoms geëksperimenteer word met datagedrewe algoritmes, wat dalk tot beter resultate kan lei vir Afrikaanse klemtoontoekenning en grafeem-na-foneem-omsetting.

SLEUTELWOORDE: AFRIKAANS, KLEMTOONTOEKENNING, GRAFEEM-NA-FONEEM-OMSETTING, G2P, KLEMTOONREËLS, KLEMTOON, NATUURLIKETAALVERWERKING, REKENAARLINGUISTIEK, REËLGEBASEERD.

INHOUDSOPGAWE

<i>VOORWOORD</i>	<i>i</i>
<i>ABSTRACT</i>	<i>ii</i>
<i>OPSOMMING</i>	<i>iv</i>
<i>INHOUDSOPGAWE</i>	<i>vi</i>

HOOFSTUK 1..... *1*

INLEIDING..... *1*

1.1	PROBLEEMSTELLING	1
1.2	DOELSTELLINGS	3
1.3	SENTRALE TEORETIESE STELLING.....	3
1.4	NAVORSINGSMETODE	4
1.4.1	Die ontwikkeling van nuwe klemtoonreëls vir Afrikaans.....	4
1.4.2	Die ontwikkeling van toetsdata	5
1.4.3	Die implementering van reëls.....	6
1.4.4	Die toetsproses en evaluasie van die klemtoontoekenner.....	6
1.4.4.1	Fase 1: Evaluasie van die klemtoontoekenner	7
1.4.4.2	Fase 2: Evaluasie van die grafeem-na-foneem-omsetter	7
1.4.4.3	Fase 3: Die invloed van morfologiese analise op klemtoontoekenning	7
1.4.4.4	Fase 4: Die invloed van klemtoon op grafeem-na-foneem-omsetting	8
1.5	ONTPLOOIING	8

HOOFSTUK 2..... *10*

KLEMTOONREËLS VIR AFRIKAANS..... *10*

2.1	INLEIDING	10
2.2	DIE VOORSTELLING VAN KLEMTOONREËLS	10

2.2.1	<i>Liniêre Klemtoonanalise</i>	11
2.2.2	<i>Die Metriese Raamwerk.....</i>	12
2.2.3	<i>Metriese bome</i>	13
2.3	<i>KLEMTOONREËLS VIR AFRIKAANS</i>	20
2.3.1	<i>Bestaande klemtoonreëls vir Afrikaans</i>	20
2.3.2	<i>Faktore wat bydra tot klemtoontoekenning.....</i>	23
2.3.2.1	<i>Die gewig van sillabes</i>	23
2.3.2.2	<i>Oop en geslote lettergrepe.....</i>	25
2.3.2.3	<i>Die sjwa</i>	26
2.3.2.4	<i>Affikse.....</i>	29
2.3.3	<i>Klemtoonreëls vir Afrikaanse simplekse woorde</i>	34
2.3.3.1	<i>Tweelettergrepige woorde</i>	34
2.3.3.2	<i>Drielettergrepige woorde.....</i>	37
2.3.3.3	<i>Vier- en meerlettergrepige woorde</i>	40
2.4	<i>SAMEVATTING</i>	42
HOOFSTUK 3		44
'N REËLGEBASEERDE BENADERING TOT KLEMTOONTOEKENNING..		44
3.1	<i>INLEIDING</i>	44
3.2	<i>DIE BENADERING TOT KLEMTOONTOEKENNING</i>	45
3.2.1	<i>Datagedrewe Benaderings</i>	45
3.2.1.1	<i>Neurale Netwerke.....</i>	45
3.2.1.2	<i>Besluitnemingsbome</i>	46
3.2.1.3	<i>Voorbeeldgebaseerde leer</i>	47
3.2.1.4	<i>Steunvektormasjiene</i>	49
3.2.1.5	<i>Versteekte Markovmodelle.....</i>	50
3.2.1.6	<i>N-gramme.....</i>	52

3.2.1.7	<i>Maksimale entropie (ME)</i>	52
3.2.2	<i>Die Linguistiese Benadering</i>	53
3.3	<i>STAPPE VIR DIE ONTWIKKELING VAN 'N REËLGEBASEERDE KLEMTOONTOEKENNER</i>	56
3.4	<i>SAMEVATTING</i>	57
HOOFSTUK 4		59
'n KLEMTOONTOEKENNER VIR AFRIKAANS		59
4.1	<i>INLEIDING</i>	59
4.2	<i>DIE KLEMTOONTOEKENNER</i>	59
4.2.1	<i>Die argitektuur van die klemtoontoeckenner</i>	60
4.2.1.1	<i>Invoer en uitvoer</i>	60
4.2.1.2	<i>Lettergreepverdeling en morfologiese analise</i>	61
4.2.1.3	<i>Die klemtoontoeckenner</i>	62
4.2.1.4	<i>Die grafeem-na-foneem-omsetter</i>	64
4.3	<i>EVALUASIE VAN DIE KLEMTOONTOEKENNER</i>	67
4.3.1	<i>Die toetsproses</i>	67
4.3.2	<i>Evaluasie</i>	73
4.3.2.1	<i>Die eerste evaluasieproses</i>	73
4.3.2.1.1	<i>Die vergelyking en evaluasie van die klemtoontoeckenners</i>	75
4.3.2.1.2	<i>Evaluasie van KTT 2009 op lopende teks</i>	77
4.3.2.1.3	<i>Vergelyking van KTT 2009 met ander klemtoontoeckenners</i>	78
4.3.2.2	<i>Die evaluasie van die grafeem-na-foneem-omsetter</i>	80
4.3.2.3	<i>Die invloed van morfologiese analise op klemtoontoeckenning</i>	81
4.3.2.4	<i>Die invloed van klemtoontoeckenning op grafeem-na-foneem-omsetting</i>	82
4.4	<i>SAMEVATTING</i>	84
HOOFSTUK 5		86
SLOT		86
5.1	<i>SAMEVATTING</i>	86

5.2	<i>GEVOLGTREKKING</i>	89
5.3	<i>AANBEVELINGS</i>	90
	<i>BIBLIOGRAFIE</i>	93
	ADDENDUM A	<i>A-I</i>

Hoofstuk 1

INLEIDING

1.1 PROBLEEMSTELLING

Hierdie verhandeling het as tema die ontwikkeling en evaluering van 'n outomatiese reëlgebaseerde klemtoontoekenner vir 'n Afrikaanse grafeem-na-foneem-omsetter.

Daar bestaan reeds 'n Afrikaanse klemtoontoekenner wat in 2006 deur Visser, Wissing en Puttkammerⁱ ontwikkel is by die Sentrum vir Tekstegnologie (CTeXt) aan die Noordwes-Universiteit as deel van 'n grafeem-na-foneem-omsetter. 'n Reëlgebaseerde benadering is gevolg om die toekenner te ontwikkel, waar klemtoonreëls op woorde toegepas word om die korrekte sillabe van 'n woord te beklemtoon. Alhoewel die toekenner 'n akkuraatheid van 93%ⁱⁱ behaal het, is die klemtoonreëls wat gebruik is nie almal linguisties akkuraat nie, en is die meeste van die transkripsiefoute wat gemaak is deur die grafeem-na-foneem-omsetter te wyte aan klemtoon wat verkeerd toegeken is. Binne hierdie projek moet die bestaande klemtoonreëls vir Afrikaans dus hersien, verbeter en uitgebrei word, en dan geïmplementeer word in die toekenner, sodat die klemtoontoekenner uiteindelik 'n hoër akkuraatheid het. Die toekenner moet ook herbruikbaar wees, sodat dit in enige ander ontleder of grafeem-na-foneem-omsetter geïnkorporeer kan word.

Die effektiwiteit en doeltreffendheid van die toekenner is direk afhanklik van die akkuraatheid en volledigheid van die klemtoonreëls wat daarin geïmplementeer is. Dit is vanselfsprekend dat 'n toekenner onmoontlik die korrekte reël sal gebruik om klemtoon aan 'n woord toe te ken as hierdie reël nie deel is van die klemtoonreëls wat geïmplementeer is nie. Dit is dus belangrik dat

ⁱ Die grafeem-na-foneem-omsetter is ontwikkel as deel van 'n finalejaarsprojek deur E Visser (met die hulp van Prof. D P Wissing en M J Puttkammer). Dus is daar geen gepubliseerde literatuur daaroor nie.

ⁱⁱ Die klemtoontoekenner het 'n akkuraatheid van 93% op woordvlak en 98% op foneemvlak en is geëvalueer op 'n woordelys wat bestaan uit 1 000 woorde.

die klemtoonreëls wat gebruik word linguisties akkuraat is, alle moontlike reëls vir Afrikaans bevat, en ook geen reëls bevat wat nie op Afrikaans van toepassing is nie.

Klemtoontoekenning is 'n baie belangrike deel van 'n grafeem-na-foneem-omsetter, omdat klemtoon heel dikwels bepaal hoe meerlettergrepige woorde uitgespreek word. Combrink en De Stadler (1987:184) ondersteun hierdie siening deur te noem dat klemtoon in baie gevalle woorde se grammatiese kategorieë onderskei, byvoorbeeld *gésel* teenoor *gesél*. Klatt (1987:737) is van mening dat klemtoontoekenning die grootste probleem is in grafeem-na-foneem-omsetting, omdat 'n verkeerde aanduiding van klemtoon in 'n woord meestal veroorsaak dat die woord verkeerd getranskribeer word.

Daar bestaan ook 'n groot leemte in die spraaktegnologie-studieveld met die outomatiese toekenning van klemtoon. Die leemte is veral groot in Afrikaans, omdat die klemtoontoekenner wat hier ter sprake is, die enigste outomatiese klemtoontoekenner vir Afrikaans is en die betrokke toekenner nog nie vrygestel is nie.

Uit die beskikbare literatuur blyk dit dat daar al heelwat verskillende benaderings gevolg is om die probleem van outomatiese klemtoontoekenning op te los. Gillis, Daelemans, Durieux en Van den Bosch (1993) het 'n klemtoontoekenner vir Nederlands ontwikkel deur 'n voorbeeld-gebaseerde benadering te volg. Die algoritme behels dat die leerdata gestoor word. As daar dan op 'n nuwe voorbeeld afgekom word, word 'n stel verwante voorbeelde onttrek uit die geheue, en dit word dan gebruik om die nuwe voorbeeld te klassifiseer.

Die data van die eksperiment wat deur Daelemans *et al.* gedoen is, het bestaan uit 4863 meersillabige Nederlandse woorde wat verkry is uit die CELEX leksikale databasis. Die los-een-uit-metode is gevolg om die eksperiment uit te voer. Dit behels dat elke item in die datastel om die beurt gebruik word as die toetsitem en die res van die datastel is dan die leerdata. Weiss en Kulikowski (aangehaal deur Gillis *et al.*, 1993:254) is van mening dat hierdie metode die voordeel het dat dit die beste moontlike skatting gee van die ware foutskaal van die algoritme.

Daelemans (1988:133) het ook die belangrikheid van klemtoontoekenning besef in sy grafeem-na-foneem-omsetter vir Nederlands, GRAFON, deurdat hy klemtoon toegeken het aan woorde voordat fonetiese transkripsie plaasgevind het. Die klemtoon word nie outomaties toegeken nie, maar dit word onttrek vanuit 'n databasis. Dus het GRAFON baie beter resultate gehad met die omsetting van grafeme na foneme, omdat daar geen klemtoonfoute gemaak is nie. Die omsetter het 97.62% van die toetswoorde korrek getranskribeer.

Nog 'n poging tot outomatiese klemtoontoekenning is gemaak deur Church (1985). Hy het 'n reëlgebaseerde klemtoontoekenner vir Engels gebou. Klemtoonreëls is gegenereer deur na die gewig van 'n sillabe, woordsoorte, morfologie en etimologie te kyk. Die toekenner het 'n akkuraatheid van 82% behaal.

Taylor (2005) het 'n statistiese benadering, naamlik versteekte Markovmodelle, gevolg om klemtoon toe te ken in 'n grafeem-na-foneem-omsetter vir Engels. Hy noem dat, in sy grafeem-na-foneem-omsetter, 'n groot hoeveelheid van die grafeme verkeerd getranskribeer is as gevolg van klemtoon wat verkeerd toegeken is. Dié model ken korrekte klemtoon toe aan 76% van die woorde. Ander benaderings tot klemtoontoekenning word in hoofstuk 3 bespreek.

Die benadering wat in hierdie projek gebruik word, is 'n reëlgebaseerde benadering, maar voordat die klemtoonreëls geïmplementeer word, moet daar eers genoeg en akkurate reëls wees. Die doel van die projek is dus om effektiewe klemtoonreëls te vind om sodoende die bestaande klemtoontoekenner te verbeter tot 'n hoogs akkurate en herbruikbare produk.

1.2 DOELSTELLINGS

Die volgende doelstellings kan daargestel word:

- a) om klemtoonreëls vir Afrikaans te ontwikkel wat linguisties akkuraat en volledig is en wat in 'n outomatiese klemtoontoekenner geïmplementeer kan word.
- b) om doeltreffende toetsdata te genereer waarmee 'n outomatiese klemtoontoekenner geëvalueer kan word.
- c) om 'n effektiewe en hoogs akkurate klemtoontoekenner te ontwikkel sodat Afrikaanse klemtoon vinnig en doeltreffend toegeken kan word.
- d) om die klemtoontoekenner en die grafeem-na-foneem-omsetter te evalueer om vas te stel of die nuwe omsetter 'n verbetering is op die bestaande omsetter.

1.3 SENTRALE TEORETIESE STELLING

Na aanleiding van bogenoemde doelstellings, is daar twee komponente wat noodsaaklik is vir die ontwikkeling van 'n outomatiese klemtoontoekenner, naamlik 'n stel klemtoonreëls, en toetsdata vir die evaluering van die toekenner. Nog 'n komponent wat ontwikkel word, is die klemtoontoekenner waarin die klemtoonreëls geïmplementeer is. Dus word die klemtoonreëls

eerste ontwikkel, sodat dit geïmplementeer kan word in die klemtoontoekenner. Die toetsdata word daarna gebruik om die klemtoontoekenner te evalueer.

Die klemtoonreëls wat gebruik word om 'n volledige stel klemtoonreëls te genereer, is dié van Wissing (1971) en Combrinck en De Stadler (1987). Die reëls word in dié studie aangevul en, waar nodig, verbeter om sodoende te verseker dat die stel klemtoonreëls linguisties akkuraat is, alle moontlike reëls vir Afrikaans bevat, en ook geen reëls bevat wat nie op Afrikaans van toepassing is nie.

Hier word verder van die standpunt uitgegaan dat 'n reëlgebaseerde benadering tot outomatiese klemtoontoekenning die beste resultate vir Afrikaans sal oplewer. Die bestaande reëlgebaseerde klemtoontoekenner het goed gevaar met die toekenning van klemtoon in Afrikaans, en sal dus, met die verbetering van die klemtoonreëls, waarskynlik baie beter resultate vir Afrikaanse klemtoontoekenning oplewer.

Na afloop van hierdie projek word daar dus 'n volledige stel klemtoonreëls vir Afrikaans ontwikkel. Daar word ook 'n klemtoontoekenner vir Afrikaans beskikbaar gestel, wat getoets is met 10 000 tekseenhede. Hierdie 10 000 tekseenhede word ook korrek geannoteer, sodat dit in die toekoms gebruik kan word as toetsdata vir die afrigting van ander algoritmes.

1.4 NAVORSINGSMETODE

Die navorsing in die studie val in drie fases uiteen. Dit is die ontwikkeling van nuwe klemtoonreëls, die ontwikkeling van voldoende toetsdata, en die implementering van die klemtoonreëls. In die vierde afdeling is die klemtoontoekenner geëvalueer.

1.4.1 Die ontwikkeling van nuwe klemtoonreëls vir Afrikaans

Op grond van 'n literatuurstudie met betrekking tot klemtoon en klemtoonreëls in Afrikaans en ook in ander tale met soortgelyke taalstrukture, word nuwe klemtoonreëls ontwikkel. Die klemtoonreëls moet linguisties akkuraat wees en moet ook van so aard wees dat dit in die outomatiese klemtoontoekenner geïmplementeer kan word. Slegs primêre klemtoon word toegeken aan simplekse woorde en afleidings. Sekondêre klemtoon word nie toegeken nie, omdat die byvoeging daarvan min effek het op die akkuraatheid van die grafeem-na-foneemomsetter (sien afdeling 4.3.2.4). Aangesien hierdie klemtoontoekenner in toekomstige studies deel kan uitmaak van 'n teks-na-spraak-stelsel, word primêre klemtoon aan al die woorde binne 'n samestelling toegeken. Dus word meer as een primêre klemtoon aan samestellings toegeken deur die klemtoontoekenner.

Die ontwikkeling van die klemtoonreëls kan in verskillende fases opgedeel word:

In die eerste ontwikkelingsfase word reëls ontwikkel wat alle onbeklemtoonbare affikse identifiseer. Dié affikse moet uitgewys word, sodat die kern van die woord wat beklemtoon moet word makliker geïdentifiseer kan word en ook sodat dié affikse nie beklemtoon word nie. Die reëls is gegenereer deur reeds ontwikkelde klemtoonreëls vir Afrikaans in ag te neem en ook deur na Afrikaanse woordpatrone te kyk. Reëls vir die onbeklemtoonbare sjwa word ook saam met dié affiksreëls ontwikkel. Dit is algemeen bekend dat daar verskeie affikse in Afrikaans is wat hoofklemtoon dra. Die reëls wat gegenereer word om beklemtoonbare affikse te identifiseer en klemtoon daaraan toe te ken, is ook in dié fase ontwikkel.

Die reëls wat bepaal of klemtoon op die eerste of tweede lettergreep van 'n tweelettergreepige woord geplaas moet word, word in die tweede fase van ontwikkeling gegenereer. Verskeie faktore bepaal waar die klemtoon in 'n tweelettergreepige woord geplaas word, soos die gewig van 'n sillabe, of dit 'n oop of 'n geslote lettergreep is en of die woord 'n simpleks of 'n afleiding is.

In die derde fase word reëls vir drie- en meerlettergreepige simplekse ontwikkel. Die klemtoon van dié simplekse verskil baie van mekaar. Dus moet 'n deeglike studie gedoen word om reëls te ontwikkel wat die klemtoon korrek aan simplekse kan toeken.

Uitsonderings op reëls is ook geïdentifiseer tydens elke fase.

Die ontwikkeling van die klemtoonreëls moet die res van die navorsing voorafgaan, aangesien die reëls geïmplementeer moet word in die klemtoontoekenner voordat evaluering kan plaasvind.

1.4.2 Die ontwikkeling van toetsdata

Toetsdata word gegenereer vir die evalueringsfase van die projek. 'n Woordelys word ontwikkel en met die oog op evaluering word twee tekste geïdentifiseer as toetsdata.

'n Woordelys bestaande uit 10 000 Afrikaanse woorde word outomaties verdeel in woorddele en lettergrepe deur die lettergreepverdeler, Calomo, en die kompositumanaliseerder, CKarmaⁱⁱⁱ, te gebruik. Vervolgens word hierdie lys nagegaan om enige foute te korrigeer. Dié lys word ontwikkel om as toetsdata tydens die evaluasie van die klemtoontoekenner gebruik te word.

ⁱⁱⁱ Die C5 Afrikaanse Lettergreepverdeler vir Outomatiese Morfologiese Ontleding en die C5 Kompositumanaliseerder vir Robuuste Morfologiese Analise is ontwikkel by die Sentrum vir Tekstegnologie aan die Noordwes-Universiteit.

Die tekste wat gebruik word tydens die evalueringsfase is dié van 'n paar boeke van die 1983-vertaling van die Afrikaanse Bybel en die Universele Verklaring van Menseregte. Hierdie tekste moet dieselfde fases deurgaan as bogenoemde woorde, voordat dit as toetsdata gebruik kan word.

1.4.3 Die implementering van reëls

Daar word in hierdie studie gepoog om die bestaande klemtoontoekeener te verbeter deur die klemtoonreëls wat in die toekenner gebruik is te verbeter. Dieselfde benadering wat in die bestaande toekenner gebruik is, word ook in die nuwe toekenner gebruik, aangesien daar deur middel van die verbetering van klemtoonreëls gepoog word om die klemtoontoekeener te verbeter.

Die benadering wat gevolg word om reëls te implementeer, is 'n reëlgebaseerde benadering. Die reëls wat ontwikkel word in die eerste fase van hierdie studie word in die vorm van 'n stel *AS-DAN*-reëls oorgeskryf in die klemtoontoekeener deur die programmeertaal, *Perl v.5.8.8*, te gebruik. Die reëls is kontekstsensitief en word geskryf in die vorm van $X/Y/Z \rightarrow X/Y/Z$. Dit beteken dat klemtoon aan die lettergreep *Y* toegeken word (aangedui deur 'n kommapunt (;) vooraan te plaas), indien dit binne die konteks van *X* en *Z* voorkom.

1.4.4 Die toetsproses en evaluasie van die klemtoontoekeener

In dié afdeling van die studie word die klemtoontoekeener eers getoets en daarna op verskeie maniere geëvalueer.

Die toetsproses het behels dat die 10 000 tekseenhede wat verkry is, gebruik word om die toekenner te toets: Die woordelys word ingedeel in twintig lyste van 500 tekseenhede elk. Dus is daar twintig fases in die toetsproses, wat elkeen uit twee stappe bestaan. In die eerste stap word klemtoon aan die 500 tekseenhede toegeken, waarna die resultate verkry word deur die uitvoer met die hand na te gaan vir foute. Tydens die tweede stap word daar gepoog om die foute wat deur die klemtoontoekeener begaan is, te korregeer deur reëls te verbeter of weg te laat, of deur die volgorde van reëls te verander. Sodra dié twee stappe uitgevoer is, word die volgende lys tekseenhede ingevoer. Dieselfde stappe geld vir die volgende lyste tot al 10 000 tekseenhede getoets is.

Nadat die toetsproses afgehandel is, word die klemtoontoekeener geëvalueer. Eerstens word die akkuraatheid van die klemtoontoekeener bepaal. In die tweede fase word die akkuraatheid van die grafeem-na-foneem-omsetter bepaal. Daarna word die invloed van verskillende faktore op klemtoontoekeening en grafeem-na-foneem-omsetting vasgestel en bespreek. Laastens word daar

bepaal wat die invloed van klemtoontoekenning op grafeem-na-foneem-omsetting is en of dit wel van groot belang is vir 'n reëlgebaseerde grafeem-na-foneem-omsetter.

1.4.4.1 Fase 1: Evaluasie van die klemtoontoekenner

Die toetsdata word gebruik om die klemtoontoekenner te evalueer. Die bestaande toekenner word ook geëvalueer, sodat die twee klemtoontoekenners vergelyk kon word. Die toekenners word as volg geëvalueer:

1. Klemtoon word aan die lys van 10 000 woorde toegeken deur die klemtoontoekenner en die bestaande klemtoontoekenner. Die toegekende klemtoon word daarna vergelyk met die korrekte klemtoon van die toetsdata, en die akkuraatheid van die twee klemtoontoekenners word dan met mekaar vergelyk.
2. Om te bepaal hoe akkuraat die klemtoontoekenning op lopende teks is, word die Bybeltekste gebruik en ook die teks van die Afrikaanse weergawe van die Universele Verklaring van Menseregte. Dieselfde word met die tekste gedoen as by 1.

Hierdie resultate word dan vergelyk om vas te stel of die nuwe klemtoontoekenner 'n verbetering is op die bestaande toekenner.

1.4.4.2 Fase 2: Evaluasie van die grafeem-na-foneem-omsetter

Die grafeem-na-foneem-omsetter word geëvalueer deur van die woordelys gebruik te maak. Die 10 000 tekseenhede word getranskribeer deur die omsetter en daarna word die akkuraatheid van die grafeem-na-foneem-omsetter met die nuwe klemtoontoekenner vergelyk met die resultate van die omsetter met die bestaande klemtoontoekenner. Informele toetse wat gedoen is, het bepaal dat tussen 65% en 70% van alle omsettingsfoute veroorsaak is deur klemtoon wat verkeerd toegeken is. Dus is die verwagting dat die grafeem-na-foneem-omsetter se akkuraatheid heelwat sal verbeter met die gebruik van die nuwe klemtoontoekenner.

1.4.4.3 Fase 3: Die invloed van morfologiese analise op klemtoontoekenning

Die faktore wat die akkuraatheid van die klemtoontoekenner en die grafeem-na-foneem-omsetter beïnvloed, is die lettergreepverdeling van woorde en ook die morfologiese analise van woorde. Klemtoon kan nie deur die toekenner aan woorde toegeken word as die woorde nie in lettergrepe opgedeel is nie, omdat klemtoon juis aan lettergrepe toegeken word. Dus word die toets slegs op morfologies-geanaliseerde woorde gedoen. 'n Woordelys waarvan morfologiese analise van

woorde weggelaat en bygevoeg is, word gebruik en die resultate van die woordelyste word met mekaar vergelyk om die invloed van dié faktor op klemtoontoekenning te bepaal.

1.4.4.4 Fase 4: Die invloed van klemtoon op grafeem-na-foneem-omsetting

Laastens word die invloed van klemtoon op grafeem-na-foneem-omsetting bepaal. Dié evaluering bestaan uit drie toetse: In die eerste toets word die omsetter geëvalueer deur 'n woordelys te genereer waaraan klemtoon 100% korrek toegeken is. Nadat die eerste toets gedoen is, word daar in die tweede toets sekondêre klemtoon bygevoeg in die lys en resultate is dan verkry. In die eerste twee toetse word die klemtoontoekenner uit die program gehaal, omdat klemtoon reeds toegeken is aan die woorde wat ingevoer word. In die laaste toets word die klemtoontoekenner weer bygevoeg in die program en klemtoon word dan deur die klemtoontoekenner aan woorde toegeken en daarna getoets. So kan 'n goeie vergelyking getref word om te bepaal wat die invloed van klemtoon op grafeem-na-foneem-omsetting is.

1.5 ONTPLOOIING

In Hoofstuk 2 word daar eerstens aandag gegee aan die bestaande klemtoonreëls vir Afrikaans en klemtoonreëls van die Nederlandse taal word ook bestudeer. Daarna word die klemtoonreëls vir Afrikaans ontwikkel, gebaseer op dié van Combrink en De Stadler (1987), Wissing (1971) en Lubbe (1993). Uitgebreide motiverings vir die uiteindelijke stel klemtoonreëls word aangebied.

Die vereistes waaraan die klemtoontoekenner moet voldoen word bespreek in Hoofstuk 3. Verskillende benaderings tot outomatiese klemtoontoekenning word ook bekyk. Aangesien die reëlgebaseerde benadering in hierdie projek gebruik word, word spesifiek aandag gegee aan hierdie benadering. Hieruit sal dit duidelik blyk waarom die reëlgebaseerde benadering in hierdie studie gebruik word.

In Hoofstuk 4 word die implementering van die klemtoonreëls verduidelik, gevolg deur 'n in-diepte-bespreking van die toetsproses. Verder word die klemtoontoekenner en die grafeem-na-foneem-omsetter geëvalueer soos hierbo beskryf. Die evalueerproses word behandel en die resultate word bespreek. Hierdie resultate word ook vergelyk met die resultate van bestaande klemtoontoekenners vir ander tale.

'n Samevatting van die projek word dan in Hoofstuk 5 gegee. Gevolgtrekkings word gemaak oor die nuwe klemtoonreëls, die benadering wat gevolg is en die klemtoontoekenner wat ontwikkel is. Verder word daar ook voorstelle en aanbevelings gemaak vir die ontwikkeling van 'n

doeltreffende en herbruikbare klemtoontoekeener vir Afrikaans met behulp van die geannoteerde toetsdata wat gegenereer kan word met die hulpbronne wat hier ontwikkel is.

Hoofstuk 2

KLEMTTOONREËLS VIR AFRIKAANS

2.1 INLEIDING

In dié studie word 'n klemtoon-toekenner gebou wat klemtoonreëls gebruik om klemtoon aan woorde toe te ken. Dit is dus nodig om 'n volledige stel klemtoonreëls te genereer, voordat die toekenner ontwikkel kan word.

In afdeling 2.2 van hierdie hoofstuk word die verskillende benaderings vir die voorstelling van klemtoonreëls bespreek. Die beste benadering vir die voorstelling van klemtoonreëls vir hierdie studie word dan gekies. Die klemtoonreëls vir Afrikaans van Combrink en De Stadler (1987), Wissing (1971), en Lubbe (1993) word in afdeling 2.3 behandel.

Verder geskied die uitbreiding en verbetering van die bestaande Afrikaanse klemtoonreëls, asook die generering van nuwe klemtoonreëls ook in afdeling 2.3, om sodoende die nodige stel klemtoonreëls vir hierdie studie te genereer. Die klemtoonreëls vir Nederlands deur Trommelen en Zonneveld (1989) en Van der Hulst (1984) word ook gebruik as basis vir die ontwikkeling van die nuwe klemtoonreëls.

Die klemtoonreëls wat hier ontwikkel word, word dan geïmplementeer in die klemtoon-toekenner en gebruik om klemtoon aan woorde toe te ken.

2.2 DIE VOORSTELLING VAN KLEMTTOONREËLS

In dié afdeling word drie voorstellings van klemtoonreëls bespreek en vergelyk, om daardeur die beste voorstelling te vind vir die Afrikaanse klemtoonreëls wat in die studie gegenereer word.

Die drie voorstellings wat behandel word is:

- Liniêre klemtoonanalise;
- Die metriese raamwerk (*grid*);
- Metriese bome.

2.2.1 Liniêre Klemtoonanalise

Die metode wat gevolg word om liniêre woordklemtoon voor te stel, is deur Chomsky en Halle (1968) opgestel in *The Sound Pattern of English* (hierna genoem SPE), en word deur baie fonoloë gebruik om hul klemtoonreëls te ontwerp. Enkeles hiervan is Halle en Keyser (1971), Wissing (1971) en Combrink en De Stadler (1987).

Die SPE klemtoonreëls het die volgende vorm:

$$V \rightarrow [1\text{stress}] / X_Y$$

Dié reël kan as volg gedefinieer word: Hoofklemtoon word aan 'n vokaal V toegeken, indien dit in die konteks X_Y voorkom, waar X aan V se linkerkant staan en Y aan die regterkant. Die taak van die klemtoonreël is om aan die vokaal 'n klemtoonwaarde te gee, hetsy hoofklemtoon, byklemtoon of geen klemtoon. Aan die einde van die afleiding kry die vokale wat beklemtoon word dan die kenmerk, $[1\text{stress}]$.

SIMBOOL	BESKRYWING
V	Enige vokaal, insluitende y
C	Enige konsonant, y uitgesluit
C_0	Nul of meer konsonante
C_m^n	Minstens m en hoogstens n konsonante
X_Y	Die konteks waarbinne V voorkom
$[1\text{stress}]$	Hoofklemtoon
$\left[\begin{smallmatrix} -\text{tense} \\ V \end{smallmatrix} \right]$	'n Ongespanne vokaal wat 'n swak groep aandui

Tabel 1: Tegniese notasie van die SPE-reëls

Die SPE-metode bevat sterk en swak groepe, waarvan sterk groepe eerder beklemtoon sal word as swak groepe.

In die onderstaande reël word Chomsky en Halle (1968:71) se klemtoonreël vir Engelse werkwoorde weergegee:

$$V \longrightarrow [1\text{stress}] / \text{---}C_0 \left(\begin{bmatrix} -\text{tense} \\ V \end{bmatrix} \right) C_0^1]$$

waar C_0 nul of meer konsonante is en C_m^n minstens m

en hoogstens n konsonante is en $\begin{bmatrix} -\text{tense} \\ V \end{bmatrix}$ is 'n swak groep.

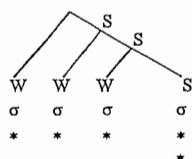
Dus beteken dié reël dat hoofklemtoon toegeken word aan die tweedelaaste vokaal van 'n woord indien die laaste vokaal ongespanne is en gevolg word deur hoogstens een konsonant, anders word die laaste vokaal beklemtoon.

Liberman en Prince (aangehaal deur Trommelen & Zonneveld, 1989:45) maak die stelling dat die SPE-klemtoonanalise die beste beskikbare teorie vir klemtoonreëls was op 'n bepaalde tyd in die ontwikkeling van generatiewe fonologie. Volgens Trommelen en Zonneveld (1989:45), word metriese fonologie as opvolger verkies bo die SPE-metode. Die redes hiervoor sal in die volgende deel bespreek word.

2.2.2 Die Metriese Raamwerk

Die rol van die metriese raamwerk is volgens Van der Hulst (1984:136) om metriese bome te interpreteer. Die raamwerk gee duidelikheid oor watter sillabes of woorddele toepaslik is vir die karakterisering van die klempatroon van 'n string.

'n Eenvoudige voorbeeld van 'n metriese raamwerk wat aan 'n boom toegeken is, sal soos volg lyk:



Figuur 1: 'n Eenvoudige metriese raamwerk

Die asteriske verteenwoordig die elemente van 'n woord wat beklemtoonbaar is. Dus, hoe hoër die graad van klemtoon, hoe meer is die asteriske.

Halle en Vergnaud (1987:227) gebruik slegs die metriese raamwerk om klemtoon voor te stel. 'n Voorbeeld van dié metriese raamwerk word deur Halle en Vergnaud (1987:44) gegee:

.	.	*	.	.	*	lyn 2				
(	*	.	*	)	(	*	.	*	)	lyn 1
(	*	*	)	(	*	)	(	*	)	lyn 0
Tennessee	Japanese									

Figuur 2: Die metriese raamwerk sonder die boomstruktuur

In Liberman en Prince se studie (Trommelen & Zonneveld, 1989:239) word weergegee dat dit die doel van die metriese raamwerk is om die voorkoms van 'n klemtoonbotsing uit te wys. 'n Klemtoonbotsing kom voor wanneer twee lettergrepe langs mekaar dieselfde klemtoonsterkte ontvang.

Die metriese boom sowel as die raamwerk word volgens Trommelen en Zonneveld (1989:240) deur Liberman en Prince gebruik. Halle en Vergnaud (1987:227) gebruik weer slegs die metriese raamwerk vir hul voorstelling van klemtoon.

2.2.3 *Metriese bome*

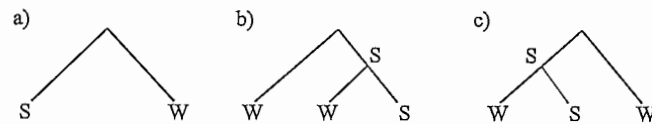
In Van der Hulst (1984:135) word twee metriese teorieë, naamlik die metriese boom en die metriese raamwerk met mekaar vergelyk. Hieruit is gevind dat daar nie 'n groot verskil is in die werking van dié teorieë nie.

Metriese bome het die volgende struktuur:



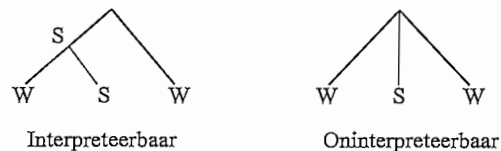
Figuur 3: Struktuur van metriese bome

Om relatiewe prominensie aan te dui, word S^o e en W 's aan konstituente toegeken. As 'n konstituent sterker is as sy susterkonstituent, word dit deur 'n S aangedui en 'n swakker konstituent word deur 'n W aangedui:



Figuur 4: Relatiewe prominensie in metriese bome

Die konstituente is dus afhanklik van mekaar om te bepaal watter een sterker is as die ander. Om die rede is dit nodig dat die S^o e en W 's altyd in pare voorkom, om die bome interpreteerbaar te maak:



Figuur 5: Die linkerkantste boom is interpreteerbaar, omdat die nodes in pare voorkom.

Die bome kan woordklemtoon of fraseklemtoon toeken. Dit hang net af of die eindpunte van die bome sillabes of woorde is. In dié studie word slegs woordklemtoon toegeken, dus sal die voorbeelde wat verder gebruik gaan word, sillabes as eindpunte hê.



Figuur 6: In klemtoontoekenning word sillabes as eindpunte vir bome gebruik

In figuur 6 kan gesien word dat die hoofklemtoon van die woord gevind word deur die pad te volg wat uitsluitlik uit S^o e bestaan. Al die ander verteenwoordig laer klemtoongrade. Bome met sillabes as eindpunte mag nooit aan beide kante ingebed wees nie. Dit kan veroorsaak dat twee konstituente dieselfde klemtoongraad het, wat dan beteken dat 'n woord op twee sillabes hoofklemtoon dra.

Die parameters wat klemtoonstelsels karakteriseer is die volgende:

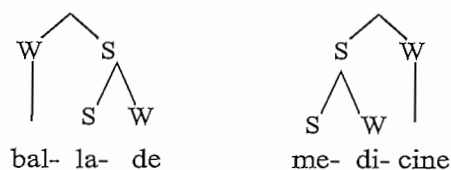
- Gebonde of ongebonde
- Kwantiteitgevoelig
- Nie-kwantiteitgevoelig

'n Taal wat gebonde voete het, se getal lettergrepe in 'n klemtoonvoet word beperk tot twee, waar dié met ongebonde voete se getal lettergrepe vry is en dus geen beperkings op het nie. Kwantiteitgevoeligheid word ook genoem swaartegevoeligheid. In dié tipe bome mag daar nie 'n vertakking in die *W*-posisie voorkom nie. Onderskeid word getref tussen swaar en ligte sillabes en dus tussen ook lang en kort sillabes. Verder word die moontlike kombinasies van die eindknope van 'n voet beperk. Indien 'n taal nie-kwantiteitgevoelig is, sal kwantiteit geen rol speel in die ligging van klemtoon nie. Dié tipe tale het 'n vasgestelde klempatroon, waar klemtoon byvoorbeeld slegs op die eerste of die laaste sillabe geleë is of dalk op elke tweede sillabe voorkom. Uit dié parameters kan vier tipe tale kom, naamlik:

- Kwantiteitgevoelig en gebonde voete
- Kwantiteitgevoelig en ongebonde voete
- Nie-kwantiteitgevoelig en gebonde voete
- Nie-kwantiteitgevoelig en ongebonde voete

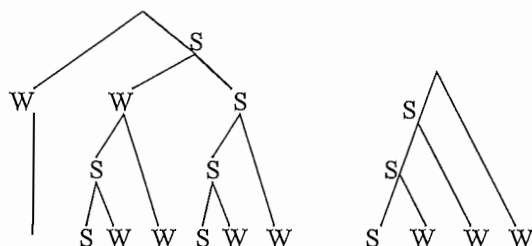
Volgens Trommelen en Zonneveld (1989:100), is Nederlands en Engels kwantiteitgevoelig en gebonde. Uit bogenoemde beskrywings van die parameters kan afgelei word dat Afrikaans in dieselfde kategorie as Engels en Nederlands val. Die volgende bome is 'n voorbeeld van elk van die verskillende parameters.

In tale wat kwantiteitgevoelig en gebonde is, word klemtoontoekenning beïnvloed deur die innerlike voorkoms van sillabes (Van der Hulst, 1984:154). Die gewig van 'n sillabe speel 'n groot rol in die tipe klemtoontoekenning.



Figuur 7: kwantiteitgevoelige en gebonde bome

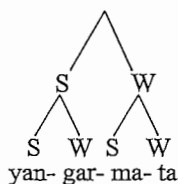
Kwantiteitgevoelige en ongebonde bome neem die volgende struktuur aan:



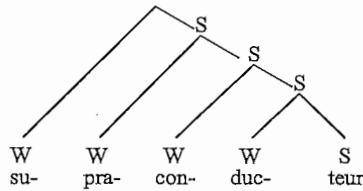
Figuur 8:Kwantiteitgevoelige en ongebonde bome

As daar vertakkende rym voorkom, sal hoofklemtoon altyd op die mees regse rym val en heel links indien daar geen vertakkende rym voorkom nie. Byklemtoon kom voor op elke eerste lettergreep sonder hoofklemtoon en ook op elke tweede vertakkende rym.

Tale wat nie-kwantiteitgevoelig en gebonde is kry klemtoon op elke ewe of onewe lettergreep. Die bome kan van links na regs toegekende *S-W* of *W-S* voete hê, of dit kan van regs na links toegekende *S-W* of *W-S* voete hê. Indien die bome van links na regs toegekende voete het, val die hoofklemtoon altyd op die eerste of tweede lettergreep van 'n woord en bome wat van regs na links toegekende voete het, kry hoofklemtoon op die laaste of tweedelaaste lettergreep van 'n woord. Die volgende voorbeelde van Trommelen en Zonneveld (1989:101) maak die begrip duidelik:



Figuur 9: Klemtoonstruktuur vir die taal Maranungku. Hoofklemtoon val altyd op die eerste lettergreep

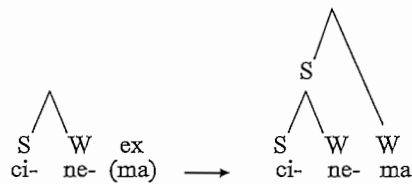


Figuur 13: Nie-kwantiteitgevoelige en ongebonde klemtoonstruktuur. Frans en Italiaans is 'n voorbeeld van dié tipe klemtoonstruktuur.

Nog 'n eienskap van metriese bome is ekstrametrisiteit. Volgens Halle en Vergnaud (1987:48) is 'n element wat ekstrametries gemerk is, onsigbaar vir die reëls wat metriese konstituente bou, slegs as

- die element die fonologiese string begin of eindig of
- as die element nie die string vorm nie.

'n Lettergreep wat 'n sjwa bevat sal as ekstrametries gemerk kan word, omdat dit geen gewig het nie en ook nie die klemtoon van die stam verander nie.



Figuur 14: Ekstrametrisiteit

Gebonde tale sal nooit bv. die derde, sesde of negende lettergreep van links of regs af, beklemtoon nie (Van der Hulst, 1984:144). Dus in 'n woord met ternêre voete wat die struktuur S-W-W het, moet die mees regse lettergreep as onsigbaar gesien word om die probleem op te los.

Van der Hulst (1984) en Trommelen en Zonneveld (1989) gebruik metriese bome om hul klemtoonreëls vir Nederlands voor te stel. Metriese fonologie verskil van die SPE-metode op verskeie maniere. In SPE is klemtoonreëls geformuleer wat formeel moontlik is, maar wat in natuurlike taal 'n baie klein waarskynlikheid het om voor te kom. Trommelen en Zonneveld (1989:40) gee 'n voorbeeld van dié klemtoonreëls:

$$V \rightarrow [1\text{stress}] / __[-\text{rond}] C_0[-\text{rond}] C_0]$$

$$V \rightarrow [+rond] / __ ((V.C_0^1) V.C_0^1)]$$

Dié reël fokus op die ongerondheid van vokale en dit is 'n vokaalrondingsreël wat werk volgens die swak-groep-beginsel. Nie een van die reëls is volgens hulle moontlik in natuurlike taal nie. Trommelen en Zonneveld (1989:40) is van mening dat die metriese bome die probleem van die moontlike voorkoms van onmoontlike klemtoonreëls aansienlik verbeter het.

In plaas daarvan om klemtoon met fonologiese kenmerke soos [-tense] uit te druk, word dit in bome en eienskappe van bome uitgedruk.

SPE se klemtoonverlagingsreël het verval en is in die metriese bome vervang deur die *S-W* verhouding tussen susterknope.

Nog 'n verskil is dat die klempatrone nie meer siklies hoef te verloop nie.

Die laaste verskil wat hier genoem gaan word, is die gebruik van lettergrepe eerder as groepe. Metriese bome het lettergrepe as eindknope, waar SPE se klemtoonreëls gebaseer is op groepe ("clusters"). Daar word dus afgelei dat SPE beweer dat lettergrepe geen rol speel in klemtoontoekenning nie. Dié studie ondersteun nie die gebruik van die groep nie, maar gebruik die lettergreep as klemtoondomein. Alhoewel Chomsky en Halle (1968:70) die groep gebruik het, het Combrink en De Stadler (1987:184) in hul liniêre benadering hul Afrikaanse klemtoonreëls gebaseer op lettergrepe. Dus is die verskil wat hier genoem word, slegs tussen metriese bome en SPE se liniêre klemtoonreëls en nie 'n beswaar teen die liniêre benadering as sulks nie.

Trommelen en Zonneveld (1989:41) noem ook dat die eienskappe van klemtoon nie moet verwys na die eienskappe van individuele segmente nie, maar eerder 'n hiërargiese ritmiese struktuur reflekteer wat sillabes, woorde, of sintaktiese frases organiseer.

Nadat al drie die metodes bespreek is, is dit duidelik dat metriese fonologie bo die liniêre reëls vir hierdie studie gekies moet word. Dus gaan metriese bome in hierdie studie gebruik word om die Afrikaanse klemtoonreëls voor te stel.

2.3 KLEMTOUNREËLS VIR AFRIKAANS

Die beskikbare Afrikaanse literatuur wat handel oor klemtoon vir Afrikaans is dié van De Villiers (1965), Wissing (1971), Combrink en De Stadler (1987) en Lubbe (1993). Uit dié literatuur is dit slegs Wissing, Combrink en De Stadler en Lubbe wat aandag skenk aan klemtoone reëls vir Afrikaans.

2.3.1 Bestaande klemtoonreëls vir Afrikaans

Beide Wissing (1971) en Combrink en De Stadler (1987) gebruik die liniêre benadering vir hul klemtoonreëls. Wissing het slegs op simplekse selfstandige naamwoorde gewerk. Combrink en De Stadler weer, het met simplekse selfstandige naamwoorde, komplekse (samestellings en afleidings), en letter- en syfername gewerk. Byklem is ook kortliks deur hulle bespreek. Lubbe(1993) maak gebruik van metriese fonologie om klemtoon toe te ken.

Ten einde hul klemtoonreëls te formuleer, onderskei hulle tussen sterk en swak groepe. Combrink en De Stadler (1987:190) stel voor dat die onderskeid tussen sterk en swak groepe eerder op 'n skaal geplaas moet word met sterk en swak groepe as twee uiterste pole. Die vokaalsterktehiërargie is een van die faktore wat in gedagte gehou moet word as sillabes op 'n skaal geplaas word.

Verder maak Wissing (1971:61) gebruik van die hulpmiddels wat deur Chomsky en Halle (1968:340) daargestel is. Die hulpmiddel wat deur Wissing (1971:61) as die belangrikste geag word is die moontlikheid om disjunktief-geordende reëls in een omvattende reël saam te vat. Volgens Chomsky en Halle (1968:342) blyk dit dat die hipotese dat reëls georden is een van die mees ondersteunde aannames van linguistiese teorie is. Die ordening van reëls maak ook betekenisvolle veralgemenings van reëls moontlik. Die hoofklemtoonreël van Wissing (1971:62) vir simplekse selfstandige naamwoorde word in figuur 15 weergegee.

$$V \rightarrow [1klem] / X_C_0 \left(\left\{ \begin{bmatrix} + \text{vok} \\ - \text{gespan} \\ + \text{vok} \\ + \text{mid} \\ - \text{voor} \\ - \text{agter} \end{bmatrix} \right\}_{C_0^1} (C_0 \left\{ \begin{bmatrix} + \text{vok} \\ - \text{gespan} \\ + \text{vok} \\ + \text{mid} \\ - \text{voor} \\ - \text{agter} \end{bmatrix} \right\}_{C_0^1}) \right)$$

Figuur 15: Die hoofklemtoonreël van Wissing (1971) vir simplekse selfstandige naamwoorde.

Dié reël is disjunktief georden, dus kan dit opgedeel word in drie klemtoonreëls wat in die volgorde *a-c* moet geskied:

- a. Indien die laaste en voorlaaste groepe swak is, sal die derdelaaste groep beklemtoon word;
- b. Indien slegs die laaste groep swak is, sal die tweedelaaste groep beklemtoon word;
- c. Finale klemtoon word toegeken aan alle woorde wat nie deur a of b bevredig is nie.

In dié reël word dieselfde notasie gebruik as SPE. Die swak groep word aangedui deur die notasie, $\begin{bmatrix} +\text{wak} \\ -\text{gespan} \end{bmatrix}$. Verder word ook aangedui dat 'n swak groep die kenmerk [+mid] het. Dus moet die ongespanne groep 'n middelvokaal bevat, wat beteken die groep moet 'n onbeklemtoonbare sjwa bevat.

Die hoofklemtoonreël vir simplekse selfstandige naamwoorde van Combrink en De Stadler (1987:193) is die volgende:

$$V \rightarrow \begin{matrix} X & & V & & Y \\ [1\text{klem}] / (\begin{bmatrix} [n \text{ sterk}] \end{bmatrix}) \begin{bmatrix} [m \text{ sterk}] \end{bmatrix} (\begin{bmatrix} [p \text{ sterk}] \end{bmatrix}) \end{matrix} \begin{matrix} N \end{matrix}$$

Figuur 16: Combrink en De Stadler (1987) se hoofklemtoonreël.

X en Y is enige getal sillabes en m, n en p dui verskeie posisies op die sillabesterktehiërargie (sien 2.3.2.1) aan en $m \geq p$, $m > n$. Dus word hoofklem toegeken aan die eerste sterkste sillabe in simplekse selfstandige naamwoorde. Combrink en De Stadler (1987:196) noem wel dat die wisselwerking tussen die dele van die sillabesterktehiërargie altyd duidelik is nie. As daar gekyk word na die sillabesterktehiërargie, dan moet die sillabe wat 'n diftong bevat altyd hoofklem dra, maar by woorde soos *luitenant* en *kruisement* is dit nie die geval nie.

Combrink en De Stadler (1987:200) bespreek ook die invloed wat affikse op klemtoon binne 'n afleiding het. 'n Afleiding word hoofsaaklik gevorm deur middel van voor- en agtervoegsels. Die meeste Afrikaanse affikse word nie beklemtoon nie, maar daar is wel 'n paar affikse wat hoofklem dra en dan ook dié wat die klemtoon in hul rigting trek. Combrink en De Stadler (1987:200) deel

die prefikse op in dié wat hoofklem dra en dié wat klemneutraal is. Die suffikse word weer in die volgende groepe opgedeel:

- die klemneutraal-suffikse, wat nie die hoofklem dra of die posisie van die hoofklem verander nie.
- die suffikse wat die hoofklem dra, en
- die suffikse wat klem in hul rigting trek en dus veroorsaak dat die hoofklem verskuif tot in die sillabe net voor die suffiks.

Die hoofklemdraende suffikse is volgens Combrink en De Stadler (1987:201) almal Romaanse suffikse. Dus beskryf hulle dit deur die volgende reël:

$$[+sill] \rightarrow [1klem] / S_{St} + \left[\begin{array}{c} [_ Y] \\ [+ Romaans_{sf}] \end{array} \right]$$

Figuur 17: Hoofklemtoon word aan klemdraende suffikse toegeken.

Die suffiks *-ies*, is die enigste uitsondering op die reël, bv. *atleties*, *parasities*, ens. Die reël van die volgende groep wat die klemtrekkende suffikse bevat, sluit dié uitsondering in:

$$[+sill] \rightarrow [1klem] / X_Y + [Z_{sf}]_A$$

waar $Z = -(e)lik, -ig, -saam, -(e)loos, -ies$

Figuur 18: Klemtrekkende suffikse word deur die reël toegeken.

Laastens bespreek Combrink en De Stadler (1987:195) ook die byklemreëls vir simplekse selfstandige naamwoorde. Vir die toekenning van byklem word voorgestel dat die hoofklemtoonreël vir simplekse selfstandige naamwoorde aangepas word om die klemtoonreël herhalend toe te pas, sodat dit opeenvolgens [1klem], [2klem], [3klem] kan toeken. Dus moet die reël eers die sterkste onbeklemtoonde sillabe soek en [1klem] daaraan toeken en daarna die volgende sterkste onbeklemtoonde sillabe soek en 'n laer klemtoongraad daaraan toeken as in die vorige toepassing.

Lubbe (1993: 78) bespreek die invloed van superswaar sillabes op klemtoon in Afrikaans. Hy noem dat slotklem in Afrikaans baie uitsonderlik is, behalwe aan woorde wat op 'n superswaar

sillabe eindig, bv. *president, vulkaan, kardoës*. Superswaar sillabes bestaan uit meer as twee segmente en het dus die sillabevorm VVK (*magneet*) of VKK (*vakant*). Dit kan ook in uitsonderlike gevalle die vorm VVKK (*akkoord*) aanneem (Lubbe, 1993:81). Die invloed van die sjwa op klemtoon in Afrikaans word in Lubbe (1993 : 87) behandel. Daar word genoem dat die sjwa gewigloos is, en dus geen klemtoon bevat nie. Dit het wel 'n invloed op klemtoon, deurdat 'n woord met 'n slotsjwa, wat voorafgegaan word deur 'n konsonant, voorfinale klemtoon ontvang (*artikel*) en 'n woord met 'n slotsjwa wat voorafgegaan word deur 'n vokaal, kry dan pre-voorfinale klemtoon (*België*).

Afrikaanse klemtoon is volgens Combrink en De Stadler (1987:201) 'n baie komplekse sisteem wat nie in een enkele klemtoonreël saamgevat kan word nie. Dus is verskeie klemtoonreëls nodig om al die klemverskynsels in Afrikaans te beskryf.

2.3.2 *Faktore wat bydra tot klemtoontoekenning*

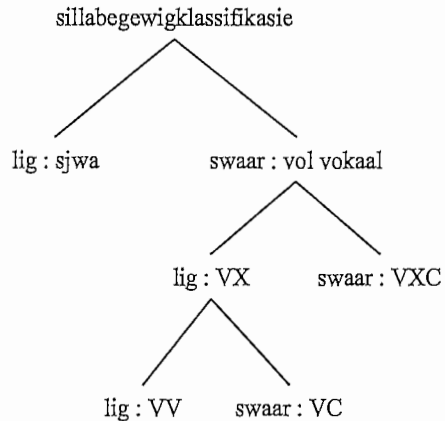
2.3.2.1 Die gewig van sillabes

Die vokaal van 'n sillabe is die element wat die belangrikste rol speel by die bepaling van 'n sillabe se gewig. Verder is die konsonante wat die sillabekern (die vokaal) omring ook in sekere gevalle 'n aanduider van of 'n sillabe swaar of lig is.

Combrink en De Stadler (1987:192) bepaal die gewig van 'n vokaal deur die vokaalhiërargie:

$$e > o > i > E > u > a > 9 > @ \\ O \\ y$$

Van der Hulst (1984:209) klassifiseer sillabegewig met die boom in figuur 19. By die eerste vertakking word uitgedruk dat die sjwa die heel ligste vokaal is en daar kan ook afgelei word dat 'n sillabe met 'n sjwa nooit hoofklemtoon kan dra nie. Die stelling word ook ondersteun deur Combrink en De Stadler (1987:192) se vokaalhiërargie, waar die sjwa ook aangedui word as die swakste vokaal.



Figuur 19: Sillabegewigklassifikasieboom van Van der Hulst (1984)

Volgens die boom is 'n vol vokaal die swaarste. 'n Sillabe wat 'n vol vokaal bevat, het die vorm VVC, dit wil sê 'n diftong of 'n lang vokaal gevolg deur 'n konsonant. Dié sillabes word geklassifiseer as superswaar. By die tweede vertakking is 'n sillabe met die vorm VXC die swaarste. Die sillabes (VCC) kom voor in woorde soos *eksperiment*, *ambulans*, *president*. As daar gekyk word na 'n woord soos *Roemeen*, waar beide sillabes uit vol vokale bestaan, sal die slotsillabe die swaarste wees, omdat die vokaal gevolg word deur 'n konsonant. Vroeër in die hoofstuk is Lubbe (1993) se siening oor swaar sillabes bekyk. Lubbe (1993:78) se reëls stem ooreen met dié van Van der Hulst (1984:209), wat hier behandel word.

Die sillabe VX is 'n ligte sillabe en kan verdeel word in die vorm VV en VC, waar $VV < VC$. Dié tipe sillabe maak nog 'n belangrike stelling: 'n Geslote sillabe is swaarder as 'n oop sillabe.

Combrink en De Stadler (1987:191) se sillabesterktehiërargie vir Afrikaans is baie dieselfde as Van der Hulst (1984:209) se sillabegewigklassifikasie vir Nederlands. Dit word soos volg opgesom:

- a) $X[+\text{diftong}] Y >$
- b) $XVK_2 >$
- c) – n. Volgens die vokaalhiërargie:

e > o > i > E > u > a > 9 > @
 O >
 y

d) Indien vokale van ongeveer gelyke sterkte:

- i. \$XVK₁\$ > \$XV\$
- ii. \$XVK_n\$ > \$XVK_{n-1}\$
- iii. \$XV[+gespan]\$ > \$XV[-gespan]\$

waar > = “sterker as” en \$ = die sillabegrens.

As die gewigskaal wat Van der Hulst (1984:209) aanneem en dié van Combrink en De Stadler (1987:191) vergelyk word, wil dit voorkom asof die reëls van sillabegewig vir Afrikaans en Nederland dieselfde is:

Van der Hulst (1984:209) se gewigskaal:

|-----|-----|-----|-----|
 @ VV VC VCC VVC
 ----- toename in gewig----->

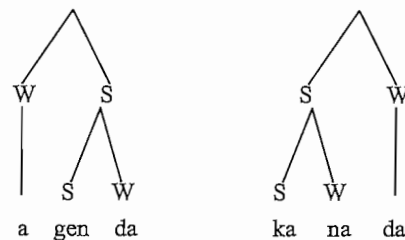
Combrink en De Stadler (1987:191) se gewigskaal volgens hul sillabesterte hiërargie:

|-----|-----|-----|-----|
 @ XV XVK₁ XVK₂ diftonge
 ----- toename in gewig----->

2.3.2.2 Oop en geslote lettergrepe

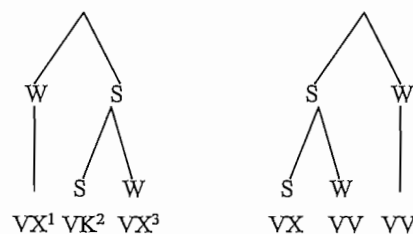
Die verskil in klemtoongedrag tussen VV- en VK-sillabes is te danke aan die feit dat VV 'n oop lettergreep is en VK 'n geslote lettergreep. Vroeër in die hoofstuk is genoem dat 'n geslote sillabe swaarder is as 'n oop sillabe en dus eerder beklemtoon sal word. Trommelen en Zonneveld (1989:147) verduidelik die verskil tussen VV- en VK-sillabes deur te noem dat VX-sillabes beide eindklemtoon kry, maar die verskil kom te voorskyn as die derdelaaste lettergreep hoofklemtoon kry. 'n Voorfinale VV-sillabe mag verbygegaan word as hoofklemtoon op die derdelaaste sillabe

val (bv. *Kanada*), maar indien die tweedelaaste sillabe VK is, mag dit nie verbygegaan word om hoofklemtoon op die derdelaaste sillabe te plaas nie (bv. *agenda*). Dus kry die tweedelaaste sillabe dan hoofklemtoon.



Figuur 20: Woorde wat eindig op die sillabevorm VK-VV sal beklemtoon word op die tweedelaaste lettergreep. Indien die sillabevorm VV-VV is, sal dit op die derdelaaste lettergreep beklemtoon word.

By finale VK kan voorfinale VV nie hoofklemtoon kry nie en by finale VX moet voorfinale VK hoofklemtoon dra, tensy die voorfinale VK 'n sjwa bevat. Die rol van die sjwa in klemtoontoekenning sal later in die hoofstuk bespreek word. Bogenoemde reëls sal dus soos volg lyk vir Afrikaans:



Figuur 21: Reëls vir Afrikaanse woorde wat eindig in die vorm VK-VX en VV-VV.

VK² mag nie 'n sjwa bevat nie.

Volgens die gewigskale van Van der Hulst (1984) en Combrink en De Stadler (1987) kan ook gesien word dat alle sillabes, vanaf superswaar tot by die VC-sillabe, geslote is en dat die oop lettergreep VV eers na al die geslote lettergrepe op die gewigskale voorkom. Dus kan daar in hierdie studie aanvaar word dat geslote sillabes eerder hoofklemtoon sal dra as oop sillabes.

2.3.2.3 Die sjwa

Die sjwa is gewigloos en sal dus nooit in 'n woord beklemtoon word nie. In Tommelen en Zonneveld (1989:142) word weergegee dat klemtoon (vir Nederlands) altyd onmiddellik voor die

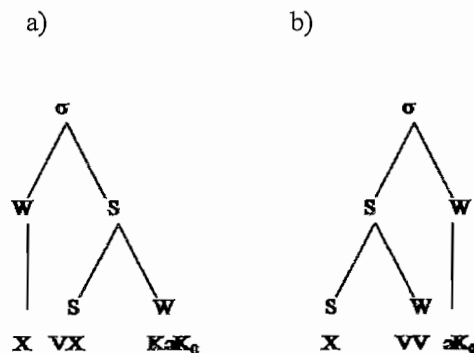
sjwa-lettergreep lê en slegs die derdelaaste lettergreep onder sekere spesifieke omstandighede kan bereik, soos *papaver*, *zigeuner*, *klavecimbel*, *reprimande*. Volgens Lubbe (1993:92) is dié stelling ook geldig vir Afrikaanse simplekse:

hipotese	miserabel	apostel	karakter	België
tabernakel	sinagoge	rabarber	helikopter	terriër
Egipte	sillabe	herinner	kabouter	Tunisië

In dié lys kan gesien word dat, in die eerste vier kolomme, die klemtoon op die lettergreep net voor die sjwa val, maar in die laaste kolom val die klemtoon op die derdelaaste lettergreep.

Lubbe (1993:92) verduidelik dié verskynsel soos volg:

As die sjwa voorafgegaan word deur minstens een konsonant, val die klemtoon op die voorlaaste lettergreep, maar as die sjwa nie voorafgegaan word deur 'n konsonant nie, val die klemtoon op die derdelaaste lettergreep. Dié reël kan vir Afrikaanse simplekse toegepas word:



Reël 1: Die sjwa-reël vir Afrikaans. Voorbeelde vir reël a) *papaver*, *pantoffel*, *kalender* en b) *Tunisië*, *terriër*, *België*.

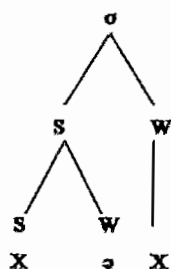
Woorde soos *offerande*, *armoede*, *oorkonde*, word uitgeskakel uit die reëls, aangesien hulle pseudo-samestellings is en nie simplekse nie.

Van der Hulst (1984:232) genereer die volgende reël vir 'n sjwa wat in die tweede lettergreep van 'n woord voorkom:

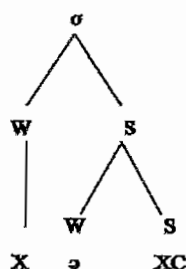
Hoofklemtoon word op die eerste lettergreep van 'n woord geplaas, indien die tweede lettergreep 'n sjwa bevat en 'n vol vokaal in die laaste lettergreep voorkom, maar as die laaste lettergreep

superswaar is (dit wil sê as die laaste lettergreep 'n VKK-sillabe of 'n diftong bevat), kry dié lettergreep hoofklemtoon.

a)



b)



Reël 2: Voorbeelde van dié reëls is a) *hottentot, kamera, algebra* en b) *harlekyn, intellek, atletiek*.

Hierdie reëls geld vir Afrikaans ook, maar het heelwat uitsonderings. Woorde wat uitsonderings op reël (a) is, is o.a. *energie, omelet, kotelet*. Uitsonderings op reël (b) sluit die volgende woorde in: *ooievaar, lessenaar, rekenaar*. As daar gekyk word na die vokaalhiërargie van Combrink en De Stadler (1987:192) kan reël (b) se uitsonderings verminder word deur by te voeg dat die eerste sillabe nie swaarder mag wees as die laaste superswaar sillabe nie. By die drie genoemde uitsonderings van reël (b), is die eerste sillabe volgens Combrink en De Stadler (1987:192) se vokaalhiërargie telkens swaarder as die laaste sillabe:

a) diftong > a; (ooi > vaar)

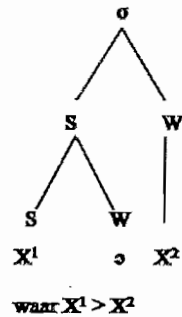
b) E > a; (les > naar)

c) e > a; (re > naar)

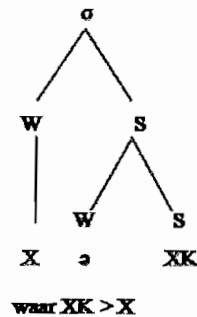
Dit is ook die geval dat *rekenaar* en *lessenaar* beide afleidings is en nie Afrikaanse simplekse nie, dus geld hulle nie werklik as uitsonderings vir die studie nie, aangesien daar aparte reëls vir afleidings ook gegenereer gaan word. By die uitsondering *energie* is dit ook die geval dat die laaste lettergreep swaarder is volgens die vokaalhiërargie as die eerste lettergreep: i > E (e < gie)

Dus kan reël 2 verander word soos volg:

a)



b)

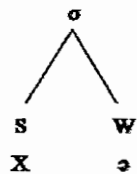


Reël 3: Die reëls vervang reël 2.

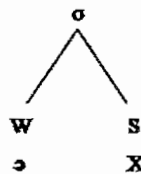
Indien geen ander reëls klemtoon kan toeken aan die uitsonderings nie, kan dié reëls verder uitgebrei word deur: 1) by reël (a) te voeg dat indien $X1 < X2$, finale klemtoon toegeken moet word; en 2) by reël (b) te voeg dat indien $XK < X$, klemtoon op die eerste lettergreep val.

Die meeste tweelettergrepige woorde het 'n sjwa op die tweede lettergreep van die woord. Die klemtoonreëls vir tweelettergrepige woorde met 'n sjwa lyk soos volg:

a)



b)



Reël 4: In 'n tweelettergrepige woord sal die lettergreep wat 'n sjwa bevat nooit beklemtoon word nie.

Voorbeelde is a) *masker, kelner, appel* en b) *persoon, debiet, mekaar*.

2.3.2.4 Affikse

Affikse is morfologiese elemente wat in normale taalgebruik nie as woorde kan funksioneer nie en dus moet dit by stamme aanleun (Combrink, 1990:21).

Die meerderheid Afrikaanse prefikse dra nie hoofklemtoon nie en verander ook nie die posisie van die hoofklemtoon nie, maar daar is wel verskeie prefikse wat hoofklemtoon dra: (Combrink en De Stadler, 1987:200)

on- : ondiens, oneer, ongemak
 a-: amoreel, asosiaal
 mis-: misverstand, misgeboorte
 wan-: wanklank, wanorde
 aarts-: aartsvyand, aartsengel

Die prefiks *her-* kry soms die klemtoon en soms nie. By woorde soos *herbenoem*, *heroorweeg*, *heruitsaai*, val die klemtoon op die prefiks, maar woorde soos *herboor*, *herbou*, *herlaai* kry nie die hoofklemtoon op die prefiks nie.

Suffikse kan weer klemtoon op drie maniere beïnvloed, naamlik

- a) Suffikse wat nie die klempatroon van die stam verander nie, soos. *-heid*, *-er* en *-loos*.
- b) Suffikse wat die hoofklem dra, soos. *-is (t)*, *-eur*, *-aal*, *-eel*, *-ent*, *-esk*.
- c) Suffikse wat die hoofklem in hul rigting trek, naamlik *-saam*, *-ig*, *-(e)lik*, *-ies*, *-end*, *-loos* en *-baar*.

In (b) is die sogenaamde klemtoongevoelige suffikse. Dié suffikse gedra hulle nie veel anders as simplekse nie, aangesien al die suffikse superswaar is, of deur klemtoonreëls as superswaar geag word.

Vergelyk die volgende woorde:

Afleiding	Simpleks
bakteriside	piramide
normaal	vulkaan
restant	basalt
benefisiaat	surrogaat

Dus kan gesê word dat die klemtoon gevoelige suffikse dieselfde klemtoonreëls sal volg as simplekse waarvan die laaste lettergreep superswaar is.

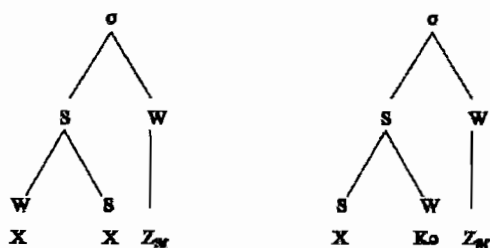
Die suffikse in (c) is almal klemtoonaantrekkend. Dit wil sê hulle trek die klemtoon in hul rigting, maar ontvang nie self hoofklemtoon nie. Die volgende woorde is voorbeelde van die verskynsel:

<i>wonderbaar</i>	<i>wonderbaarlik</i>
<i>afgod</i>	<i>afgodies</i>
<i>inneem</i>	<i>innemend</i>
<i>waarneem</i>	<i>waarneembaar</i>

Combrink en De Stadler (1987:200) se klemtoonreël vir dié verskynsel (weergegee in figuur 18 van hierdie hoofstuk) word gebruik om die volgende reël vir Afrikaans te genereer:

a)

b)



waar Z_{gr} = -ik, -ig, -saam, -loos, -ies, -end, -baar.

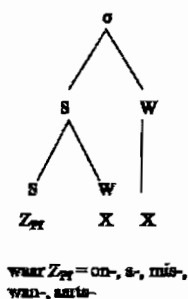
Reël 5: Woorde soos a) *atleties*, *bedagsaam* en b) *radeloos*, *afskuwelik*, *waarneembaar*, word deur dié reël behandel.

Reël 5 lui dat hoofklemtoon op die sillabe net voor die spesifieke suffikse geplaas word, indien dit nie voorafgegaan word deur 'n sjwa nie, anders word die klemtoon twee sillabes voor die suffiks geplaas.

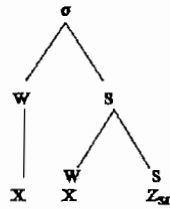
Die reël kan ook werk vir 'n getal woorde, waarvan die suffikse nie die klem patroon verander nie, naamlik:

- or: aggressor, sensor
- sie: absorpsie, refleksie
- naar: rekenaar, dorpenaar
- heid: hoeveelheid, ruheid
- ing: sendeling, beweging
- skap: geselskap, moederskap

Die reël vir Afrikaanse prefikse wat hoofklemtoon dra stel dat as 'n woord se eerste lettergreep prefiks Z_{Pf} bevat, waar $Z_{Pf} = on-, a-, mis-, wan-, aarts-$ is, dan sal die prefiks hoofklemtoon dra, anders word die hoofklemtoon van die stam nie beïnvloed nie. Afrikaanse suffikse wat hoofklemtoon dra stel dat as 'n woord 'n klemdraende suffiks Z_{Sf} bevat, dan sal die lettergreep waarin die suffiks voorkom, beklemtoon word, anders word die hoofklemtoon van die stam nie beïnvloed nie. Daar is verskeie suffikse wat klemdraend is, soos *-aat, -ent, -esk, -eus, -ief, -isme, -kasie, -urg, -teit*. Die volgende boom wat gebou is om dié reëls verteenwoordig 'n drielettergreepige woord, maar die reël sal geld vir 'n woord met enige getal lettergrepe:



Reël 6: Beklemtoonbare prefikse sal hoofklemtoon kry. Voorbeelde van woorde wat deur dié reël beklemtoon word, is *misverstand, aartsyand, wanorde*.



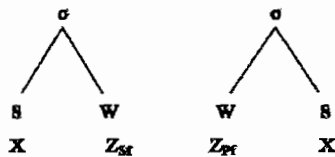
Reël 7: Beklemtoonbare suffikse sal hoofklemtoon kry. Voorbeelde is *dramaturg, fraktuur, miljoen, grotesk*.

Die volgende reëls hanteer alle ander afleidings wat nie die bogenoemde reëls bevredig nie.

Reël 8 en 9 dui aan dat enige onbeklemtoonbare prefiks, Z_{Pf} of suffiks, Z_{Sf} nie beklemtoon sal word nie. Indien 'n een lettergreep van 'n tweelettergrepige woord 'n prefiks of 'n suffiks bevat, sal die ander lettergreep wat 'n simplekse stam bevat, beklemtoon word.

a)

b)

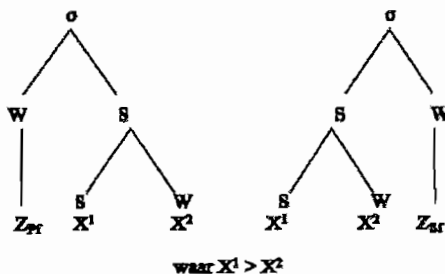


Reël 8: 'n Onbeklemtoonbare affiks sal nooit beklemtoon word nie. Voorbeelde: a) *eenheid, sending, heimlik* b) *verdryf, geskat, beveel*

In 'n drie- of meerlettergrepige woord, waarvan die begin of eindlettergreep 'n affiks is, sal deur middel van die sillabesterteiërargie wat vroeër in die hoofstuk weergegee is, bepaal word watter een van die oorblywende lettergrepe beklemtoon sal word.

a)

b)



Reël 9: Voorbeelde van dié reël is a) *verafrikaans, befoeter* en b) *hoeveelheid, direkteurskap*

Die bome in reël 9 is slegs vir twee gevalle geteken, maar 'n woord sal op enige lettergreep wat volgens die sillabestertekhiërargie die swaarste is, beklemtoon kan word, solank dit nie 'n onbeklemtoonbare affiks is nie.

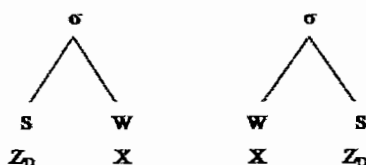
2.3.3 Klemtoonreëls vir Afrikaanse simplekse woorde

In dié afdeling gaan die klemtoonreëls vir Afrikaanse simplekse woorde gegenereer word. Alle reëls word uitgevoer onder die voorwaarde dat die sillabe wat deur die *W*-voet verteenwoordig word, nie swaarder is, volgens die sillabestertekhiërargie, as die sillabe onder die *S*-voet nie. Combrink en De Stadler (1987:193) se hoofklemtoonreël (sien figuur 16) word ook in die reëls geïnkorporeer, deurdat die reëls in die spesifieke volgorde van die sillabestertekhiërargie moet uitgevoer word, sodat daar verseker kan word dat die sterkste sillabe altyd hoofklemtoon dra. Ook word die reëls vir simplekse woorde onder die voorwaarde uitgevoer dat daar nie enige affikse in die woord teenwoordig is nie.

2.3.3.1 Tweelettergrepige woorde

In die afdeling word metriese bome gebou vir tweelettergrepige simplekse. Die reëls word hoofsaaklik afgelei van Combrink en De Stadler (1987:191) se sillabestertekhiërargie.

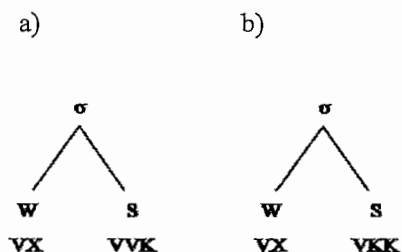
In reël 10 word aangedui dat indien 'n woord 'n diftong bevat, sal daardie lettergreep wat die diftong bevat, hoofklemtoon dra.



Reël 10: Die lettergreep wat 'n diftong bevat, sal beklemtoon word. Z_D = 'n diftong

In die bogenoemde reël word die eerste deel van Combrink en De Stadler (1987:191) se sillabestertekhiërargie weergegee, naamlik dat diftonge heel bo in die sillabestertekhiërargie lê. Erkende Afrikaanse diftonge is [ai, oi, ui, eu, Q, U, Y] soos in *baai*, *nooi*, *roei*, *leeu*, *pyl*, *uit*, *oud*. Reël 10 sal aan woorde soos *vyand*, *beitel*, *heide* inisiële hoofklemtoon toeken en sal finale hoofklemtoon toeken aan woorde soos *kony*n, *fontein*, *vallei*. Uitsonderings op die reël is *aalwyn*, *aarbe*i, *aambe*i, *arbe*id.

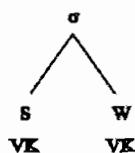
Reëls 11: Die volgende reëls plaas hoofklemtoon op die laaste lettergreep van woorde met superswaar eindsillabes naamlik VX-VVK en VX-VKK:



Reël 11: Die reël sal woorde soos *kameel, kasteel, bandiet* en b) *balans, triomf, miljard* hanteer. Uitsonderings op die reël is a) *potlood, liggaam, wierook* en b) *antwoord, klimaks, affiks*.

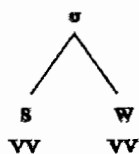
Dus sal reël 11 beteken dat indien 'n tweelettergrepige woord op VKK of VVK eindig, die laaste lettergreep beklemtoon sal word.

Reël 12: 'n Woord waarvan beide lettergrepe geslote is, sal hoofklemtoon op die eerste lettergreep kry.



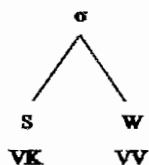
Reël 12: Tipiese woorde wat beklemtoon sal word deur die reël, is *mentor, kursus*. Uitsonderings op die reël is o.a. *kristal, karkas, sukses*.

Reël 13: Indien beide lettergrepe van 'n tweelettergrepige woord, oop is, sal die woord op die eerste lettergreep beklemtoon word.



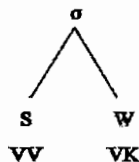
Reël 13: Dié reël sal woorde soos *tuba, baba, ghoeroe* beklemtoon. Uitsonderings op die reël is *kopie, taboe, idee*.

Reël 14: 'n Woord waarvan die eerste lettergreep toe is en die tweede oop is, sal hoofklemtoon dra op die eerste lettergreep.



Reël 14: Voorbeelde vir dié reël is *kombi, vodka, firma*. Uitsonderings op die reël is woorde soos *moskee, elpee*.

Reël 15: Indien die eerste lettergreep van 'n woord oop is en die tweede toe is, sal hoofklemtoon op die eerste lettergreep val.

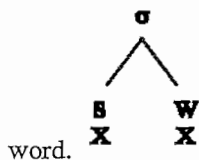


Reël 15: Woorde wat deur die reël beklemtoon sal word is *piesang, radar, robot*. Uitsonderings is *baron, program, hotel*.

Daar kan dus 'n veralgemening gemaak word deur te sê dat as 'n tweelettergrepige woord nie 'n superswaar finale sillabe bevat nie, dit hoofklemtoon op die eerste sillabe sal kry.

Daar kan uit reëls 12-15 afgelei word dat indien 'n tweelettergrepige woord nie 'n superswaar eindsillabe het nie, dit op die eerste lettergreep beklemtoon sal word. Dus sal een reël ontwikkel word om reëls 12-15 te vervang.

Reël 16: Indien reël 10 en 11 nie 'n woord bevredig nie, sal die woord beklemtoon word op die eerste lettergreep. Dit is dus noodsaaklik dat die reëls in die volgorde 10, 11, 16 uitgevoer

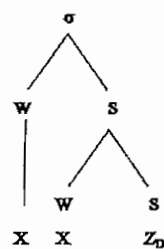


Reël 16: Enige tweelettergrepige woord wat nie op 'n superswaar sillabe eindig nie, sal beklemtoon word op die eerste lettergreep.

2.3.3.2 Drielettergrepige woorde

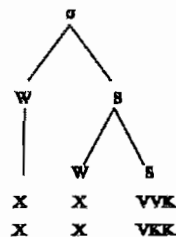
Die reëls vir drielettergrepige woorde wat hier onder gegenereer gaan word, is slegs vir woorde wat nie 'n sjwa bevat nie, dus kan van al die onderstaande reëls gesê word dat $V \neq @$.

Reël 17: Indien 'n diftong in 'n drielettergrepige simpleks voorkom, sal daardie lettergreep wat die diftong bevat, beklemtoon word. Die boom is slegs vir een geval gebou, maar kan gewysig word, sodat enige lettergreep wat 'n diftong bevat, beklemtoon kan word. Indien 'n drielettergrepige woord meer as een diftong bevat, sal die linkerkantste diftong hoofklemtoon kry.



Reël 17: 'n Lettergreep wat 'n diftong bevat, sal beklemtoon word.

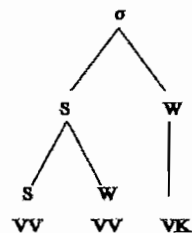
Reël 18: Indien 'n drielettergrepige woord eindig op 'n superswaar sillabe, sal daardie laaste lettergreep beklemtoon word.



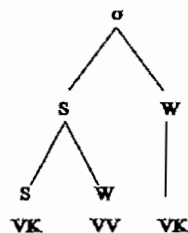
Reël 18: Woorde soos *paragraaf*, *pelikaan*, *dinamiet* en *dividend*, *testament*, *predikant* sal deur die reël beklemtoon word.

Reël 19: Indien 'n drielettergrepige woord se laaste twee lettergrepe die vorm VV-VK aanneem, sal die derdelaaste lettergreep beklemtoon word.

a)



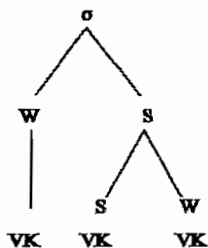
b)



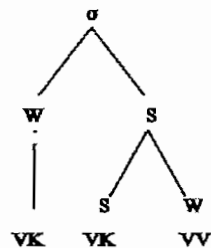
Reël 19: Woorde soos a) *stadium, monitor, maraton* en b) *kalsium, alkohol, albatros* sal deur die reël beklemtoon word. Uitsonderings vir die reëls is woorde soos a) *salaris, kabinet, krokodil* en b) *almanak, diktator, messias*

Reël 20: Indien 'n drielettergrepige woord enige van die vorms van -af hieronder aanneem, sal dit beklemtoon word op die tweedelaaste lettergreep.

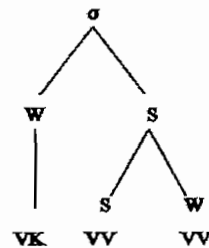
a)



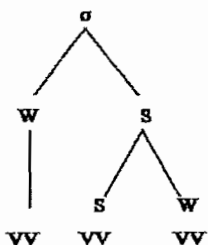
b)



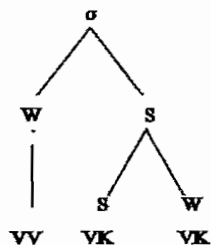
c)



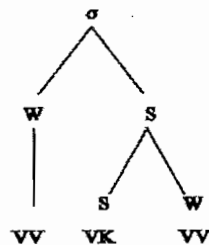
d)



e)



f)

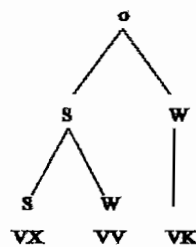


Reël 20: Voorbeelde vir die reëls is *gorilla, kollega, kanarie, elektron, kwitansie*.

Daar is heelwat uitsonderings op dié reëls, maar voorbeelde dui oorheersend aan dat hoofklemtoon op die tweedelaaste lettergreep val.

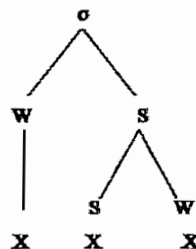
As bogenoemde reëls vir drielettergrepige woorde in die volgorde van reël 17 - 20 uitgevoer word, kan die reëls verkort word deur reël 19a-b te vervang met reël 21 en reël 20a-f te vervang met reël 22.

Reël 21: 'n Drielettergrepige simplekse woord waarvan die laaste twee lettergrepe die vorm VV-VK het, sal klemtoon dra op die eerste lettergreep.



Reël 21: Dié reël vervang reël 19 a en b.

Reël 22: Enige drielettergrepige simplekse woord wat nie bevredig word deur enige van reëls 17, 18 of 21 nie, se tweedelaaste lettergreep word beklemtoon.



Reël 22: Dié reël vervang reël 20a-f slegs as die reëls in volgorde van 17-20 uitgevoer word.

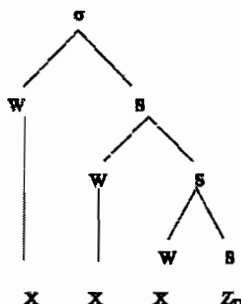
Hierdie twee reëls, veral reël 22 mag slegs gebruik word, indien dit in die volgorde, 17, 18, 21,22 uitgevoer word en steeds, soos vroeër genoem, geen sjwa bevat nie.

Enige drielettergrepige simplekse wat 'n sjwa bevat, sal beklemtoon word deur die reëls vir woorde wat 'n sjwa bevat (Reël 1 en 3).

2.3.3.3 Vier- en meerlettergrepige woorde

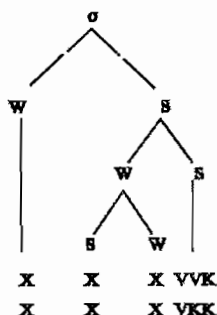
Daar is baie min vier- en meerlettergrepige simplekse in Afrikaans, aangesien die meerderheid woorde van dié grootte samestellings of afleidings is. Die eerste twee reëls is weereens dié wat 'n diftong bevat of op 'n superswaar sillabe eindig.

Reël 23: Enige lettergreep in 'n simplekse woord wat 'n diftong bevat, sal beklemtoon word. Indien meer as een diftong in 'n woord voorkom, sal die mees linkerkantste lettergreep wat 'n diftong bevat hoofklemtoon kry.



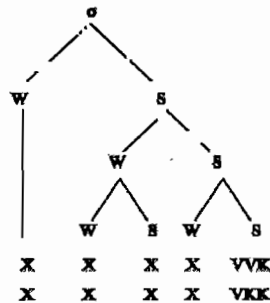
Reël 23: Enige lettergreep wat 'n diftong bevat, sal beklemtoon word.

Reël 24: Soos vroeër in die hoofstuk verduidelik, kry woorde met 'n superswaar eindsillabe 'n klemtoon op daardie sillabe. Enige woord met 'n slotlettergreep met die vorm VVK of VKK sal op daardie lettergreep beklemtoon word.



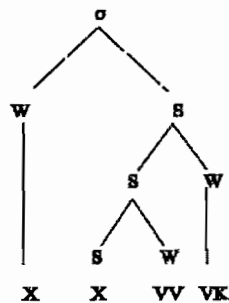
Reël 24: Die reël sal woorde soos *eksperiment*, *infinities* beklemtoon.

Reël 25: 'n Boom wat 'n vyflettergrepige woord met 'n superswaar eindsillabe (VVK of VKK) verteenwoordig lyk soos volg:



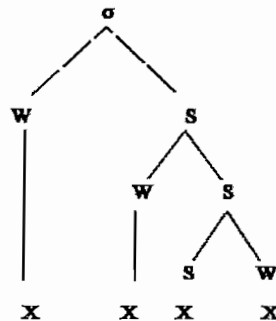
Reël 25: Vyflettergrepige woorde met superswaar eindsillabes is nie volop nie. 'n Voorbeeld van so 'n woord is *universiteit*.

Reël 26: 'n Vier- of meerlettergrepige woord waarvan die laaste twee lettergrepe die vorm VV-VK bevat, sal beklemtoon word op die dertelaaste lettergreep.



Reël 26: Die reël sal woorde soos *akkordeon*, *ammonium* beklemtoon.

Reël 27: Die volgende reël het dieselfde funksie as reël 22. Dit ken hoofklemtoon toe aan die tweedelaaste lettergreep van vier- of meerlettergrepige woorde, wat nie bevredig word deur die boonste drie reëls nie.



Reël 27: Die reël hanteer woorde soos *diafragma*, *akademie*, *evangelie*, *panorama*, *referendum*. Uitsonderings op dié reël is woorde soos *kategorie*, *individu*, *filosofie*, *marionet*.

Al die bome wat in die afdeling vir vierlettergrepige woorde gebou is, geld vir enige simplekse wat vier- of meerlettergrepig is.

2.4 SAMEVATTING

Die eerste doelstelling van hierdie projek is om 'n stel klemtoonreëls vir Afrikaans te ontwikkel wat gebruik gaan word om 'n reëlgebaseerde klemtoontoekenner te bou. In dié hoofstuk is die verskillende benaderings tot die ontwikkeling van klemtoonreëls bespreek. Daar is twee benaderings bespreek, naamlik die liniêre klemtoonanalise (ook genoem die SPE-metode) en die metriese klemtoonanalise, naamlik metriese bome en metriese raamwerke.

Die klemtoonreëls wat in dié hoofstuk ontwikkel is, is gebaseer op die Afrikaanse klemtoonreëls van Wissing (1971), Combrink en De Stadler (1987) en Lubbe (1993) en ook op die Nederlandse klemtoonreëls van Van der Hulst (1984) en Trommelen en Zonneveld (1989). Die metriese bome is gekies as benadering vir die generering van Afrikaanse klemtoonreëls vir hierdie studie.

Die uiteindelijke stel klemtoonreëls bestaan uit 19 reëls wat onderverdeel kan word in vyf kategorieë. Dié kategorieë word in tabel 2 uiteengesit.

In die volgende hoofstuk word die vereistes ondersoek waaraan die klemtoontoekenner moet voldoen. Verskillende benaderings tot outomatiese klemtoontoekenning word ook bespreek. Spesifieke aandag word gegee aan die reëlgebaseerde benadering, aangesien dié benadering in die projek gebruik word. Hieruit sal dit duidelik blyk waarom die reëlgebaseerde benadering in hierdie studie gebruik is.

KATEGORIE	AANTAL REËLS
Tweelettergrepige simplekse	3
Drielettergrepige simplekse	4
Vier- en meerlettergrepige simplekse	4
Sjwa-reëls	3
Affiksreëls	5

Tabel 2: Uiteensetting van klemtoonreëls vir Afrikaans

In Hoofstuk 4 word die toetsproses in diepte bespreek. Daar word ook aangetoon hoe die klemtoontoekenner ontwikkel word deur die klemtoonreëls te gebruik wat in hierdie hoofstuk ontwikkel is. Daarna word die klemtoontoekenner en die grafeem-na-foneem-omsetter geëvalueer. Die evalueerproses sal beskryf word en die resultate sal bespreek word. Hierdie resultate word ook vergelyk met die resultate van bestaande klemtoontoekenners vir ander tale. Sodoende word die akkuraatheid en effektiwiteit van die klemtoonreëls bepaal.

Hoofstuk 3

'N REËLGEBASEERDE BENADERING TOT KLEMTONTOKENNING

3.1 INLEIDING

Om 'n klemtoontekenner te ontwikkel, word 'n stel klemtoonreëls benodig, asook 'n geskikte algoritme. In die vorige hoofstuk is 'n volledige stel klemtoonreëls gegenereer wat gebruik kan word om klemtoon aan woorde toe te ken. Hierdie hoofstuk handel oor die benadering wat gevolg word om die Afrikaanse klemtoontekenner te ontwikkel.

Daar is al heelwat verskillende benaderings gevolg om klemtoontekenners te ontwikkel. Die benaderings wat gevolg is sluit in datagedrewe benaderings en linguistiese reëls. Datagedrewe benaderings, soos Versteekte Markovmodelle en voorbeeldgebaseerde benaderings, gebruik reeds geannoteerde data om 'n klemtoontekenner outomaties te "leer" hoe om die korrekte klemtoon aan woorde toe te ken. Binne die linguistiese benadering word klemtoonreëls deur kenners vir 'n spesifieke taal geskryf. Dié reëls word dan geïmplementeer en outomaties toegepas op woorde om sodoende die korrekte klemtoon aan woorde toe te ken.

In hierdie hoofstuk word daar gefokus op twee verskillende benaderings tot klemtoontokenning. In 3.2.1 word al die verskillende datagedrewe benaderings bespreek, en die linguistiese benadering word in 3.2.2 behandel. Daar sal ook gekyk word na 'n paar klemtoontekenners wat ontwikkel is met behulp van hierdie benaderings. In 3.2.3 sal die stappe uiteengesit word wat gevolg word om die klemtoontekenner suksesvol te bou.

3.2 *DIE BENADERING TOT KLEMTONTOKENNING*

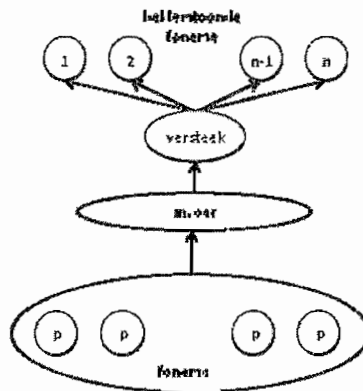
Die benaderings wat gevolg kan word om klemtoon aan woorde toe te ken, kan in twee kategorieë verdeel word, naamlik 'n linguistiese benadering of 'n datagedrewe benadering. Die verskil tussen die linguistiese en die datagedrewe tegnieke lê nie in die werking van die tegnieke nie, maar in die manier waarop die reëls gegenereer word. By die linguistiese benadering word die reëls handgemaak deur 'n linguïst en by die datagedrewe benadering word reëls outomaties gegenereer vanuit die data. Die voordeel van die linguistiese benaderings is dat linguistiese kennis direk toegepas kan word, waarenteen datagedrewe metodes weer baie maklik geleer kan word en nie taalafhanklik is nie (Reichel *et al*, 2008:159). In dié afdeling gaan verskillende benaderings tot klemtoontokenning bespreek word.

3.2.1 *Datagedrewe Benaderings*

3.2.1.1 *Neurale Netwerke*

Kunsmatige neurale netwerke poog om die menslike sensus van spraak na te boots, deur inligting op dieselfde manier te verwerk (Rojas, 1996:3). Die neurale netwerk bestaan volgens Gallant (1993:3) uit 'n stel eenhede en 'n stel eenrigting-datakonneksies wat al die eenhede verbind. Sodra invoer aan die netwerk gegee word, ondersoek die eenhede dit en bereken dan 'n aktiveerwaarde vir die invoer. Die aktiveerwaarde word dan aangestuur deur die konneksies na die eenhede toe. Elke konneksie het 'n gewig, wat dan gemeet word aan die aktiveerwaarde wat aan die invoer toegeken is. So word 'n uitvoer dan verkry.

Reichel *et al.* (2008) gebruik die neurale netwerk metode om klemtoon toe te ken aan Engelse woorde binne 'n grafeem-na-foneem-stelsel. Die invoervlak bevat die eerste tien foneme van die woord. Die uitvoervlak bevat dieselfde hoeveelheid nodes as die foneme in die invoervlak. Die uitvoernode met die hoogste waarde bepaal die posisie van die klemtoon binne die foneemstelsel. Dus sal die eerste foneem van die woord beklemtoon word as die eerste node die hoogste waarde het. Die voordele van neurale netwerke is dat dit verskillende databasisse kan hanteer en daar is baie verskillende leeralgoritmes beskikbaar. Die nadele van die algoritme is dat daar geen ontwerpriglyne bestaan nie, en die akkuraatheid van die stelsel maak staat op die groot stelsel afgrigtingsdata wat nie altyd beskikbaar is nie. Verder het dit ook nie die vermoë om te redeneer nie, dit verduidelik nie die uitvoer nie en dit is nie maklik aanpasbaar nie (Alkoffash, Bawaneh en Rabea, 2008:611).



Figuur 22: Die neurale netwerkmodel van Reichel (2008)

Arciuli en Thompson (2006) het ook die neurale netwerk gebruik om klemtoon aan Engelse woorde toe te ken. Hulle het hoofsaaklik met tweelettergrepige werkwoorde en selfstandige naamwoorde gewerk. Die netwerk is ontwerp om 'n klempatroon aan 'n woord toe te ken deur na woordbegin en wordeinde te kyk. 'n Woordbegin strek vanaf die begin van die woord tot by die eerste vokaal of groep vokale. Die wordeinde strek vanaf die letter na die laaste vokaal tot by die einde van die woord. Die neurale netwerk is afgerig op 10134 woorde en het 'n akkuraatheid van 86% behaal.

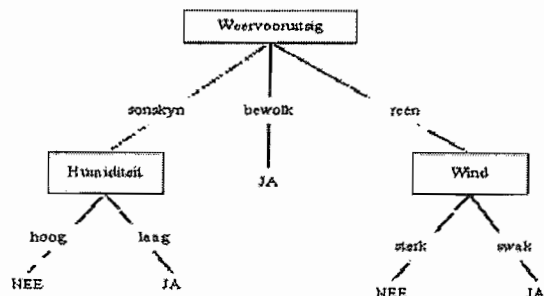
3.2.1.2 Besluitnemingsbome

'n Besluitnemingsboom klassifiseer probleme deur dit vanaf die wortel tot by die blaarnode uit te sorteer. Elke node in die boom is 'n toets vir een van die eienskappe van die probleem. Wanneer 'n probleem aan die boom as invoer gegee word, begin dit die probleem oplos deur by die wortelnode te begin. Die eienskappe van die probleem word dan getoets. Die tak wat die eienskappe die beste pas is dan die een waarlangs die probleem gestuur word. Op die volgende vlak van die boom word dieselfde proses dan herhaal tot 'n finale antwoord bereik word.

In Mitchell (1997:54) word genoem dat besluitnemingsbome die beste werk vir probleme met die volgende eienskappe:

- Gevalle word verteenwoordig deur vasgestelde eienskappe. In figuur 23 kan die eienskappe van die weervoorsig slegs *sonnig*, *bewolk* of *reën* wees.
- Die uitvoer van die probleem het afsonderlike waardes bv. *ja* of *nee*

- Onderskeidende beskrywings is nodig
- Die afrigtingsdata kan foute bevat
- Die afrigtingsdata is nie volledig nie en sommige van die waardes ontbreek.



Figuur 23: 'n Besluitnemingsboom vir die geval, SpeelTennis. Die vraag wat hier gevra word is of die weervoorsigte gunstig is om te gaan tennis speel (Mitchell, 1977:53).

'n Groot voordeel van besluitnemingsbome is dat dit maklik aangepas kan word vir nuwe data. Dit is ook 'n algoritme wat maklik is om te verstaan en hanteer data wat teks bevat baie goed. (Alkoffash, Bawaneh & Rabea, 2008:611).

Sef *et al.* (2002) het besluitneembome gebruik om klemtoon toe te ken aan Sloweense woorde. Eerstens is 'n boom gebruik om die leksikale klemtoon van elke vokaal te bereken. Daarna is die klemtoon van die hele woord bereken volgens die hoeveelheid beklemtoonde vokale en die lengte van die woord. As meer as een vokaal gekies word om klemtoon te ontvang, word een van die gekose vokale ewekansig gekies.

Besluitnemingsbome is ook in 'n studie deur Reichel *et al.* (2008) gebruik om klemtoon aan woorde toe te ken. Die posisie van klemtoon is bepaal deur die woordeinde in ag te neem. In hul studie is woordsoorte, affikse, getal lettergrepe, die gewig van die laaste vier lettergrepe, asook die posisie van die lettergreep binne die woord gebruik om klemtoon toe te ken.

3.2.1.3 Voorbeeldgebaseerde leer

Die benadering is daarop gebaseer dat wanneer mense 'n nuwe taalfenomeen teëkom, hulle nie reëls onttrek uit hul vorige ondervindings om die fenomeen te klassifiseer nie, maar 'n

herinnering van 'n soortgelyke fenomeen oproep om die nuwe een te klassifiseer (Daelemans en Van den Bosch, 2005:4).

Die algoritme “leer” deur slegs gegewe afrigtingsdata te stoor. Aan elke patroon wat teëgekomp word, word 'n kategorie toegeken, deur te kyk of die patroon al vroeër teëgekomp is in die afrigstel. Wanneer 'n nuwe item teëgekomp word, word 'n stel voorbeelde wat eenders is as die nuwe item, in die geheue opgesoek. Die voorbeelde word dan gebruik om die een voorbeeld wat die naaste aan die nuwe item is te bereken en so die nuwe item te klassifiseer (Gillis *et al*, 2000:82).

Die voorbeeldgebaseerde tegniek is 'n raamwerk en metodologie vir toenemende gekontroleerde masjienleer. Voorbeeldgebaseerde leer is gebaseer op die hipotese dat meerderheid intelligente gedrag voorkom uit die onmiddellike gebruik van gestoorde voorvalle wat vroeër ondervind is. Die prestasie van die metode hang geheel en al af van die keuse van afrigitems wat in die geheue gehou moet word (Gillis *et al*, 1993:454). Die kenmerk wat voorbeeldgebaseerde leer onderskei van ander datagedrewe benaderings, is die feit dat geen eksplisiete uittreksels gedurende die afrigtingsfase gegenereer word op die basis van die afrigvoorbeelde nie (Daelemans, 1993:164). Die algoritme benodig nie 'n deskundige om reëls te maak nie, en het die vermoë om te redeneer. Een nadeel van die metode is dat die voorspelling van nuwe data beperk word deur die data wat dit reeds bemerk het (Alkoffash, Bawaneh en Rabea, 2008:611).

Die klemtoontoekenner wat Gillis *et al.* (1993) vir die Nederlandse taal gebou het deur voorbeeldgebaseerde leer te gebruik, beklemtoon 96.43% van alle woorde wat beklemtoon word op die tweedelaaste lettergreep korrek, en 86.84% van alle woorde wat klemtoon kry op die derdelaaste lettergreep.

Die voordele van die metode is volgens Mitchell (1997:231) die volgende:

- Die metode kan verskillende benaderings vind vir elke aparte geval wat lei tot die uitvoer van die probleem.
- Dit kan meer komplekse, simboliese verteenwoordigings van voorbeelde gebruik as ander metodes.

Mitchell (1997:231) noem ook 'n paar nadele van die metode:

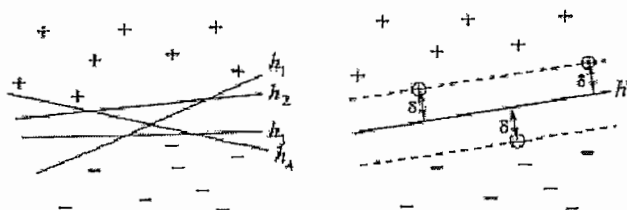
- Die koste om nuwe items te klassifiseer, is baie hoog. Alle berekeninge vind plaas tydens klassifisering, eerder as om dit te doen wanneer die items die eerste keer teëgekom word.
- Dit oorweeg alle moontlike eienskappe vir die items, wanneer dit poog om eenderse afrigvoorbeelde uit die geheue te onttrek. Dus bestaan daar die moontlikheid dat die items wat werklik eenders is, baie ver uitmekaar kan voorkom.

3.2.1.4 Steunvektormasjiene

Die idee van 'n steunvektormasjien (SVM) is om die invoervektore x te pas in 'n hoë-dimensionele terrein Z , deur een of ander nie-liniêre passing, om sodoende 'n optimale hipervlak te konstrueer (Vapnik, 2000:138).

Volgens Joachims (2002:36) is die SVM gebaseer op die strukturele-risiko-minimaliseerbeginsel vanuit die statistiese leerteorie. 'n Hipotese moet gevind word waarvoor die laagste waarskynlikheid vir foute gewaarborg moet word vir 'n gegewe afrigvoorbeeld. Die SVM moet 'n hipervlak vind wat die positiewe en negatiewe lede van 'n gegewe klas van mekaar skei.

Veronderstel gegewe afrigtingsdata word geskei deur minstens een hipervlakke h (sien figuur 24). Dit sal beteken dat alle positiewe afrigvoorbeelde aan die een kant van die hipervlak is en al die negatiewe afrigvoorbeelde aan die ander kant is. Daar is oor die algemeen verskeie hipervlakke wat die afrigtingsdata sonder foute verdeel. Die SVM kies die hipervlak met die grootste speling δ , dit wil sê die hipervlak wat die verste weg is van al die naaste afrigvoorbeelde (Fletcher, 2009). Die speling δ is die afstand vanaf die hipervlak tot by die naaste afrigvoorbeelde. Die voorbeelde naaste aan die gekose hipervlak word steunvektore genoem. Die gekose hipervlak word dan gebruik wanneer nuwe data geklassifiseer word (Joachims, 2002:37).



Figuur 24: 'n Binêre klassifikasieprobleem in twee dimensies (Joachims, 2002: 37)

Die klemtoontoekenning wat deur Dou *et al.* (2009) gedoen is vir Engels, gebruik die SVM tegniek. Die klemtoontoekenstelsel pas 'n woord aan die beklemtoonde ekwivalent van die woord. Hul toekenproses het drie fases:

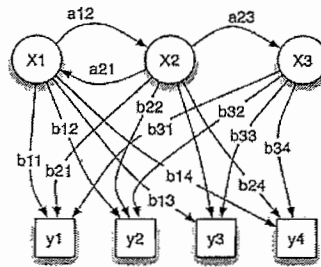
- Eerstens word woorde aan substringe, s , gepas. Die substringe is die basiese eenhede in die reeks.
- Daarna word 'n spesifieke klempatroon, t , gekies vir elke substring-reeks. 'n SVM word gebruik om die moontlike patrone vir elke reeks te rangskik.
- Laastens word die klempatroon gebruik om die beklemtoonde vorm van die woord te genereer.

Die toekenner is afgerig met ongeveer 60 000 Engelse woorde en het primêre en sekondêre klemtoon korrek toegeken aan 93.5% van die woorde waarop dit geëvalueer is.

3.2.1.5 Versteekte Markovmodelle

Versteekte Markovmodelle word in verskeie natuurlike taaltoepassings gebruik. Dit word die meeste gebruik in spraakherkenningstelsels, maar word tog ook vir ander toepassing, soos klemtoontoekenning, gebruik. Versteekte Markovmodelle (VMMs) bestaan uit 'n stel eindige toestande, waarvan elkeen geassosieer word met 'n waarskynlikheidsdistribusie. Om van een toestand na 'n ander te beweeg, is 'n stel oorgangswaarskynlikhede nodig. In 'n spesifieke toestand kan 'n uitkoms gegenereer word, na gelang van die geassosieerde waarskynlikheidsverdeling. Slegs die uitkoms is aan 'n eksterne waarnemer sigbaar. Al die ander toestande is versteek. (Warakagoda, 1996)

Die algoritme werk deurdat dit die waarskynlikste opeenvolging van klempatrone vind, na gelang van die waarskynlikhede van die model. Die waarskynlikheid dat 'n klempatroon vir 'n gegewe woord voorkom word bereken deur die n vorige klempatrone ook in ag te neem wanneer 'n klempatroon aan 'n woord toegeken word (Taylor, 2005).



Figuur 25: 'n Versteekte Markovmodel

X = toestande

y = moontlike waarnemings

a = toestandoorgangswaarskynlikhede

b = uitkomswaarskynlikhede

'n VMM word volgens Rabiner (1989:260) gekenmerk deur die volgende elemente:

- Die getal toestande van die model, N .
- Die getal waarneemsimbole, M , per toestand.
- Die toestandoorgangswaarskynlikheid $A = \{a_{ij}\}$
- 'n Waarskynlikheidsverspreiding in elkeen van die toestande
- Die eerste toestandverspreiding $\pi = \{\pi_i\}$.

Die voordele en nadele van die VMM word uiteengesit deur Valsan *et al.* (2002:261):

Voordele:

- Kragtige leer- en ontsyfermetodes
- Die vermoë om opeenvolgings goed te kan hanteer.
- Buigbare topologie vir statistiese fonologie en sintaksis.

Nadele:

- Swak onderskeiding tussen modelle
- Onrealistiese aannames moet gemaak word om 'n VMM teorie aan te voer, naamlik die onafhanklikheid van agtereenvolgende invoervektore en opeenvolgende toestande.

Taylor (2005) het 'n grafeem-na-foneem-omsetter vir Engels gebou deur VMMe te gebruik. Om dit beter te laat werk, het hy ook 'n klemtoontoekenner met VMMe gebou. In die leksikon wat hy gebruik het, het een vokaal presies een lettergreep aangedui, dus is alle konsonante is uit die grafeme verwyder. Die vokale word dus die observasies en die klempatrone is die versteekte toestande. 'n Afrigtingstel van ongeveer 110 000 Engelse woorde is gebruik en die model het 'n akkuraatheid van 76% behaal.

3.2.1.6 N-gramme

'n N-gram model is 'n waarskynlikheidsmodel wat voorspel wat die volgende item in 'n reeks sal wees. Die model gebruik die vorige $n-1$ lettergrepe om die waarskynlikheid te bereken dat die volgende lettergreep beklemtoon sal wees.

Demberg *et al.* (2007) ken klemtoon toe aan Duitse woorde deur n-gramme te gebruik. Dit bereken die mees waarskynlike volgorde van woordklemtoonetikette vir 'n gegewe lettergreepvolgorde. Een groot probleem wat Demberg *et al.* (2007:98) ondervind het met klemtoontoekenning deur middel van n-gramme, is dat die konteks-venster nie die hele woord insluit nie. Dit veroorsaak dat die algoritme nie weet of daar reeds klemtoon aan 'n spesifieke woord toegeken is buite die konteks-venster nie. Die fout het daartoe gelei dat byna 20% van die woorde inkorrekt beklemtoon is.

3.2.1.7 Maksimale entropie (ME)

Maksimale entropie is 'n statistiese raamwerk wat uitkomst teken aan elke voorkoms, deur gebruik te maak van die voorkoms se geskiedenis of die konteks waarin dit staan. Dit bereken die waarskynlikheid, $p(o|h)$, vir enige o vanaf al die moontlike uitkomstesruimtes, O , en vir alle h vanaf al die moontlike historiese uitkomstesruimtes, H . Die berekening van $p(o|h)$ in 'n ME hang af van 'n stel binêre kenmerke om voorspellings te maak oor die uitkomst.

Die beginsel van Maksimale entropie bepaal dat wanneer daar slegs gedeeltelike inligting oor die moontlike uitkomstes is, die verdeling $p(a,b)$ gekies moet word wat die onsekerheid oor die ontbrekende inligting maksimaliseer (Jaynes, 1957, aangehaal deur Ratnaparkhi, 1977:2).

Volgens Barros en Weiss (2006:178), is die benadering van maksimale entropie bekend daarvoor dat dit meerduidigheidsprobleme in natuurliketaalverwerking kan oplos. Die doel van ME is om 'n funksie $X \rightarrow Y$ te bereken, wat 'n objek $x \in X$ vir 'n klas $y \in Y$ kan voorspel. Y verteenwoordig die voorafbepaalde klasse vir elke taak van die voorspellingsprobleem. Tydens klemtoontoeëkennung word daar met 'n binêre klassifikasie gewerk, waar die klas waar is vir 'n beklemtoonde lettergreep en vals is vir 'n onbeklemtoonde lettergreep. Die klemtoontoeëkennung van Barros en Weiss (2006) het 'n akkuraatheid behaal van 85.57%, op 'n toetskorpus van 550 woorde.

3.2.2 Die Linguistiese Benadering

Die linguistiese benadering (ook genoem die kennisgebaseerde benadering of reëlgebaseerde benadering) behels hoofsaaklik dat 'n stel reëls deur 'n linguïst opgestel word en dan in 'n program geïmplementeer word in die vorm van AS-DAN reëls. Die stelsel is dan veronderstel om gegewe data eenders te behandel as wat die linguïst sou.

Ligêza (2006:92) noem dat 'n reëlgebaseerde stelsel se reëls die volgende vorm aanneem:

reël : (*vereistes*) $\rightarrow$ (*besluit*)

waar (*vereistes*) 'n formule is wat bepaal wanneer 'n reël gebruik kan word (en *besluit* is) die uitwerking van die toepassing van die reël. Meeste reëlgebaseerde stelsels het 'n hele aantal vereistes en slegs een besluit:

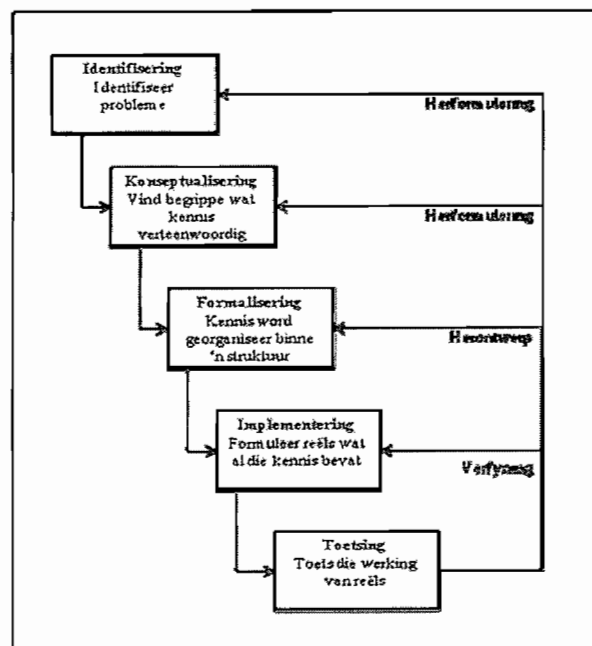
$p_1 \text{ EN } p_2 \text{ EN } \dots \text{ EN } p_n \rightarrow h$

Kennis word dus deur 'n deskundige vasgevang in die vorm van menslike logika, en word dan geprogrammeer deur gebruik te maak van een of ander programmeertaal (Leondes, 2000:2).

Daar is ook verskeie strategieë waarvolgens reëls uitgevoer kan word. Freeman-Hargis (2002) beskryf 'n paar van die strategieë:

- Eerste toepaslik: Die reëls word in 'n spesifieke volgorde uitgevoer. Die strategie laat toe dat kontrole uitgeoefen kan word van watter reëls voorkeur kry.
- Willekeurig: Die reëls word willekeurig uitgevoer. In dié tipe strategie kan daar ook 'n waarskynlikheid aan elke reël gekoppel word, sodat reëls met 'n hoër waarskynlikheid eerste uitgevoer word.
- Mees spesifiek: Die reël met die meeste voorwaardes word eerste gekies en uitgevoer.
- Mins resente gebruik: Elke reël kry 'n tydstempel om te bepaal wanneer dit die laaste keer uitgevoer is. Dié strategie werk goed wanneer dit nodig is dat al die reëls ten minste een keer uitgevoer moet word.
- Beste reël: Elke reël moet 'n gewig kry wat sal help bepaal of dit in ag geneem moet word bo die ander reëls. Die reël met die beste uitkomst vir 'n spesifieke probleem sal dan gekies word volgens sy gewig.

Ligêza (2006:237) noem dat 'n reëlgebaseerde stelsel uit vyf stappe bestaan:



Figuur 26: Die reëlgebaseerde lewensiklus

- Identifisering: Die tipe probleme wat die stelsel moet kan oplos moet eerstens geïdentifiseer word. Dit sluit ook die identifisering in van die data waarop die stelsel moet werk, asook die hulpbronne wat beskikbaar is om die probleme te help oplos.
- Konseptualisering: Die hoofontwerp van die stelsel moet uitgewerk word. Alle begrippe moet gevind word wat die kennis verteenwoordig.
- Formalisering: Die organisering van die kennis geskied in hierdie fase. Reëls word gegenereer en reggemaak vir die implementeerfase wat volg.
- Implementering: 'n Vooruitgekose programmeertaal word gebruik om die reëls wat in die vorige stap gegenereer is om te sit in 'n uitvoerbare program.
- Toetsing: Die resultate van die implementering word geëvalueer en moontlike foute word tydens dié stap geïdentifiseer en gekorrigeer.

Die voordele van die algoritme is dat dit die uitvoer verduidelik, sodat 'n gebruiker kan sien hoe 'n spesifieke uitvoer gevind is. Dit is geskik vir klein datastelle en dit is 'n maklik verstaanbare algoritme. Een groot nadeel van die algoritme is die feit dat dit nie aanpasbaar is nie (Alkoffash, Bawaneh en Rabea, 2008:611). Dus sal die stelsel wat in die studie ingebou word taal-spesifiek wees en nie vir ander tale gebruik kan word nie.

Braga en Coelho (2008:179) het 'n reëlgebaseerde klemtoontoekeener vir Europese en Brasiliaanse Portugees gebou om 'n teks-na-spraak-stelsel te verbeter. Die toekeener bevat 31 reëls wat gebaseer is op die analise van die laaste grafeme van elke woord. Die reëlgebaseerde stelsel het baie goeie resultate opgelewer. Dit het 'n akkuraatheid van 99.59% vir Europese Portugees, 99.6% vir Brasiliaanse Portugees en 98.52% vir Galisies behaal. Volgens Braga en Coelho (2008:182), kan dié benadering maklik aangepas word vir ander tale. Rustullet *et al.* (2008) het Braga en Coelho (2008) se benadering gevolg om 'n reëlgebaseerde klemtoontoekeener vir die Katalaanse taal te bou. Hulle het ook, soos Braga en Coelho (2008), baie goeie resultate behaal met die reëlgebaseerde benadering.

McPeters en Tharp (1983) het die klemtoonreëls van Liberman en Prince (1977) gebruik om klemtoon toe te ken in 'n teks-na-spraak stelsel. Geen verandering is aangebring aan die reëls nie. Primêre klemtoon is korrek toegeken aan 73% van die toetswoorde, en sekondêre klemtoon het slegs 53% behaal. McPeters en Tharp (1983:197) noem dat linguistiese teorieë 'n stewige

basis aan kunsmatige intelligensiestelsels kan verskaf, waarop mensetaaltegnologie-komponente gebou kan word.

Church (1985) se klemtoonreëls fokus op die gewig van die sillabe, woordsoorte, morfologie en etimologie om klemtoon aan woorde toe te ken. Die klemtoontoekenner het 'n akkuraatheid van 82% behaal.

In dié studie word die reëlgebaseerde benadering gevolg om 'n klemtoontoekenner vir Afrikaans te ontwikkel. Die vorige klemtoontoekenner het 'n goeie akkuraatheid van 93% behaal met 'n stel reëls wat nie linguisties korrek is nie. Dit is dus moontlik om die klemtoontoekenner heelwat te verbeter deur 'n volledige en linguisties korrekte stel reëls te implementeer. Die reëlgebaseerde benadering het goeie resultate opgelewer vir verskeie tale. Die resultate wat in dié afdeling genoem is, dui daarop dat 'n hoogs-akkurate klemtoontoekenner met 'n reëlgebaseerde benadering wel moontlik is, alhoewel die toekenner dan taal-spesifiek sal wees. In 'n reëlgebaseerde stelsel is dit ook baie maklik om verbeteringe en verstellings aan die reëls aan te bring.

3.3 *STAPPE VIR DIE ONTWIKKELING VAN 'N REËLGEBASEERDE KLEMTOONTOKENNER*

In die vorige afdeling is die vyf stappe genoem wat gevolg moet word om 'n reëlgebaseerde stelsel te bou. Dié vyf stappe sal nou beskryf word vir die konstruering van 'n reëlgebaseerde klemtoontoekenner.

- Identifisering van die probleem: Die doel van die studie is om 'n klemtoontoekenner te bou wat binne 'n grafeem-na-foneem-stelsel klemtoon aan tekseenhede toeken. Die klemtoontoekenner moet spesifiek vir Afrikaans ontwerp word. Om die klemtoontoekenner te bou, word 'n stel klemtoonreëls vir Afrikaans benodig. Die hulpbronne wat gebruik kan word om die toekenning van klemtoon te vergemaklik, is die kompositumanaliseerder, CKarma, en die lettergreepverdeler, Calomo. Lettergreepverdeling en morfologiese analise is noodsaaklike komponente wat die taak van klemtoontoekenning vergemaklik.
- Konseptualisering van die hoofontwerp: Eerstens sal data, wat as invoer gebruik gaan word, opgedeel word in tekseenhede waarvan alle leestekens verwyder is. Daarna sal lettergreepverdeling en morfologiese analise op die tekseenhede gedoen word.

Hierdie eerste drie stappe is nie deel van die klemtoontoekenner nie, en sal vooraf gedoen word deur eksterne hulpbronne te gebruik. Volgende sal die tekseenhede as invoer gegee word aan die klemtoontoekenner, wat dan klemtoon sal toeken aan die tekseenhede deur die geïmplementeerde reëls te gebruik. Elke woord word afsonderlik in die klemtoontoekenner ingevoer en word dan gepas aan die reëls binne die klemtoontoekenner. Die reël wat die tekseenheid die beste pas, sal dan op die tekseenheid toegepas word en so klemtoon daaraan toeken. Om te verhoed dat daar aan 'n tekseenheid meer as een keer klemtoon toegeken word, sal die reëls so geskryf word dat sodra 'n reël klemtoon aan 'n woord toeken, dit onmiddellik uitgevoer word en nie nog aan ander reëls gepas word nie. Die uitvoer vanaf die klemtoontoekenner sal dan direk na die grafeem-na-foneem-omsetter gestuur word as invoer, waar dit dan getranskribeer sal word in foneme.

- Formalisering van die reëls: In hoofstuk 2 is 'n stel klemtoonreëls vir Afrikaans reeds ontwikkel. Dit bestaan uit 19 reëls, wat tydens die volgende stap in die program geïmplementeer gaan word. Die reëls is onderverdeel in klasse. Dit bestaan uit reëls vir simplekse woorde en vir afleidings. Reëls moet só geformuleer word dat daar nie teenstrydighede kan voorkom nie.
- Implementering van die reëls: Die reëls sal geïmplementeer word deur die programmeertaal Perl te gebruik. Wanneer reëls geïmplementeer word moet die volgorde van die reëls in ag geneem word. Die geskrewe reëls moet ook so min moontlik verander word tydens die implementeringsproses. Die implementeringsproses sal in die volgende hoofstuk meer volledig bespreek word.
- Toetsproses: Die program sal op 'n lys van 10 000 tekseenhede getoets word. Die lys sal onderverdeel word in twintig lyste van 500 woorde elk. Nadat elke lys getoets is, sal foute wat in die program mag voorkom, eers gekorrigeer word en reëls sal weggelaat of bygevoeg word, soos nodig. Die toetsproses sal ook in die volgende hoofstuk in detail bespreek word.

3.4 *SAMEVATTING*

In die hoofstuk is die benaderings bespreek wat gevolg kan word om 'n klemtoontoekenner te bou. Die benaderings is in twee afdelings opgedeel, naamlik 'n datagedrewe benadering en 'n reëlgebaseerde benadering. Die verskil tussen datagedrewe benaderings en die reëlgebaseerde

benadering is dat datagedrewe benaderings “leer” vanaf ’n datastel, waar die reëlgebaseerde benadering afhanklik is van ’n deskundige om ’n stel reëls te genereer wat dan geïmplementeer word in ’n program. Die datagedrewe benadering bestaan uit verskeie algoritmes. In dié hoofstuk is slegs ’n paar daarvan bespreek, naamlik neurale netwerke, besluitnemingsbome, voorbeeldgebaseerde leer, steunvektormasjiene, versteekte Markovmodelle, n-gramme, en maksimale entropie.

Die doel van die studie is om ’n hoogs-akkurate klemtoontoekeener te ontwikkel wat die akkuraatheid van die reëlgebaseerde grafeem-na-foneem-omsetter sal verbeter. Die benadering wat gekies is om hierdie doel te bereik, is die reëlgebaseerde benadering. Die bestaande klemtoontoekeener het ’n akkuraatheid van byna 90% behaal met ’n stel reëls wat nie linguisties korrek is nie, dus is die verwagting dat ’n reëlgebaseerde klemtoontoekeener met ’n stel reëls wat linguisties akkuraat is, ’n baie hoër akkuraatheid kan behaal. Alhoewel dit ’n taalspesifieke toepassing gaan wees, sal dit steeds tot groot voordeel strek vir Afrikaanse taaltegnologietoepassings wat klemtoontoekening nodig. Die programmeertaal Perl gaan gebruik word vir die implementering van die klemtoonreëls. Die rede waarom dié taal gekies is, word in die volgende hoofstuk bespreek.

In die volgende hoofstuk word die ontwikkeling van die klemtoontoekeener bespreek. Daar sal ’n beskrywing gegee word van die programmeertaal wat gebruik word, asook ’n beskrywing van die argitektuur van die program. Die implementeringsproses en die toetsproses sal daarna beskryf word. Die resultate sal aan die einde van die hoofstuk weergegee word.

Hoofstuk 4

'n KLEMTOONTOEKENNER VIR AFRIKAANS

4.1 INLEIDING

In hoofstuk 3 is die benaderings ondersoek wat moontlik gevolg kan word vir die ontwikkeling van 'n klemtoontoekenner. Die linguistiese of reëlgebaseerde benadering is gekies vir die ontwikkeling van die klemtoontoekenner. In hierdie hoofstuk word die implementeringsproses beskryf, asook die proses om toetsdata te genereer.

In hoofstuk 2 is klemtoonreëls gegenereer wat geïmplementeer word in die program. Sodra die reëls geïmplementeer is, word die toetsproses begin. Die toetsproses word in twintig fases opgedeel, waartydens foute wat deur die klemtoontoekenner gemaak word, uitgewys en gekorrigeer word.

Nadat die reëls met 10 000 tekseenhede getoets is, sal die program geëvalueer word. Die 10 000 tekseenhede waarmee die klemtoontoekenner getoets is, word dan vergelyk met 'n weergawe waarvan die klemtoon korrek aan die tekseenhede toegeken is en die akkuraatheid van die klemtoontoekenner word dan outomaties bereken. Daarna word die klemtoontoekenner ook geëvalueer met verskillende tekste. Die verskillende komponente wat die klemtoontoekenner en die grafeem-na-foneem-omsetter beïnvloed word ook geëvalueer en bespreek.

In dié hoofstuk sal die werking van die klemtoontoekenner eerstens bespreek word. Daarna sal die toetsproses in detail bespreek word. Volgende sal 'n verduideliking gegee word van die evalueerproses, asook die bespreking en interpretering van die relevante resultate. Die bestaande klemtoontoekenner sal met dieselfde stel van 10 000 tekseenhede geëvalueer word om vergelykende resultate op te lewer.

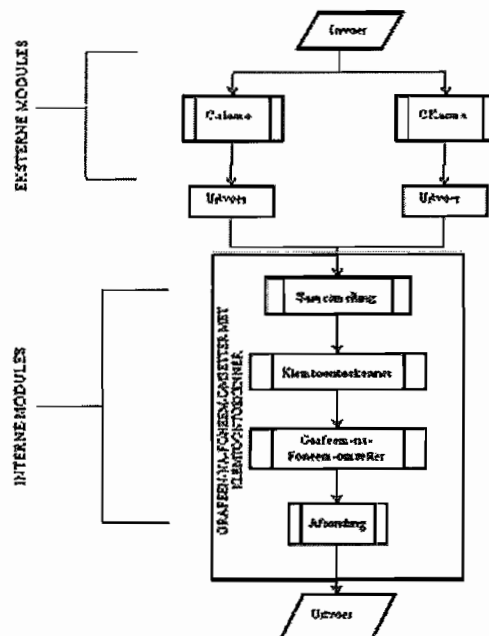
4.2 DIE KLEMTOONTOEKENNER

Die klemtoontoekenner is ontwikkel met die programmeertaal Perl (Practical Extraction and Reporting Language), 'n programmeertaal wat gebruik word om willekeurige teks deur te gaan en

inligting daarvan te onttrek. Dit is in 1987 deur Larry Wall ontwikkel en bevat die beste eienskappe van o.a. C, awk, sed, sh en BASIC (The Perl Directory, 2009). Een groot voordeel van Perl is die feit dat dit gratis is en versoenbaar is met enige bedryfstelsel. Nog 'n voordeel van Perl is sy vermoë om teks te manipuleer deur middel van patroonpassing (pattern matching) en reëlmatige uitdrukkings (regular expressions) (Turner, 2008). Die argitektuur van die klemtoontoekeener word hierna bespreek.

4.2.1 Die argitektuur van die klemtoontoekeener

In dié studie maak die klemtoontoekeener deel uit van 'n grafeem-na-foneem-omsetter vir Afrikaans. Die program waarin die klemtoontoekeener voorkom, bestaan uit verskeie interne en eksterne modules. Al die modules word hieronder bespreek:



Figuur 27: Argitektuur van die grafeem-na-foneem-omsetter

4.2.1.1 Invoer en uitvoer

Invoer: Die invoer bestaan uit enige hoeveelheid tekseenhede.

Uitvoer vanaf die eksterne modules: Die uitvoer wat vanaf die eksterne modules kom, is dan die invoer vir die interne modules. Die uitvoer moet wel eers met die hand nagegaan word, om seker

te maak dat al die tekseenhede korrek verdeel is. Die uitvoer is in die vorm van 'n csv-lêer, waarvan die eerste kolom die oorspronklike tekseenhede bevat. Die tweede kolom bevat al die tekseenhede wat in lettergrepe verdeel is en die derde kolom bevat al die tekseenhede wat morfologies opgedeel is. Die uitvoer vanaf dié modules lyk soos volg:

	A	B	C
1	aangesig	aan*ge*sig	aan+gesig
2	aangestuur	aan*ge*stuur	aan+gestuur
3	aangeteken	aan*ge*te*ken	aan+geteken
4	aangetrek	aan*ge*trek	aan+getrek

Figuur 28: Uitvoer vanaf eksterne modules

Uitvoer: Die uitvoer is 'n csv-lêer, netsoos die invoer-lêer. In die eerste kolom is die oorspronklike tekseenhede. Die saamgesmelte CKarma- en Calomo-data is in die tweede kolom. Die tekseenhede waaraan klemtoon toegeken is, is in die derde kolom en die getranskribeerde tekseenhede in die vierde kolom:

	A	B	C	D
1	woord	CalomoCKarmaMerge	Klem	Foneties
2	aangesig	aan+ge*sig	;aan*ge*;sig	anx@s@x
3	aangestuur	aan+ge*stuur	;aan*ge*;stuur	anx@styr
4	aangeteken	aan+ge*te*ken	;aan*ge*;te*ken	anx@tek@n
5	aangetrek	aan+ge*trek	;aan*ge*;trek	anx@trEk

Figuur 29: Uitvoer

In kolom C van figuur 29, lyk dit of die toekenner primêre en sekondêre aan woorde toeken. In die voorbeeld is al die woorde samestellings wat deur CKarma verdeel word en dus kry elke aparte woord van die samestelling primêre klemtoon. Woorde wat nie deur CKarma verdeel word nie, sal slegs een lettergreep bevat wat beklemtoon word.

4.2.1.2 Lettergreepverdeling en morfologiese analise

Calomo: In hoofstuk 1 is genoem dat die lettergreepverdeler gebruik word om toetsdata te genereer vir die program. In hoofstuk 2 kan gesien word dat woordklemtoon telkens toegeken word aan 'n lettergreep in 'n woord. Dit is dus baie belangrik dat die tekseenheid wat ingevoer word, in lettergrepe verdeel is, sodat die klemtoontoekenner kan onderskei tussen die lettergrepe en sodoende klemtoon aan die korrekte lettergreep kan toeken. J.J McCarthy (1979:443) noem ook dat die sillabestruktuur die meeste gebruik word van al die taalstrukture, om te bepaal waar 'n

woord se klemtoon lê. Die lettergreepverdelers oftewel woordafbrekers Calomo word gebruik om alle woorde wat in die klemtoontoekenner ingevoer moet word, in lettergrepe te verdeel. Die woord *sonnestelsel* sal geanaliseer word as *son*ne*stel*sel*. Die asteriske dui die posisie aan waar die lettergrepe verdeel word.

CKarma: Die kompositumanaliseerder speel ook 'n kritieke rol in die toekenning van klemtoon aan 'n groot hoeveelheid Afrikaanse tekseenhede is samestellings. Dus is dit van groot belang dat die tekseenhede wat ingevoer word, eers verdeel word in tekseenhede, sodat klemtoon aan die verskillende tekseenhede van 'n samestelling toegeken kan word en sodoende die akkuraatheid van die klemtoontoekenning en die grafeem-na-foneem-omsetting kan verhoog. CKarma neem 'n teksstring as invoer en gee 'n geanaliseerde string, bestaande uit onafhanklike konstituenten en valensiemorfeme, as uitvoer. As 'n woord soos *sonnestelsel* byvoorbeeld geanaliseer word, sal die uitvoer *son_ne+stelsel* wees. Die onafhanklike konstituent word geskei met 'n plusteken, en die onderstreep dui die begin van 'n valensiemorfeem aan. Die klemtoontoekenner benodig nie inligting oor valensiemorfeme nie, dus word alle onderstrepe uit die data verwyder voor dit as invoer aan die toekenner gegee word. In die evaluasieproses sal meer duidelikheid verkry word oor die invloed van die kompositumanaliseerder op die klemtoontoekenning.

Samesmelting: In dié module word die laaste twee kolomme van die invoer lêer, naamlik dié vanaf Calomo en dié vanaf CKarma, saamgesmelt tot een kolom. Die invoer vanaf Calomo en CKarma word binne die program saamgevoeg, deur die woorde in kolom B en C (fig.7) met mekaar te vergelyk en slegs 'n sterretjie (*) met 'n plusteken (+) te vervang waar daar 'n (+) in kolom C voorkom.

4.2.1.3 Die klemtoontoekenner

Die invoer vir dié module is die CKarma- en Calomo-data wat saamgesmelt is. Indien 'n woord verdeel word deur CKarma, sal dit as twee aparte woorde aan die klemtoontoekenner gestuur word. Dus sal elke woord sy eie primêre klemtoon kry. Sodra beide woorde beklemtoon is, sal hulle dan weer aanmekaar gevoeg word soos in kolom C van figuur 29.

Die klemtoontoekenner bestaan uit 'n getal AS-DAN reëls, wat gebaseer is op die klemtoonreëls wat in hoofstuk 2 gegenereer is. Dit is belangrik om reëls te orden, sodat die reëls met die hoogste prioriteit eerste uitgevoer kan word en om die teenstrydigheid van reëls te voorkom (Han & Kamber, 2006:319). Die klemtoontoekenner gebruik beide tegnieke. Eerstens word die reëls

in klasse verdeel, in terme van die hoeveelheid lettergrepe wat 'n tekseenheid bevat. Dus is die module verdeel in reëls vir:

- eenlettergrepige tekseenhede,
- tweelettergrepige tekseenhede,
- drielettergrepige tekseenhede, en
- vier- of meerlettergrepige tekseenhede.

In elke klas word die reëls dan georden sodat die prefiks-reëls heel eerste uitgevoer word, dan die reëls vir suffikse en die sjwa. Laastens word die reëls vir Afrikaanse simplekse uitgevoer. Dié volgorde verseker dat die tekseenhede wat by die laaste stel reëls, naamlik die simplekse reëls, uitkom, wel almal simplekse stamme is en geen affikse bevat nie.

In hoofstuk 2 is gemeld dat affikse onbeklemtoonbaar, klemtoonaantrekkend, of klemdraend kan wees. Dit is dus noodsaaklik dat die klemtoontoekenner onderskei tussen die verskillende klasse affikse. Die module herken affikse deurdat dit vooraf as veranderlikes geklassifiseer is. Die lyste veranderlikes is gebaseer op die affiks-lyste wat Combrink (1990:37) geskep het. Die program kan dus affikse in 'n lys opsoek en dit dan pas by 'n tekseenheid. Die meeste meerlettergrepige tekseenhede in Afrikaans, afgesien van samestellings, bestaan uit 'n prefiks en/of 'n suffiks en nog 'n een- of tweelettergrepige simpleks. As die affikse dan herken word, is dit hoofsaaklik 'n een- of tweelettergrepige woord wat oorbly om geanaliseer (ontleed) te word. Sodanig word die kanse vir die klemtoontoekenner aansienlik verminder om klemtoon aan die verkeerde lettergreep toe te ken. Op dié manier kan die klemtoontoekenner 'n woord soos *bekragtig* verdeel in *be* + *krag* + *tig*, waar *krag* dan die stam sal wees wat beklemtoon word en *be-* en *-tig* sal herken word as onbeklemtoonbare affikse, uit 'n lys wat dié tipe affikse vooraf identifiseer.

Slegs primêre klemtoon word deur die klemtoontoekenner aan tekseenhede toegeken. Sekondêre klemtoon in Afrikaanse simplekse is baie onreëlmatig en kan dus nie suksesvol deur middel van reëls toegeken word nie. Vroeër in die hoofstuk is reeds genoem dat samestellings wel primêre klemtoon kry op elke deel wat deur CKarma verdeel word. Die akkuraatheid van die grafeem-na-foneem-omsetter word nie baie beïnvloed deur die toekenning van sekondêre klemtoon nie. Dus is daar besluit dat sekondêre klemtoon vir die doeleindes van hierdie studie uitgelaat gaan word. Tydens die evaluasieproses is die invloed van sekondêre klemtoon op die akkuraatheid van die

grafeem-na-foneem-omsetter getoets. Die resultate het gewys dat die akkuraatheid van die omsetter maar met 0.6% gestyg het met die byvoeging van sekondêre klemtoon. Later in die hoofstuk sal die toets in diepte bespreek word.

4.2.1.4 Die grafeem-na-foneem-omsetter

Die grafeem-na-foneem-omsetter is reeds in 'n vorige studie ontwikkel en word hier bespreek slegs as agtergrond. Die module bestaan net soos die klemtoontoekenner uit 'n getal AS-DAN reëls, wat tekseenhede waaraan klemtoon toegeken is, omsit in foneme wat uitgedruk word in die fonetiese alfabet, SAMPA (Speech Assessment Methods Phonetic Alphabet).

Daar is 'n paar veranderinge aangebring aan die SAMPA foneemstel, omdat daar uitsluitlik een simbool per klank gebruik moet word. In die tabel 3 word die veranderinge aangedui.

SAMPA SIMBOOL	GEWYSIGDE SIMBOOL	VOORBEELD IN SAMPA	ORTOGRAFIESE VOORBEELD
a:	a	man	maan
e:	e	nem	neem
o:	o	lop	loop
y:	y	fyr	vuur
ɹ:	ɹ	d2r	deur
@i	Q	pQl	peil / pyl
9y	Y	Yt	uit
9u	U	Ut	oud
tS	C	Cali	tjalie

Tabel 3: SAMPA-veranderinge

Die module is ook in klasse verdeel, naamlik:

- Woorddele,
- Diakritiese tekens,
- Vokale,
- Konsonante.

In die eerste deel van die module word woorddele getranskribeer. Die klas kan ook beskryf word as 'n klas uitsonderings, aangesien die woorddele wat hier getranskribeer word nie korrek omgesit kan word in foneme as dit volgens die grafeem-na-foneem-reëls getranskribeer word nie. In die vorige studie waar dié module ontwikkel is, is 'n toets gedoen om te bepaal of hierdie klas reëls noodsaaklik is vir die module. In die toets is 'n stel van 1 000 ewekansig gekose tekseenhede onttrek uit 'n lys van 10 000 tekseenhede. Eers is die akkuraatheid van die module verkry deur die reëls wat woorddele transkribeer weg te laat. Daarna is die reëls weer bygevoeg. Die akkuraatheid van die module het, met die byvoeging van die reëls, met 3,3% gestyg.

Die tweede klas reëls transkribeer alle diakritiese tekens (akuut, gravis, deelteken en kappie). Dit bestaan uit al die moontlike vorms waarin diakritiese tekens kan voorkom, bv. *oë, ië, eë, ê, ô*.

Vokale word in die derde klas getranskribeer. Elke vokaal kan op 'n hele aantal maniere uitgespreek word. Die grafeem-na-foneem-omsetter is afhanklik van lettergrepe en klemtoon om te bepaal wat die uitspraak van 'n spesifieke vokaal gaan wees. As voorbeeld kan die woord *flenterklere* gebruik word. Die woord word verdeel deur die lettergreepverdeler en die kompositumanaliseerder as *flen*ter + kle*re*. Dus gaan die module se invoer die twee tekseenhede *;flen*ter* en *;kle*re* wees. 'n Sjwa kan voorkom in die volgende vorme: *-e, -el, -em, -en, -er, -es*. Dus sal die */e/* eerste lettergreep van *flenter* as 'n sjwa getranskribeer word as die nie beklemtoon is nie. Dus sal die */e/* in *flen* as 'n foneem */E/* getranskribeer word, omdat dit beklemtoon is en in 'n toe lettergreep lê. Die tweede lettergreep is ook geslote, maar onbeklemtoon. Dus sal dié */e/* as 'n sjwa getranskribeer word. In die tweede woord, *klere*, is die eerste en die tweede lettergreep oop. Aangesien die eerste lettergreep beklemtoon word, word dit as 'n lang */e/* getranskribeer en die tweede lettergreep, wat sonder klemtoon is, word dan as 'n sjwa getranskribeer.

Die reëls binne die program wat dié voorkomste transkribeer lyk soos volg:

- Om *flen* te transkribeer na *fEn*:

```
if($token =~ /^(\.)(\*)?(\^)(\$V)(e)(\$K)(\*)$/) {
```

```
$token = $1.$2.$3.$4.'E'.$6.$7;}
```

waar \$K = enige konsonant, \$V = enige vokaal.

Dié reël kan gelees word dat as daar 'n /e/ binne 'n beklemtoonde lettergreep voorkom, wat nie voorafgegaan word deur 'n vokaal nie en gevolg word deur 'n konsonant, die /e/ getranskribeer sal word as die foneem /E/.

- Om *ter* te transkribeer na *t@r*:

```
if($token =~ /^(\.)(\*)?(\^)(\$K*)(e)(\$S)?(\*)$/) {
```

```
$token = $1.$2.$3.$4.'@'.$6.$7;}
```

waar \$K = enige konsonant, \$S = enige sonorante konsonant is.

Dié reël vereis dat 'n /e/ nie binne 'n beklemtoonde lettergreep mag wees nie en mag nie voorafgegaan word deur enigiets anders as nul of meer konsonante nie. Die /e/ moet ook gevolg word deur nul of een sonorante konsonant (*l, m, n, r, s*). Dan sal dit as 'n /@/ (sjwa) getranskribeer word.

Die reëls wat *klere* transkribeer na *kler@* is soortgelyk aan bogenoemde reëls met die verskil dat dit vir 'n /e/ in 'n oop lettergreep geskryf word, met ander woorde 'n lettergreepverdeling of woordeinde sal op die /e/ volg.

Klemtoon het 'n baie groot invloed op grafeem-na-foneem-omsetting. In die vorige studie, waar die ou klemtoontoekenner ontwikkel is, is die invloed van klemtoon op grafeem-na-foneem-omsetting ook getoets. Die toets is op 'n duisend ewekansig gekose tekseenhede gedoen. Die grafeem-na-foneem-omsetter se akkuraatheid, met die klemtoon vanaf die vorige klemtoontoekenner wat 90% akkuraat is, was 89.4%. Die klemtoon is toe aangepas, sodat dit 100% korrek is, en weer op dieselfde stel tekseenhede getoets. Die grafeem-na-foneem-omsetter het met dié toets 'n akkuraatheid van 96.7% behaal. Dus is klemtoontoekenning baie belangrik vir die omsetting van grafeme na foneme. Alhoewel die meeste grafeem-na-foneem-omsetters

klemtoon gebruik is daar wel verskeie grafeem-na-foneem-omsetters, soos die van Davel (2005) en Van den Bosch en Daelemans (1993) wat nie klemtoontoekenning gebruik nie, maar tog die probleem van grafeem-na-foneem-omsetting met sukses oplos.

In die laaste klas word konsonante getranskribeer. Klemtoon en lettergrepe het geen invloed op die transkripsie van konsonante nie. Slegs 'n paar konsonante, soos die /g/, /ng/, /ch/, /w/ en /v/, word verander na hul korrekte foneem.

Alhoewel die afrondingsmodule apart is van die grafeem-na-foneem-omsetter, is dit eintlik deel van die module. Sodra die tekseenhede deur die klemtoontoekenner en die grafeem-na-foneem-omsetter geprosesseer is, word al die tekseenhede wat deur die kompositumanaliseerder opgedeel is, weer aanmekaar geheg. Baie tekseenhede word dan verkeerd getranskribeer, omdat daar dan dubbel-konsonante in voorkom waar die twee tekseenhede verbind is, bv.:

$$\text{Op} + \text{pAt} \rightarrow \text{OppAt}$$
$$\text{IEs@nar} + \text{rek@nar} \rightarrow \text{IEs@narrek@nar}$$

In hierdie module word al dié gevalle gesoek en gekorrigeer. Dit is die laaste module van die program, dus word dit gedruk in die uitvoerlêer soos by figuur 29.

4.3 *EVALUASIE VAN DIE KLEMTOONTOEKENNER*

Dié afdeling bestaan uit twee afdelings, naamlik die toetsproses en die evaluasie van die klemtoontoekenner. In die eerste afdeling word toetsdata as invoer aan die klemtoontoekenner gegee en die toekenner word dan met dié data verbeter. In die tweede afdeling word dieselfde data en nog ander tekste gebruik om die toekenner te valueer.

4.3.1 *Die toetsproses*

Die toetsproses sal uit 20 fases bestaan. Elke fase bestaan weer uit twee dele. In die eerste stap word 500 tekseenhede aan die klemtoontoekenner as invoer gegee. Klemtoon word dan aan dié 500 tekseenhede toegeken. Tydens die tweede stap word die 500 tekseenhede waaraan klemtoon toegeken is, per hand gekorrigeer en die akkuraatheid van die toets word so bepaal. Daarna word daar gepoog om die foute wat begaan is, reg te maak en sodoende die program te verbeter. Hierdie proses word herhaal met 'n volgende 500 tekseenhede, totdat al 10 000 tekseenhede

getoets is. So word 'n volledig geannoteerde lys dan saamgestel wat in die evaluasieproses gebruik word.

Die toetsdata bestaan uit verskillende tekseenhede. Dit is saamgestel uit simplekse, komplekse, en afleidings. Alle leenwoorde en eiename is uit die lys verwyder, sodat dit suiwer uit Afrikaanse tekseenhede bestaan. Die volgende tabel toon hoe die lys saamgestel is, in terme van die persentasie simplekse, komplekse en samestellings, asook die getal lettergrepe waaruit die tekseenhede bestaan.

TIPE	PERSENTASIE	GETAL TEKSEENHEDE
Eenlettergrepig	11.4 %	1144
Tweelettergrepig	38%	3797
Simplekse	14%	532
Komplekse	22%	835
Afleidings	64%	2430
Drielettergrepig	30.2%	3020
Simplekse	8%	242
Komplekse	35%	1057
Afleidings	57%	1721
Vier- en meerlettergrepig	20.4%	2039
Simplekse	34%	693
Komplekse	46%	938
Afleidings	20%	408

Tabel 4: Samestelling van woordelys

Die lys is in alfabetiese volgorde gerangskik en is daarna verdeel in twintig woordelyste van 500 tekseenhede elk. Na elke fase is die lyste per hand nagegaan en foute wat deur die program gemaak is, is geïdentifiseer en gekorrigeer waar moontlik.

Die meeste foute wat voorgekom het, is foute wat gemaak is rondom affikse. 'n Woord soos *bewe* of *gewel* is byvoorbeeld telkens beklemtoon op die tweede lettergreep, omdat die eerste lettergreep die vorm van 'n prefiks aanneem. Die fout is oorkom deur die volgorde van reëls te verander, sodat die reël 4 eerste uitgevoer word. Die reël lui dat as 'n tweelettergrepige se laaste lettergreep 'n sjwa bevat, die woord beklemtoon sal word op die eerste lettergreep. Dus sal die woord *bewe* (**be** (*prefiks*) - *we* (*sjwa*)), op die eerste lettergreep beklemtoon word. Dié reël het wel uitsonderings, soos *vertel*, *gebel*, *gestel*. Daar is ook affikse bygevoeg en weggeneem uit die lyste wat prefikse en suffikse definieer, om sodoende die beste resultate te verkry.

Reëls 21 en 26 is uit die program verwyder, omdat dit nie korrekte klemtoon aan woorde toegeken het nie. Die reëls kom daarop neer dat indien 'n drie- of meerlettergrepige woord se laaste twee lettergrepe die vorm VV-VK aanneem (dus 'n oop lettergreep gevolg deur 'n toe lettergreep), sal die lettergreep beklemtoon word wat die twee lettergrepe voorafgaan. Die reëls sou woorde soos *horison*, *monitor*, *leksikon*, korrek beklemtoon, maar woorde soos *almanak*, *salaris*, *karamel*, *krokodil*, *diagram*, *notaris*, het die toekenner verkeerd beklemtoon. Daar is tydens die toetsproses bevind dat laasgenoemde woorde in die meerderheid is, maar omdat dié tipe woorde nie reëlmatige klemtoon het nie, is besluit om die vokaalhiërargie van Combrink en De Stadler (1987:192) vir die taak in te span.

Daar is nog 'n reël bygevoeg vir die voorkoms van die woorduitgange *-da*, *-na*, *-sie*, *-ium*, *-ika*, *-ia*, *-ea*. Woorde met dié woorduitgange word beklemtoon op die lettergreep wat die woorduitgang voorafgaan, behalwe as die voorafgaande lettergreep 'n sjwa bevat, dan kry die lettergreep wat die sjwa-lettergreep voorafgaan die klemtoon. Dié woorduitgange kan vergelyk word met die klemtoonaantrekkende suffikse. In die onderstaande tabel word voorbeelde van dié woorde gelys. Dié reël geld nie net vir afleidings nie, maar ook vir simplekse wat op dié vorm eindig.

<i>-da</i>	<i>-na</i>	<i>-sie</i>	<i>-ium</i>	<i>-ika</i>	<i>-ia</i>	<i>-ea</i>
agenda	Afrika	abduksie	akwarium	Afrika	absentia	area
anakonda	arena	burokrasie	repetorium	Amerika	ammonia	kornea
Andromeda	ballerina	definisie	stadium	anestetika	bulimia	tragea
Bermuda	bandana	eleksie	adverbium	basilika	fantasia	hakea
estrada	Botswana	graduasie	kriterium	biotika	gloria	asalea
gouda	gimkana	advertensie	stramonium	dioptrika	hortensia	koglea
hasiënda	hiëna	anneksasie	seminarium	dinamika	Kenia	protea
jakaranda	hosanna	hipotensie	planetarium	elektronika	magnolia	
lambda	iguana	demensie	sporangium	estetika	personalia	
veranda	ikebana	abduksie	sirkonium	farmaseutika	sepia	
Oeganda	indoena	provinsie	presidium	grafika	urania	
propaganda	konsertina	responsie	osmium	japonika	tibia	
werda	lantana	produksie	natrium	logika	pizzeria	
Uitsonderings						
Kanada	patina				Samaria	alinea
hadida	alumina				Sagaria	Hosea
	platina					Judea
	stamina					Korea
						Galilea

Tabel 5: Voorbeelde van woorduitgange *-da*, *-na*, *-sie*, *-ium*, *-ika*, *-ia*, *-ea*.

In hoofstuk 2 is die vokaalhiërargie bespreek. Nadat reël 11 vir tweelettergrepige woorde, reël 18 vir drielettergrepige woorde en reël 24 of 25 vir vier- of meerlettergrepige woorde onderskeidelik uitgevoer is, word die vokaalhiërargie eers gebruik om klemtoon aan simplekse stamme toe te ken. Die vokaalhiërargie wat geïmplementeer word lyk soos volg: e>o>i>E>y>u>O>a>9.

Die toetsproses het die volgende resultate opgelewer:

TOETS	GETAL TEKSEENHEDE	GETAL FOUTE	AKKURAATHEID
T1	500	73	85.4%
T2	500	49	90.2%
T3	500	57	88.6%
T4	500	56	88.8%
T5	500	51	89.8%
T6	500	33	93.4%
T7	500	49	90.2%
T8	500	55	89%
T9	500	48	90.4%
T10	500	43	91.4%
T11	500	56	88.8%
T12	500	75	85%
T13	500	50	90%
T14	500	41	91.8%
T15	500	40	92%
T16	500	16	96.8%
T17	500	41	91.8%
T18	500	21	95.8%
T19	500	38	92.4%
T20	500	21	95.8%

Tabel 6: Die resultate van Toets 1-20

In die tabel is dit duidelik dat die klemtoontoeckenner verbeter het na afloop van die toetse. Die uitskieters wat hier gesien word, is te wyte aan die samestelling van die woordelyste. Waar een- en tweelettergrepige tekseenhede in die meerderheid is, was die akkuraatheid van die toekenner heelwat hoër as by die ander toetse.

Die onderstaande tabel gee 'n duideliker beeld, ten opsigte van die verhouding van die hoeveelheid lettergrepe waaruit die woorde binne elke toets bestaan:

TOETS	PERSENTASIE TEKSEENHEDE WAT ELKE LETTERGREEPKLAS VERTEENWOORDIG			
	Eenlettergrepig	Tweelettergrepig	Drielettergrepig	Vier- en meerlettergrepig
T1	3.60%	30.80%	38.20%	27.40%
T2	4.40%	35.20%	36.80%	23.60%
T3	16.80%	38.60%	27.20%	17.40%
T4	13.60%	31.80%	29.60%	25.00%
T5	11.20%	40.20%	28.00%	20.60%
T6	4.40%	44.60%	28.40%	22.60%
T7	16.20%	39.00%	27.80%	17.00%
T8	8.60%	30.60%	34.00%	26.80%
T9	13.20%	35.00%	28.60%	23.20%
T10	19.40%	39.40%	24.80%	16.40%
T11	10.00%	33.60%	35.40%	21.00%
T12	8.80%	29.00%	32.20%	30.00%
T13	6.60%	31.60%	36.40%	25.40%
T14	14.80%	33.40%	28.20%	23.60%
T15	17.40%	40.20%	24.20%	18.20%
T16	19.80%	44.80%	20.00%	15.40%
T17	15.20%	39.60%	25.20%	20.00%
T18	5.80%	33.60%	37.00%	23.60%
T19	10.80%	35.60%	36.00%	17.60%
T20	26.20%	33.00%	26.80%	14.00%

Tabel 7: Die samestelling van Toets 1-20

In die volgende afdeling volg 'n beskrywing van die evaluasieproses van die klemtoontoekenner en die relevante resultate word ook bespreek.

4.3.2 *Evaluasie*

In hoofstuk 1 is die evaluasieproses bespreek. Daar is genoem dat die evaluasieproses uit vier fases sal bestaan. In die eerste fase sal die 10 000 tekseenhede wat gebruik is as toetsdata weer getoets en geëvalueer word. Die toets sal ook op die vorige klemtoontoeckenner gedoen word. Hierdie resultate sal vergelyk word om vas te stel of die nuwe klemtoontoeckenner 'n verbetering is op die ou klemtoontoeckenner. Die foute wat die toekenner tydens die evaluering maak, sal ook bespreek word. Daarna sal die klemtoontoeckenner se akkuraatheid op lopende teks bepaal word, deur die teks van die Afrikaanse weergawe van die Universele Verklaring van Menseregte en twee tekste uit die 1983 vertaling van die Afrikaanse Bybel te gebruik.

Tydens die tweede evaluasiefase sal die grafeem-na-foneem-omsetter se akkuraatheid op dieselfde teks as fase 1 geëvalueer word. In die derde fase sal 'n toets gedoen word om die belangrikheid van die kompositumanaliseerder te bepaal. Daarna sal die aandag gevestig word op die invloed wat klemtoontoeckenning op grafeem-na-foneem-omsetting het.

4.3.2.1 **Die eerste evaluasieproses**

In Tabel 5 is die resultate van die toetsproses weergegee. Tydens die twintig toetse is gevind dat daar sekere foute is, wat deurentyd deur die klemtoontoeckenner gemaak word. Dié foute word hieronder bespreek:

- Die woorduitgange *-is* en *-is (t)*: Aangesien die onbeklemtoonde *-is* meer algemeen voorkom as die beklemtoonde *-is (t)*, word laasgenoemde deur die toekenner geklassifiseer as onbeklemtoonbaar. Enige woord met die woorduitgang *-is*, sal op die voorlaaste lettergreep beklemtoon word. Dus sal tekseenhede soos *pianis*, *futuris*, *bloemis*, altyd verkeerd beklemtoon word op die voorlaaste lettergreep, maar tekseenhede soos *begrafnis*, *sekretaris*, *kermis*, sal op die korrekte lettergreep beklemtoon word.
- Die woorduitgang *-ie*: In Combrink (1990:39) se agtervoegsellys verskyn daar drie *ie*-agtervoegsels. Die eerste *-ie* soos in *biologie*, *strategie*, *teorie*, waar die agtervoegsel klemdraend is en dan die tweede en derde *-ie* soos in *onnie* en *koninkie*, waar die agtervoegsel onbeklemtoonbaar is. Laasgenoemde agtervoegsels is die mees algemene vorm van die *-ie*. Dus is die reëls wat dié agtervoegsels hanteer, geskryf vir die

onbeklemtoonbare tweede en derde klas en sal die toekenner tekseenhede wat in die eerste klas val, nie korrek beklemtoon nie.

- Die sjwa: Die sjwa kom algemeen voor in die volgende vorm: *-e, -el, -em, -en, -er, -es*. Die onbeklemtoonbaarheid van die sjwa, veroorsaak dat tekseenhede wat eindig op 'n /e/ soortgelyk aan die sjwa, nie op daardie lettergreep beklemtoon sal word nie, byvoorbeeld, *sisteme, kombers, visuele, proses, tabel*.
- Tekseenhede wat voorvoegsels bevat, word in die toekenner heel eerste behandel. Dit gee aanleiding tot die probleem dat woord soos *bewend, versus* en *besig*, op die tweede lettergreep beklemtoon gaan word i.p.v. op die eerste lettergreep. Daar is gepoog om 'n reël by te voeg wat lui dat as daar 'n voorvoegsel voorkom wat direk gevolg word deur 'n agtervoegsel, die voorvoegsel beklemtoon moet word. Hierdie reël het nie gewerk nie, omdat dit tekseenhede soos *berig, berus, beskik, vermeld, verkies*, verkeerd beklemtoon het.
- Agtervoegsels word dikwels voorafgegaan deur 'n lettergreep wat 'n sjwa bevat, byvoorbeeld *vermeerdering, gelerig, dopeling, geneties*. Dan is daar afleidings wat dieselfde vorm as bogenoemde voorbeelde het, maar wat nie 'n sjwa bevat nie: *tarivering, stigmatisering, kleinserig*. Die reël wat hierdie tipe tekseenhede hanteer, lui dat as 'n agtervoegsel voorafgegaan word deur 'n lettergreep met 'n sjwa (*-e, -el, -en, -em, -er, -es*), die gepaste lettergreep wat die sjwa-lettergreep voorafgaan, beklemtoon sal word. Dus sal die tekseenhede in die eerste voorbeeld korrek beklemtoon word, maar die tekseenhede in die tweede voorbeeld sal hanteer word asof dit 'n sjwa bevat en dus nie korrek beklemtoon word nie.
- Prefikse wat soms beklemtoon word en soms nie: Klemtoon word verkeerd toegeken aan woorde soos *onkruid, onvermoë, herkoms*, deur die klemtoon nie op die prefiks te plaas nie, omdat die tipe prefikse wat in dié kategorie val, eerder nie beklemtoon sal word nie, soos *onaktief, onbybels, oneg, herboor, herhaal, herinner*.

Die resultate in Tabel 5 kan nie vergelyk word met mekaar nie, omdat foute reggemaak is na elke toets. Dus gaan die toekenner weer op dieselfde 10 000 tekseenhede getoets word om finale resultate te verkry. Daarna sal die toekenner ook op verskillende tekste getoets word om resultate van die klemtoontoekenner op lopende teks te verkry.

4.3.2.1.1 Die vergelyking en evaluasie van die klemtoontoekeners

Die bestaande klemtoontoekenner en die klemtoontoekenner wat in hierdie studie ontwikkel is, is geëvalueer deur beide toekeners te toets op dieselfde stel van 10 000 tekseenhede. Om te verseker dat beide toekeners aan dieselfde standarde voldoen, is 'n evaluasie-woordelys opgestel, sodat die uitvoer outomaties getoets kon word om sodoende die resultate te verkry.

In Tabel 8 is die resultate van die Klemtoontoekenner wat in dié studie ontwikkel is (hierna genoem KTT 2009). Die getal foute en die akkuraatheid van die toekenner is na elke 1 000 tekseenhede verkry.

GETAL TEKSEENHEDE	GETAL FOUTE	AKKURAATHEID
1 000	105	89.50%
2 000	185	90.75%
3 000	244	91.87%
4 000	338	91.55%
5 000	414	91.72%
6 000	527	91.22%
7 000	617	91.19%
8 000	673	91.59%
9 000	736	91.82%
10 000	791	92.09%

Tabel 8: Resultate van KTT 2009 op 10 000 tekseenhede

In die bostaande tabel kan daar gesien word dat KTT 2009 'n finale akkuraatheid van 92% behaal het op 'n stel van 10 000 tekseenhede. Tabel 4 toon dat 45.6% van al die tekseenhede in die woordelys afleidings is. Verder is 28.3% van die woordelys komplekse, wat ook samestellende afleidings insluit. Dus kan daar gesê word dat die meerderheid van die tekseenhede afleidings is. In die vorige afdeling is die foute bespreek wat gemaak is deur KTT 2009 tydens die toetsproses.

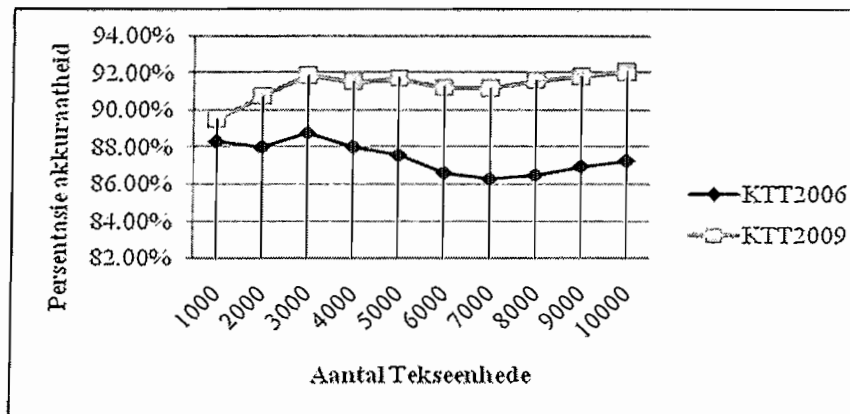
Al die foute is foute wat gemaak word op afleidings. Dus kan daar tot die gevolgtrekking gekom word dat KTT 2009 baie beter sal vaar op 'n woordelys waarvan die afleidings minder is.

In Tabel 9 word die resultate van die bestaande klemtoontoeckenner (hierna genoem KTT 2006) weergegee.

GETAL TEKSEENHEDE	GETAL FOUTE	AKKURAATHEID
1 000	117	88.30%
2 000	240	88.00%
3 000	337	88.77%
4 000	479	88.03%
5 000	622	87.56%
6 000	802	86.63%
7 000	959	86.30%
8 000	1079	86.51%
9 000	1174	86.96%
10 000	1274	87.26%

Tabel 9: Resultate van KTT 2006 op 10 000 tekseenhede

Met 'n finale akkuraatheid van 87.26% op dieselfde stel tekseenhede as KTT 2009, is dit duidelik dat KTT 2006 'n heelwat swakker toekenner is as KTT 2009. 'n Voordeel wat KTT 2006 bo KTT 2009 gehad het, is dat die toekenner 'n ingeboude uitsonderingslys het. Dit het veroorsaak dat 290 van die tekseenhede outomaties korrek beklemtoon is. In die onderstaande grafiek is 'n vergelyking getrek tussen KTT 2009 en KTT 2006. Daar kan ook gesien word dat die persentasie akkuraatheid van die twee klemtoontoeckenners op geen punt kruis nie. Dus is daar glad nie 'n vergelyking te tref tussen die prestasies van die twee toekenners nie. Daar kan dus tot die gevolgtrekking gekom word, dat KTT 2009 'n groot verbetering is op die bestaande KTT 2006.



Grafiek 1: Vergelyking van KTT 2006 en KTT 2009

4.3.2.1.2 Evaluasie van KTT 2009 op lopende teks

Die tekste wat gebruik is om dié evaluasie te doen, is die van die Verenigde Verklaring van Menseregte (hierna VVVM) en twee tekste uit die Bybel. Die tekste uit die Bybel is geneem uit die eerste vyf hoofstukke van die boeke Job en Openbaringe. Om die toets te doen, is daar van die VVVM en die Bybeltekste woordelyste gemaak, wat die frekwensie van elke woord in die teks aandui. Alle leestekens is verwyder uit die tekste. Verder is geen verandering gemaak aan die tekste nie. Die VVVM-tekste bestaan uit 1397 tekseenhede. As daar 'n woordelys van dié teks gemaak word, bestaan dit slegs uit 413 tekseenhede. Die Bybeltekste bestaan onderskeidelik uit 2252 en 2980 tekseenhede, waarvan 596 en 581 in woordelysvorm voorkom.

Die drie tekste is afsonderlik getoets. Die tipe teks wat in die VVVM gevind is, is baie dieselfde as die 10 000 woordelys waarmee die vorige toetse gedoen is. Dit is ook duidelik in die resultate wat behaal is deur die toekenner op woordvlak. Die toekenner se akkuraatheid op die Bybeltekste is weer heelwat hoër, omdat die tekseenhede wat in die Bybel voorkom, baie eenvoudiger is as dié van ander tekste.

In Tabel 10 kan die samestelling van die tekste beskou word:

TEKS	PERSENTASIE TEKSEENHEDE WAT ELKE LETTERGREEPKLAS VERTEENWOORDIG			
	Eenlettergrepig	Tweelettergrepig	Drielettergrepig	Vier- en meerlettergrepig
VVVM	26.5%	32.7%	22.3%	18.5%
Bybelteks 1 (Job	48.3%	38.3%	10%	3.4%
Bybelteks 2 (Op.	45%	38.5%	12.3%	4.2%

Tabel 10: Samestelling van tekste

Omtrent die helfte van die Bybeltekste bestaan uit eenlettergrepige tekseenhede. Dit is dus vanselfsprekend dat die akkuraatheid van die toekenner op die Bybeltekste soveel hoër as dié van die VVVM en ook, op woordvlak, as dié van die woordelys. Die VVVM gee dus 'n meer akkurate beeld van die akkuraatheid van KTT 2009, aangesien die samestelling daarvan meer neig na dié van die woordelys wat in die vorige toets gebruik is. Die akkuraatheid van die tekste op woordvlak is ook gemeet, om sodoende 'n meer objektiewe meting van die akkuraatheid van die klemtoontoekenner te verkry. In die onderstaande tabel word die resultate van die drie toetse weergegee:

TEKSBEKRYWING	GETAL TEKSEENHEDE IN TEKS	GETAL TEKSEENHEDE	AKKURAATHEID			
			Lopende Teks		Woordvlak	
			Getal Foute	Persentasie korrek	Getal Foute	Persentasie korrek
VVVM	1397	413	51	96.3%	30	92.7%
Bybelteks 1 (Job1-5)	2252	596	44	98.05%	28	95.3%
Bybelteks 2 (Op. 1-5)	2980	581	56	98.12%	33	94.3%

Tabel 11: Resultate van tekste

4.3.2.1.3 Vergelyking van KTT 2009 met ander klemtoontoekenners

Daar is verskeie klemtoontoekenners vir ander tale ontwikkel. Taylor (2005) het 'n Engelse klemtoontoekenner ontwikkel deur gebruik te maak van Versteekte Markovmodelle. Die model

ken klemtoon toe met 'n akkuraatheid van 76%. Nog 'n klemtoontoekenner is deur Church (1985) ontwikkel vir Engels. 'n Reëlgebaseerde benadering is gevolg en die toekenner het 'n akkuraatheid van 82% behaal. Ander statistiese metodes soos neurale netwerke en steunvektormasjiene is ook al gebruik vir die toekenning van klemtoon:

KLEMTOONTOKENNER	TAAL	BENADERING	GETAL	AKKURAATHEID
Taylor (2005)	Engels	Versteekte	±110 000 tekseenhede	76%
Arciuli en Thompson (2006)	Engels	Neurale Netwerk	10 134 tekseenhede	86%
Dou <i>et al.</i> (2009)	Engels	SVM	±60 000 tekseenhede	93.5%
Braga en Coelho (2008)	Portugees	Reëlgebaseerd	n.v.t.	99%
McPeters en Tharp (1983)	Engels	Reëlgebaseerd	n.v.t.	73%
Church (1985)	Engels	Reëlgebaseerd	n.v.t.	82%
KTT 2009	Afrikaans	Reëlgebaseerd	10 000 tekseenhede	92%

Tabel 12: Vergelyking van klemtoontoekenner vir verskillende tale

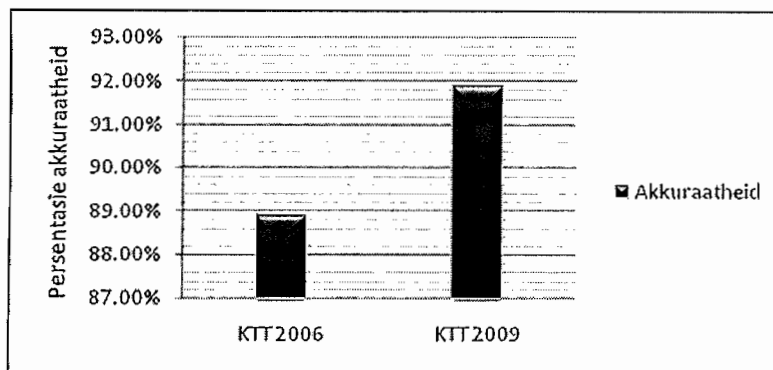
Die klemtoontoekenner vir Portugees vaar die beste van al die klemtoontoekenner wat hierbo genoem is. Vir dié taal was dit 'n goeie keuse om die reëlgebaseerde benadering as algoritme te kies, omdat Portugees 'n sillabe-tyd-taal (syllable-timed language) is (Bond, 2001). Dit beteken dat die sillabes van Portugees almal omtrent ewe lank uitgespreek word. Volgens Coetzee en Wissing (2007:266) het sillabe-tyd-tale 'n baie eenvoudige sillabestruktuur, waar klemtoontydte, soos Afrikaans en Engels, 'n komplekse sillabestruktuur het. Volgens Schutz (aangehaal deur Bond, 2001) is die klempatroon van Portugees baie voorspelbaar, in so mate dat 70% van alle Portugeese woorde klemtoon kry op die voorlaaste lettergreep. Dan word daar ook aksenttekens in die taal gebruik wat klemtoon aandui.

Die klemtoontoekenner wat in hierdie studie ontwikkel is, se akkuraatheid vergelyk baie goed met die klemtoontoekenner wat vir ander tale ontwikkel is. Die klemtoontoekenner wat deur Church (1985) ontwikkel is, het ook 'n reëlgebaseerde benadering gevolg en die reëls het ook gefokus op die gewig van sillabes en morfologie, net soos KTT 2009. Die klemtoontoekenner wat hier ontwikkel is, het dus baie goeie resultate behaal as dit vergelyk word met ander klemtoontoekenner.

4.3.2.2

Die evaluasie van die grafeem-na-foneem-omsetter

Die grafeem-na-foneem-omsetter is op dieselfde woordelys as die klemtoontoekenners getoets. Die eerste toets is gedoen met KTT 2006 as klemtoontoekenner. Die tweede toets is gedoen met KTT 2009 as klemtoontoekenner. Die omsetter is ook op dieselfde 10 000 tekseenhede as bovermelde toetse geëvalueer en die akkuraatheid is ook outomaties verkry, deur die uitvoer outomaties te vergelyk met die lys wat tydens die toetsproses ontwikkel is. Die onderstaande grafiek toon die akkuraatheid wat die grafeem-na-foneem-omsetter behaal het met die onderskeie klemtoontoekenners.



Grafiek 2: Resultate van die grafeem-na-foneem-omsetter

Die grafeem-na-foneem-omsetter het met 3 persentasie punte verbeter met die implementering van KTT 2009.

Die foute wat gemaak word deur die grafeem-na-foneem-omsetter, is meestal vokale wat verkeerd getranskribeer word. Die omsetter maak veral foute met die toekenning van die sjwa. Vroeër in die hoofstuk is die tipiese foute wat KTT 2009 maak, bespreek. Baie van die foute wat deur die grafeem-na-foneem-omsetter gemaak word, word veroorsaak deur die foute wat deur die klemtoontoekenner gemaak word.

Tabel 13 bevat voorbeelde van die foute wat deur die grafeem-na-foneem-omsetter gemaak is, as gevolg van die foute wat deur KTT 2009 gemaak is. As KTT 2009 nie foutiewe klemtoon aan die tekseenhede in Tabel 13 toegeken het nie, sou die grafeem-na-foneem-omsetter al die tekseenhede korrek getranskribeer het. In die volgende twee fases van die evaluasie gaan die faktore bespreek word wat klemtoontoekenning en grafeem-na-foneem-omsetting beïnvloed. Die

invloed van klemtoontoekenning op grafeem-na-foneem-omsetting sal in die vierde fase verder bespreek word.

TIPE FOUT	WOORD	INKORREK BEKLEEMTOON	INKORREK GETRANSKRIBEER
Woorduitgang <i>-ie</i>	simpatie	sim*;pa*tie	s@mpati
	fotografie	fo*to*;gra*fie	fotoxrafi
Die sjwa	kombers	;kom*bers	kOmb@rs
	sisteme	;sis*te*me	s@st@m@
	kermis	ker*;mis	k@rm@s
Prefiksfout	versus	ver*;sus	f@rs9s
	besig	be*;sig	b@s@x
	bewend	be*;wend	b@v@nt
Suffiksfout	kleinserig	;klein*se*rig	klQns@r@x
	stigmatisering	;stig*ma*ti*se*ring	st@xmAtis@r@N

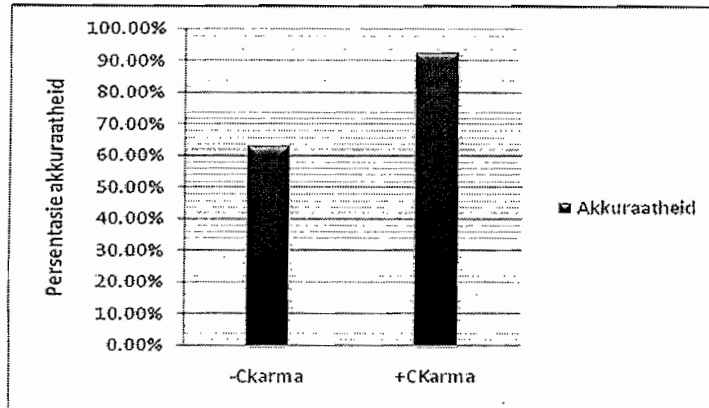
Tabel 13: Grafeem-na-foneem-omsetterfout

4.3.2.3 Die invloed van morfologiese analise op klemtoontoekenning

Die kompositumanaliseerder, CKarma, verdeel samestellings in woordeboekinskrywings. Dit is 'n baie belangrike komponent van die klemtoontoekenner. In dié fase word toetse gedoen om die belangrikheid van hierdie komponent vir klemtoontoekenning uit te stip.

Ten einde die toets te doen, is 'n lys van 1 000 tekseenhede ewekansig gekies uit die woordelys wat vir die vorige evaluering gebruik is. Om dié tekseenhede ewekansig te onttrek, is 'n program in Perl geskryf, wat dit outomaties kan onttrek. Die invoer vanaf CKarma is weggelaat uit die lys tekseenhede en klemtoon is aan die tekseenhede toegeken. Hierna is die akkuraatheid van die klemtoontoekenner outomaties verkry. Nadat die eerste toets gedoen is, is die invoer vanaf CKarma weer bygevoeg. Klemtoon is weer aan die tekseenhede toegeken en die akkuraatheid is

weer outomaties verkry. Die resultate wat verkry is uit die twee toetse kan in grafiek 3 beskou word.



Grafiek 3: Die invloed van morfologiese analise op klemtoon

Die eerste staaf dui die eerste toets aan waar morfologiese analise verwyder is uit die invoer. Dit het 'n akkuraatheid van 63.2% behaal. Die tweede staaf dui weer die tweede toets aan waar morfologiese analise weer bygevoeg is. Met slegs 72 foute wat gemaak is, het dié toets 92.8% akkuraatheid behaal. Die toekenner waarvan morfologiese analise bygevoeg is, is omtrent 30% meer akkuraat as die een waarvan die morfologiese analise weggelaat is.

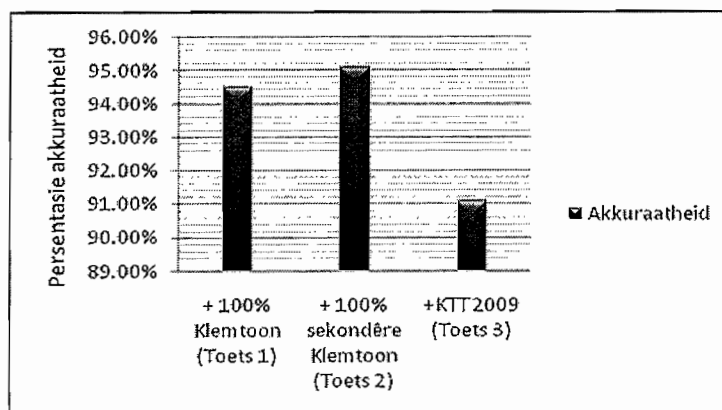
Morfologiese analise is baie belangrik vir KTT 2009 om 'n hoë akkuraatheid te behaal, omdat dit slegs primêre klemtoon aandui en dus slegs een lettergreep per woord beklemtoon. Indien 'n woord soos *koepletsonnet* beklemtoon word deur KTT 2009 en dit is nie verdeel deur die kompositumanaliseerder as *koeplet+sonnet* nie, sal dit slegs *koep*let*son*;net* as uitvoer gee, in plaas van *koep*;let*son*;net*. Die woord sal dan verkeerdelik getranskribeer as *kupl@tsOnEt*. Vir KTT 2009 is dit dus van kritieke belang dat morfologiese analise deel uitmaak van die klemtoontoekenproses.

4.3.2.4 Die invloed van klemtoontoekenning op grafeem-na-foneem-omsetting

In dié afdeling word die grafeem-na-foneem-omsetter geëvalueer, om vas te stel wat die invloed van die klemtoontoekener daarop is. 'n Nuwe lys van 1 000 tekseenhede is verkry op dieselfde manier as vir die vorige evaluasiefase. Daar is drie verskillende toetse uitgevoer in dié fase:

- Toets 1: Klemtoon is per hand toegeken aan die 1 000 tekseenhede, om te verseker dat dit 100% korrek is. Die klemtoontoekenner is uitgesluit uit die grafeem-na-foneem-omsetter, sodat dit geen veranderings kon aanbring nie aan die invoer wat reeds beklemtoon is. Daarna is grafeem-na-foneem-omsetting gedoen op die 1 000 tekseenhede. Die uitvoer is outomaties getoets om die akkuraatheid te bepaal.
- Toets 2: Sekondêre klemtoon is per hand toegeken aan dieselfde 1 000 tekseenhede, wat in Toets 1 gebruik is. Die klemtoontoekenner is weereens uitgesluit uit die grafeem-na-foneem-omsetter. Daarna is grafeem-na-foneem-omsetting gedoen op die 1 000 tekseenhede. Die akkuraatheid van die omsetter is outomaties bepaal.
- Toets 3: In dié toets is dieselfde 1 000 tekseenhede as vir die vorige twee toetse gebruik. Klemtoon is verwyder uit die 1 000 tekseenhede en die klemtoontoekenner is weer ingesluit by die grafeem-na-foneem-omsetter. Klemtoontoekenning en grafeem-na-foneem-omsetting is gedoen en die akkuraatheid van die grafeem-na-foneem-omsetting is weereens outomaties verkry.

Nadat die drie toetse gedoen is, is die resultate met mekaar vergelyk, om sodoende die invloed van klemtoontoekenning op grafeem-na-foneem-omsetting te bepaal. In die volgende grafiek word die resultate met mekaar vergelyk.



Grafiek 4: Die invloed van klemtoon op grafeem-na-foneem-omsetting

Die eerste staaf toon die resultate van Toets 1 aan. Met 100% toegeken klemtoon het die grafeem-na-foneem-omsetter 94.5% van al die tekseenhede korrek getranskribeer. Die resultate van Toets 1 is 3.4% hoër as dié van Toets 3. Dit is dus duidelik dat klemtoontoekenning 'n baie

groot rol in grafeem-na-foneem-omsetting speel. Hoe beter die klemtoontoekenning is, hoe beter vaar die grafeem-na-foneem-omsetter.

Met die byvoeging van sekondêre klemtoon, in Toets 2, het die akkuraatheid met slegs 0.6% gestyg. Vroeër in die hoofstuk is genoem dat KTT 2009 geen sekondêre klemtoon aan tekseenhede toeken nie, omdat sekondêre klemtoon in Afrikaans baie onreëlmatig is. Uit die resultate wat uit Toets 2 verkry is, is dit dus duidelik dat sekondêre klemtoon nie 'n baie groot rol speel in die akkuraatheid van die grafeem-na-foneem-omsetter nie en dit is dit dus nie noodsaaklik dat KTT 2009 sekondêre klemtoontoekenning moet doen nie. Indien sekondêre klemtoon bygevoeg word, is die kans baie goed dat die akkuraatheid van die toekenner kan daal, omdat die waarskynlikheid dan groter is dat die sekondêre en/of primêre klemtoon verkeerd toegeken is.

Die klemtoontoekenner kan moontlik met nog meer reëls verbeter word, maar dit sal nooit 'n akkuraatheid van 100% behaal nie. As 'n groter hoeveelheid toetsdata verkry kan word, kan die toekenner op meer tekseenhede getoets word en so kan die reëls dan nog verbeter word, om moontlik 'n hoër akkuraatheid te kry.

4.4 SAMEVATTING

In die hoofstuk is die toetsproses en die evaluasieproses beskryf. Die werking van die program waarbinne die klemtoontoekenner voorkom en die komponente waaruit die program bestaan, is ook ten volle bespreek.

In afdeling 4.3 is die toetsproses bespreek. Die klemtoonreëls wat in hoofstuk 2 gegeneer is, is in die program geïmplementeer en gebruik om klemtoon aan 10 000 tekseenhede toe te ken. Die lys tekseenhede is met die hand gekorrigeer om uiteindelik 'n lys van 10 000 tekseenhede te kon hê wat gebruik is vir evaluasiedoeleindes. Die toetsproses het bestaan uit 20 fases van 500 tekseenhede elk, bestaan, waartydens die klemtoontoekenner getoets is. Die uitvoer van elke fase is telkens met die hand gekorrigeer, en foute wat deur die klemtoontoekenner gemaak is, is telkens gekorrigeer waar moontlik.

Nadat die toetsproses voltooi is, is die klemtoontoekenner geëvalueer met dieselfde 10 000 tekseenhede, waar dit 'n akkuraatheid van 92.09% behaal het. 'n Vergelykende toets is gedoen om vas te stel of KTT 2009 'n verbetering is op sy voorganger, KTT 2006. Die toetse het bevestig dat

die nuwe klemtoontoekenner 4.8% beter vaar as KTT 2006 op 'n stel van 10 000 tekseenhede. Die nuwe klemtoontoekenner is ook geëvalueer op verskillende tekste, naamlik Die Verenigde Volke se Handves van Menseregte en twee verskillende tekste uit die 1983 vertaling van die Afrikaanse Bybel. Op die lopende teks van die Verenigde Volke se Handves van Menseregte het die toekenner 'n akkuraatheid van 96.3% behaal, wat op woordvlak 'n resultaat behaal het van 92.7%.

Die grafeem-na-foneem-omsetter, waarbinne die klemtoontoekenner werk, is ook geëvalueer. Dit het 'n akkuraatheid van 91.9% behaal. In 'n vergelykende toets is daar vasgestel dat die akkuraatheid van die omsetter heelwat verbeter het met die implementering van KTT 2009. Dit het 3% beter gevaar as die omsetter met KTT 2006.

Die invloed van morfologiese analise op die klemtoontoekenner is ook getoets. Uit die evaluasie blyk dit dat morfologiese analise 'n baie belangrike komponent in KTT 2009 is. Die klemtoontoekenner kan nie 'n hoë akkuraatheid behaal sonder dié komponent nie.

Toetse is ook gedoen om die invloed van klemtoontoekenning op grafeem-na-foneem-omsetting te bepaal. Eerstens is 'n toets gedoen deur aan 1 000 tekseenhede per hand 100% klemtoon toe te ken en dan as invoer aan die grafeem-na-foneem-omsetter te gee. Daarna is aan dieselfde tekseenhede ook sekondêre klemtoon toegeken. In die derde toets is dieselfde tekseenhede gebruik en is klemtoon daaraan toegeken deur KTT 2009. Dit is dan ook deur die grafeem-na-foneem-omsetter getranskribeer. Die resultate wat verkry is, het aangedui dat korrekte klemtoon baie belangrik is vir die akkuraatheid van die grafeem-na-foneem-omsetting. Sekondêre klemtoon het nie 'n groot verbetering in die grafeem-na-foneem-omsetting aangebring nie en dit is ook die rede waarom KTT 2009 nie gemoeid is met sekondêre klemtoon nie. Die grafeem-na-foneem-omsetter is 3.4% meer akkuraat met klemtoon wat 100% korrek is, as wanneer KTT 2009 dit toeken.

Die akkuraatheid van KTT 2009 is vergelyk met dié van klemtoontoekenners wat ontwikkel is vir ander tale. Om die akkuraatheid van die klemtoontoekenner nog te verbeter, sal dit nodig wees om meer tekseenhede te verkry, sodat die klemtoontoekenner met meer tekseenhede getoets kan word om die reëls verder te verbeter.

Hoofstuk 5

SLOT

5.1 SAMEVATTING

Teks-na-spraakstelsels is van groot belang vir die huidige samelewing. Dit word hoofsaaklik gebruik in toepassings wat mense met verskeie gebreke help om in die samelewing aan te pas, soos byvoorbeeld skermlesers vir mense met verswakte sig, leesprobleme, en ook spraakprobleme. Verder is die taaltegnologietoepassing ook al gebruik om geskrewe nuusberigte om te sit na 'n klankproduksie, soos *Ananova*. Een van die kerntegnologieë wat deel uitmaak van 'n teks-na-spraakstelsel, is grafeem-na-foneem-omsetting, wat dan ook klemtoontoekenning insluit. Klemtoontoekenning het in die afgelope paar jaar 'n oplewing gehad in die spraaktegnologie-wêreld, en is meestal ontwikkel om grafeem-na-foneem-omsetting te verbeter. In hoofstuk 1 is aangedui dat daar verskeie hulpbronne benodig word om 'n suksesvolle reëlgebaseerde klemtoontoekenner te ontwikkel, naamlik 'n stel klemtoonreëls vir Afrikaans, 'n algoritme, en 'n stel geannoteerde toetsdata. Die bestaande hulpbronne was nie voldoende vir die studie nie, en moes verbeter word. Dus is die volgende doelstellings vir die studie daargestel om hulpbronne te ontwikkel en te verbeter:

- a) om klemtoonreëls vir Afrikaans te ontwikkel wat linguisties akkuraat en volledig is en wat in 'n outomatiese klemtoontoekenner geïmplementeer kan word.
- b) om doeltreffende toetsdata te genereer waarmee 'n outomatiese klemtoontoekenner geëvalueer kan word.
- c) om 'n effektiewe en hoogs akkurate klemtoontoekenner te ontwikkel sodat Afrikaanse klemtoon vinnig en doeltreffend toegeken kan word.

In hoofstuk 2 is 'n stel klemtoonreëls vir Afrikaans ontwikkel. Besonderhede is verstrekk oor die verskillende metodes wat gevolg kan word om klemtoonreëls voor te stel. Drie voorstellings,

naamlik die liniêre klemtoonanalise, die metriese raamwerk en die metriese bome is bespreek en vergelyk. Die beste metode, naamlik metriese bome, is vir die doel gekies, omdat dit die moontlike voorkoms van onmoontlike klemtoonreëls baie beter hanteer as liniêre reëls. Dit is ook 'n baie eenvoudiger en meer leesbare voorstelling van klemtoonreëls as die liniêre reëls en die metriese raamwerk. In dié afdeling is ook vasgestel dat Afrikaans, soos Engels en Nederlands, kwantiteitgevoelig en gebonde is, en dat klemtoontoekenning dus beïnvloed word deur die innerlike voorkoms van sillabes.

Besonderhede oor die klemtoonreëls van Wissing (1971) en Combrink en De Stadler (1987) is verstrek, en die reëls van Van der Hulst (1984) en Trommelen en Zonneveld (1989) wat vir Nederlands ontwikkel is, is ook in diepte bespreek. Die faktore wat bydra tot klemtoontoekenning is die gewig van sillabes, oop en geslote lettergrepe, die sjwa, en ook die voorkoms van affikse in woorde.

Die gewig van sillabes is die belangrikste faktor wat in die studie gebruik is om klemtoon aan woorde toe te ken. In die hoofstuk is vasgestel dat gewig van die sillabe bepaal word deur die vokaal van 'n sillabe en die konsonante wat die sillabekern omring. Die sillabesterktehiërargie van Combrink en De Stadler (1987) en Van der Hulst (1984) is gebruik om die Afrikaanse klemtoonreëls te ontwikkel. Die manier waarop oop en geslote lettergrepe asook die posisie van die sjwa die klemtoon beïnvloed, is ook breedvoerig bespreek. Daar is ook in dié afdeling genoem dat die volgorde van reëls noodsaaklik is vir die doeltreffendheid van die klemtoonreëls.

Tydens die ontwikkeling van nuwe klemtoonreëls vir Afrikaans, is verskillende fases gevolg: In die eerste fase is die reëls vir woorde wat affikse bevat, ontwikkel. Daar is aangetoon dat affikse beklemtoonbaar of onbeklemtoonbaar kan wees, en dan is daar ook affikse wat in sommige gevalle beklemtoon word en in ander gevalle nie. Daar is onderskeid getref tussen prefikse en suffikse en die manier waarop die affikse klemtoon beïnvloed. In dié fase is daar ook reëls ontwikkel vir woorde wat 'n sjwa bevat. Die lettergreep net voor die sjwa-lettergreep word meestal beklemtoon, en daar is slegs sekere omstandighede waar dié reël nie geld nie. Hierna is reëls gegenereer wat klemtoon aan tweelettergrepige simplekse toeken. Die algemene reël vir tweelettergrepige simplekse stamme is dat dié woorde beklemtoon word op die eerste lettergreep, behalwe as die laaste lettergreep 'n superswaar sillabe bevat. In die derde fase is klemtoonreëls ontwikkel wat drielettergrepige simplekse stamme hanteer, en in die laaste fase is die reëls vir vier- en meerlettergrepige simplekse stamme behandel.

In hoofstuk 3 is die verskillende benaderings wat al gevolg is om klemtoon toe te ken, bespreek. Nadat 'n duidelike oorsig gegee is van die verskillende benaderings, is dit duidelik gemaak waarom die reëlgebaseerde benadering gekies is vir die doeleindes van die projek. Die benadering voldoen ook aan al die vereistes hierbo genoem en is dus geskik vir Afrikaanse klemtoontoekenning. Die benadering is al verskeie kere vir die taak van klemtoontoekenning vir verskillende tale ingespan en het telkens goeie resultate behaal. Die benadering het dus 'n hoë akkuraatheidsyfer en die programmeertaal Perl wat in die studie gebruik word is maklik aanleerbaar. Die program is ook gebruikersvriendelik met 'n lae hulpbronverbruik. Die program is ook herbruikbaar deurdat dit gebruik kan word om geannoteerde data te verskaf vir ander toepassings wat klemtoontoekenning en getranskribeerde teks benodig.

Nadat die benadering tot klemtoontoekenning gekies is, is die klemtoonreëls wat ontwikkel is, geïmplementeer. Hierdie proses is breedvoerig in hoofstuk 4 bespreek. Die argitektuur van die program en die werking van elk van die modules van die program is verduidelik. Vervolgens is die toetsproses beskryf. Die toetsproses het uit twintig fases bestaan. Tydens elke fase is die program met 500 tekseenhede getoets, waarna die uitvoer met die hand gekontroleer is en foute wat deur die program gemaak is, is reggemaak deur reëls weg te neem, by te voeg, of die volgorde van reëls in die program te verander. Nadat al twintig fases uitgevoer is, was daar 'n woordelys bestaande uit 10 000 korrek-geannoteerde tekseenhede. Hierna is die evaluasie van die klemtoontoekenner gedoen. Dit het uit vier fases bestaan: Eerstens is die woordelys wat in die toetsdata gegenereer is gebruik om die ou en die nuwe klemtoontoekenner met mekaar te vergelyk. Die KTT2009 het met 'n akkuraatheid van 92% byna 5% beter gevaar as KTT2006. Uit dié resultate is dit duidelik dat die nuwe klemtoonreëls 'n groot verbetering is op die klemtoontoekenner. Die toekenner is ook op lopende teks getoets en het 'n akkuraatheid van 96.3% op die VVVM teks behaal en 98% behaal op Bybeltekste. Die hoë persentasie wat behaal is op die Bybeltekste is te danke aan die feit dat die teks meestal uit een- en tweelettergreppige woorde bestaan. Nadat dié toetse gedoen is, is die akkuraatheid van die grafeem-na-foneem-omsetter bepaal en nog 'n vergelykende evaluasie is gedoen om te sien of die akkuraatheid verbeter het met die gebruik van die nuwe klemtoontoekenner. Die grafeem-na-foneem-omsetter het 'n akkuraatheid van 91.9% behaal met KTT2009, wat 3% hoër is as die akkuraatheid van die omsetter met KTT2006. In die derde fase is die invloed van morfologiese analise op klemtoontoekenning bepaal, en laastens is die invloed van klemtoontoekenning op grafeem-na-foneem-omsetting bepaal. Daar is gevind dat die faktore 'n groot bydrae lewer tot die

akkuraatheid van die omsetter, en aan die hand van die toetse is daar ook gewys waarom sekondêre klemtoon nie in die program toegeken word nie.

5.2 *GEVOLGTREKKING*

Die toekenner voldoen aan die vereistes wat gestel is, deurdat dit 'n goeie akkuraatheid behaal het, die program is maklik om mee te werk, die taal waarin dit geskryf het is eenvoudig om aan te leer, en die hulpbrongebruik van die toekenner is laag. Die toekenner is herbruikbaar deurdat dit nuwe afrigtingsdata vir volgende projekte maklik, vinnig en akkuraat kan genereer.

Tydens die toetsproses is die foute wat gemaak word deur die klemtoontoekenner uitgewys en is daar gepoog om die reëls te verbeter en is daar wel 'n paar veranderinge aangebring aan die reëls: Reël 21 en 26 wat drie- en meerlettergrepige tekseenhede hanteer wat eindig in die vorm VV-VK, is uit die klemtoonreëlstel verwyder, omdat die reëls te veel uitsonderings gehad het. Die vokaalhiërargie (e>o>i>E>y>u>O>a>9) is gebruik in die plek van die reëls wat uitgehaal is.

Die tipe foute wat die klemtoontoekenner maak, word breedvoerig in hoofstuk 4 bespreek, maar 'n veralgemening kan gemaak word deur te sê dat die toekenner hoofsaaklik foute maak met woorde wat affikse bevat of wat lettergrepe bevat wat dieselfde vorm as 'n affiks aanneem, soos byvoorbeeld *verwers* en *verwerf*. Beide woorde begin met 'n lettergreep *ver-*, wat dit onmoontlik maak vir die klemtoontoekenner om te bepaal of dit 'n affiks is of nie.

Tydens die evaluering van die klemtoontoekenner is gevind dat die klemtoonreëls wat in die studie ontwikkel is effektief is vir die toekenning van Afrikaanse klemtoon. As die toekenner in 'n latere stadium met nuwe data getoets word, mag nuwe foute moontlik na vore kom wat kan veroorsaak dat verdere wysigings van die reëls nodig mag wees.

Die benadering wat vir die doeleindes van die projek gekies is, naamlik die reëlgebaseerde benadering, het baie goed gegaan met die toekenning van klemtoon. Die programmeertaal Perl beskik oor die vermoë om teks te manipuleer deur middel van reëlmatige uitdrukkings, wat dit baie maklik gemaak het om die klemtoonreëls te implementeer in die program. Die reëlgebaseerde benadering wat in dié studie vir klemtoontoekenning gebruik word, blyk suksesvol te wees as daar gekyk word na die resultate wat dit opgelewer het tydens die evaluasieproses. Die resultate wat opgelewer is, is ook baie vergelykbaar met die klemtoontoekenning wat vir ander tale ontwikkel is. Deur 'n reëlgebaseerde benadering te volg,

is ook vasgestel dat klemtoon in Afrikaans nie van so aard onreëlmatig is dat dit nie deur 'n stel reëls bepaal kan word nie.

Een groot nadeel van die reëlgebaseerde benadering is dat die toekenner taalspesifiek is, en alhoewel dit gewysig kan word vir ander tale, sal die wysiging 'n groot taak wees. Dus voldoen dit nie ten volle aan die laaste vereiste, naamlik aanpasbaarheid nie. Die klemtoonstruktuur van Afrikaans en Nederlands het nie werklik groot verskille nie (Van Huyssteen en Pilon, 2009). Die reëls wat Van der Hulst (1984) en Trommelen en Zonneveld (1989) vir Nederlands ontwikkel het, het sonder om veel veranderings aan te bring vir Afrikaans ook gewerk, dus sal die klemtoontoekenner maklik gewysig kan word vir Nederlands.

5.3 AANBEVELINGS

Aan die einde van die studie kan die volgende aanbeveling gemaak word om 'n klemtoontoekenner vir Afrikaans te verbeter:

- 'n Etiketteerder wat affikse annoteer kan 'n goeie toevoeging tot 'n klemtoontoekenner wees. Die etiketteerder moet in staat wees om affikse te annoteer in terme van hul invloed op klemtoon, met ander woorde of dit klemdraend, nie-klemdraend, of klemtoonaantrekkend is. Die etiketteerder moet verkieslik 'n eksterne module wees, sodat foute gekorrigeer kan word voordat klemtoon aan die woorde toegeken word. Die annotering van affikse is nodig om te voorkom dat konstituente wat die vorm van 'n affiks aanneem, maar nie 'n affiks is nie, foutiewelik herken word as 'n affiks, soos in die voorbeelde *verwers* en *verwerf*.
- Dit is ook van belang vir klemtoontoekenning om te weet onder watter grammatiese kategorie 'n spesifieke woord val. Deur 'n woordsoortetiketteringsmodule by te voeg kan die kanse dat 'n woord soos *gésel* verkeerdlik as *gesél* beklemtoon word, uitgesluit word. Nog 'n voorbeeld van dié probleem is die woorde *onderwérp* en *ónderwerp*. Die tipe module sal ook moontlik die probleem met afleidings wat verkeerd beklemtoon word kan oplos, wat dan die behoefte aan 'n affiks-etiketteerder kan uitskakel. Aangesien daar reeds 'n woordsoortetiketteerder vir Afrikaans bestaan wat deur Pilon (2005) ontwikkel is, kan die aanbeveling met min moeite getoets word voordat 'n affiks-etiketteerder vir die taak ontwikkel word.
- Die afgelope twee dekades is masjienleertegnieke in die menslike-taaltegnologie studieveld al hoe meer gebruik. Die grootste probleem met dié benadering is dat dit groot stalle afrigtingsdata benodig wat nie altyd beskikbaar is nie. Met die ontwikkeling van

dié reëlgebaseerde klemtoontoeckenner sal die generering van afrigtingsdata vir 'n datagedrewe klemtoontoeckenner en grafeem-na-foneem-omsetter vir Afrikaans vinnig en maklik wees. Die eksterne modules, CKarma en Calomo, het beide 'n hoë akkuraatheid, dus sal die korrigering van die data wat van dié modules af kom waarskynlik nie nodig wees met die ontwikkeling van nuwe afrigtingsdata nie, omdat die uitvoer vanaf die klemtoontoeckenner gekorrigeer moet word voordat dit as afrigtingsdata gebruik kan word. Die bogenoemde aanbevelings moet ook in gedagte gehou word met die ontwikkeling van 'n datagedrewe klemtoontoeckenner. Die klemtoontoeckenner wat in dié studie ontwikkel is, het 'n hoë akkuraatheid behaal. Dus is dit moontlik dat 'n datagedrewe benadering met 'n groter hoeveelheid afrigtingsdata maklik ontwikkel kan word tot 'n hoogs-akkurate klemtoontoeckenner. In die studie is vasgestel dat Afrikaanse klemtoon met 'n reëlgebaseerde stelsel 'n baie hoë akkuraatheid kan behaal. Dus bestaan die moontlikheid om die beste van die twee wêreldes te kombineer en dus 'n hibriediese stelsel te bou, wat dan waarskynlik die beste metode vir Afrikaanse klemtoontoeckenning kan wees.

In die studie is daar ook 'n paar probleme ondervind met die grafeem-na-foneem-omsetter waarbinne die klemtoontoeckenner ontwikkel is. Die probleme met hul moontlike oplossings word vervolgens bespreek:

- In hoofstuk 4 is genoem dat die laaste module waardeur die tekseenhede gaan voordat dit uitgevoer word, die tekseenhede afrond, sodat daar nie twee van dieselfde foneme langs mekaar sal voorkom in 'n woord nie, soos byvoorbeeld die woord *volkklanke* sal getranskribeer word as *fOIkklANk@*. Binne die laaste module sal die transkripsie verander word na *fOIkLANk@*. Alhoewel dit die korrekte fonetiese transkripsie van die woord is, kan dit nou moontlik verkeerdelik uitgespreek word as *vol-klanke*. Dit is dus nodig dat die transkripsie van so aard verander word dat dit alle dubbelsinnigheid kan uitskakel, soos om alle samestellings te skei met 'n onderstreep, byvoorbeeld *fOIk_kLANk@*.
- Die grafeem-na-foneem-omsetter is ontwikkel om 'n woord slegs in sy kanoniese vorm te transkribeer, wat veroorsaak dat woorde nie altyd foneties getranskribeer word nie. Om die probleem op te los is dit waarskynlik nodig om 'n foneem-na-foon-omsetter te bou om die kanoniese vorms om te sit na die gebruiklike vorms van woorde. In die studie is die akkuraatheid van die omsetter op eiename nie getoets nie, maar so module sal moontlik

baie noodsaaklik wees vir die fonetiese transkripsie van eiename. Die vraag wat hieruit kan ontstaan is: Wat is korrekte fonetiese transkripsie? 'n Woord soos *professor* word kanonies getranskribeer as *profEsOr* en die fonetiese transkripsie kan dan wees *prufEsOr* of *prufEs@r*. 'n Woord soos *praat* kan ook verskillend uitgespreek word, soos byvoorbeeld *prO:t* of *prat*. 'n Moontlike oplossing vir die verskynsel is om die verskillende uitsprake streeksgebonde te maak en dus die uitspraak te verander na gelang van die streek waar dit gebruik sal word.

Ter afsluiting: Die studie het 'n suksesvolle en akkurate klemtoontoekeenner vir Afrikaans opgelewer, met bevredigende resultate wat die akkuraatheid van die grafeem-na-foneem-omsetter waarbinne dit gebou is, heelwat verbeter het. Dit kan 'n bydrae lewer tot die mensetaaltegnologie-studieveld, deurdat nuwe afrigtingsdata vir 'n datagedrewe klemtoontoekeenner of grafeem-na-foneem-omsetter vir Afrikaans met min moeite gegenereer kan word, wat dan 'n groot bydrae kan lewer tot Afrikaanse teks-na-spraak-stelsels.

BIBLIOGRAFIE

- Alkoffash, M.S., Bawaneh, M.J., & Al Rabea, A.I. 2008. Which Software Cost Estimation Model to Choose in a Particular Project. *Journal of Computer Science*, 4(7):606-612.
- Arciuli, J. & Thompson, J. 2006. Improving the assignment of lexical stress in text-to-speech systems. (In *Proceedings of the 11th Australian International Conference on Speech Science & Technology*. Auckland, December p.296-300)
- Barros, M.J. & Weiss, C. 2006. Maximum Entropy motivated Grapheme-to-Phoneme, Stress and Syllable boudary prediction for Portuguese Text-to-Speech. IV *Jornadas en Tecnologia del Habla*. 177 – 182.
- Bond, K. 2001. Karen's Linguistic Issues : *Free Resources for theachers and students of English*. [Web:] <http://www3.telus.net/linguisticissues/pronunciation.html> [Datum van gebruik: 20 Okt. 2009].
- Braga, D. & Coelho, L. 2008. Automatic Word Stress Marker for Portuguese TTS. V *Jornadas en Tecnologia del Habla*, 179-182, 2008.
- Chomsky, N. & Halle, M. 1968. *The Sound Pattern of English*. New York : Harper & Row.
- Church, K. 1985. Stress Assignment in Letter to Sound Rules for Speech Synthesis. (In *Proceedings of the 23rd annual meeting on Association for Computational Linguistics*. Chicago p.246-253).
- Coetzee, A.W. & Wissing, D.P. 2007. Global and local durational properties in three varieties of South African English. *Linguistic Review*, 24 : 263-289.
- Combrink, J.G.H. & De Stadler, J.G. 1987. *Afrikaanse Fonologie*. Johannesburg : Macmillan.
- Combrink, J.G.H. 1990. *Afrikaanse Morfologie*. Kaapstad : Academica.
- Daelemans, W. 1988. Grafon : A Grapheme-to-Phoneme Conversion System for Dutch. (In *Proceedings Twelfth International Conference on Computational Linguistics*, p.133-138).

-
-
- Daelemans, W., Gillis, S., Durieux, G. & Van Den Bosch, A. 1993. Learnability and Markedness in Data-Driven Acquisition of Stress. (In Ellison & Scobbie, eds. *Computational Phonology*. Edinburgh: Centre of Cognitive Study, University of Edinburgh. p. 157-178).
- Daelemans, W. & Van Den Bosch, A. 2005. *Memory-Based Language Processing*. Cambridge : Cambridge University Press.
- Davel, M.H. 2005. *Pronunciation Modelling and Bootstrapping*. (Proefskrif – D.Phil.) Pretoria: UP.
- De Villiers, M. 1965. *Afrikaanse Klankleer : Fonetiek, Fonologie en Woordbou*. Kaapstad: Gothic.
- Demberg, V., Schmid, H., & Mohler, G. 2007. Phonological Constraints and Morphological Preprocessing for Grapheme-to-Phoneme Conversion. (In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Prague, Junie p.96-103).
- Dou, Q., Bergsma, S., Jiampojarn, S. & Kondrak, G. 2009. A Ranking Approach to Stress Prediction to Letter-to-Phoneme Conversion. (In *Proceedings of the 47th Annual Meeting of the ACL*. Singapore, Augustus p.118 – 126).
- Fletcher, T. 2000. *A very basic SVM tutorial for complete beginners*. [Web:] <http://www.tristanfletcher.co.uk/SVM%20Explained.pdf> [Datum van gebruik: 3 Apr. 2009.]
- Freeman-Hargis, J. 2002. *Rule-based systems and identification trees*. [Web:] <http://ai-depot.com/Tutorial/RuleBased.html> [Datum van gebruik: 3 April. 2009]
- Gallant, S.I. 1993. *Neural Network Learning and Expert Systems*. Cambridge: MIT.
- Gillis, S., Daelemans, W., Durieux, G., & Van Den Bosch, A. 1993. Learnability and Markedness: Dutch Stress Assignment. (In *Proceedings 15th conference of the cognitive science society*, Belgium, 1993. 452-457).

-
- Gillis, S., Daelemans, W., Durieux, G. 2000. 'Lazy Learning': A Comparison of Natural and Machine Learning of Word Stress. (In Broeder, P & Murre, J., eds. *Models of Language*, Oxford: Oxford University Press P.76 – 99.)
- Halle, M. & Keyser, S.J. 1971. *English Stress : Its Form, Its Growth, and Its Role in Verse*. New York : Harper & Row.
- Halle, M. & Vergnaud, J. 1987. *An Essay on Stress*. Cambridge : MIT.
- Han, J & Kamber, M. 2006. *Data Mining : Concepts and Techniques*. San Diego : Elsevier. 703p
- International Standardization Organization. 1991. *Software Engineering – Product Quality*. ISO 9126. [Web:] <http://www.issco.unige.ch/projects/ewg96/node13.html> [Datum gebruik: 2008-09-03].
- Jenkin, K.L. & Scordilis, M.S. 1994. *Automatic methods for syllable stress classification in continuous speech*. [Web:] <http://www.assta.org/sst/SST-94-Vol-11/cache/SST-94-VOL2-Chapter12-p19.pdf> [Datum van gebruik: 6 Apr. 2009].
- Joachims, T. 2002. *Learning to Classify Text using Support Vector Machines*. Dordrecht : Kluwer Academic
- Kempen, W. 1969. *Samestelling, Afleiding en Woordsoortelike Meerfunksionaliteit in Afrikaans*. Kaapstad : Nasou.
- Klatt, D.G. 1987. Review of Text-to-Speech Conversion for English. *Journal of Acoustical Society of America, September*. 82 (3):737-793.
- Leondes, C.T. (Red) 2000. *Knowledge-based systems: techniques and applications*. San Diego: Academic Press. 1449p.
- Ligeza, A. 2006. *Logical Foundations of Rule-based Systems*. Berlin : Springer. 309p.
- Lubbe, H.J. 1993. Klem in Afrikaans en die superswaar rym. *South African Journal of Linguistics*, November. 18 : 78-99.
-

-
- Marincic, D., Tusar, T., Gams, M. & Sef, T. 2009. Analysis of Automatic Stress Assignment in Slovene. *Informatica*, 20 (1) : 35-50.
- Mccarthy, J.J. 1979. On Stress and Syllabification. *Linguistic Inquiry Summer*.10(3): 443-465.
- McPeters, D.L. & Tharp, A.L. 1983. Application of the Liberman-Prince Stress Rules to Computer Synthesized Speech. (In *Proceedings of the first conference on Applied natural language processing*. p.192-197).
- Mitchell, T. 1997. *Machine Learning*. London : McGraw-Hill.
- Rabiner, L.R. 1989. A Tutorial on Hidden Markov Models and Selected Applications on Speech Recognition. (In *Proceedings of the IEEE*, 77(2) : 257-286.)
- Pilon, S. 2005. *Outomatiese Afrikaanse Woordsoortettketering*. (Verhandeling – MA). Potchefstroom: Noordwes-Universiteit
- Ratnaparkhi, A. 1997. *A Simple Introduction to Maximum Entropy Models for Natural Language Processing*. (Technical report). Philladelphia: University of Pennsylvania
- Reichel, U., Pfitzinger, R. & Hain, H.U. 2008. English grapheme-to-phoneme conversion and evaluation. *Speech and Language Technology*, 11:159-166.
- Rojas, R. 1996. *Neural Networks : a systematic introduction*. Berlin : Springer..
- Sef, T., Skrjanc, M. & Gams, M. 2002. Automatic Lexical Stress Assignment of Unknown Words of Highly Inflected Slovenian Language. (In SOJKA, P., KOPECEK, I. & PALA, K., eds. *Text, speech, and dialogue* : 5th International Conference. Berlin : Springer. 165-172.
- Taylor, P. 2005. *Hidden Markov Models for Grapheme to Phoneme Conversion*. [Web:] http://mi.eng.cam.ac.uk/~pat40/eurospeech05_form04.pdf [Datum van gebruik: 26 Febr 2007].
- Perl 2009. The Perl Directory, *About Perl*. [Web:] <http://www.perl.org/about.html> [Datum van gebruik: 5 Sept. 2009].
-

-
-
- Trommelen, M. & Zonneveld, W. 1989. *Klemtoon en Metrische Fonologie*. Muiderberg : Dick Coutinho.
- Turner, J. 2008. *You Used Perl to Write What?* [Web:]
http://www.cio.com/article/175450/You_Used_Perl_to_Write_WHAT [Datum van gebruik: 5 Sept. 2009].
- Valsan, Z., Gavat, I. & Sabac, B. 2002. Statistical and Hybrid Models for Speech Recognition in Romanian. *International Journal of Speech Technology*, July. 5 : 259-268.
- Van Den Bosch, A. & Daelemans, W. 1993. Data-Oriented Methods for Grapheme-to-Phoneme Conversion. *Proceedings of EACL*, August. 45-53.
- Van Der Hulst, H. 1984. *Syllable Structure and Stress in Dutch*. Dordrecht : Foris. 260 p.
- Van Huyssteen, G.B. & Pilon, S. 2009. Rule-based Conversion of Closely-related Languages: A Dutch-to-Afrikaans Converter. (In *Proceedings of the 2006 Conference of the Pattern Recognition Association of South Africa*. Stellenbosch).
- Vapnik, V.N. 2000. *The Nature of Statistical Learning Theory*, 2nd ed. New York: Springer.
- Warakagoda, N. 1996. Hidden Markov Models. [Web:]
<http://jedlik.phy.bme.hu/~gerjanos/HMM/node2.html> [Datum van gebruik: 3 Apr. 2009].
- Wissing, D.P. 1971. *Fonologie en Morfologie van die Simplekse Selfstandige Naamwoord in Afrikaans: 'n Transformasioneel-Generatiewe Beskrywing*. Amsterdam : Buijten & Schipperheijn.

ADDENDUM A

KLEMTTOONREËLS VIR AFRIKAANS

I. TEGNIESE NOTASIE SOOS GEBRUIK IN REËLS

<i>TEGNIесе NOTASIE</i>	<i>BESKRYWING</i>
S	Sterk konstituent
W	Swak konstituent
V	Enige vokaal
K	Enige konsonant
ə	’n Sjwa
VV	’n Oop lettergreep
VK	’n Geslote lettergreep
X	Enige letter / groep letters
K ₀	Nul of meer konsonante
Z _{Sf}	’n Suffiks
Z _{Pf}	’n Prefiks
Z _D	’n Diftong
VVK/VKK	Superswaar syllables met die vorm VVK of VKK.

Tabel 14: Tegniese notasie wat gebruik word in die klemtoonreëls.

II. GEWYSIGDE SAMPA FOONSTEL

SAMPA SIMBOOL	SLEUTELWOORD IN SAMPA	ORTOGRAFIES
KONSONANTE		
Plosiewe		
p	pQp	pyp
b	bAl	bal
t	tOt	tot
d	dAk	dak
k	kOk	kok
g	gOlf	gholf
Frikatiewe		
f	fEl	vel, fel
v	wEt	wet
s	s9s	sus
z	zulu	Zulu
S	Sampu	sjampoe
Z	Zanr@	genre
x	xEk	gek
h	hOk	hok
C	CEk	tjek
Nasale		
m	mEns	mens
n	nat	naat
N	ANs	angs

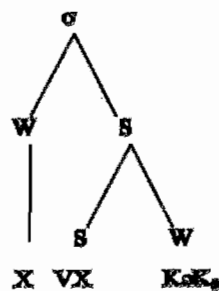
Likwide		
l	lEl	lel
r	rOk	rok
Half-vokale		
j	jAs	jas
w	twe	twee
VOKALE		
i	sin	sien
e	skre	skree
a	mat	maat
u	mut	moet
o	rok	rook
y	fur	vuur
2	l2n	leun
E	mEt	met
A	rAt	rat
O	mOt	mot
@	v@t	wit
9	n9l	nul
Primêre diftonge		
Q	pQl	pyl, peil
Y	rYt	ruit
U	sÛt	sout

Tabel 15: Gewysigde SAMPA foonstel

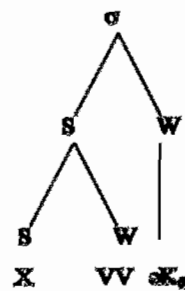
III. REËLS VIR WOORDE WAT 'n SJWA BEVAT (Sien 2.3.2.3 bl.27) :

Reël 1: a) As die sjwa-lettergreep voorafgegaan word deur minstens een konsonant val die klemtoon op die voorlaaste lettergreep, byvoorbeeld *papawer*, *pantoffel*, maar b) as die sjwa voorafgegaan word deur 'n vokaal, sal die derdelaaste lettergreep beklemtoon word, byvoorbeeld *terriër*, *België*:

a)

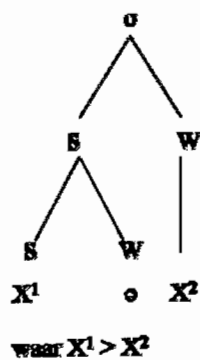


b)

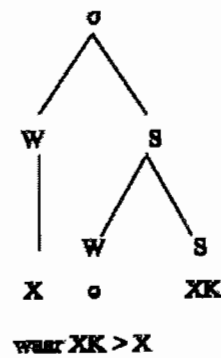


Reël 3: As 'n drielettergrepige woord 'n sjwa in die tweede lettergreep bevat, sal daar volgens die vokaalhërargie bepaal word watter een van die laaste of eerste lettergreep beklemtoon word, byvoorbeeld *hottentot*, *algebra*, *harlekyn*, *atletiek*:

a)

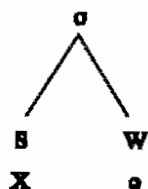


b)

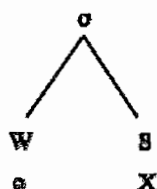


Reël 4: Die lettergreep wat 'n sjwa bevat in 'n tweelettergrepige woord, sal nie beklemtoon word nie, byvoorbeeld *masker*, *mekaar*:

a)



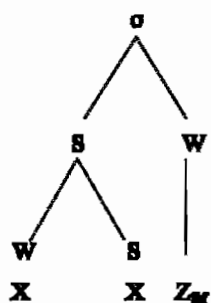
b)



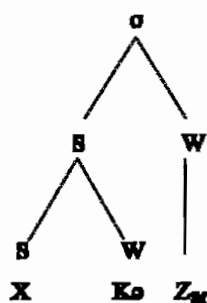
IV. REËLS VIR AFLEIDINGS (Sien 2.3.2.4 bl.30):

Reël 5: a) Indien 'n suffiks klemtoonaantrekkend is en nie voorafgegaan word deur 'n lettergreep wat 'n sjwa bevat nie, sal die voorafgaande lettergreep beklemtoon word byvoorbeeld *atleties*, *bedagsaam*, b) anders sal die derdelaste lettergreep beklemtoon word, byvoorbeeld *radeloos*, *waarneembaar*:

a)

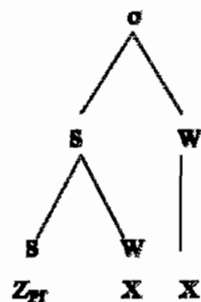


b)



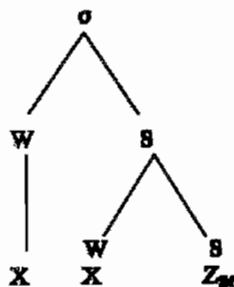
waar Z_{af} = -ik, -ig, -saam, -loos, -ies, -end, -baar.

Reël 6: Indien 'n prefiks beklemtoonbaar is, sal daardie prefiks beklemtoon word, byvoorbeeld *mis*verstand, *aarts*vyand, *wan*orde:



waar Z_{Pr} = *on-*, *a-*, *mis-*,
wan-, *aarts-*

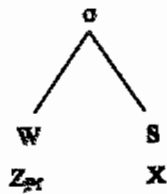
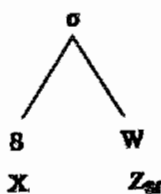
Reël 7: Indien 'n suffiks klemdraend is, sal daardie suffiks beklemtoon word, byvoorbeeld *dramaturg*, *grotesk*:



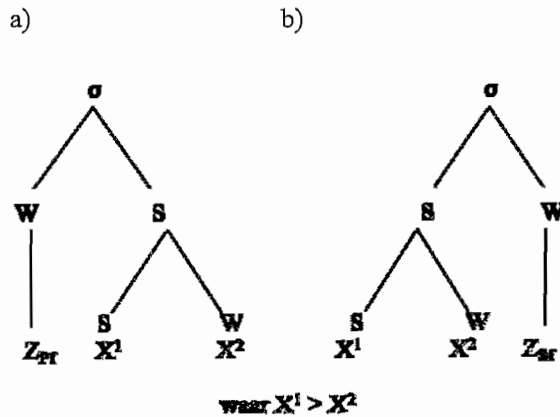
Reël 8: Enige onbeklemtoonbare affiks sal nie beklemtoon word nie, byvoorbeeld *een*heid, *sen*ding, *ver*dryf, *ges*kat:

a)

b)



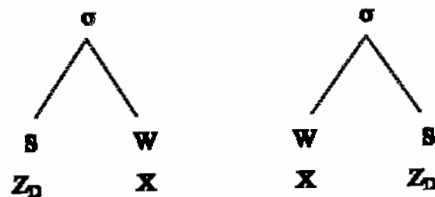
Reël 9: In drie- of meerlettergrepe woorde wat onbeklemtoonbare affikse bevat, sal volgens die vokaalhiërargie bepaal word watter van die lettergrepe wat nie affikse is nie, beklemtoon sal word byvoorbeeld, *verafrikaans*, *direkteurskap*:



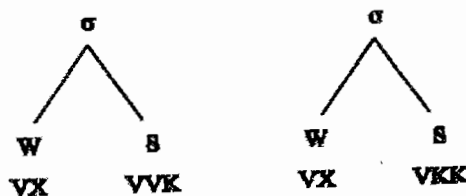
V. REËLS VIR AFRIKAANSE SIMPLEKSE:

i. Tweelettergrepe simplekse (Sien 2.3.3.1 bl.35):

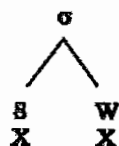
Reël 10: Enige lettergreep wat 'n diftong bevat sal beklemtoon word, byvoorbeeld *asgaai*, *toernooi*:



Reël 11: Enige eindlettergreep met die vorm VVK of VKK sal beklemtoon word, byvoorbeeld *kameel*, *bandiet*, *triomf*, *miljard*:

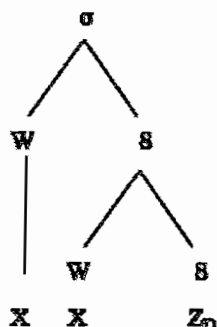


Reël 16: Aan 'n tweelettergrepige woord wat nie deur bogenoemde reëls bevredig is nie, sal klemtoon toegeken word volgens die vokaalhiërargie: (e>o>i>E>y>u>O>a>9), anders sal dit beklemtoon word op die eerste lettergreep:

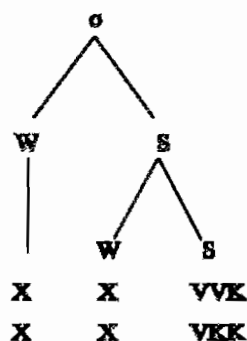


ii. *Drielettergrepige simplekse (Sien 2.3.3.2 bl.38):*

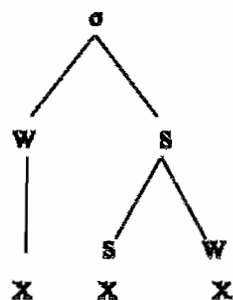
Reël 17: Enige lettergreep wat 'n diftong bevat sal beklemtoon word:



Reël 18: Enige eindlettergreep met die vorm VVK of VKK sal beklemtoon word, byvoorbeeld *pelikaan*, *predikant*:

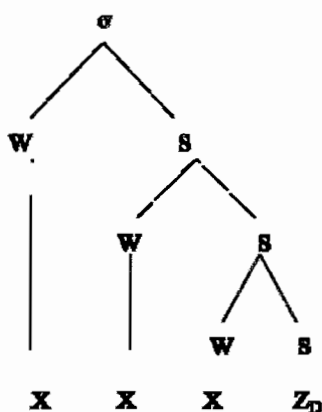


Reël 22: Aan 'n drielettergrepige woord wat nie deur bogenoemde reëls bevredig is nie, sal klemtoon toegeken word volgens die vokaalhiërargie: (e>o>i>E>y>u>O>a>9), anders sal dit beklemtoon word op die tweedelaaste lettergreep:

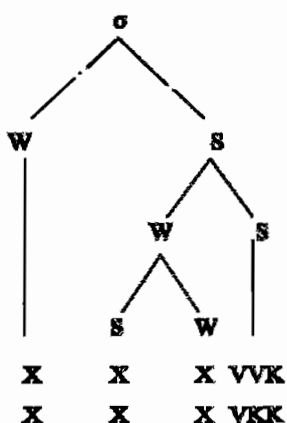


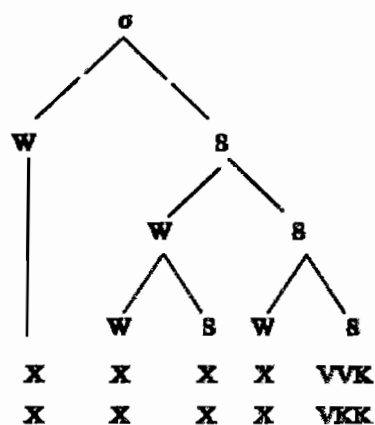
iii. Vier- en meerlettergrepige simplekse (Sien 2.3.3.3 bl.41):

Reël 23: Enige lettergreep wat 'n diftong bevat sal beklemtoon word:



Reëls 24 en 25: Enige eindlettergreep met die vorm VVK of VKK sal beklemtoon word, byvoorbeeld *eksperiment*, *universiteit*:





Reël 27: 'n Vier- of meerlettergreepige woord wat nie deur bogenoemde reëls bevredig is nie, sal aan klemtoon toegeken word volgens die vokaalhiërargie: (e>o>i>E>y>u>O>a>9), anders sal dit beklemtoon word op die tweedelaaste lettergreep, byvoorbeeld *diafragma*, *akademie*:

