Hoofstuk 5

Analise van gebruiksgebaseerde voorbeelde van deelwoorde

5.1 Inleiding

In praktiese analises, soos die annotering van korpusdata, blyk dit soms moeilik te wees om te onderskei tussen verskillende vorme van die werkwoord en hulle funksies. Die fokus van hierdie hoofstuk is om te ondersoek hoe deelwoorde in die Afrikaanse deel van die *National Centre for Human Language Technology*–korpus (NCHLT-korpus) in terme van lemmatisering en woordsoortetikettering (WS-etikettering) hanteer is. Die derde doelstelling van die studie, is hier ter sprake: om te bepaal of die insigte wat tot dusver verkry is in die beskrywings en konseptualiserings van die deelwoord alternatiewe lemmas en WS-etikette van deelwoorde in die korpus sal hê.

In afdeling 5.2 word ’n kort oorsig oor die NCHLT-projek gegee, aangesien die NCHLT-korpus (CTexT, 2013a) in hierdie studie gebruik word vir die analise van geannoteerde gebruiksvoorbeelde van deelwoorde. Dit is nodig om daarvan kennis te neem dat die voorbekendstellingsweergawe van die NCHLT-data gebruik is vir die analise van annoterings; dit geld die protokoldokumente, die korpus en die toetsteks.

In afdeling 5.3 kom die insigte wat vir die fyner onderskeid tussen verskillende werkwoordvorme ingewin is, aan bod. Met betrekking tot sowel lemmatisering as WS-etikettering gaan grys gebiede en vae grense, byvoorbeeld tussen predikatiewe voltooide deelwoorde en passiefvorme van die werkwoord, nie deug nie. ’n Klinkklare onderskeid tussen verskillende werkwoordvorme is nodig om bruikbare lemmas en WS-etikette daar te stel. Dit is nodig om eers genoegsaam tussen verskillende vorme van die werkwoord te onderskei, voor die NCHLT-protokoldokumente vir lemmatisering en WS-etikettering bespreek word.

Die twee vlakke van annotering waarmee hierdie studie gemoeid is, lemmatisering en WS-etikettering, word daarna afsonderlik van naderby beskou. Die interpretasie van die NCHLT-protokolle vir beide lemmatisering en WS-etikettering word in verband gebring met die standaarde en riglyne daargestel deur die *Expert Advisory Group on Language*

Engineering Standards (EAGLES, 1996) en die *Corpus Gesproken Nederlands* (CGN) (Van Eynde, 2004). Deurdat etiketstelle in verskillende tale sover moontlik voldoen aan die EAGLES-standaarde (1996), verseker dit dat etiketstelle herkenbaar en herbruikbaar in 'n internasionale konteks is (Van Eynde, 2004:5). Nie net is die CGN ook op die EAGLES-riglyne geskoei nie, maar kan die CGN – vanweë ons noue taalverwantskap – taalspesifieke insig in die lemmatisering en WS-etikettering van deelwoorde in Afrikaans bied.

Die NCHLT-protokol vir lemmatisering (CTexT, 2013b) kom in afdeling 5.4 onder die loep. Die riglyne vir lemmatisering word gemeet aan die insigte wat tot dusver oor die deelwoord verkry is om sodoende probleme in die protokoldokument uit te wys en voorstelle vir korreksies aan die protokol te maak, of om probleme uit die weg te ruim. Die protokoldokument vir lemmatisering word ook vergelyk met die riglyne vir lemmatisering soos dit vervat is in die CGN-dokument (Van Eynde, 2004:4; 26-27).

Nadat die NCHLT-protokol vir lemmatisering bespreek is, kom die NCHLT-protokol vir etikettering aan die beurt in afdeling 5.5). Die WS-etiketstel wat binne die NCHLT-projek gebruik is, is die WS-etiketstel wat deur Pilon (2005) vir Afrikaans daargestel is. Die wyse waarop sy WS-etikette vir deelwoorde hanteer het, word gemeet aan die nuwe insigte wat in die vorige hoofstukke oor die deelwoord ingewin is. Die EAGLES-standaarde en veral die CGN-riglyne vir WS-etikettering (Van Eynde, 2004) sal deurgaans as verdere riglyne gebruik word.

Die uiteindelike doel met die hoofstuk is om voorstelle te maak oor hoe deelwoorde beter in die NCHLT-korpus gelemmatiseer en geëtiketteer kan word. Die wyse waarop die deelwoord geannoteer word, behoort vir fyner sowel as growwe annoterings waar en bruikbaar te wees.

5.2 Die NCHLT-projek

Die NCHLT-projek is 'n projek van CTexT (Noordwes-Universiteit) wat oor vier jaar gestrek het (2010 tot 2013) en wat deur die Departement van Kuns en Kultuur van die Suid-Afrikaanse regering befonds is. Dit het ten doel om 50 000 tekseenhede (*tokens*) vir tien van die amptelike landstale (Engels uitgesluit) op vier vlakke te annoteer. Hierdie vier vlakke sluit in tekseenheid-identifikasie, lemmatisering, WS-etikettering en

morfologiese analise. Annotasies word gedoen met behulp van LARA2 (*Lexicon Annotation and Regulation Assistant version 2.0*).

Vir die doel van hierdie studie word slegs die wyse waarop die deelwoord binne die NCHLT-projek hanteer is met betrekking tot lemmatisering en WS-etikettering betrek. Die totale aantal Afrikaanse tekseenhede wat in die NCHLT-projek geannoteer is, is 60 318 eenhede (CTexT, 2013a). Hierdie groottotaal is verdeel in twee subkorpusse, naamlik die NCHLT-proefkorpus van 55 484 tekseenhede (verder genoem die NCHLT-korpus) en die NCHLT-toetsteks van 5 834 tekseenhede (verder genoem die NCHLT-toetsteks) wat vir kwaliteitskontrole aangewend is. Laasgenoemde word in hierdie hoofstuk ingespan vir die analise van hoe die deelwoord in die Afrikaanse deel van die NCHLT-projek hanteer is. Indien daar nie voorbeelde ter illustrasie in die toetsteks gevind word nie, sal voorbeelde in die volledige NCHLT-korpus gesoek word.

In die NCHLT-projek is twee protokolle deur CTexT saamgestel vir onderskeidelik lemmatisering (CTexT, 2013b) en WS-etikettering (CTexT, 2013c). Die doel van die protokolle is om prosedures en verduidelikings te formuleer en daardeur annoteerders van die korpus te rig en te lei in die onderskeie vlakke van annotering. Beide protokolle maak aanspraak op verantwoordelike bestuur, deurdat duidelike stappe uitgespel is vir enige wysigings of voorstelle aan die protokolle. Die protokolle word daarom nie as statiese dokumente gesien nie, maar as lewende dokumente wat kan verander soos nuwe insigte bereik word ter verbetering van die protokolle en annotering. Verder word prosedures of riglyne verskaf wat annoteerders sal help om akkurate lemmas en WS-etikette aan te bring. In die hoofteks van elk van die protokolle, sowel as in onderskeie bylae, word gedetailleerde verduidelikings en voorbeelde gegee.

5.3 Die afgrensing van deelwoorde

Binne die CGN stel Van Eynde (2004:65) dat die aanduiding van ’n WS-etiket die enigste verpligte eienskap is waaraan ’n WS-etiketstel volgens die EAGLES-standaarde moet voldoen. Die grootste en onmiddellike probleem in terme van hierdie verpligte standaard, is om te bepaal onder watter oorhoofse woordklas die deelwoord moet val: as ’n vorm van die werkwoord of as ’n adjektief.

Die bestaande riglyn binne die NCHLT-korpus hanteer deelwoorde as adjektiewe. Myns insiens is hierdie riglyn ontoereikend, aangesien die deelwoord as ’n vorm van die werkwoord in die eerste plek onder werkwoorde geklassifiseer behoort te word. In die konseptuele karakterisering van die deelwoord vanuit ’n kognitiewe gebruiksgebaseerde beskrywingsraamwerk (vergelyk Hfst. 4), is dit bevestig dat die verbale karakter van die deelwoord behoue bly wanneer dit optree in ander woordklasfunksies. Wanneer ’n deelwoord dus as ’n adjektief in ’n sin gebruik word, behoort dit van gewone adjektiewe onderskei te word as ’n ánder tipe adjektief met ’n verbale aard (dus ’n deelwoord).

Binne die CGN-dokument word ’n soortgelyke redenasie gevoer oor funksiewisseling van woordsoorte. Van Eynde argumenteer: “[e]en woord als *maandag* bijv. wordt vaak in bijwoordelijke functies gebruikt, zoals in *ik heb hem maandag nog gesproken*, maar is qua woordsoort een substantief, en wordt bij de tagging dan ook niet als bijwoord maar als substantief behandeld” (Van Eynde, 2004:7). Bykomend bepaal CGN oor die adjektiwiese gebruik van die deelwoord dat “[h]et onderscheid tussen deelwoord en adjectief wordt in de sectie over de werkwoorden toegelicht” (Van Eynde, 2004:19). CGN etiketteer deelwoorde volledig as werkwoorde (Van Eynde, 2004:31; 76-77).

Ter ondersteuning van die keuse om deelwoorde volledig onder werkwoorde te hanteer, kan eerstens aangevoer word dat deelwoorde nie uitsluitlik in ’n adjektiwiese funksie optree nie, maar dit kan ook in ’n bywoordelike of ’n voorsetselfunksie optree. Dit gebeur wel dat deelwoorde volkome geleksikaliseer het as adjektiewe of voorsetsels, en in sulke gevalle sal dié woorde dan bloot as adjektiewe of voorsetsels gelemmatiseer en geëtiketteer word. In die volgende afdeling oor lemmatisering (vergelyk 5.4) word sulke gevalle in detail verduidelik.

’n Verdere argument ter stawing daarvan dat deelwoorde onder werkwoorde geklassifiseer behoort te word, is dat dit logies min sin maak dat een vorm van die werkwoord (die PK-vorme in die verledetyd- en passiefkonstruksie) wel onder werkwoorde geklassifiseer word, maar dat ’n ander vorm van die werkwoord as ’n adjektief geklassifiseer word. Indien die bestaande etiketstel gevolg word om verskillende vorme van die werkwoord onder verskillende woordklasse te resorteer, impliseer dit dat die werkwoordvorme in die voorbeeldsinne hieronder totaal

verskillende etikette sal kry. Alhoewel die sinstrukture byna identies voorkom, roep (59a) meer die gebeurtenis op (’n agentlose passiefvorm), terwyl (59b) meer ’n toestand oproep (voltooide deelwoordvorm).

(59a) *Sy het al haar huishoudelike takies afgehandel: die bed is **opgemaak** en die stoep is **gevee**.*

(59b) *Haar huis is altyd netjies. Die bed is **opgemaak** en die stoep is **gevee**.*

By die karakterisering van deelwoorde op die fonologiese en semantiese pool, staan die verbale karakter van deelwoorde op die voorgrond. By lemmatisering staan die verbale karakter van deelwoorde ook op die voorgrond, maar by die toekenning van WS-etikette, tree die funksie van die woord op die voorgrond. By lemmatisering word deelwoorde hanteer soos werkwoorde en by WS-etikettering word deelwoorde hanteer volgens die funksie waarin dit voorkom: adjektiewe, bywoorde of voorsetsels.

5.4 Lemmatisering binne die NCHLT-projek

Met betrekking tot die annotering van ’n korpus, verduidelik Van Eynde (2004:3) dat “[d]e eerste stap in de taalkundige ontsluiting van het corpus ... de toekenning van tags en lemmata aan de eenheden [behelst] ...”. Die eerste vlak van annotering wat bespreek word, is dan ook die korrekte lemmatisering per tekseenheid (woordinskrywing). ’n Lemma is ’n “woord, woorddeel of woordgroep wat in ’n woordeboek of ander naslaanwerk opgeneem word as ’n onderwerp vir verklaring en (of) behandeling; ... sodanige woord, ... [en die] woorddeel of woordgroep [het] die status van leksikale eenheid ...” (*eWAT*, 2009). ’n Lemma is dus die ongeflekteerde vorm van ’n woord, ’n trefwoord in ’n woordeboek (CTexT, 2013a:1). Die lemmatisering geskied op ’n woord-vir-woord-basis en gevolglik word elke eenheid apart gelemmatiseer.

In die volgende onderafdelings word verwysings na deelwoorde in die protokoldokument vir lemmatisering (CTexT, 2013b) bespreek. Eers word algemene riglyne vir lemmatisering wat ook vir deelwoorde geld, bespreek (vergelyk 5.4.1 tot 5.4.4) en dan word die riglyne spesifiek vir deelwoorde een-vir-een bespreek (vergelyk 5.4.5.1 tot 5.4.5.5). Vir duidelikheid en maklike verwysing, word die relevante dele uit die protokol telkens in figure by die toepaslike deel weergegee. Uittreksels uit die

protokol word telkens in 'n skoon blok gegee, en voorstelle ter verbetering van die uittreksel word in 'n grys blok gegee waarin die voorstel duidelik met wit gemerk word.

5.4.1 Lemmatisering van werkwoorde

Binne die hoofteks van die protokol onder die subopskrif '*Verbs*' (CTexT, 2013b:3) word riglyne gegee vir die lemmatisering van werkwoorde (vergelyk Figuur 25). Punte 8 tot 11 in Figuur 25 gee voorbeelde van watter deel van werkwoordvorme (verskillende fleksie-affikse) om watter redes verwyder word om lemmas te lewer. Die lemmas van hierdie vyf eenhede (vergelyk punte 8 tot 11 in Figuur 25) sal dus *drink*, *skop*, *skreeu*, *meganiseer* en *breek* wees. Wat deelwoorde spesifiek betref (vergelyk punte 10 en 11, Figuur 25), word die werkwoordbasis as lemma gegee. Daar is egter nie 'n verwysing na die passiefkonstruksie in die lemmaprotokol nie. Die insig waartoe in die studie gekom is, is dat die verledetyd- en passiefvorm konsekwent dieselfde vorm van die werkwoord is en ook konsekwent dieselfde gelemmatiseer sal word. Vir volledigheidsonthalwe, kan die lemmaprotokol aangepas word deur die passiefvorm van die werkwoord saam met die verledetydvorm te meld (vergelyk punt 9, Figuur 26).

Verbs
8. Infinitive *-e* (e.g. *iets te drink**e***)
9. Past tense (e.g. ***ge**skop*)
10. Present participle (e.g. *skreeu**end***)
11. Past participle *ge-...-t/-d* (e.g. ***ge**meganiseer**d**;* also strong past participles like ***gebroke***)

Figuur 25: Uittreksel uit die lemmaprotokol (CTexT, 2013b:3)

In die CGN (Van Eynde, 2004:4) word 'n verdere riglyn gegee wat moontlik tot die lemmaprotokol toegevoeg kan word. Indien die stam van die woord nie 'n bestaande woord is nie, word die geflekteerde vorm (soos wat dit as werkwoord gebruik word) as die lemma gegee. In (60) is die *-niet* in *geniet* nie 'n herkenbare woord nie en daarom word die lemma as *geniet* gegee.

(60) *'n Hele aantal sake het reeds aandag **geniet*** (NCHLT-toetsteks, eenheid 4096)
Tekseenheid: *geniet* (*ge*-lose verledetydvorm van die werkwoord)
Lemma: *geniet*

Aansluitend hierby kan ook bygevoeg word dat wanneer deelwoorde gevorm word van afgeleide werkwoorde (dié gevorm met *ge-*, *be-*, *her-*, *er-*, *ver-* en *ont-*), word slegs die deelwoorduitgange verwyder (vergelyk Figuur 25, punte 10 en 11) en die afgeleide

werkwoordbasis is die lemma. Die volgende drie voorbeelde hiervan (vergelyk 61a tot 61c), is in die NCHLT-toetsteks (CTexT, 2013a) gevind. Die sinskonteks, die tekseenheid onder bespreking en die lemma word telkens gegee.

(61a) ... *die* ***besturende*** *direkteur* ... (NCHLT-toetsteks, eenheid 408)
Tekseenheid: *besturende* (onvoltooide deelwoord)
Lemma: *bestuur*

(61b) ... *tot 'n* ***verenigde*** *staat* ... (NCHLT-toetsteks, eenheid 1687)
Eenheid: *verenigde* (swak voltooide deelwoord)
Lemma: *verenig*

(61c) ... *'n* ***beslote*** *korporasie* ... (NCHLT-toetsteks, eenheid 389)
Eenheid: *beslote* (sterk voltooide deelwoord)
Lemma: *besluit*

Figuur 26 stel voor hoe Figuur 25 verbeter kan word. Die witgemerkte dele wys waar die verbeterings aangebring is.

Verbs
8. Infinitive *-e* (e.g. *iets te drink**e***)
9. Verb form in past tense / passive constructions (e.g. ***ge****skop*)
   - If the stem of the construction form is not an existing verb, the inflected form should be given as the lemma, e.g. *geniet* (and not *niet* that is not a recognised verb).
   - Derived verbs (with prefixes *ge-*, *be-*, *her-*, *er-*, *ver-* and *ont-*) remain as they are.
10. Present participle (e.g. *skreeu**end***)
11. Past participle *ge-...-t/-d* (e.g. ***ge****meganiseer****d****;* also strong past participles like ***gebroke***)

Figuur 26: Voorstel ter verbetering van die lemmaprotokol (CTexT, 2013b:3)

5.4.2 Lemmatisering van deadjektiwiese naamwoorde

In die bylaaggedeelte van die lemmatiseringsprotokol, onder die subopskrif '*Nouns*' (CTexT, 2013b:9), word die volgende riglyn gegee (vergelyk Figuur 27) ten opsigte van deadjektiwiese naamwoorde, wat ook vir deelwoorde kan geld.

- Note cases of deadjectival nouns, derived using the ***-e*** morpheme – these should not be confused with the attributive form of the adjective. For example, *domme* in *die* ***domme*** *swaap* should be lemmatised as *dom*, because the ***-e*** is being used attributively. *Dommes* (*die lede van die span is 'n spul* ***dommes***) should not be lemmatised as *dom*, but as *domme*. Remember that the part of speech category should not be changed.

Figuur 27: Uittreksel uit die lemmaprotokol (CTexT, 2013b:9)

Die riglyne in die CGN verskil egter van die bostaande riglyn uit die lemmaprotokol (vergelyk Figuur 27). Volgens die CGN word alle substantiewe wat in ander funksies gebruik word (soos onder andere nominaal gebruikte adjektiewe of deelwoorde) as daardie woordsoort hanteer (Van Eynde, 2004:13). So word *het geschrevene* en *een gekwetste* binne die CGN as voorbeelde gegee van voltooide deelwoorde wat nominaal gebruik word (Van Eynden, 2004:31; 76). Aansluitend by die CGN is dit ook nie ongekend vir Afrikaanse bronne om na deadjektiwiese naamwoorde as deelwoorde te verwys nie (Kempen, 1982:471; Van Schoor, 1983:219-220; Du Toit, 1986:138).

Myns insiens is die bostaande riglyn uit die NCHLT-lemmaprotokol (vergelyk Figuur 27) korrek. Indien die riglyn uit die NCHLT-lemmaprotokol toegepas word op deadjektiwiese naamwoorde waarvan die basis ’n deelwoord is, besef ’n mens weer eens dat ’n deelwoord in werklikheid nooit in ’n naamwoordelike funksie gebruik kan word nie, maar dat ’n naamwoord wel van ’n deelwoord gevorm kan word. Só ’n naamwoord, afgelei van ’n deelwoord, is dan ’n afleiding en word gewoon as naamwoord hanteer. Alhoewel dit moontlik is dat die naamwoord en die deelwoord presies dieselfde vorm vertoon, is hulle morfologiese bou verskillend. Die twee voorbeelde in (62a) en (62b) uit die NCHLT-toetsteks (CTexT, 2013a) illustreer die verskil duidelik.

(62a) ... *kopieë van die* ***volgende*** *dokumente* ... (NCHLT-toetsteks, eenheid # 293)
Eenheid: *volgende* (attributiewe onvoltooide deelwoord)
Lemma: *volg*

(62b) *Doen die* ***volgende*** *in die afdeling* ... (NCHLT-toetsteks, eenheid # 53)
Eenheid: *volgende* (selfstandige naamwoord)
Lemma: *volgende*

In die toetsteks (CTexT, 2013b) is daar vyf voorbeelde waar naamwoorde van deelwoorde gevorm is. Tabel 11 bevat ’n foutanalise, sowel as voorstelle vir korreksies in die plek van foutiewe lemmas soos dit tans in die toetsteks voorkom. Tekseenheid 386 is die enigste van die vyf eenhede wat korrek gelemmatiseer is en moet dus bly soos wat dit is. Die ander vier eenhede se lemmas is nie korrek volgens die riglyn in die protokol hierbo nie (vergelyk Figuur 27) en voorstelle vir nuwe lemmas is in Tabel 11 aangedui. Let verder daarop dat tekseenheid 2725 deel vorm van ’n eienaam, naamlik *Eskom Beherend*.

Tabel 11: Lemmas van naamwoorde met deelwoordbasisse

Eenheid #	Eenheid	Lemma tans	Nuwe lemma	Sinskonteks
53	volgende	volg	volgende	Doen die **volgende** in die afdeling ...
71	volgende	volg	volgende	Verstrek die **volgende** in die afdeling ...
386	nie-ingesetene	nie-ingesetene	✓	'n nie-inwonende individu ('n **nie-ingesetene**)
1592	uitgewekenes	uitwyk	uitgewekene	... en die terugkeer van **uitgewekenes**.
2725	Beherend	Beheer	Beherend	... Eskom **Beherend** sal funksioneer.

5.4.3 Lemmatisering van eenhede met 'n ontkenningsprefiks of deelsgenitiewe *-s*

In die hoofteks van die NCHLT-lemmaprotokol onder '*Adjectives*' (CTexT, 2013b:11), word twee riglyne gegee wat die lemmatisering van deelwoorde raak, die ontkenningsprefiks (vergelyk Figuur 28) en die deelsgenitiewe *-s* (vergelyk Figuur 29). Die eerste hiervan betrek die ontkenningsprefiks wat in die lemma van adjektiewe behoue bly.

- Note that morphemes that express negation **should not be removed**.
  onmoontlike *onmoontlik*
  nie-vervangbare *nie-vervangbaar*

Figuur 28: Uittreksel uit die lemmaprotokol (CTexT, 2013b:11a)

Die riglyn hierbo (vergelyk Figuur 28) geld ook vir woorde met ontkenningsprefikse aan deelwoordbasisse, byvoorbeeld *onvermoeid* of *onvoldoende* (vergelyk Tabel 12). Op dieselfde wyse as wat dit die geval is by adjektiewe, bly ontkenningsprefikse behoue in die lemmas van deelwoorde, maar indien daar 'n addisionele attributiewe *-e* in die woord is, word dit verwyder. Die ontkenningsprefiks wat in die lemma behoue bly, is 'n afleidingsaffiks wat 'n adjektief van die deelwoord aflei, en gevolglik verloor die deelwoord hiérdeur sy verbale aard (as ons in gedagte hou dat deelwoorde gelyktydig werkwoordelik en adjektiwies is). Dit verklaar moontlik die rede waarom *nie-ingesetene* (tekseenheid 386) aanvanklik in Tabel 11 korrek gelemmatiseer is. Die implikasie vir woordklasse van die lemmas waarvan die ontkenningsprefikse behoue bly, is dat sulke tekseenhede ook nie meer as deelwoorde geëtiketteer sal word nie, maar gewoon as adjektiewe.

In die toetsteks (CTexT, 2013b) is daar nege voorbeelde van eenhede met 'n ontkenningsprefiks (agt voorbeelde indien 'nie-ingesetene' wat reeds as naamwoord

hanteer is, nie ingereken word nie). Al hierdie woorde is, sover dit die ontkenningsaffiks betref, korrek gelemmatiseer (vergelyk Tabel 12). Tekseenheid 5807 se lemma is in Tabel 12 vir volledigheidsonthalwe gekorrigeer met betrekkings tot die riglyn wat ons reeds in Figuur 25 hanteer het, nie met betrekking tot die ontkenningsprefiks nie.

Tabel 12: Lemmas van woorde met ontkenningsmorfeme

Eenheid #	Eenheid	Lemma tans	Nuwe lemma	Sinskonteks
382	nie-inwonende	nie-inwonend	✓	'n **nie-inwonende** individu
386	nie-ingesetene	nie-ingesetene	✓	'n nie-inwonende individu ('n **nie-ingesetene**)
1029	onvermoeide	onvermoeid	✓	... Suid-Afrikaners se **onvermoeide** stryd.
1484	onvermoeid	onvermoeid	✓	... en **onvermoeid** bepleit het.
1551	onverpoosd	onverpoosd	✓	... wat haar **onverpoosd** vir verandering beywer het.
1786	onvermoeide	onvermoeid	✓	Julle **onvermoeide** en heldhaftige opofferings ...
4537	onvoldoende	onvoldoende	✓	... **onvoldoende** infrastruktuur.
5807	onbesonge	onbesonge	onbesing	... aan al ons helde en heldinne, besonge en **onbesonge**, bekend en **onbekend** ...
5811	onbekend	onbekend	✓	

'n Tweede nota (vergelyk Figuur 29) in die NCHLT-lemmaprotokol onder '*Adjectives*' (CTexT, 2013b:11) wat moontlik by die deelwoord kan voorkom, behels die deelsgenitiewe *-s*. Indien 'n deelwoord met 'n deelsgenitiewe *-s* voorkom, word die lemma daarsonder weergegee.

Partitive genitive
-s in: moois, leliks
Lemmas: mooi, lelik

Figuur 29: Uittreksel uit die lemmaprotokol (CTexT, 2013b:11b)

In die toetsteks (CTexT, 2013a) is daar nie 'n enkele voorbeeld hiervan nie. Die voorbeeld in (63) is 'n selfuitgedinkte voorbeeld ter illustrasie.

(63) *Daar is iets **voorspellends** in sy optrede*.
Eenheid: *voorspellends* (onvoltooide deelwoord)
Lemma: *voorspel*

5.4.4 Lemmatisering van deelwoorde met partikelwoorde as basis

Die gedeelte in die NCHLT-lemmaprotokol (vergelyk Figuur 30) wat die riglyn gee van hoe deelwoorde met partikelwerkwoorde as basis hanteer moet word, word in die bylaag onder riglyne vir die verledetydkonstruksie gegee (CTexT, 2013b:12). Wanneer voltooide deelwoorde van partikelwerkwoorde gevorm word, kom die *-ge-* tussen die dele: tussen die partikel en die werkwoord. Die lemmas van sulke deelwoorde is die basisvorm van die partikelwerkwoord.

- Compound verbs, consisting of two independent parts and that have a past tense morpheme (like ***aange-*** (*aangedryf*) and ***opge-*** (*opgelaai*)), are lemmatised by removing the ***-ge-*** part of the word. *Aangedryf* and *opgelaai* are lemmatised respectively as *aandryf* and *oplaai.*

Figuur 30: Uittreksel uit die lemmaprotokol (CTexT, 2013b:12)

In die toetsteks (CTexT, 2013) is daar vier voorbeelde waar dié riglyn vir deelwoorde geld. Drie van die vier voorbeelde is korrek gelemmatiseer. Die een foutiewe eenheid is verkeerdelik gelemmatiseer op grond van 'n vorige riglyn wat reeds bespreek is, naamlik *uitgewekenes* (eenheid 1592), 'n naamwoord in die meervoud (vergelyk Figuur 27 en Tabel 11). Die lemma van *uitgewekenes* was gegee as *uitwyk*, maar in Tabel 11 is die riglyn vir naamwoorde met deelwoordbasisse in Figuur 27 gevolg en is die lemma as *uitgewekene* gekorrigeer: die naamwoordvorm is behou en die meervoud *-s* verwyder. Aangesien deelwoorde van partikelwerkwoorde baie produktief is (vergelyk Tabel 9 in afdeling 4.3.1), behoort hierdie riglyn (vergelyk Figuur 30) herhaal te word in die bylaaggedeelte van die protokol wat spesifieke riglyne vir die lemmatisering van deelwoord gee (vergelyk Figuur 32, CTexT, 2013b:13). Tabel 13 illustreer die drie eenhede wat korrek gelemmatiseer is volgens die riglyn in Figuur 30.

Tabel 13: Lemmas van deelwoorde gevorm van deeltjiewerkwoorde

Eenheid #	Eenheid	Lemma tans	Nuwe lemma	Sinskonteks
443	opgelei	oplei	✓	... oproepsentrumpersoneel wat goed **opgelei** is ...
2264	uitgebreide	uitbrei	✓	... die **Uitgebreide** Openbare Werke-program ...
2478	vrygestel	vrystel	✓	... arbeidstatistiek wat Dinsdag **vrygestel** is, dui ...

5.4.5 Algemene nota oor die lemmatisering van deelwoorde

In die NCHLT-lemmaprotokol (CTexT, 2013b:13-15) word 'n samevattende riglyn gegee vir die lemmatisering van deelwoorde (vergelyk Figuur 31). Aangesien hierdie deel

spesifiek fokus op die lemmatisering van deelwoorde, word dit punt-vir-punt beskou en op die NCHLT-toetsteks van toepassing gemaak.

Participles: General

- The following default principle apply: **if a verbal analysis is possible, apply such analysis**.
  - In other words, when it is possible and it makes sense to identify a verb in a participle, identify such verb as the lemma of the participle form.
  - To general paraphrasing tests could be applied:
    - Present participles: "X wat Y"
      - *lopende water* — *water wat **loop***
    - Past participles: "X wat *ge-*/*ø*-Y is"
      - *geregistreerde student* — *student wat geregistreer is*
      - *begunstigde persone* — *persone wat begunstig is*
      - *betrokke onderwerp* — *onderwerp wat betrek is*

Figuur 31: Uittreksel uit die lemmaprotokol (CTexT, 2013b:13a)

In die toetsteks (CTexT, 2013a) is 51 eenhede as deelwoorde gemerk waarvan 'n natuurlike verbale lesing deur parafrasering gekry kan word, en 52 eenhede is as deelwoorde gemerk wat in 'n passiefkonstruksie voorkom. Die 52 eenhede wat in die passiefkonstruksie voorkom, behoort nie as deelwoorde gemerk te word nie, maar as 'n perifrastiesekonstruksievorm van die werkwoord (vergelyk afdeling 5.4.1 en Figuur 26). Die 51 eenhede met 'n natuurlike verbale lesing bestaan uit 28 onvoltooide deelwoorde, 16 swak voltooide deelwoorde en 7 sterk voltooide deelwoorde. Vergelyk die verbale paragraserings van 'n voorbeeld van elke tipe deelwoord uit die toetsteks hieronder: 'n onvoltooide deelwoord in (64a), 'n swak voltooide deelwoord in (64b) en 'n sterk voltooide deelwoord in (64c).

(64a) ... *die leiers van die **regerende** party en* ... (NCHLT-toetsteks, eenheid 1262)
Lemma: *regeer*
Verbale parafrase: die party wat **regeer**

(64b) ... *oorspronklike of **gewaarmerkte** kopieë van* ... (NCHLT-toetsteks, eenheid 289)
Lemma: *waarmerk*
Verbale parafrase: kopieë wat **gewaarmerk is**

(64c) ... *aan die **betrokke** ondersoekeenheid oordra* ... (NCHLT-toetsteks, eenheid 856)
Lemma: *betrek*
Verbale parafrase: die ondersoekeenheid wat **betrek is**

Figuur 31 kan aangepas word (vergelyk Figuur 32) deur die riglyn wat in Figuur 30 gegee is, te herhaal in die gedeelte van die bylaag wat spesifiek riglyne vir die deelwoord gee (CTexT, 2013b:13).

Participles: General

- The following default principle apply: **if a verbal analysis is possible, apply such analysis**.
  - In other words, when it is possible and it makes sense to identify a verb in a participle, identify such verb as the lemma of the participle form.
  - To general paraphrasing tests could be applied:
    - Present participles: "X wat Y"
      - *lopende water* *water wat **loop***
    - Past participles: "X wat ge-/ø-Y is"
      - *geregistreerde student* *student wat ge**registreer** is*
      - *begunstigde persone* *persone wat **begunstig** is*
      - *betrokke onderwerp* *onderwerp wat **betrek** is*
  - Compound verbs, consisting of two independent parts and that have a past tense morpheme (like ***aange***- (aangedryf) and ***opge***- (opgelaai)), are lemmatised by removing the -***ge***- part of the word. *Aangedryf* and *opgelaai* are lemmatised respectively as *aandryf* and *oplaai*.

Figuur 32: Voorstel ter verbetering van die lemmaprotokol (CTexT, 2013b:13a)

In die toetsteks (CTexT, 2013a) is daar ook deelwoorde waarvan 'n verbale parafrase nie natuurlik is nie (vergelyk Tabelle 14, 15 en 16). Die punte in die riglyn in Figuur 33 wys juis na sulke deelwoorde wat nie gelemmatiseer word nie, maar wat in die deelwoordvorm bly, terwyl slegs die attributiewe -*e* verwyder word indien dit voorkom. Vervolgens word elkeen van hierdie vyf riglyne apart bespreek en op die toetsteks van toepassing gemaak.

- The following classes of participles **are not lemmatised** (i.e. the participle forms remain as they are; remove attributive -*e* where possible).
  - Lexicalised adjectives
  - Participles with adjectival prefixes
  - Compounding derivations/synthetic compounds
  - Prepositions
  - Participles with noun stems

Figuur 33: Uittreksel uit die lemmaprotokol (CTexT, 2013b:13b)

5.4.5.1 Geleksikaliseerde adjektiewe

Die protokoldokument (CTexT, 2013b:13) lys 17 voorbeelde van deelwoord wat geleksikaliseer het as adjektiewe (vergelyk Figuur 34). Die redenasie in die protokoldokument is dat die adjektiwiese lesing van hierdie eenhede dominant is en

daarom word hulle bloot as adjektiewe hanteer tydens lemmatisering. Hierdie eenhede word dan ook dienooreenkomstig as adjektiewe geëtiketteer.

○ **Lexicalised adjectives**	
▪ *ingewikkelde*	*ingewikkeld*
▪ *uitnemende*	*uitnemend*
▪ *uitstekende*	*uitstekend*
▪ *uitstaande (sake)*	*uitstaande*
▪ *ooglopende*	*ooglopend*
▪ *bereid*	*bereid*
▪ *verwante, aanverwante, nouverwante*	*verwant, aanverwant, nouverwant*
▪ *geleë*	*geleë*
▪ *bekende*	*bekend*
▪ *opsienbarende*	*opsienbarend*
▪ *gereelde*	*gereeld*
▪ *spannende*	*spannend*
▪ *opgewekte (musiek)*	*opgewek*
▪ *uitgelese (gaste)*	*uitgelese*
▪ *opwindende*	*opwindend*
▪ *besliste (houding)*	*beslis*
▪ *allerhande*	*allerhande*

Figuur 34: Uittreksel uit die lemmaprotokol (CTexT, 2013b:13c)

Die term ‘leksikalisasie’ kan, onder andere, ook gebruik word om te verwys na “a historical process by which, e.g. a former suffix becomes an independent lexical unit” (Matthews, 1997:206). Mens kan dus redeneer dat aangesien leksikalisasie ’n proses is, dit ook relatief kan wees: wat een persoon as geleksikaliseerd ervaar (’n suiwer adjektiwiese lesing van ’n woord), verskil van ’n volgende (’n verbale lesing van die woord). Daarom is so ’n lys in die protokoldokument (vergelyk Figuur 34) onontbeerlik vir annoteerders.

Van die 17 voorbeelde in die protokollys, kom ses tekseenhede in die toetsteks voor. Vier van die ses tekseenhede in die toetsteks is altesaam 12 keer korrek as adjektiewe gelemmatiseer volgens die voorskrif van die protokol (vergelyk Tabel 14), naamlik: *bereid*, *bekend*, *besliste* en *geleë*. Die ander twee tekseenhede in die toetsteks wat nie as adjektiewe hanteer is nie, maar soos deelwoorde gelemmatiseer is, word ook in Tabel 14 aangetoon: *uitstaande* en *uitstekende*. In Tabel 14 is die lemma van *uitstekende* bloot gekorrigeer aangesien ’n *uitstekende poging* nie ’n verbale parafrase van ‘’n poging wat uitsteek’ toelaat nie. Mens sou kon redeneer dat ’n poging wat uitsteek een is wat bo ander pogings uitstaan om sodoende ’n verbale interpretasie te regverdig. Dit sou ’n

figuurlike denksprong vra, en aangesien ***uitstekende*** *poging* in werklikheid direk as 'n *baie goeie poging* interpreteer word, bevestig dit eerder *uitstekend* as 'n geleksikaliseerde adjektief. Dieselfde argument geld vir ***uitstaande*** *leierskap*.

Tabel 14: Geleksikaliseerde adjektiewe as deelwoorde gemerk

Eenheid #	Eenheid	Lemma tans	Nuwe lemma	Sinskonteks
5776	bereid	bereid	✓	... is dit 'n ideaal waarvoor ek **bereid** is om te sterf.
507	bekend	bekend	✓	... dat informante se identiteit nie **bekend** sal word nie ....
1516 2132	besliste	beslis	✓	... groot moed en **besliste** leierskap aan die dag lê ...
110	geleë	geleë	✓	... die voorstad waarin die onderneming **geleë** is ...
4105	uitstaande	uitstaan	uitstaande	… te danke aan sy **uitstaande** leierskap …
5521	uitstekende	uitsteek	uitstekend	… vir hul **uitstekende** poging.
5267	onderskeie	onderskeie	✓	... in hul **onderskeie** begrotingsredes ...
5108	oorlede	oorlede	✓	... haar **oorlede** man se pensioen ...
Korpus 37 998	gebore	gebore	✓	... die demokratiese grondwet wat uit enorme opofferings **gebore** is ...

In die toetsteks is verdere voorbeelde geïdentifiseer van deelwoorde wat as adjektiewe geleksikaliseerd is. Twee tekseenhede is korrek in die toetsteks as geleksikaliseerde adjektiewe hanteer, naamlik *onderskeie* en *oorlede*, en behoort slegs by die protokollys gevoeg te word (vergelyk Figuur 35).

- **Lexicalised adjectives**
  - *ingewikkelde* — *ingewikkeld*
  - *uitnemende* — *uitnemend*
  - *uitstekende* — *uitstekend*
  - *uitstaande (sake)* — *uitstaande*
  - *ooglopende* — *ooglopend*
  - *bereid* — *bereid*
  - *verwante, aanverwante, nouverwante* — *verwant, aanverwant, nouverwant*
  - *geleë* — *geleë*
  - *bekende* — *bekend*
  - *opsienbarende* — *opsienbarend*
  - *gereelde* — *gereeld*
  - *spannende* — *spannend*
  - *opgewekte (musiek)* — *opgewek*
  - *uitgelese (gaste)* — *uitgelese*
  - *opwindende* — *opwindend*
  - *besliste (houding)* — *beslis*
  - ~~*allerhande*~~ — ~~*allerhande*~~
  - *onderskeie*
  - *is gebore*
  - *is oorlede*

Figuur 35: Voorstel ter verbetering van die lemmaprotokol (CTexT, 2013b:13)

Daar is verder ook nie ’n voorbeeld van *gebore* in die NCHLT-toetsteks nie, maar wel in die NCHLT-korpus (tekseenhede 28 259, 28 281 en 37 998). Al drie annoterings van *gebore* in die korpus is as oorganklike werkwoorde geëtiketteer, en indien hulle as geleksikaliseerde adjektiewe in die protokollys opgeneem word, behoort hulle dienooreenkomstig as adjektiewe geëtiketteer te word.

Een van die voorbeelde in die protokollys, naamlik *allerhande* (vergelyk Figuur 34), oortuig nie as ’n voorbeeld van ’n deelwoord wat as ’n adjektief geleksikaliseer het nie. In Figuur 35 is *allerhande* wit gemerk en doodgetrek aangesien dit glad nie ’n deelwoord is nie. Hierdie voorbeeld behoort bloot geskrap te word. Al hierdie voorstelle vir toevoegings tot die protokollys is wit gemerk en in Figuur 35 aangedui.

5.4.5.2 Deelwoorde met adjektiwiese prefikse

Indien deelwoorde prefikse bevat wat adjektiveerders is, word hulle ook as adjektiewe hanteer (vergelyk Figuur 36).

- **Participles with adjectival prefixes**
  - *aartsingewikkeld*
  - *ongerep, ongeleë, onbeduidend, onvoldoende, onaangemeld, onvermoeid*
  - *nie-ingesetene, nie-inwonend, nie-bestaande, nie-erkende*
  - *self-geskrewe*
  - *eersvolgend*

Figuur 36: Uittreksel uit die lemmaprotokol (CTexT, 2013b:13d)

Die adjektiwiese prefikse in Figuur 36 sluit in *aarts-*, *nie-*, *on-*, *self-* en *eers-*. Daar is geen voorbeelde in die NCHLT-toetsteks van woorde wat met *aarts-*, *self-* of *eers-* begin nie, maar in die volledige NCHLT-korpus is daar een voorbeeld van ’n deelwoord met *self-* wat wel soos ’n adjektief gelemmatiseer is (vergelyk (65)). Die ontkenningsprefikse *nie-* en *on-* is reeds vroeër hanteer (vergelyk afdeling 5.4.3). Alhoewel die riglyn in die protokol vir ontkenningsprefikse reeds onder adjektiewe gegee is (CTexT, 2013b:12), is dit sinvol dat dit weer hier gegee word (CTexT, 2013b:13 uitgebeeld in Figuur 36).

(65) ... *sluit ’n* ***selfgeadresseerde*** ... *A4-koevert in* ... (NCHLT-korpus, eenheid 19 176)
Lemma: *selfgeadresseerd* (attributiewe adjektief)

5.4.5.3 Deelwoorde in samestellende afleidings

In ’n samestellende afleiding, wanneer ’n deelwoordmorfeem op ’n woordgroep van toepassing is, word dit nie gelemmatiseer nie. Die deelwoordmorfeem bly behoue in die lemma, en slegs die attributiewe *-e* word verwyder indien dit voorkom. Die lemmaprotokol (CTexT, 2013b:13-14) lys 12 voorbeelde as riglyn vir die lemmatisering van sulke woorde (vergelyk Figuur 37).

- **Compounding derivations/synthetic compounds**
  - sodoende (wat so doen) — sodoende
  - onderstaande (wat onder staan) — onderstaande
  - bostaande (wat bo staan) — bostaande
  - doeltreffende (wat die doel tref) — doeltreffend
  - ondergenoemde (wat onder genoem is) — ondergenoemde
  - bogenoemde (wat bo genoem is) — bogenoemde
  - nouspannende (wat nou span) — nouspannend
  - voormelde (wat voor vermeld is) — voormelde
  - voornoemde (wat voor genoem is) — voornoemde
  - hoogaangeskrewe (wat hoog aangeskryf is) — hoogaangeskrewe
  - regsgebaseerde (wat op die reg gebaseer is) — regsgebaseerd
  - doelgerigte (wat op ’n doel gerig is) — doelgerig

Figuur 37: Uittreksel uit die lemmaprotokol (CTexT, 2013b:13-14)

Daar is vier gevalle in die NCHLT-toetsteks van deelwoordmorfeme in samestellende afleidings. Al vier is korrek gelemmatiseer (vergelyk Tabel 15).

Tabel 15: Deelwoorde in samestellende afleidings

Eenheid #	Eenheid	Lemma	Nuwe lemma	Sinskonteks
225	onderstaande	onderstaande	✓	... die **onderstaande** inligting ...
1004	sodoende	sodoende	✓	... en **sodoende** ... gedenk.
2852	tuisgebaseerde	tuisgebaseerd	✓	... **tuisgebaseerde** versorging ...
2356	gemeenskapsgebaseerde	gemeenskapsgebaseerd	✓	... **gemeenskapsgebaseerde** sorg

5.4.5.4 Deelwoorde as voorsetsels

Volgens die riglyne in die protokol (vergelyk Figuur 38) behoort deelwoorde wat in ’n voorsetselfunksie gebruik word, as voorsetsels hanteer en gemerk te word. Myns insiens behoort daar eers gekyk te word of die deelwoorde in die voorsetselfunksies ’n natuurlike verbale parafrase lewer. Indien dit ’n natuurlike verbale parafrase lewer, soos byvoorbeeld *insluitende* – ‘wat insluit’, behoort die lemma *insluit* te wees, maar indien die verbale parafrase nie natuurlik is nie, soos byvoorbeeld *gedurende*, behoort

die lemma *gedurende* te wees. Dit is soms baie moeilik om te bepaal of sommige deelwoordvorme 'n bywoord van skakeling of 'n voorsetsel is.

- **Prepositions**
  - gedurende
  - ingevolge
  - rakende
    - Note that *rakende* should be lemmatised as *raak* in the context of *rakende hoeke*.

Figuur 38: Uittreksel uit die lemmaprotokol (CTexT, 2013b:14a)

Binne die toetsteks is *gedurende* vyf keer (tekseenhede 1655, 1841, 2172, 5697, en 4180) en *insluitende* twee keer (tekseenhede 406 en 4044) as voorsetsels gebruik. Deur verbale omskakelings blyk dat 'n sambreelriglyn nie vir alle voorsetsels kan geld nie. 'n Verbale parafrase is nodig om te bepaal of die betrokke eenheid as 'n voorsetsel of as 'n deelwoord in 'n bywoordelik funksie gelemmatiseer en geëtiketteer behoort te word. Ek stel voor dat alle voorbeelde waar 'n verbale parafrase ongrammatikaal is, as voorsetsels hanteer word en dat alle voorbeelde waar 'n natuurlike verbale parafrase verkry word, as bywoorde hanteer word (vergelyk Figuur 39). Vergelyk die verbale parafrases hieronder in (66a) en (66b).

(66a) ***Gedurende*** *die loop van die jaar* ... (NCHLT-toetsteks, eenheid 1655)
Verbale analise: *wat duur die jaar / *die jaar wat duur
Lemma: *gedurende* (geleksikaliseerde voorsetsel)

(66b) ... *alle direkteure,* ***insluitende*** *die* ... (NCHLT-toetsteks, eenheid 406)
Verbale analise: wat **insluit** die besturende direkteur
Lemma: insluit (deelwoord in 'n bywoordelik funksie)

In die bostaande analises is dit duidelik dat *gedurende* 'n onnatuurlike verbale parafrase lewer en dus altyd as voorsetsel hanteer behoort te word. Hierteenoor lewer *insluitende* wel 'n natuurlik verbale parafrase en behoort dit dus eerder as 'n deelwoord hanteer te word. Tabel 16 onderskei tussen deelwoorde as bywoorde van skakeling (met 'n verbale parafrase) en voorsetsels (sonder 'n verbale analise).

Tabel 16: Onderskeid tussen deelwoorde as bywoorde van skakeling en voorsetsels

Eenheid #	Deelwoorde as bywoorde van skakeling	Lemma	WS-etiket
NCHLT-korpus eenheid 42 244	... samesprekings voer *aangaande* wetgewing ... ... wat wetgewing **aangaan**	aangaan	bywoord
NCHLT-korpus eenheid 22 316	... mag 'n uniform dra *afhangende* van ... ... mag 'n uniform dra wat **afhang** van ...	afhang	bywoord
NCHLT-korpus eenheid 48 349	... hierdie instansie ... *betreffende* hul doktrines ... ... hierdie instansie ... **wat betref** hul doktrines ...	betref	bywoord
NCHLT-toetsteks eenheid 406	... direkteure, *insluitende* die besturende direkteur ... ... direkteure, **wat insluit** die besturende direkteur ...	insluit	bywoord
eWAT	... wetgewing *rakende* 'n saak ... ... wetgewing **wat raak** 'n saak ...	raak	bywoord
Eenheid #	**Deelwoorde as voorsetsels**	**Lemma**	**WS-etiket**
NCHLT-toetsteks eenheid 1 655	*Gedurende* die loop van hierdie jaar ... *wat duur die loop van hierdie jaar ...	gedurende	voorsetsel
eWAT	*Ingevolge* 'n wetlike bevoegdheid ... **Wat involg** 'n wetlike bevoegdheid ...	ingevolge	voorsetsel

Daar is slegs 'n handjievol deelwoorde wat funksioneer as bywoorde van skakeling of as voorsetsels en kan daarom akkuraat geannoteer word. Tabel 16 is gevolglik opgestel uit die NCHLT-korpus sowel as die NCHLT-toetsteks om 'n duideliker riglyn in die protokol neer te lê. Die e*WAT* is geraadpleeg vir voorbeeldsinne by *rakende* en *ingevolge* (uit die protokollys, vergelyk Figuur 38), aangesien die korpusse nie gebruiksvoorbeelde hiervan opgelewer het nie. Die onderskeid tussen deelwoorde as bywoorde en voorsetsels in Tabel 16 is natuurlik nie 'n volledige lys nie, maar sluit al die voorbeelde in wat in die NCHLT-projek voorkom (uit beide korpusse), of waarna in die protokol verwys word. Figuur 39 reflekteer die voorstelle ter verbetering van die protokoldokument.

- **Prepositions**
  - gedurende
  - ingevolge
  - ~~rakende~~
    - ~~Note that *rakende* should be lemmatised as *raak* in the context of *rakende hoeke*.~~
- It is not always easy to determine whether a participle functions as a preposition or as an adverb. In such cases:
  - If a natural verbal reading is possible, treat it as an adverb and lemmatise the adverbial participle.
  - If the verbal reading is ungrammatical, treat it as a preposition and do not lemmatise the participial preposition.

Figuur 39: Voorstel ter verbetering van die lemmaprotokol (CTexT, 2013b:14a)

5.4.5.5 Deelwoorde wat van naamwoorde gevorm is

In Hoofstuk 2 is aangetoon dat Kempen (1982:141-143) na woorde soos *bebaard* en *getiteld* as 'preverbale' verwys, terwyl in Hoofstuk 3 genoem is dat Booij (2002:77) daarna as pseudodeelwoorde verwys. In navolging van die argumente daar, is ek dit eens met die riglyn in die protokol dat hierdie woorde nie as deelwoorde hanteer behoort te word nie. Figuur 40 bevat 'n lys pseudodeelwoorde wat bloot as adjektiewe hanteer moet word (CTexT, 2013b:14).

- **Participles with noun stems**, and where the noun could not be interpreted as a verb (conversion)
  - getiteld
  - gevleueld
  - gemiddeld
  - geletterd
  - gesyferd
  - getand
  - bebaard
  - bebloed
  - beblaard
    - Note cases such as *ontmande* and *beboste* where *ontman* and *bebos* are the respective lemmas.

Figuur 40: Uittreksel uit die lemmaprotokol (CTexT, 2013b:14b)

Buiten *gemiddelde* wat twee keer in die NCHLT-korpus voorkom, is nie een van die ander voorbeelde in die protokollys (vergelyk Figuur 40) in die korpus gevind nie. In beide gevalle waar *gemiddelde* voorkom (NCHLT-korpus, eenhede 2799 en 3371) is dit korrek gelemmatiseer volgens die voorskrifte van die protokol.

5.4.6 Samevatting oor die protokol vir lemmatisering

Die protokol vir lemmatisering (CTexT, 2013a:14-15) sluit af met 'n samevatting van die manier waarop onvoltooide, swak voltooide en sterk voltooide deelwoorde gelemmatiseer behoort te word. Al die riglyne in hierdie samevatting is alreeds vroeër hanteer. Alle oorblywende deelwoorde in die toetsteks wat gevolglik nie deel uitmaak van een van die spesiale uitsluitklasse nie, is deelwoorde met 'n natuurlike verbale interpretasie. Al hierdie deelwoorde is korrek gelemmatiseer in die toetsteks.

Korrekte lemmatisering is uiters waardevol as 'n eerste vlak van annotering. As die lemmas korrek geannoteer word, maak dit die keuse van woordsoortetikettering as 'n volgende vlak van annotering veel makliker.

5.5 Woordsoortetikettering binne die NCHLT-projek

WS-etikettering behels die toekenning van leksikale en morfologiese kenmerke aan woorde in 'n spesifieke gebruikskonteks (Van Eynde, 2004:4). Van Eynde (2004:3) sit die vereistes uiteen waaraan die CGN se WS-etiketstel streef om te voldoen. Dit behels:

- WS-etikette moet inligting bevat wat ooreenstem met algemene gebruik (soos uiteengesit in die *ANS* (Geerts, *et al*., 1984)) om sodoende woordontleding te ondersteun;
- WS-etikette moet so na as moontlik aansluit by die heersende internasionale EAGLES-standaarde;
- Elke tekseenheid moet 'n eie lemma en 'n eie WS-etiket hê;
- WS-etikette moet as 'n geskikte basis dien vir hoër vlakke van taalkundige annotasie, soos byvoorbeeld vir sintaktiese analise; en
- Die notering moet oorsigtelik, kompak en maklik leesbaar wees.

Die tweede protokoldokument (CTexT, 2013c) binne die NCHLT-projek is opgestel vir WS-etikettering. Net soos met die protokoldokument vir lemmatisering, is die doel van die protokoldokument vir WS-etikettering ook om annoteerders te rig en te lei om die Afrikaanse korpus van die NCHLT-projek korrek te annoteer – hier net op woordsoortvlak. Die WS-etiketstel sowel as die protokoldokument vir WS-etikettering (CTexT, 2013c) is grootliks geskoei op Pilon (2005) se studie oor outomatiese Afrikaanse WS-etikettering. Daarom sal daar gereeld in die analise wat volg na Pilon se studie verwys word. Die protokoldokument vir WS-etikettering is, net soos wat dit die geval is vir die protokoldokument vir lemmatisering, ook 'n lewende dokument wat kan verander deur die regte prosedure te volg.

Die WS-etiketstel vir Afrikaans (CTexT, 2013c) maak glad nie voorsiening vir 'n aparte deelwoordetiket nie, aangesien deelwoorde slegs op grond van hul funksie in die sin geëtiketteer word. Die protokoldokument (CTexT, 2013c:1) verklaar dat 'n woordsoort "[t]he function that a word fulfils in a sentence (i.e. in a grammatical context)" behels, en dit word bepaal deur die morfosintaktiese gedrag van die woord in 'n spesifieke konteks (CTexT, 2013c:1). Die gevolg hiervan is dat tekseenhede soos *betroubare* (eenheid 2669, 'n gewone adjektief) en *gebaseer* (eenheid 1177, 'n voltooide deelwoord) beide

geëtiketteer is met die WS-etiket ASA: adjektief/stellend/attributief (vergelyk Tabel 18).

Tabel 17: Deelwoorde geëtiketteer as adjektiewe

Eenheid #	Eenheid	Lemma	WS-etiket	Sinskonteks
2669	betroubare	betroubaar	ASA	Om **betroubare** kragtoevoer te verseker, ...
1166	gebaseer	baseer	ASA	... wat tans in Londen **gebaseer** is.

Myns insiens is só 'n hantering van deelwoorde prakties, maar arm in die sin dat die werkwoordelike karakter van die deelwoord, waarvoor teoretiese sowel as beskrywende bevestiging in hoofstukke 3 en 4 gekry is, in die etikettering verlore gaan. Tog strook die etikettering van deelwoorde in die NCHLT-projek met wat Van Eynde (2004:4) gestel het, naamlik dat 'n WS-etiket die spesifieke gebruikskonteks van die woord moet reflekteer.

Omdat die bestaande WS-etiketstel glad nie deelwoordetikette insluit nie, is dit voor die hand liggend dat die protokoldokument vir etikettering ook nie verdere inligting en leiding bevat met betrekking tot deelwoorde nie (CTexT, 2013c).

5.5.1 Algemene beginsels van die WS-etiketteringsprotokol

Volgens die WS-etiketteringsprotokol (CTexT, 2013c:4) is die WS-etiketstel van die NCHLT-projek gebaseer op drie bronne, naamlik:

- Die EAGLES-standaarde (1996);
- CGN-dokument (Van Eynde, 2004); en
- *Outomatiese Afrikaanse Woordsoortetikettering* (Pilon, 2005).

Pilon (2005:21) regverdig die bovermelde bronne en verduidelik dat WS-etikette rekening behoort te hou met internasionaal erkende standaarde en bestaande etiketstelle. Deur rekening te hou met die EAGLES-standaarde se bestaande etiketgleuwe wat gereserveer word vir sekere eienskappe, verseker 'n mens die toeganklikheid en herbruikbaarheid van 'n etiketstel. Die CGN-etiketstel is óók op die EAGLES-standaarde gebaseer, en daarom kan dit volgens Pilon (2005:22) vir Afrikaans om drie redes voordelig wees om daarmee rekening te hou, te wete dat: (i) CGN van die nuutste en mees gevorderde WS-etiketstelle is; (ii) CGN 'n hoë vlak van spesifisiteit het;

en (iii) aangesien Nederlands en Afrikaans so nou verwant is aan mekaar, dit taalverwante insig vir Afrikaanse WS-etikette kan bied.

Van nader beskou is die WS-etiketstel van die NCHLT-projek identies aan die WS-etiketstel wat Pilon (2005) voorstel. Alhoewel haar voorstel ook op die EAGLES-standaarde gebaseer is en ook die CGN-etiketstel in ag geneem het, verskil dit radikaal in terme van ’n deelwoordetiket. Hierdie wanpassing tussen die Afrikaanse WS-etiketstel en ander WS-etiketstelle noop ’n vergelyking tussen die verskillende WS-etiketstelle om vas te stel waar en hoekom dit verskil (vergelyk afdeling 5.5.2). EAGLES (1996) het ook ’n parallelle WS-etiketstel vir Nederlands, EAGLES:Dutch (1996), wat ook in die vergelyking ingesluit word.

5.5.2 Vergelyking van werkwoordetikette tussen EAGLES, EAGLES:Dutch, CGN en NCHLT

Van Eynde (2004:65) verduidelik die EAGLES-standaarde as ’n drievoudige onderskeid wat gemaak word tussen verpligte eienskappe en waardes; aanbevole eienskappe en waardes; en opsionele taalspesifieke toevoegings. Die verpligte eienskap van EAGLES betrek slegs een eienskap, naamlik die woordklas wat geëtiketteer moet word; 13 waardes (woordklasse) word gespesifiseer. Omdat die deelwoord ’n vorm van die werkwoord is, word spesifiek die aanbevole eienskappe en waardes van werkwoorde in Tabel 18 met mekaar vergelyk om te bepaal waar die werkwoordetiket in Afrikaans van die ander werkwoordetikette verskil. In die vyfde kolom bied ek ’n voorstel aan vir waar Pilon se eienskappe en waardes moontlik aangepas kan word (die grysgemerkte verskille tussen die laaste twee kolomme by die aanbevole eienskappe en waardes in gleuwe (iv) en (v) word ná die tabel verduidelik).

Tabel 18: Vergelyking van die verpligte en aanbevole eienskappe en waardes tussen EAGLES (1996), EAGLES: Dutch (1996), die CGN (Van Eynde, 2004:65-66) en dié vir Afrikaans (Pilon, 2005:39)

Verpligte kenmerk				
EAGLES	**EAGLES: DUTCH**	**CGN**	**Pilon**	**Voorstel**
VERBS	VERBS	WERKWOORDEN	WERKWOORDE	WERKWOORDE
Aanbevole eienskappe met waardes				
EAGLES	**EAGLES: DUTCH**	**CGN**	**Pilon**	**Voorstel**
(i) Person 1. First 2. Second 3. Third	**Person** *first* *second* *third*	'Person' het nie 'n ekwivalent in die CGN-etiketstel nie. Onderskeid word verreken onder voornaamwoorde.	**(i) Persoon** 1. eerste 2. tweede 3. derde	**(i) Persoon** 1. eerste 2. tweede 3. derde
(ii) Gender 1. Masculine 2. Feminine 3. Neuter		'Gender' het nie 'n ekwivalent in die CGN-etiketstel nie aangesien Nederlandse werkwoorde geen variasie in genus vertoon nie.	**(ii) Genus** 1. manlik 2. vroulik 3. onsydig	**(ii) Genus** 1. manlik 2. vroulik 3. onsydig
(iii) Number 1. Singular 2. Plural	**Number** *singular* *plural*	'Number' word hanteer ondervoornaamwoorde (PVAGR) en naamwoorde (GETAL-N) vir nominaal gebruikte buigbare vorme.	**(iii) Getal** 1. enkelvoud 2. meervoud	**(iii) Getal** 1. enkelvoud 2. meervoud
(iv) Finiteness 1. Finite 2. Non-finite	**Verb-Form** *infinitive* *pres participle* *past participle* *finite* *imperative*	'Finiteness' korrespondeer met die onderskeid tussen persoonsvorme en buigbare vorme. Dit word onder 'Verb form/mood' hanteer (sien WVORM).	**(iv) Voltooidheid** 1. voltooid 2. onvoltooid	**(iv) Finietheid** 1. finiete ww. 2. infiniete ww.
(v) Verb form/mood 1. Indicative 2. Subjunctive 3. Imperative 4. Conditional 5. Infinitive 6. Participle 7. Gerund 8. Supine	**Mood (finite verbs)** *indicative* *subjunctive*	**Verb form/mood** **WVORM** infinitief deelwoord **PVTIJD** teenwoordige tyd verlede tyd	**v) Ww-vorm** 1. indikatief 2. subjunktief 3. imperatief 4. kondisioneel 5. infinitief 6. partikel 7. gerund 8. supine 9. -ing-vorm	**v) Ww.-vorm / wyse** 1. indikatief 2. subjunktief 3. imperatief 4. kondisioneel 5. infinitief 6. deelwoord 7. gerundium 8. supine

EAGLES	EAGLES: DUTCH	CGN	Pilon	Voorstel
(vi) Tense 1. Present 2. Imperfect 3. Future 4. Past	**Tense** *present* *past*	'Tense' korrespondeer met die onderskeid tussen teenwoordige en verlede tyd. Dit word onder werkwoordvorm hanteer (sien PVTIJD).	**(vi) Tyd** 1. teenwoordig 2. imperfektief 3. toekomend 4. verlede 5. gemarkeerd 6. ongemarkeerd	**(vi) Tempus** 1. teenwoordig 2. imperfektum 3. toekomend 4. verlede
(vii) Voice 1. Active 2. Passive		'Voice' is nie relevant vir Nederlands nie omdat dit nie morfologies gemarkeer word nie.	**(vii) Modus** 1. aktief 2. passief	**(vii) Passiwiteit** 1. aktief 2. passief
(viii) Status 1. Main 2. Auxiliary	**Verb-Type** *full verb* *auxiliary* *modal* *impersonal*	**Status** hoofdwerkwoorden hulpwerkwoorden	**(viii) Status** 1. hoof 2. mede 3. hulp	**(viii) Status /tipe** 1. hoof 2. hulp

Onder die verpligte kenmerk 'VERB', is daar agt aanbevole eienskappe met waardes vir elke eienskap. Die eerste drie eienskappe, (i) 'Person', (ii) 'Gender' en (iii) 'Number', lewer glad nie 'n probleem op nie. In intermediêre etikette vir deelwoorde sal hiérdie waardes telkens 0 wees aangesien dit nie op deelwoorde van toepassing is nie.

By die vierde eienskap, (iv) 'Finiteness', verskil Pilon van die ander deurdat haar interpretasie aspektualiteit betrek (wat nie een van die aanbevole eienskappe is nie), terwyl die gleuf eerder gereserveer is vir finietheid. Dit is ook by hierdie gleuf waar die deelwoord 'n eerste waarde ontvang, want deelwoorde en infinitiewe is voorbeelde van infiniete werkwoorde. Indien Pilon se bedoeling was om te verwys na deelwoorde, sou dit kon werk in die sin wat EAGLES:Dutch dit gedoen het. Die waardes (1) 'voltooid' en (2) 'onvoltooid' kon die tipes deelwoorde benoem, maar dan behoort 'infinitief' ook by haar as 'n waarde gespesifiseer te word en die gleuf behoort anders benoem te word.

Tog is dit by die vyfde eienskap, (v) 'Verb form/mood', waar die Afrikaanse interpretasie in terme van die deelwoord tekortskiet. By eienskap (v), waarde (5) is 'Participle' vertaal met 'partikel' en indien 'n mens wou vasstel waar en hoekom die Afrikaanse interpretasie van die werkwoordetiket van ander werkwoordetikette verskil, is dit presies hier en as gevolg van 'n per abuise vertaalfout. Dit is nie duidelik

wat Pilon met die laaste waarde, (9) '*-ing*-vorm', bedoel nie, en dit mag wees dat die 'gerundium' bedoel word wat reeds in waarde (7) gelys is. Hoe dit ookal sy, waarde (9) hoort nie daar nie. Die enigste twee waardes wat by die vyfde eienskap relevant is vir Afrikaans, is (5) 'infinitief' en (6) 'deelwoord'; dit mag ook wees dat (7) 'gerundium' vir Afrikaans relevant is, maar ek is nie self daarvan oortuig nie en laat dit vir eers daar, aangesien dit buite die skopus van hierdie studie val.

Reeds in terme van eienskappe (iv) en (v) waar die Afrikaanse eienskappe en waardes nie met die ander etiketstelle klop nie, skakel dit 'n sinvolle intermediêre etiket vir die deelwoord in Afrikaans met die bestaande werkwoordetiket uit. Daar is nog verdere verskille op te merk tussen Pilon se eienskappe en waardes vir die werkwoord en ander etiketstelle (vergelyk Tabel 18), maar so 'n analise val buite die omvang van hierdie studie.

In terme van hierdie wanpassing tussen die verskillende etiketstelle, moet dit in gedagte gehou word dat hierdie etiketstelle vir Engels en Nederlands opgestel is en in Engels en Nederlands kán die deelwoordvorm van die werkwoord in 'n werkwoordfunksie voorkom, terwyl dit in Afrikaans nooit die geval is nie. Die vraag kan nou ontstaan hoe die ander etiketstelle die deelwoord hanteer om voorsiening te maak daarvoor dat die deelwoord nie noodwendig altyd in 'n werkwoordfunksie voorkom nie (soos wat dit dan altyd die geval in Afrikaans is).

Ter illustrasie kan ons kyk hoe die EAGLES:Dutch-werkwoordetiket byvoorbeeld 'n deelwoord sou hanteer wat in 'n adjektiwiese funksie gebruik word, byvoorbeeld *die* ***huilende*** *babas*. Hoe sou die intermediêre etiket van *huilende* volgens die aanbevole eienskappe en waardes lyk?

As daar na die aanbevole eienskappe gekyk word, sal die eerste drie eienskappe met 0 (nul) gemerk word, aangesien hulle nie op 'n deelwoord in 'n adjektiwiese funksie van toepassing is nie. Eienskap (iv) 'Finiteness' sal met die waarde 2 'Non-finite' gemerk word, en dan sal eienskap (v) 'Verb form/mood' met die waarde 6 'Participle' ook gemerk word. Verder sal die eienskappe (vi) 'Tense', (vii) 'Voice' en (viii) 'Status' ook almal met 'n 0 gemerk word, want ook nie een van hulle is van toepassing op 'n deelwoord in 'n adjektiewe funksie nie. Hiervolgens sal die intermediêre etiket van *huilende* dan V00026000 wees.

Buiten dat die meeste waardes by die werkwoordeienskappe in die intermediêre etiket 0 sal wees (wat dit alreeds van ander werkwoorde onderskei), is daar nie ’n ander aanbevole eienskap wat ’n deelwoord se adjektiwiese (of ander) funksie erken nie. Wanneer die opsionele eienskappe en waardes van die werkwoordetiket bestudeer word (vergelyk Tabel 19), word daar in die EAGLES:Dutch ’n verdere eienskap gestel ‘Use (non-finite)’ waar die verbale, adjektiwiese of naamwoordelike funksies van die deelwoord gemerk kan word. Die opsionele eienskappe en waardes is, anders as die aanbevole eienskappe en waardes, gleuwe waar taalspesifieke eienskappe en waardes toegevoeg kan word en waar die gleuwe nie so spesifiek gereserveer is vir sekere eienskappe en waardes nie.

In die laaste kolom in Tabel 19, waarin ek voorstelle maak, is hierdie eienskap as (xiv) ‘Gebruik (infiniete)’ voorgestel met die waardes (1) ‘adjektief attributief’, (2) ‘adjektief predikatief’, (3) ‘bywoord’, (4) ‘voorsetsel’ en (5) ‘naamwoord’. Die funksie van die deelwoord kan dus maklik met hierdie eienskap en waardes gespesifiseer word. Daar is ’n vraagteken agter ‘voorsetsel’ by waarde (4), aangesien daar in die afdeling oor lemmatisering (vergelyk 4.5.4.5) geargumenteer is dat ’n deelwoord in ’n voorsetselfunksie as ’n voorsetsel gemerk word. Waarde (5) ‘naamwoord’ is nie by hierdie eienskap gevoeg vir die deelwoord se onthalwe nie, maar is voorlopig daar vir gevalle soos ’n gerundium of ’n infinitief wat moontlik in ’n naamwoordfunksie gebruik kan word.

Tabel 19: Vergelyking van die opsionele eienskappe en waardes tussen EAGLES (1996), EAGLES: Dutch (1996), die CGN (Van Eynde, 2004:65-66) en dié vir Afrikaans (Pilon, 2005:39)

Opsionele eienskappe				
EAGLES	**EAGLES: DUTCH**	**CGN**	**Pilon**	**Voorstel**
			(ix) Aspek 1. perfektief 2. imperfektief	**(ix) Tipe deelwoord** 1. voltooid 2. onvoltooid
	Separability *separable*		**(x) Skeibaarheid** 1. skeibaar 2. onskeibaar	**(x) Skeibaarheid** 1. skeibaar
			(xi) Refleksiwiteit 1. reflektief 2. onreflektief	**(xi) Gemarkeerdheid** 1. gemarkeerd 2. ongemarkeerd

EAGLES	EAGLES: DUTCH	CGN	Pilon	Voorstel
	Main-Verb Func *intransit* *transit* *reflex*		**(xii) Mede-ww.** 1. het 2. is	**(xii) Hoofww.-funksie** 1. oorganklik 2. onoorganklik 3. koppelwerkwoord 4. voorsetselwerkwoord
	Auxiliary *hebben* *zijn* *hebben or zijn*		**(xiii) Mww-funksie** 1. primêr 2. modaal	**(xiii) Hulpww.-funksie** 1. hulpww. van tyd 2. hulpww. van vorm 3. hulpww. van modaliteit
	Use (non-finite) *Verbal* *Adjectival* *nominal*		**(xiv) Ww-tipe** 1. oorganklik 2. onoorganklik 3. koppel 4. voorsetsel 5. hulp-modaliteit 6. hulp-tyd 7. hulp-aspek 8. hulp-modus	**(xiv) Gebruik (infiniete)** 1. adjektief attributief 2. adjektief predikatief 3. bywoord 4. voorsetsel (?) 5. naamwoord
	Inversion *Inverted*			
	Word order separable verbs *main clause* *sub-clause*			
	Politeness			

Die voorstelle wat ek in Tabel 18 en Tabel 19 gemaak het (in navolging van EAGLES:Dutch), is slegs ter illustrasie van waar ek vermoed die gleuwe vir genoegsame deelwoordetikettering verlore gegaan het in die Afrikaanse werkwoordetiket van Pilon en die NCHLT. Indien die opsionele eienskappe en waardes in die intermediêre etiket bygereken word volgens wat ek voorstel, sal eienskap (ix) 'Tipe deelwoord' met die waarde 2 gemerk word, die volgende vier eienskappe met 0 aangesien hulle nie op die deelwoord van toepassing is nie, en laastens die eienskap (ivx) 'Gebruik (infiniete)' met 'n waarde 1 gemerk word. 'n Volledige intermediêre etiket vir *huilende* in *die **huilende** babas*, sal dan V00026000200001 wees.

Dit sou dus moontlik wees om 'n deelwoord bevredigend te etiketteer met ander bestaande WS-etiketstelle sodat die interne verbale karakter sowel as die funksie daardeur gereflekteer word. Behalwe vir die praktiese oorweging om deelwoorde te etiketteer volgens die funksie waarbinne hulle gebruik word, kon ek nie enige ander redenasie vind waarom Pilon (2005) of die NCHLT-projek (2013) gekies het om deelwoorde nie onder werkwoorde te hanteer nie. Nieteenstaande die rede(s) vir die oorweging om die deelwoord volledig onder funksie te hanteer, wil dit voorkom asof die

hele werkwoordetiket vir Afrikaans verdere wetenskaplike ondersoek verdien, maar val buite die fokus van hierdie studie.

5.5.3 Die etikettering van die deelwoord in die NCHLT-projek

Dit is reeds genoem dat die deelwoord in die NCHLT-projek volledig hanteer is volgens die funksie waarin dit gebruik word (vergelyk afdeling 5.3). So is deelwoorde wat as adjektiewe funksioneer as adjektiewe geëtiketteer, en deelwoorde wat as bywoorde funksioneer as bywoorde geëtiketteer. Ofskoon 'n wanpassing uitgewys is in die vorige afdeling (vergelyk afdeling 5.5.2) tussen die werkwoordetiket vir Afrikaans en die ander internasionaal erkende etiketstelle, klop die hantering van deelwoorde in Afrikaans met hoe die CGN (Van Eynde, 2004:3-4) WS-etikettering definieer, naamlik as die toekenning van WS-etikette binne 'n spesifieke gebruikskonteks. Die CGN bied ook verdere regverdiging hiervoor in die vierde vereiste waarna dit streef om te voldoen, naamlik dat die WS-etikette as 'n geskikte basis moet dien vir hoër vlakke van taalkundige annotasie, soos byvoorbeeld vir sintaktiese analise (Van Eynde, 2004:3). Aansluitend hierby, is daar reeds in hierdie studie uitgewys dat die deelwoordvorm van die werkwoord in beide Engels en Nederlands ook in werkwoordfunksies optree, terwyl dit nooit die geval in Afrikaans is nie.

Die deelwoord is in hierdie studie ook telke male beskryf as 'n transkategoriale kategorie (Booij, 2002:79), 'n woord waarvan die morfologie 'n intermediêre karakter vertoon (Langacker, 1987:145; 2008a:119-120). Dit is derhalwe 'n vorm van die werkwoord wat, sonder om sy verbale karakter te verloor, as 'n ander woordklas optree. In afdeling 5.3 is gesuggereer dat, afhangend van 'n mens se perspektief, verskillende karaktertrekke van die deelwoord op die voorgrond gaan tree. Wanneer 'n mens vanuit 'n fonologiese, morfologiese of semantiese hoek na deelwoorde kyk, tree die verbale karakter van die deelwoord op die voorgrond – in KG-terme sou mens dit kon stel dat die werkwoordbasis in die kolliggedeelte is. Wanneer die deelwoord egter beskou word vanuit 'n WS-etiketteringsperspektief, tree die funksie van die woord in 'n spesifieke sintaktiese konteks op die voorgrond, wat juis Pilon (2005) en die NCHLT-projek (2013) se hantering van die deelwoord bruikbaar maak.

'n Laaste regverdiging van die manier waarop die deelwoord in die projek geëtiketteer is, lê daarin dat etiketstelle verskillende grade van granulariteit (Van Eynde, 2004:5)

kan vertoon. ’n Growwe wyse van WS-etikettering sou etikettering impliseer waar slegs die woordklasse vir die etikette gebruik word (d.w.s. slegs die V vooraan die etiket word gebruik wat die tekseenheid etiketteer as ’n werkwoord). So ’n growwe WS-etikettering lewer beperkte granulariteit. ’n Fyner WS-etikettering, daarenteen, sou etikettering wees met die volledige etiket (dus met verskillende onderskeidende waardes) wat ’n hoë vlak van granulariteit lewer. In terme van beperkte granulariteit (growwe WS-etikettering) sou die bestaande etikettering van deelwoorde volgens funksie (Pilon, 2005; CTexT, 2013c), beter resultate lewer as wanneer deelwoorde met ’n werkwoordetiket hanteer sou word. Tydens growwe WS-etikettering sal daar met die bestaande WS-etikette geen vreemde taalstrukture na vore kom nie. Vergelyk hoe growwe etikettering van sin (67a) sou lyk indien die deelwoord as ’n V vir werkwoord (vergelyk (67b)) of ’n B vir bywoorde (vergelyk (67c)) geëtiketteer sou word. Die growwe etikettering van (67b) pas nie in die taalmodel van Afrikaans waar twee werkwoorde op dié wyse langs mekaar staan nie. Die growwe etikettering van (67c) daarenteen, lewer glad nie ’n probleem vir ’n taalmodel of enige verdere sintaktiese ontleding nie.

(67a) *Hy loop **singend** die deur uit*

(67b) P V **V** D N S

(67c) P V **B** D N S[47]

5.5.4 Analise van die etikettering van die deelwoord in die NCHLT-toetsteks

Die analise van die etikettering van deelwoorde in die NCHLT-toetsteks volg aanvanklik dieselfde volgorde as waarin die riglyne vir die lemmatisering hanteer is. Daarna word die oorblywende deelwoorde in die NCHLT-toetsteks beskou. In die tabelle in hierdie afdeling word daar telkens ’n alternatiewe etiket langs die bestaande etiket gelys om te illustreer hoe die WS-etiket sou verskil indien die deelwoord ’n eie deelwoordetiket onder werkwoorde gehad het (soos geïllustreer met die EAGLES:Dutch WS-etiketstel in afdeling 5.5.2). Die afkortings wat vir die alternatiewe WS-etikette gebruik word, volg die patroon van die gleuwe eienskappe en waardes wat in die werkwoordetiket vir die deelwoord geld. ‘V’ staan vir die verpligte kenmerk ‘werkwoord’ en ‘D’ staan vir die

[47] P = voornaamwoord, V = werkwoord, B = bywoord, D = determineerder, N = naamwoord, S = voorsetsel

aanbevole eienskap ‘deelwoord’. Die twee verdere opsionele eienskappe word aangetoon met ‘V’ (voltooid) of ‘O’ (onvoltooid) in terme van die eienskap ‘Aspek’, en ‘AA’ (adjektief, attributief), ‘AP’ (adjektief, predikatief), ‘B’ (bywoord) of ‘S’ (voorsetsel) in terme van die eienskap ‘Gebruik’.

Tabel 20 illustreer die etikettering van naamwoorde van deelwoordbasisse (vergelyk ook afdeling 5.4.2). Die bestaande afkortings vir die NCHLT-etikette moet gelees word as: NA (naamwoord, abstrak), NSE (naamwoord, soortnaam, enkelvoud), NSM (naamwoord, soortnaam, meervoud), NEE (naamwoord, eienaam, enkelvoud). Vier uit die vyf is sonder twyfel korrek. Die laaste tekseenheid, *Beherend*, is problematies aangesien dit ’n deelwoord is wat in hierdie spesifieke geval deel uitmaak van ’n eienaam. Dit sal meer sin maak om hierdie tekseenheid eerder te etiketteer as NEE.

Tabel 20: Etikettering van naamwoorde met deelwoordbasisse

Eenheid #	Eenheid	NCHLT-etiket		Alternatiewe etiket	Sinskonteks
53	volgende	NA	✓	VDON	Doen die **volgende** in die afdeling ...
71	volgende	NA	✓	VDON	Verstrek die **volgende** in die afdeling ...
386	nie-ingesetene	NSE	✓		’n nie-inwonende individu (’n **nie-ingesetene**)
1592	uitgewekenes	NSM	✓		... en die terugkeer van **uitgewekenes**.
2725	Beherend	NSE	NEE		... Eskom **Beherend** sal funksioneer.

Tabel 21 illustreer die korrekte etikettering van al die tekseenhede van woorde van deelwoordbasisse met ontkenningsmorfeme (vergelyk ook afdeling 5.4.3). Die ontkenningsmorfeem maak dat die woorde nie meer as deelwoorde gereken word nie, en daarom sou die alternatiewe etikette nie van die NCHLT-etikette verskil nie. Die korthandafkortings vir die etikette in Tabel 21, moet gelees word as: ASA (adjektief, stellend, attributief), ASP (adjektief, stellend, predikatief), NSE (naamwoord, soortnaam, enkelvoud), en BS (bywoord, stellend).

Tabel 21: Etikettering van woorde van deelwoordbasisse met ontkennings-morfeme

Eenheid #	Eenheid	NCHLT-etiket		Alternatiewe etiket	Sinskonteks
382	nie-inwonende	ASA	✓		’n **nie-inwonende** individu
386	nie-ingesetene	NSE	✓		’n ... individu (’n **nie-ingesetene**)
1029	onvermoeide	ASA	✓		... Suid-Afrikaners se **onvermoeide** stryd.
1484	onvermoeid	BS	✓		... en **onvermoeid** bepleit het.

1551	onverpoosd	BS	✓		... wat haar **onverpoosd** vir verandering beywer het.
1786	onvermoeide	ASA	✓		Julle **onvermoeide** en heldhaftige opofferings ...
4537	onvoldoende	ASA	✓		... **onvoldoende** infrastruktuur.
5807	onbesonge	ASP	✓		... aan al ons helde en heldinne, besonge en **onbesonge**, bekend en **onbekend** ...
5811	onbekend	ASP	✓		

In Tabel 22 word aangetoon dat beide deelwoorde van partikelwerkwoorde korrek geëtiketteer is volgens die NCHLT-voorskrifte, maar dat die alternatiewe etiketterings sal verskil by albei woorde. Die afkortings vir die etikette in die tabel moet gelees word as: ASA (adjektief, stellend, attributief) en ASP (adjektief, stellend, predikatief. Die korthandafkortings vir die alternatiewe etikette, moet gelees word as: VDVAA (werkwoord, deelwoord, voltooid, adjektief, attributief) en VDVAP (werkwoord, deelwoord, voltooid, adjektief, predikatief).

Tabel 22: Etikettering van deelwoorde van partikelwerkwoorde

Eenheid #	Eenheid	NCHLT-etiket		Alternatiewe etiket	Sinskonteks
2264	uitgebreide	ASA	✓	VDVAA	... die **Uitgebreide** Openbare Werke-program ...
4898	toegewyd	ASP	✓	VDVAP	... staatsamptenare wat **toegewyd** en bekwaam is ...

Al die geleksikaliseerde adjektiewe wat in die lemmaprotokol gelys is (vergelyk afdeling 5.4.5.1), die deelwoorde met adjektiwiese prefikse (vergelyk afdeling 5.4.5.2), sowel as deelwoorde in samestellende afleidings (vergelyk afdeling 5.4.5.3), is almal korrek as adjektiewe geëtiketteer volgens die voorskrifte van die NCHLT-projek. ’n Alternatiewe deelwoordetiket sou ook nie by een van hierdie tekseenhede ’n verskil gemaak het nie.

Tabel 23 som die WS-etikette op wat gegee en voorgestel is vir deelwoorde in ’n bywoordelike en voorsetselfunksie (vergelyk ook afdeling 5.4.5.4). Die afkortings vir die bestaande WS-etikette in die tabel moet gelees word as: SVS (voorsetsel) en BS (bywoord, stellend). Die afkorting vir die alternatiewe WS-etiket moet gelees word as: VDOB (werkwoord, deelwoord, onvoltooid, bywoord). Omdat dit nie altyd maklik is om tussen deelwoorde in ’n voorsetselfunksie en dié in ’n bywoordfunksie te onderskei nie, is daar in afdeling 5.4.5.4 ’n voorstel gemaak om in gevalle waar ’n verbale parafrase natuurlik is, dit as bywoorde te merk, en die waar ’n verbale parafrase nie natuurlik is

nie, as voorsetsels te merk. Hiervolgens sou ek drie items volgens die NCHLT-voorskrifte anders etiketteer (vergelyk slegs een regmerkie in Tabel 23). Aangesien deelwoorde wat nie meer ’n natuurlike verbale parafrase toelaat nie as voorsetsels gemerk word, sal daar nie ’n ander WS-etiket vir voorsetsels in die alternatiewe kolom wees nie.

Tabel 23: Etikettering van deelwoorde wat as voorsetsels of bywoorde optree

Eenheid #	Eenheid	NCHLT-etiket		Alternatiewe etiket	Sinskonteks
406 4044	insluitende	SVS	BS	VDOB	... direkteure, **insluitende** die besturende direkteur ...
4064	bestaande	SVS	BS	VDOB	... **bestaande** uit nege provinsies ...
1655	gedurende	SVS	✓	SVS	**Gedurende** die loop van hierdie jaar ...
1004	sodoende	BS	SVS	SVS	... en **sodoende** ’n waterskeidingsoomblik ...

Buiten die tekseenhede wat alreeds in hierdie afdeling behandel is, is daar nog 48 deelwoorde in die NCHLT-toetsteks wat vergelyk kan word. Vier hiervan is sterk verlede deelwoorde (*betrokke*, *beslote*, *verbonde* en *besonge*) wat nie ’n verskil in etikettering gaan oplewer nie, aangesien sterk werkwoorde soos geleksikaliseerde adjektiewe hanteer word. In agt tekseenhede kom die deelwoord in samestellende afleidings voor (*onderstaande*, *bykomende*, *deurslaggewende*, *welvarende*, *arbeids-absorberende*, *toonaangewende*, *gemeenskapsgebaseerde*, *tuisgebaseerde*), met die gevolg dat ook by hierdie groep geen verskil sal wees nie.

Tabel 24 wys die verskillende WS-etikette wat ’n alternatiewe werkwoordetiket sou lewer vir werkwoorde wat gevorm is met *ge-*, *be-*, *ver-* en *-eer*. Die afkortings vir die bestaande WS-etikette in die tabel moet gelees word as: ASA (adjektief, stellend, attributief) en VVHOG (werkwoord, verledetyd, hoofwerkwoord, onskeibaar, oorganklik). Die afkortings vir die alternatiewe etikette stel voor: werkwoord, deelwoord, voltooid, attributief en dan wissel die laaste letter afhangend of dit attributief of predikatief gebruik is.

Tabel 24: WS-etikette wat ’n alternatiewe werkwoordetiket sou lewer

Eenheid #	Eenheid	NCHLT-etiket		Alternatiewe etiket	Sinskonteks
182	geregistreerde	ASA	✓	VDVAA	... die **geregistreerde** naam ...
1166	gebaseer	VVHOG	ASP	VDVAP	... wat tans in London **gebaseer** is ...
1687	verenigde	ASA	✓	VDVAA	... tot ’n **verenigde** staat gelei.

4657					
3111	bepaalde	ASA	✓	VDVAA	... wat vir 'n **bepaalde** uitkoms ...
3118 3313	gedetailleerde	ASA	✓	VDVAA	... 'n **gedetailleerde** leweringsooreenkoms ...
3510	bekwame	ASA	✓	VDVAA	... en **bekwame** werksmag te verseker ...
3605	gekwalifiseerde	ASA	✓	VDVAA	... om die aantal **gekwalifiseerde** Wiskunde- en Wetenskaponderwysers ...

Ten slotte word die oorblywende deelwoorde beskou waarvan al die tekseenhede korrek geëtiketteer is volgens die voorskrifte van die NCHLT-projek. Tabel 25 illustreer hoe 'n alternatiewe werkwoordetiket 'n fyner aspektuele onderskeid sou kan lewer tussen verskillende deelwoorde, maar steeds die funksie aantoon waarbinne dit gebruik word. Die bestaande WS-etiket (ASA) dui aan: adjektief, stellend, attributief; en die twee alternatiewe WS-etikette wat hier voorgestel word, verskil van mekaar op grond van die aspektuele onderskeid: voltooid of onvoltooid. VDVAA moet lees 'werkwoord, deelwoord, voltooid, adjektief, attributief' en VDOAA moet lees 'werkwoord, deelwoord, onvoltooid, adjektief, attributief'.

Tabel 25: Fyner onderskeid wat 'n alternatiewe werkwoordetiket sou lewer

Eenheid #	Eenheid	NCHLT-etiket		Alternatiewe etiket	Sinskonteks
289	gewaarmerkte	ASA	✓	VDVAA	... **gewaarmerkte** kopieë ...
1464	onderhandelde	ASA	✓	VDVAA	... strewe na 'n **onderhandelde** skikking ...
2686	geïntegreerde	ASA	✓	VDVAA	... **geïntegreerde** hulpbronplan ...
3951	gesteelde	ASA	✓	VDVAA	... om **gesteelde** goedere te koop.
4662	ontwikkelde	ASA	✓	VDVAA	... wat die **ontwikkelde** wêreld ...
408	besturende	ASA	✓	VDOAA	... die **besturende** direkteur ...
446	ondersoekende	ASA	✓	VDOAA	... in **ondersoekende** onderhoudvoering ...
1262	regerende	ASA	✓	VDOAA	... van die **regerende** party ...
1431	uitmuntende	ASA	✓	VDOAA	... 'n **uitmuntende** voorbeeld van ...
1806	oorblywende	ASA	✓	VDOAA	... die **oorblywende** jare van my lewe ...
2658	mededingend	ASA	✓	VDOAA	... dat ons spoornetwerk **mededingend** is ...
2710	stygende	ASA	✓	VDOAA	... teen **stygende** elektrisiteitspryse ...
2971	onderskeidende	ASA	✓	VDOAA	Die **onderskeidende** kenmerk ...
3007	Uitvoerende	ASA	✓	VDOAA	... die Uitvoerende Gesag en die ...
4367	Omvattende	ASA	✓	VDOAA	**Omvattende** Landelike- ontwikkelings-program ...
4534	lekkende	ASA	✓	VDOAA	... water deur **lekkende** pype ...
4731	bindende	ASA	✓	VDOAA	... 'n wettig **bindende** verdrag ...
5561	volgende	ASA	✓	VDOAA	... vir die **volgende** paar maande ...

5.7 Samevatting

Binne die NCHLT-projek (CTexT, 2010-2013), waar daar ongeveer 50 000 woorde in elk van die tien van die amptelike landstale op vier vlakke geannoteer is, het die annotering van deelwoorde in Afrikaans baie vrae opgelewer. Die doel van hierdie hoofstuk was om te bepaal of ’n groter insig in wat die deelwoord in Afrikaans behels, enige alternatiewe lemma- of woordsoortannotering binne die NCHLT-projek sou teweegbring.

Die belangrikste voorstelle wat vir die lemmaprotokol gemaak is, is om (i) te onderskei tussen die PK-vorme van die werkwoord (die verledetyd- en passiefkonstruksies) en die deelwoordvorm; (ii) om die riglyn vir die lemmatisering van deeltjiewerkwoorde te herhaal by die riglyn vir deelwoorde; (iii) om nog geleksikaliseerde adjektiewe tot die lys in die protokol by te voeg; (iv) en om duideliker te onderskei tussen wanneer ’n deelwoord as ’n voorsetsel of as ’n bywoord optree.

Die protokoldokument vir WS-etikettering is volgende in oënskou geneem. Daar bestaan geen WS-etiket vir deelwoorde in die NCHLT se WS-etiketstel nie, en gevolglik is daar weinig oor die deelwoord in die protokoldokument. Daarom is die WS-etiketstel vir Afrikaans vergelyk met EAGLES en die CGN om vas te bepaal hoekom Afrikaans, anders as die ander WS-etiketstelle, nie deelwoorde spesifiek etiketteer nie. Twee moontlike antwoorde is verkry: aan die een kant is ’n vertaalfout in die WS-etiketstel vir Afrikaans uitgewys wat moontlik daartoe bygedra het dat daar nooit ’n aparte waarde vir ’n deelwoord in die WS-etiketstel opgeneem is nie. ’n Ander moontlikheid is dat daar vir die Afrikaanse WS-etiketstel besluit is om suiwer op grond van die funksie van ’n tekseenheid te etiketteer. Wat ookal die beweegrede was, is daar aangetoon dat dit moontlik sou wees om ’n intermediêre etikette vir deelwoorde te skep wat in pas is met ander WS-etiketstelle sonder om die funksie waarin die deelwoord optree, prys te gee.

Ten slotte is die wyse waarop deelwoorde binne die NCHLT-projek hanteer is, oorweeg en ook vergelyk met ’n alternatiewe wyse van WS-etikettering. Hierdie alternatiewe wyse om deelwoorde te hanteer, is ’n voorstel wat ontwerp is in lyn met EAGLES:Dutch. Die wyse waarop die NCHLT-voorskrifte deelwoorde hanteer (etiketteer hulle volgens funksie) maak natuurlik sin uit ’n praktiese oorweging – wat ’n WS-soortetiketteerder per definisie is. Tog, behalwe in gevalle waar growwe etikettering gedoen word, beloof

die voorgestelde alternatiewe eienskappe en waardes om deelwoorde ryker te etiketteer sonder om die funksie waarin die deelwoord optree, te versaak.