Benoemde-entiteitherkenning vir Afrikaans

G. D. Matthew

Benoemde-entiteitherkenning vir Afrikaans

G. D. Matthew

Skripsie voorgelê ter gedeeltelike nakoming van die vereistes vir die graad *Magister Artium in Algemene Taal- en Literatuurwetenskap* aan die Vaaldriehoek-kampus van die Noordwes-Universiteit

Studieleier: Me. S. Pilon

Mede-Studieleier: Prof. J.C. Roux

Oktober 2013

Bedankings

Ek wil graag die volgende persone bedank:

- My studieleier Suléne Pilon en my mede-studieleier Prof. Justus Roux vir al die geduld en raad en waarsonder ek nie my skripsie sou kon voltooi het nie.
- Prof. Jan-Louis Kruger wat altyd gewillig was om my ’n af daggie of twee te gee as ek aan my skripsie wou werk.
- Dankie aan my vriende en familie wat altyd gewillig was om my te help al het hulle nie rêrig geweet waaroor my skripsie gaan nie.
- Prof. Bertus Van Rooy wat altyd gewillig was om hulp te verleen hetsy t.o.v. kennis of die reël van finansiële ondersteuning.
- Dan wil ek God die eer gee wat my die vermoë gegee het om tot hier te kom.
- Laastens, indien daar iemand is wie ek uit gelaat het, BAIE DANKIE.

Opsomming

Benoemde-entiteitherkenning vir Afrikaans

Die Grondwet van Suid-Afrika vereis van die regering om alle inligting in die tien inheemse tale van Suid-Afrika (uitsluitende Engels), beskikbaar te stel. Daarom het die regering die inligting wat alreeds vir tien tale bestaan, vrylik aan die publiek beskikbaar gestel en 'n poging word ook aangewend om die hoeveelheid inligting wat beskikbaar is in hierdie tale te vermeerder (Groenewald & Du Plooy, 2010). Hierdie bekendstelling van inligting help dan ook om Krauwer (2003) se idee te volg waar 'n inventaris voorgestel word vir die minimale aantal taalverwante hulpbronne wat nodig is vir 'n taal om kompeterend op die vlak van navorsing en onderrig wat bekend staan as die "Basic Language Resource Kit" (BLARK). Aangesien meeste van die tale in Suid-Afrika hulpbronskaars is, is dit in die beste belang vir die kulturele groei van die land, om vir elk van die inheemse Suid-Afrikaanse tale 'n BLARK te ontwikkel.

In Hoofstuk 1 word die noodsaaklikheid vir die ontwikkeling van 'n implementeerbare benoemde-entiteitherkenner (BEH) vir Afrikaans bespreek deur eerstens te verwys na die Grondwet van Suid-Afrika (Republic of South Africa, 2003) se taalbeleid. Tweedens word die idee van 'n BLARK (Krauwer, 2003) vir Suid-Afrikaanse tale bespreek wat gevolg word deur 'n oudit wat fokus op die aantal hulpbronne en verspreiding van mensliketaaltegnologie vir al elf Suid-Afrikaanse tale (Sharma Grover *et al*., 2010). Sharma Grover *et al*. (2010) bevestig dat daar 'n tekort aan teksgebaseerde-hulpmiddels vir Afrikaans is. Hierdie studie fokus dan om die behoefte aan teksgebaseerde-hulpmiddels te bevredig, deur te fokus op die ontwikkeling van 'n benoemde-entiteitherkenner vir Afrikaans.

In Hoofstuk 2 word 'n beskrywing gegee van wat 'n entiteit en 'n benoemde entiteit is. Verder in die hoofstuk word die proses van tegnologieherwinning verduidelik met behulp van ander studies waar die idee van tegnologieherwinning suksesvol toegepas is (Rayner *et al.*, 1997). Laastens word verskille tussen Afrikaanse- en Nederlandse benoemde entiteite bespreek. Hierdie verskille is vervolgens in drie kategorieë verdeel, naamlik: identiese kognate, nie-identiese kognate en onverwante entiteite.

Hoofstuk 3 begin met 'n beskrywing van *Frog* (Van den Bosch *et al*., 2007), die Nederlandse BEH wat in hierdie studie gebruik is, en die funksies en werking van die benoemde-entiteitsherkenningskomponent daarvan. Daarna volg 'n beskrywing van die Afrikaans-na-Nederlands-omskakelaar (A2DC) (Van Huyssteen & Pilon, 2009) en laastens word die verskillende eksperimente wat uitgevoer is, uiteengesit.

Die studie bestaan uit ses eksperimente waarvan die eerste is om te bepaal wat die resultate van *Frog* op die Nederlandse data is. Die tweede eksperiment evalueer die effektiwiteit van *Frog* op onveranderde (rou) Afrikaanse data. Die volgende twee eksperimente evalueer die resultate van *Frog* op vernederlandsde data. Die laaste twee eksperimente evalueer die effektiwiteit van *Frog* op rou en vernederlandsde Afrikaanse data met die byvoeging van gazetteers (of te wel naamlyste) as deel van die preprosesseringstap.

Ter samevatting word ondermeer vergelykings getref tussen die benoemde-entiteitherkenner vir Afrikaans wat in hierdie studie ontwikkel is en die benoemde-entiteitherkenningsafdeling wat in Puttkammer (2006) se tekseenheididentifiseerder vir Afrikaans. Daar word ook ten slotte ’n paar voorstelle vir toekomstige navorsing voorgestel.

SLEUTELWOORDE:

BENOEMDE-ENTITEITHERKENNING, FROG, BLARK, GAZETTEERS, AFRIKAANS, ENTITEITE, INLIGTINGONTTREKKING, TEGNOLOGIEHERWINNING.

Summary

Named Entity Recognition for Afrikaans

According to the Constitution of South Africa, the government is required to make all the information in the ten indigenous languages of South Africa (excluding English), available to the public. For this reason, the government made the information, that already existed for these ten languages, available to the public and an effort is also been made to increase the amount of resources available in these languages (Groenewald & Du Plooy, 2010). This release of information further helps to implement Krauwer's (2003) idea that there is an inventory for the minimal number of language-related resources required for a language to be competitive at the level of research and teaching. This inventory is known as the "Basic Language Resource Kit" (BLARK). Since most of the languages in South Africa are resource scarce, it is of the best interest for the cultural growth of the country, that each of the indigenous South African languages develops their own BLARK.

In Chapter 1, the need for the development of an implementable named entity recogniser (NER) for Afrikaans is discussed by first referring to the Constitution of South Africa's (Republic of South Africa, 2003) language policy. Secondly, the guidelines of BLARK (Krauwer, 2003) are discussed, which is followed by a discussion of an audit that focuses on the number of resources and the distribution of human language technology for all eleven South African languages (Sharma Grover, Van Huyssteen & Pretorius, 2010). In respect of an audit conducted by Sharma Grover *et al*. (2010), it was established that there is a shortage of text-based tools for Afrikaans. This study focuses on this need for text-based tools, by focusing on the development of a NER for Afrikaans.

In Chapter 2 a description is given on what an entity and a named entity is. Later in the chapter the process of technology recycling is explained, by referring to other studies where the idea of technology recycling has been applied successfully (Rayner *et al*., 1997). Lastly, an analysis is done on the differences that may occur between Afrikaans and Dutch named entities. These differences are divided into three categories, namely: identical cognates, non-identical cognates and unrelated entities.

Chapter 3 begins with a description of *Frog* (van den Bosch *et al*, 2007), the Dutch NER used in this study, and the functions and operation of its NER-component. This is followed by a description of the Afrikaans-to-Dutch-converter (A2DC) (Van Huyssteen & Pilon, 2009) and finally the various experiments that were completed, are explained.

The study consists of six experiments, the first of which was to determine the results of Frog on Dutch data. The second experiment evaluated the effectiveness of *Frog* on unchanged (raw)

Afrikaans data. The following two experiments evaluated the results of *Frog* on “Dutched” Afrikaans data. The last two experiments evaluated the effectiveness of *Frog* on raw and “Dutched” Afrikaans data with the addition of gazetteers as part of the pre-processing step.

In conclusion, a summary is given with regards to the comparisons between the NER for Afrikaans that was developed in this study, and the NER-component that Puttkammer (2006) used in his tokeniser. Finally a few suggestions for future research are proposed.

KEY WORDS:

NAMED ENTITY RECOGNITION, FROG, BLARK, GAZETTEERS, AFRIKAANS, ENTITIES, INFORMATION EXTRACTION, TECHNOLOGY RECYCLING.

Inhoudsopgawe

Lys van tabelle

Lys van figure

1. Inleiding

1.1 Kontekstualisering

Volgens die Grondwet van Suid-Afrika word daar van die regering vereis om alle inligting in die tien inheemse tale van Suid-Afrika (uitsluitende Engels), beskikbaar te stel. Daarom het die regering die inligting wat alreeds vir tien tale bestaan (soos dokumente en literatuur), vrylik aan die publiek beskikbaar gestel en 'n poging word ook aangewend om die hoeveelheid inligting wat beskikbaar is in hierdie tale te vermeerder (Groenewald & Du Plooy, 2010).

Volgens Krauwer (2003) bestaan daar 'n inventaris van taalverwante hulpbronne wat nodig is vir 'n taal om kompeterend op die vlak van navorsing en onderrig te wees, wat bekend staan as die "Basic Language Resource Kit" (BLARK). Krauwer (2003) stel verder dat BLARK se inhoud kan verskil ten opsigte van die behoefte van die gegewe taal, maar dat die BLARK van die taal aan 'n infrastruktuur moet voldoen wat help om hulpbronne te bestuur, te onderhou en te versprei. Aangesien die meeste van die tale in Suid-Afrika hulpbronskaars is, is dit in die beste belang vir die kulturele groei van die land, om vir elk van die inheemse Suid-Afrikaanse tale 'n BLARK te ontwikkel. Volgens Krauwer (2003) val BLARK komponente gewoonlik onder die volgende drie kategorieë:

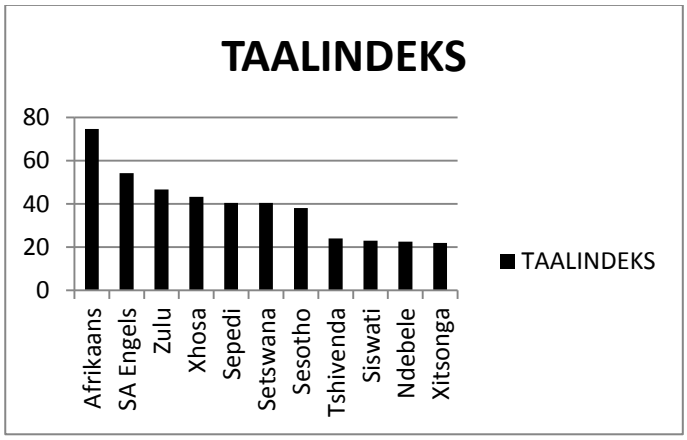
- **Standaarde**: Hier word na die spesifieke standaarde wat deur 'n taal gebruik gaan word verwys, byvoorbeeld die skryfwyse van 'n fonetiese alfabet en die standaard van die annotasie van woordsoortetikettering (byvoorbeeld "N" dui 'n selfstandige naamwoord aan).
- **Data of Hulpbronne**: Hierdie kategorie bestaan uit al die hulpbronne of grammatikas wat taalverwant is, byvoorbeeld tesourie, geskrewe en gesproke korpora, terminologie versamelings en enkel- en meertalige woordeboeke.
- **Kerntegnologieë of modules**: Hierdie kategorie bestaan uit programme wat as boustene vir toepassings gebruik kan word. Hierdie programme dien as die kern van 'n spesifieke toepassing. 'n Woordsoortetiketteerder staan byvoorbeeld as 'n kerntegnologie bekend omdat dit nooit as toepassing aangebied word nie, maar 'n belangrike onderdeel van toepassings soos spel- en grammatikatoetsers, masjienvertaalsisteme en inligtingonttrekkingsisteme is.

Ter voorbereiding vir die ontwikkeling van 'n BLARK vir al elf amptelike tale van Suid-Afrika, is 'n oudit in 2010 (Sharma Grover *et al*., 2010) voorgestel om te bepaal watter inheemse tale 'n behoefte het aan watter tipe hulpbronne in terme van standaarde, data en kerntegnologieë. Om die hulpbronskaarsheid van die tale te bepaal, is 'n mensliketaaltegnologie (MTT) taalindeks

saamgestel. Vir die skep van die taalindeks was daar gekyk na die totale MTT aktiwiteit vir elke taal asook na die vordering en toeganklikheid van elke taal se MTT hulpbronne en toepassings (Sharma Grover *et al*., 2010) en word in Figuur 1 en Tabel 1 weergegee. Tabel 1 gee die volwassenheidsindeks en toeganklikheidsindeks vir elke taal soos bepaal ten opsigte van die modules, data en toepassings van elke taal. Die volwassenheidsindeks en toeganklikheidsindeks word dan gekombineer om 'n taalindeks vir elke taal te bereken. Uit Figuur 1 kan afgelei word dat Afrikaans die beste toegerus is ten opsigte van hulpbronne en die verspreiding daarvan. Afrikaans word dan gevolg deur Suid-Afrikaanse Engels, IsiZulu en IsiXhosa. Aan die stertpunt van die grafiek lê die tale wat die minste hulpbronne besit en vrystel, wat onder andere SiSwati, Ndebele, Xitsonga en Tshivenda is.

TAAL	VOLWASSENHEIDSINDEKS	TOEGANGKLIHEIDSINDEKS	TAALINDEKS
Afrikaans	37.9	36.7	74.6
SA Engels	26	28.2	54.2
Zulu	21.7	25	46.7
Xhosa	20.9	22.3	43.2
Sepedi	18.1	22.3	40.4
Setswana	18.5	21.9	40.4
Sesotho	17.7	20.4	38.1
Tshivenda	11.9	12.1	24
Siswati	11.6	11.4	23
Ndebele	11.5	11	22.5
Xitsonga	10.9	11	21.9

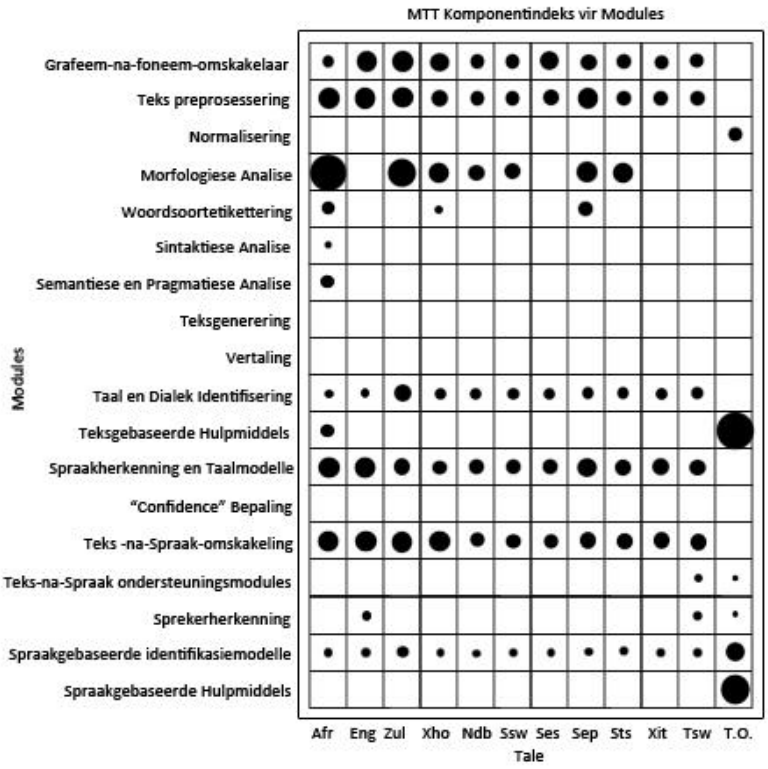
Tabel 1: Volwassenheidsindeks, Toeganklikheidsindeks en Taalindeks vir Suid-Afrikaanse tale



Figuur 1: MTT taalindeks vir Suid-Afrika

Ter opvolging hiervan is 'n MTT komponentindeks ontwikkel wat voorgestel word met 'n borrelgrafiek (Figuur 2). Die grote van die borrel dui die hoeveelheid aktiwiteit wat ten opsigte van data, modules en toepassings wat vir elke taal, plaasvind. Die komponentindeks word bereken op min of meer dieselfde manier as die taalindeks, maar fokus op die spesifieke items van 'n taal

en word bereken deur die sommasie van die volwassenheidsindeks en die toeganklikheidsindeks per item van ’n taal. Die volwasssenheidsindeks word bereken deur die sommasie van die aantal komponente in die “onder ontwikkeling” fase van toepassings, die aantal alfa weergawes, die aantal beta weergawes en die aantal vrygestelde weergawes vir elke taal (Sharma Grover, 2009). Die toeganklikheidsindeks word bepaal deur die aantal komponente in die “ongespesifiseerd” verspreidingsgroep, die aantal komponente wat nie toegangklik is nie, die aantal komponente wat vir navorsing en onderwys beskikbaar is, die aantal komponente wat kommersieel beskikbaar is en die aantal komponente wat vir navorsing, onderwys en kommersieel beskikbaar is vir elke taal te sommeer (Sharma Grover, 2009). Die doel van hierdie indeks is om aan navorsers ’n idee te gee oor die noodsaaklikheid vir ontwikkeling ten opsigte hulpbronne vir die onderskeie Suid-Afrikaanse tale (Sharma Grover *et al*., 2010).



Figuur 2: MTT komponentindeks vir modules

Gegewe die bogenoemde figuur blyk dit dat daar ’n groot aanvraag is na die ontwikkeling van hulpmiddels vir teks in Afrikaans. Dit bring dan mee dat daar in hierdie studie gefokus word op die ontwikkeling van ’n benoemde-entiteitherkenner vir Afrikaans aangesien benoemde-entiteitherkenners belangrik is vir teksprosessering veral in inligtingonttrekkingsisteme wat fokus op filtrering van groot hoeveelhede data.

Gegewe die groot hoeveelhede data wat vandag in digitale formaat beskikbaar (ook toenemend vir inheemse Suid-Afrikaanse tale) is daar ’n behoefte aan tegnieke of prosedures wat relevante inligting uit sodanige data kan onttrek. Een so ’n tegniek of prosedure is deur gebruik te maak

van ’n inligtingonttrekkingsisteem. Die doel van ’n inligtingonttrekkingsisteem is om die ongestruktureerde data binne tekste, op een of ander bruikbare manier, te struktureer (Jurafsky & Martin, 2010: 725). Een van die kerntegnologieë wat in ’n inligtingonttrekkingsisteem nodig is, is ’n benoemde-entiteitherkenner (BEH). Benoemde-entiteitherkenning het te make met die identifisering en kategorisering van entiteite wat inligting met betrekking tot persoonsname, pleknaam, organisasiename, bedrae en datums bevat (Jurafsky & Martin, 2010: 726).

Benoemde-entiteitherkenners speel egter nie net ’n belangrike rol in inligtingonttrekkingsisteme nie, maar is ook belangrik binne ander toepassingsvelde soos:

- vraag-en-antwoordsisteme Jurafsky en Martin (2010:783);
- die identifikasie van proteïenstringe (Jurafsky en Martin (2010:757); en
- tekseenheididentifiseerders (Puttkammer, 2006).

Verskeie tegnieke is al gebruik om benoemde-entiteitherkenningsisteme te ontwikkel met ondermeer handgeskrewe reëls of reëlmatige uitdrukkings (Cuzerzan en Yarowsky, 2002), masjienleeralgoritmes (Malouf, 2002) en kombinasies van hierdie tegnieke (hibriede sisteme) (Carreras *et al.* 2002). Volgens Nadeau en Sekine (2006) is die neiging die afgelope vyf na tien jaar om meer masjienleeralgoritmes in plaas van reëlmatige uitdrukkings te gebruik, omdat die afrigting van masjienleeralgoritmes baie vinniger is as om reëls vanuit data af te lei. In ’n poging om BEH-sisteme te ontwikkel het verskillende navorsers al van verskeie masjienleertegnieke gebruik gemaak (Nadeau & Sekine, 2006), waaronder:

- besluitnemingsbome (Sekine, 1998);
- versteekte Markov modelle (HMM’s) (Bikel *et al*., 1998);
- maksimum entropie (Borthwick *et al*., 1999; Skut & Brants, 2002);
- ondersteuningsvektormasjiene (“Support Vector Machines”) (Asahara & Matsumoto, 2003);
- voorwaardelike willekeurige velde (“Conditional random fields”) (MacCullum & Li, 2003).

Puttkammer (2006) het ’n BEH vir Afrikaans ontwikkel as deel van ’n tekseenheididentifiseerder. Hierdie BEH gebruik die *k*-Naastebuurpuntalgoritme in kombinasie met ’n besluitnemingsboom as klassifiseringsmodel. Hierdie tipe klassifiseringsalgoritme (*k*-Naastebuurpuntalgoritme) word afgerig met groot hoeveelhede data (1 607 sinne en 40 906 woorde waarvan 3 068 benoemde entiteite is) wat hoofsaaklik saamgestel is uit e-posboodskappe, webbladsye, tydskrifte en koerantartikels wat op die internet beskikbaar is (Puttkammer, 2006:11). Die afrigtingsdata het altesaam uit 224 eienskappe vir elke woord bestaan wat die sewe tekseenhede van elke fokuswoord (drie tekseenhede voor en drie tekseenhede na die woord) en 32 ander eienskappe wat

die woord help identifiseer, insluit (Puttkammer. 2006:11). Die afrigtingsdata van die *k*-Naastebuurpuntalgoritme word dan volgens bepaalde eienskappe in klusters of groepe verdeel. Wanneer ongesiene of vreemde data deur die sisteem geklassifiseer moet word, word die data ten opsigte van eienskappe van 'n element binne elke kluster, geklassifiseer met betrekking tot *k* wat die aantal elemente aandui wat vergelyk moet word (Puttkammer, 2006:68).

Soos met alle "lui" leermetodes word prosessering van data by die *k*-Naastebuurpuntalgoritme vertraag totdat 'n nuwe klassifisering gedoen word. Dit lei dan daartoe dat die verwerkingsproses van die sisteem baie stadig kan wees (Puttkammer, 2006:53). Die *k*-Naastebuurpuntalgoritme word om hierdie rede ook met 'n algemene besluitnemingsboomalgoritme (IGTREE) gekombineer. Volgens Puttkammer (2006:71) is 'n besluitnemingsboom "'n masjienleeralgoritme wat sekere veralgemenings oor afrigtingsdata maak, deur die data te groepeer volgens eienderse eienskappe. Hierdie groepering kan dan metafories as 'n omgekeerde boom voorgestel word." IGTREE is bekend daarvoor dat dit die klassifikasieproses van die k-Naastebuurpuntalgoritme bespoedig, maar nog steeds die akkuraatheid daarvan behou (Van Den Bosch *et al*., 2007). Gegewe die feite dat daar 'n groot hoeveelheid eienskappe benodig word vir identifikasie en klassifikasie van woorde (224 per woord) sowel as die feit dat die masjienleeralgoritme al hierdie inligting in die geheue moet stoor asook die feit dat die *k*-Naastebuurpuntalgoritme van "lui" leermetodes gebruik maak (Puttkammer, 2006:66), kan die gevolgtrekking gemaak word dat Puttkammer se tekseenheididentifiseerder (wat die BEH insluit) nie implementeer kan word as 'n toepassing nie.

Groenewald en Du Plooy (2010) het 'n eenvoudige tipe BEH vir Afrikaans ontwikkel in 'n poging om konfidensiële inligting in teksdata te identifiseer en dan woorde en frases wat konfidensieel van aard is met ander, willekeurig-geselekteerde woorde en frases, van dieselfde woordsoorttipe te vervang en staan as 'n Anonimiseerder bekend (Groenewald & Du Plooy, 2010: 2). Die grootste verskil tussen die Anonimiseerder en 'n normale BEH is die feit dat die Anonimiseerder slegs die inligting wat konfidensieel van aard is (byvoorbeeld persoonsname, geldbedrae, ensovoorts) identifiseer en met ander, soortgelyke inligting vervang uit voorafopgestelde gazetteers (of te wel naamlyste) (Groenewald & Du Plooy, 2010: 3). Die benoemde entiteite word dan met die hulp van reëlmatige uitdrukkings ("regular expressions") geïdentifiseer en word dan in die gazetteers opgesoek. Indien die benoemde entiteit in een van die gazetteers gevind word, word die benoemde entiteit met 'n gepaste etiket geannoteer (Groenewald & Du Plooy, 2010: 3). In 'n normale BEH word alle entiteite (soos datums, plekname, organisasiename, ensovoorts) geïdentifiseer en van een of ander etiket voorsien deur middel van 'n masjienleeralgoritme waarna dit in 'n gepaste kategorie geplaas word. Aangesien die Anonimiseerder die entiteite met ander willekeurig-geselekteerde entiteite vervang, dit van gazetteers gebruik maak om die vervanging

te doen en dit ’n beperkte etiketstel het, is dit onvoldoende om as ’n BEH vir Afrikaans of in ’n inligtingonttrekkingsisteem gebruik te word.

1.2. Probleemstelling

Alhoewel daar al navorsing gedoen is oor die ontwikkeling van ’n BEH vir Afrikaans, bestaan daar tans nie ’n effektiewe, vinnige en implementeerbare BEH vir Afrikaans nie. Die feit dat Afrikaans ’n hulpbronskaarstaal is, moet in gedagte gehou word wanneer ’n BEH vir Afrikaans ontwikkel word. Aangesien Afrikaanse hulpbronne en kerntegnologieë nie geredelik beskikbaar is nie, kan die ontwikkeling van ’n BEH vir Afrikaans ’n stadige en duur proses wees.

Een manier waarop die ontwikkeling van kerntegnologieë bespoedig kan word, is deur van tegnologieherwinning gebruik te maak. Tegnologieherwinning is ’n proses waartydens die hulpbronne van ’n soortgelyke of nabyverwante taal (L1) gebruik word om ’n ander taal (L2) se data te analiseer of te annoteer (Rayner *et al*., 1997). Gewoonlik is een van die tale hulpbronryk (bv. Nederlands) en die ander hulpbronskaars (bv. Afrikaans). Dit laat die hulpbronskaars taal toe om voordeel te trek uit die feit dat die ander taal meer hulpbronne het, wat weer help met die ontwikkeling van kerntegnologieë vir die hulpbronskaars taal (Rayner *et al*., 1997:1).

Uit vorige studies (Pilon *et al*., 2010; Van Huyssteen & Pilon, 2009) blyk dit dus dat tegnologieherwinning belowende resultate lewer vir Afrikaans wanneer Nederlandse kerntegnologieë in die herwinningsproses gebruik word. Vir hierdie studie gaan die benoemde-entiteitherkenner wat deur Van den Bosch (2007) ontwikkel is, *Frog*, gebruik word in die tegnologieherwinningeksperimente aangesien, net soos die BEH wat deur Puttkammer ontwikkel is, dit deel uitmaak van ’n groter natuurliketaalprosesseringsisteem.

1.3. Navorsingsdoelwitte

Gegewe die bogenoemde, het hierdie studie vyf navorsingsdoelwitte, naamlik om:

a. Nederlandse en Afrikaanse benoemde entiteite te vergelyk om sodoende vas te stel watter ortografiese verskille tussen die twee tale die effektiwiteit van ’n Nederlandse BEH op Afrikaanse data kan beïnvloed;
b. Nederlandse data deur die Nederlandse BEH te annoteer sodat die resultate daarvan vergelyk kan word met die resultate van die Afrikaanse data;
c. Afrikaanse data deur ’n Nederlandse BEH te laat annoteer om sodoende die presisie, herroeping en *f*-telling van die Nederlandse BEH in die annotasie van Afrikaanse data te bereken;

d. die Afrikaanse data wat deur die Nederlandse BEH geannoteer is; krities te analiseer om sodoende vas te stel watter pre- en/of post-prosesseringstappe nodig is om die resultate te verbeter; en
e. om pre- en/of post-prosesseringstappe te implementeer om die resultate van die Nederlandse BEH op Afrikaanse data te verbeter om sodoende ’n effektiewe, vinnige en implementeerbare BEH vir Afrikaans te ontwikkel.

1.4. Sentrale Teoretiese Stelling

Tegnologieherwinning is nog nooit gebruik in die ontwikkeling van ’n Afrikaanse BEH nie, maar gegewe die goeie resultate wat reeds deur tegnologieherwinning verkry is, vir ander kerntegnologieë (Pilon *et al*., 2010), word die veronderstelling gemaak dat ’n Nederlandse BEH waarskynlik goeie resultate (dit wil sê resultate wat goed vergelyk met die resultate wat op Nederlandse data verkry is) op Afrikaanse data sal behaal. Daar word ook aangeneem dat die verskille tussen Afrikaanse en Nederlandse benoemde-entiteite van so ’n aard is dat dit sal moontlik wees om die resultate van die Afrikaanse afrigtingsdata met pre- en/of post-prosessering te verbeter. Die verbeteringe behoort van so ’n aard te wees dat die uiteindelike Afrikaanse BEH meer doeltreffend en effektief sal wees (wat betref prosesseringstyd, presisie, herroeping, *f*-telling en akkuraatheid) as reeds bestaande Afrikaanse BEH’s.

1.5. Metodologie

In hierdie studie gaan die Nederlandse BEH in *Frog,* wat deur Van Den Bosch *et al.* (2007) ontwikkel is, gebruik word om Afrikaanse data te annoteer. Daar gaan pre- en/of post-prosessering op die data toegepas word in ’n poging om die resultate wat op die rou Afrikaanse data verkry is, te verbeter.

1.5.1. Breë Benadering

Aangesien daar tans geen implementeerbare benoemde-entiteitherkenner vir Afrikaans is nie en omdat dit ’n groot hoeveelheid hulpbronne sal verg om ’n benoemde-entiteitherkenner van nuuts af vir Afrikaans te ontwikkel, gaan ’n eksperimentele ontwikkeling in hierdie studie gebruik word. Eksperimentele ontwikkeling dui op die aanneming, kombinering, vorming en gebruik van bestaande wetenskaplike, tegnologiese-, besigheids- of ander relevante inligting en vaardighede met die doel om planne te produseer of ontwerpe te skep vir nuwe, veranderde of verbeterde produkte, prosesse of dienste (volgens InnoviSCOP, 2006).

Die eksperimentele ontwikkeling gaan gevolg word deur ’n bestaande Nederlandse BEH te gebruik en die resultate daarvan dan te verbeter deur van pre- en/of post-prosesseringstappe gebruik te maak.

Aangesien dit nog nie duidelik is wat hierdie pre- en /of post-prosesseringstappe gaan wees nie en aangesien die inligting nog nie beskikbaar is nie, is dit noodsaaklik om van eksperimentele ontwikkeling gebruik te maak om hierdie inligting te bekom.

1.5.2. Literatuurstudie

In hierdie afdeling gaan in die eerste plek spesifiek gefokus word op Afrikaanse en Nederlandse benoemde-entiteite om vas te stel watter ortografiese verskille tussen hierdie benoemde-entiteite bestaan en hoe dit moontlik die resultate van die Nederlandse BEH op Afrikaanse data gaan beïnvloed. Die geïdentifiseerde verskille sal ’n aanduiding gee van die tipe pre- en/of post-prosessering wat gaan nodig wees om die effektiwiteit van die Nederlandse BEH op Afrikaanse data te verbeter.

Tegnologieherwinning gaan ook bestudeer word, aangesien dit die benadering is wat gebruik gaan word om ’n BEH vir Afrikaans te ontwikkel. Volgens Rayner *et al.* (1997) is tegnologieherwinning nie ’n nuwe konsep nie. Die basiese idee is, dat indien die twee tale L2 en L1 genoeg verwantskappe met mekaar toon, dit makliker sal wees om die sagteware of kerntegnologieë, wat van toepassing is op L1, te verander om aan die behoeftes van L2 te voldoen eerder as om sagteware of kerntegnologieë van nuuts af vir L2 te skep (Rayner *et al*., 1997:2).

1.5.3. Die Ontwikkelingsproses

Die ontwikkeling van die Afrikaanse benoemde-entiteitherkenner sal in 6 stappe geskied wat hieronder uiteengesit word.

1. Die Nederlandse BEH word gebruik om rou Afrikaanse data te annoteer.
2. Die afvoer van die eerste stap word geëvalueer en geanaliseer om vas te stel watter pre- en/of post-prosesseringstappe noodsaaklik is om die resultate van die Nederlandse BEH te verbeter.
3. Na aanleiding van inligting wat verkry is uit die analise in stap 2, word pre- en/of post-prosesseringsmodules ontwikkel om die akkuraatheid van die Nederlandse BEH op Afrikaanse data te verbeter.
4. Die pre- en/of post-prosesseringsmodules sal dan op die Afrikaanse data toegepas word voordat die Nederlandse BEH weer gebruik sal word om die aangepaste Afrikaanse data te annoteer.

5. Die afvoer van die proses wat in stap 4 toegepas is, sal dan geëvalueer word volgens internasionale benoemde-entiteitherkenningspraktyke, naamlik deur gebruik te maak van presisie ("precision"), herroeping ("recall") en *f*-telling ("*f*-score").
6. In die laaste stap sal die resultate van die Nederlandse BEH op die rou en aangepaste Afrikaanse data onderskeidelik, met mekaar vergelyk word om te bepaal tot hoe 'n mate die pre- en/of post-prosessering die resultate beïnvloed het. Die uiteindelike Afrikaanse BEH sal ook vergelyk word met bestaande Afrikaanse BEH's in terme van prosesseringstyd, presisie, herroeping, *f*-telling en akkuraatheid.

Presisie dui op die verhouding tussen die aantal entiteite wat korrek geïdentifiseer is teenoor die totale aantal entiteite wat geïdentifiseer is (1). Herroeping dui op die verhouding tussen die aantal entiteite wat korrek geïdentifiseer is teenoor die aantal entiteite wat geïdentifiseer moet word (2). ***f***-telling verteenwoordig die harmoniese gemiddeld tussen presisie en herroeping. (3) (Manning *et al*, 2009).

$$\text{Presisie} = \frac{\text{Die aantal entiteite wat korrek geïdentifiseer is}}{\text{Totale aantal entiteite wat geïdentifiseer is}} \quad (1)$$

$$\text{Herroeping} = \frac{\text{Die aantal entiteite wat korrek geïdentifiseer is}}{\text{Die aantal entiteite wat geïdentifiseer moet word}} \quad (2)$$

$$\text{F-Telling} = \frac{2 \times \text{Presisie} \times \text{Herroeping}}{\text{Presisie+ Herroeping}} \quad (3)$$

1.5.4. Dataverwerking

Nadat *Frog* die toevoerdata ontvang, word die teks in woorde verdeel en elke woord op 'n aparte lyn geplaas. Die woorde word dan een vir een deur *Frog* geïdentifiseer en geklassifiseer volgens die ses voorafbepaalde kategorieë (ORG, PER, LOC, MISC, EVE en PRO). Alhoewel *Frog* konteks-sensitief is in terme van die klassifikasies, word elke woord steeds apart van 'n etiket voorsien. Dit dra daartoe by dat daar entiteite kan voorkom wat gedeeltelik korrek geklassifiseer is. Vals-positiewe identifikasies (woorde wat as entiteite geëtiketteer is, maar nie entiteite is nie) moet eerder in evaluasies gepenaliseer word as entiteite wat net gedeeltelik geklassifiseer word aangesien valse-positiewe identifikasies baie meer nadelig kan wees vir 'n sisteem. (Marrero *et al*. 2009).Daarom gaan elke woord wat deel uitmaak van 'n benoemde entiteit afsonderlik behandel word.

'n Voorbeeld van so 'n gedeeltelike klassifikasie is te sien in die geval van "President Thabo Mbeki" waar die etikette "O_B-PER_I-PER" toegeken is. Die regte etiket is veronderstel om "B-PER_I-PER_I-PER" te wees. Die eerste etiket sal dan as 'n verkeerdelike klassifikasie beskou word, maar die ander twee etikette word as korrek aanvaar.

'n Verwarringsmatriks is gebruik om die verskillende kombinasies van foute (byvoorbeeld 'n entiteit of gedeelte van 'n entiteit wat as 'n PER geëtiketteer moes word, maar as 'n ORG geëtiketteer is), of te wel 'n PER-> ORG kombinasie fout).

Volgens die literatuur uit die domein van benoemde-entiteitherkenning word meeste van die sisteme en eksperimente ten opsigte van *f*-tellings met mekaar vergelyk (Tjong Kim Sang, 2002). Daar is gepoog om 'n korpus saam te stel wat vergelykbaar is met Puttkammer (2006) se sisteem, maar as gevolg van 'n beperking op tyd en hulpbronne en ander eksterne faktore, kon dit nie verwesenlik word nie.[1] Die poging om statistiese analises op die verskillende eksperimente toe te pas het ook nie gewerk nie. Aangesien die aantal benoemde entiteite in hierdie studie baie min was (310), kon daar ook nie tienvoudige kruivalidasie op die data toegepas word nie (toets vir statistiese beduidendheid) omdat 'n tiende van die data (die toets data) nie voldoende resultate sou lewer nie

[1] Pogings om toegang tot die BEH van Puttkammer te verkry, met die oog op 'n direkte vergelyking was onsuksesvol.

2. Literatuurstudie

2.1. Inleiding

Gegewe die eksponensiële uitbreiding van data wat in digitale formaat beskikbaar is, word tegnieke of prosedures benodig om deur hierdie data te filtreer sodat relevante inligting daaruit onttrek kan word. Vir hierdie doeleindes kan 'n inligtingonttrekkingsisteem (IOS) gebruik word. Die IOS se hoofdoel is om die inhoud van tekste te struktureer (Jurafsky & Martin, 2010) en gevolglik word hierdie gestruktureerde data (byvoorbeeld entiteite wat volgens sekere kategorieë geannoteer is) ook meer toeganklik en bruikbaar.

Een van die kerntegnologieë wat nodig is in 'n IOS, is 'n benoemde-entiteitherkenner. Benoemde-entiteitherkenning (BEH) fokus op die identifisering en kategorisering van entiteite wat inligting soos name, bedrae, persentasies en datums insluit (Jurafsky & Martin, 2010: 726). Volgens Nadeau en Sekine (2006) is die term "benoemde-entiteit" die eerste keer tydens die sesde "Message Understanding Conference" (MUC-6) in 1995 gebruik. Die hooffokus van die MUC-6 was inligtingonttrekking (IO) en aangesien BEH 'n sub-kategorie van inligtingonttrekking is, is dit ook tydens die kongres bespreek.

Volgens Van Huyssteen (2000: 52; ter verwysing na Langacker, 1987) word 'n entiteit gedefinieer as enigiets waarna verwys kan word vir analitiese doeleindes. Dit sluit konkrete dinge, verhoudings, sensasies en waardes in. Volgens Puttkammer (2006:22) verwys die begrip "entiteit" in "benoemde entiteit" meestal na 'n spesifieke ding (aansyn). Met betrekking tot die konsep "benoemde" in "benoemde entiteit" verwys Puttkammer (2006) na Van Huyssteen (2000:53) wat aanvoer dat "taal 'n simboliese tekensisteem is waar betekenis toegeken word op konsensusbasis." Volgens Puttkammer beteken dit dus dat "die entiteit aan 'n enkele aansyn veranker is, hetsy deur 'n naamgewingsritueel (soos by persoonsname), 'n outoritêre instelling (soos pleknname wat deur pleknaamkomitees bereël is), 'n registrasieproses (soos besigheids- en pleknname) of konvensie (soos by titels)". Puttkammer (2006:25) wys dan daarop dat die definisie vir 'n benoemde entiteit soos volg daar sal uitsien:

> "'MIV Benoemde entiteit is 'MIV aansyn wat binne die konseptuele ruimte aan 'MIV enkele instansiëring veranker word deur middel van konvensie, 'MIV geïnstitusionaliseerde proses of 'MIV outoriteit en waarvan die skryfwyse of wetlik, of deur een of ander outoriteit bepaal word."

In ’n studie deur Desmet en Hoste (2010) is drie klassifiseringsraamwerke gekombineer, ’n *k*-Naastebuurpuntalgoritme (“*k*-nearest neighbour”), voorwaardelike-willekeurigevelde (“conditional random fields”) en ’n ondersteuningsvektormasjien (“support vector machine”). Elke klassifiseringsraamwerk word met die afvoerdata van die vorige klassifiseringsraamwerke afgerig om sodoende die slegte eienskappe van elk van die vorige klassifiseringsraamwerke uit te kanselleer. Hierdie BEH is vir Nederlands ontwikkel en het ’n *f*-telling (“*f*-measure”) van 0.83 behaal.

In ’n ander studie deur Black en Vasilakopoulos (2002) is ’n BEH vir Spaans en Nederlands ontwikkel. Die sisteem bestaan uit ’n transformasiegebaseerde-leermetode (TL) en ’n eenvoudige besluitnemingsboom-induksieskema (BIS). Die besluitnemingsboom, wat met verskeie eienskappe geïnduseer is, word gebruik om die klas (of kategorie) van die benoemde-entiteit uit ’n onbekende klas, te bepaal. Vir Spaans het die sisteem ’n *f*-telling van 0.80 behaal en vir Nederlands ’n *f*-telling van 0.82 behaal.

Malouf (2002) het gepoog om ’n taalonafhanklike BEH te ontwikkel en dit op Spaans en Nederlands getoets. Die sisteem maak gebruik van ’n tipe waarskynlikheidsetikettering (“probabilistic tagging”). Gegewe ’n stel opeenvolgende woorde, word daar gepoog om die ooreenstemmende patroon van etikette, binne die bestaande woordeskat van die etikette, te soek. Tipiese etikette sluit in “B” (dui die begin van entiteit aan), “I” (dui aan dat die entiteit nog ’n gedeelte bevat) en “O” (dui aan dat dit nie ’n benoemde-entiteit is nie) wat dan aan die entiteite toegeken word. Daar is ook van ’n versteekte Markov-model (“Hidden Markov Model”) en maksimum entropie gebruikgemaak om die waarskynlikheid van ’n woord tussen ander woorde te bepaal. Op hierdie manier is die parameters verfyn om die beste akkuraatheid te behaal. Die sisteem het vir Spaans ’n *f*-telling van 0.73 en vir Nederlands ’n *f*-telling van 0.70 behaal.

Om die prosedures en implementerings wat in die vorige hoofstuk genoem is, in perspektief teenoor mekaar te stel, moet daar ’n afsonderlike, maar wel volledige analise en uiteensetting van elementêre gedeeltes van elke denkwyse gedoen word. In afdeling 2.2. fokus die bespreking op tegnologieherwinning en hoe dit gebruik kan word om ’n benoemde-entiteitherkenner vir Afrikaans te ontwikkel. In afdeling 2.3 gaan ’n volledige analise gedoen word ten opsigte van die verskille tussen Afrikaanse en Nederlandse benoemde entiteite. In afdeling 2.4 sal ’n samevatting van die hoofstuk verskaf word.

2.2. Tegnologieherwinning

Scannell (2006:1) beweer dat die beginsel van tegnologieherwinning baie goed kan werk vir hulpbronskaarstale (soos byvoorbeeld Afrikaans), veral wanneer die teikentaal (L1) hulpbronryk is (soos byvoorbeeld Nederlands). Volgens Rayner *et al*. (1997: 2) behels tegnologieherwinning

die ontwikkeling van hulpbronne vir hulpbronskaarstale (L2) deur die herontwerp of verandering van kerntegnologieë van hulpbronryke tale (L1). Rayner *et al*. (1997) stel dat indien die twee tale L1 en L2 genoeg ooreenkomste tussen mekaar toon, dit makliker sal wees om die sagteware wat van toepassing is op L1 te verander om aan die behoeftes van L2 te voldoen as om sagteware van nuuts af vir L2 te skep (Rayner *et al*., 1997:2).

In 'n studie deur Pilon *et al*. (2010) is 'n Afrikaanse-woordsoortetiketteerder ontwikkel, deur 'n Nederlandse-woordsoortetiketteerder en omgeskakelde teks (wat omgeskakel is, deur 'n Afrikaans-na-Nederlands-omskakelaar (A2DC), sien afdeling 3.3) te gebruik. Die sisteem het 'n akkuraatheid (presisie) bo 90% behaal, wat ook voorheen deur Pilon (2005), met 'n soortgelyke sisteem, verkry is deur slegs 10 000 woorde handmatig te annoteer vir afrigtingsdata.

In 'n studie deur Villazón-Terrazas *et al*. (2010) word 'n metode voorgestel om ontologieë uit nie-ontologiese hulpbronne te ontwikkel deur gebruik te maak van 'n veranderde herontwikkelingsmodel wat gewoonlik vir sagteware-ontwikkeling gebruik word. Ontologie verwys gewoonlik na 'n spesifieke stel objekte wat verkry is deur die analise van 'n enkele domein (Jurafsky & Martin, 2006:616). 'n WordNet (Fellbaum, 1998) word ook gebruik om die verwantskappe tussen die nie-ontologiese hulpbronterme te bepaal. Nie-ontologiese hulpbronne (NOH) is kennishulpbronne wat semanties nog nie deur 'n ontologie geformaliseer is nie. Die vier vlakke van sagteware-ontwikkeling (wat standaard in die praktyk gebruik word) word dan stelselmatig verander sodat dit gebruik kan word om ontologieë te skep. Hierdie studie is 'n voorbeeld van tegnologieherwinning aangesien dit die aanpassing van ontwikkelingsmetodes van een domein genoodsaak het sodat die ontwikkelingsmetodes in 'n ander domein gebruik kon word.

Martinovic (2008) het 'n Serwiese inligtingonttrekkingsisteem (IOS) ontwikkel deur van 'n bestaande Engelse IOS gebruik te maak. Die sisteem bestaan uit die EBART-teksversameling (Ebart, 2010) (bestaande uit Serviese nuusartikels) en die SMART-onttrekkingsisteemalgoritme (Salton, 1971). Die prosessering van die teksversameling bestaan uit drie fases, naamlik:

- die omskakeling van nie-ASCII Serviese letters (*ć*, *č*, *đ*, *š* en *ž*) na ooreenstemmende ASCII-voorstellings (*cx*, *cy*, *dx*, *sx* en *zx*);
- stopwoordverwydering; en
- woordkonflikhantering.

Weens die kompleksiteit van die Serviese taal (Martinovic, 2008) is twee algoritmes spesifiek ontwerp om te onderskei tussen die vorms van die verskeie woordsoorte (dit wil sê voorkoming van woordkonflik), naamlik die Uitputbare Konflikalgoritme (UKA) en die Rudimentêre Konflikalgoritme (RKA) (Martinovic, 2008:14). Die doel van die UKA is om die ooreenkomste tussen

woorde te vind ten opsigte van geslag en komplekse alliterasies. Die RKA is geskep deur die UKA te vereenvoudig na die mees basiese reëls. Hierdie nuwe RKA is vry van oortollige normalisasie omdat die reëls wat voorheen probleme gegee het, nie meer daarin van toepassing is nie. Die volgende tabel (Tabel 2) dui die presisie van elke eksperiment van Martinovic (2008) met verskillende dokumentgroottes aan.

Aantal Dokumente	Basisvlak Presisie	1ste Algoritme Gem. Presisie	2de Algoritme gem. Presisie
5	0.698	0.745	0.823
10	0.601	0.676	0.754
15	0.520	0.633	0.701
20	0.478	0.576	0.644
Navraag Gem.	0.574	0.658	0.730
% toename in onttrekking	-	14.5%	27.2%

Tabel 2: Presisie van elke eksperiment met verskillende dokumentgroottes.

Rayner *et al*. (1997) stel twee metodes voor wat gebruik kan word om linguistiese inligting, soos grammatikas, leksikons en oordragreëls ("transfer rules") vir masjienvertaalsisteme wat nabyverwante tale prosesseer, te gebruik. Die eerste benadering begin deur 'n funksionele grammatika en leksikon aan L1 (die eerste taal) te gee. Die tweede benadering is gebaseer op vertaling tussen twee nabyverwante tale. Nabyverwante tale verwys na twee of meer tale waarvan daar klein verandering (hetsy sintakties, morfologies of ortografies) tussen die tale voorkom (Van Huyssteen & Pilon, 2009). Dit gee verder aanleiding dat die tegnologie van een taal na 'n ander taal oorgedra kan word deur minimale veranderinge aan die tegnologie aan te bring. Dit kan ook, in meeste van die gevalle, baie nuttig wees omdat dit baie duur ten opsigte van tyd en insameling van data sal wees om kerntegnologieë of modules vir 'n taal van nuuts af te ontwikkel (Rayner *et al*., 1997).

'n Eksperiment is gedoen om gesproke taal van Sweeds-na-Frans te vertaal deur eers van Sweeds-na-Engels en dan Engels-na-Frans te vertaal. In 'n tweede eksperiment is daar van Engels na Sweeds en dan van Sweeds na Deens vertaal. Die resultate vir die verskeie eksperimente word in Tabel 3 en Tabel 4 voorgestel (Rayner *et al.,* 2007*)*.

	SWE -> FRA	ENG ->SWE
Volledig aanvaarbaar	29.4%	56.5%
Onnatuurlike styl	16.3%	7.75%
Klein sintaktiese foute	15.2%	11.75%
Groot sintaktiese foute	2.0%	4.75%
Gedeeltelike vertalings	7.0%	8.75%
Gemors	22.9%	5.0%
Swak vertaling	7.0%	4.0%
Geen vertaling	0.2%	1.5%

Tabel 3: Vertaling van Sweeds na Frans en Engels na Sweeds vir ongesiende spraak-data

Uit Tabel 3 blyk dit dat dit veel moeiliker vir Rayner was om van Sweeds na Frans te vertaal (29.4% volledige vertaling) teenoor die vertaling van Engels na Sweeds (56.6%). Vir die Engels-na-Deensvertaling was 52.5% daarvan volledig (Tabel 4).

	ENG -> DE
Volledig aanvaarbaar	52.5%
Onnatuurlike styl	0.4%
Klein sintaktiese foute	24.4%
Groot sintaktiese foute	0.7%
Gedeeltelike vertalings	0.0%
Gemors	0.9%
Swak vertaling	10.7%
Geen vertaling	10.3%

Tabel 4: Vertaling van Engels na Deense data met verteenwoordigende data

Pilon *et al*. (2010) gebruik 'n Nederlandse woordsoortetiketteerder (WSE) vir die annotering van onveranderde (rou) en vernederlandsde Afrikaanse data. Die doel van hierdie studie was om die effek van vernederlandsde data ten opsigte van tegnologieherwinning te bepaal. Dieselfde data wat gebruik is om die Nederlands-na-Engels-masjienvertalingssisteem in die METIS II-projek (Vandeghinste *et al*., 2006) te evalueer, is in hierdie studie gebruik. Vir hierdie eksperiment is Afrikaanse vertalings van METIS II- data, deur *Tadpole* (Van den Bosch, 2007) geannoteer. Die afvoer van *Tadpole* is met 'n goudstandaard vergelyk en presisie herroeping en *f*-telling is bepaal.

Daarna is Afrikaanse data met 'n woord-vir-woordvertaalsisteem verander sodat die data meer na Nederlands "lyk". Na die omskakeling is die vernederlandsde data deur *Tadpole* geannoteer en die annotasies is weer met die goudstandaard vergelyk om presisie, herroeping en *f*-telling te bepaal (Tabel 5) (Pilon *et al*., 2010).

	Resultate vir rou Afrikaanse data			**Resultate vir vernederlandsde data**		
	Presisie	**Herroeping**	***f*-telling**	**Presisie**	**Herroeping**	***f*-telling**
N	0.54	0.86	0.67	0.67	0.91	0.77
ADJ	0.61	0.73	0.66	0.64	0.78	0.7
V	0.86	0.61	0.71	0.89	0.62	0.73
NUM	1	0.79	0.88	0.97	0.76	0.86
PRON	0.34	0.55	0.42	0.84	0.88	0.86
ART	0.16	0.01	0.02	0.95	1	0.97
PREP	1	0.81	0.9	0.99	0.99	0.99
CONJ	0.65	0.59	0.62	0.96	0.86	0.91
ADV	0.64	0.85	0.73	0.78	0.7	0.74
INTERJ	0	0	0	0	0	0
SPEC	0.43	0.74	0.54	0.2	0.07	0.1

Tabel 5: Presisie, herroeping en f-telling vir etikette van die rou en vernederlandste data

Die elf woordsoort-kategorieë in Tabel 5 word soos volg beskryf (Pilon *et al.*, 2010):

- N -> Selfstandige naamwoord;
- ADJ -> Adjektief;
- V -> Werkwoord;
- NUM -> Nommers;
- PRON -> Eiename;
- ART -> Artikels;
- PREP -> Voorsetsels;
- CONJ -> Voegwoorde;
- ADV -> Bywoorde;
- INTERJ -> Tussenwerpsel; en
- SPEC -> Spesiale tekseenhede.

Die akkuraatheid van die vernederlandsde Afrikaanse data was 80.6% in vergelyking met die 62.6% vir die rou data.

Uit die voorbeelde wat verskaf is blyk dit tog moontlik om tegnologieherwinning te gebruik om die ontwikkeling van hulpbronskaarstale te bespoedig. Alhoewel tegnologieherwinning nog nooit gebruik is in die ontwikkeling van ’n Afrikaanse BEH nie, dui die goeie resultate wat reeds met behulp van tegnologieherwinning verkry is vir ander kerntegnologieë (Pilon *et al.*, 2010), daarop dat ’n Nederlandse BEH waarskynlik goeie resultate, dit wil sê resultate wat goed vergelyk met die resultate wat op Nederlandse data verkry is, op Afrikaanse data sal behaal. Verder behoort dit ook moontlik te wees om die resultate van die Nederlandse BEH op Afrikaanse data te verbeter deur middel van pre- en/of post-prosesseringstegnieke. Voordat ’n Nederlandse BEH egter vir Afrikaans aangepas kan word, moet daar deeglike kennis geneem word van die ooreenkomste en verskille tussen Afrikaanse en Nederlandse benoemde entiteite. Hierdie inligting sal gebruik word om te bepaal watter pre- en/of post-prosesseringstappe nodig gaan wees om die resultate van die Nederlandse BEH op Afrikaanse data te verbeter.

2.3. Afrikaans en Nederlandse benoemde-entiteite

Uit ’n studie deur Van Huyssteen en Pilon (2009) is gevind dat daar klein morfologiese en ortografiese verskille tussen Afrikaans en Nederlandse woorde bestaan wat dit geskik maak vir die proses van tegnologieherwinning. Daar is drie tipes verhoudings tussen Nederlandse en Afrikaanse benoemde entiteite geïdentifiseer. Die entiteite kan identiese kognate, nie-identiese kognate of onverwant wees. Elkeen van hierdie verhoudings sal vervolgens in meer detail bespreek word.

2.3.1. Identiese kognate

Identiese kognate kan geïdentifiseer word as daardie woorde wat nie ortografies of semanties verander wanneer dit van een taal na ŉ ander taal vertaal word nie. Hierdie kategorie kan verder in twee sub-kategorieë verdeel, naamlik “Onveranderd” en “Nederlandse-spelvorme”. Onder die sub-kategorie “Onveranderd” ressorteer benoemde-entiteite wat by vertaling geen preprosessering nodig sal hê wanneer die Afrikaanse data met die Nederlandse BEH geannoteer word nie. Dit impliseer dat hulle onveranderd sal bly. Voorbeelde van hierdie woorde is die volgende:

- Plekname: Afghanistan, Athene, Nederland, Japan, ensovoorts.
- Persoonname: Abdul, Leopold, Benjamin, Willem, ensovoorts.
- Organisasiename: Virgin, Astra, Discovery, Heineken, ensovoorts.
- Getalle: een, twee, drie en vier (Getalle soos “vyf” hoort nie in hierdie kategorie nie , aangesien die Nederlandse spelvorm “vijf” van Afrikaans verskil.)

Onder die ander sub-kategorie, “Nederlandse-spelvorme”, word entiteite gekategoriseer wat voorkom of dit sistematies verskil het (sien 2.3.2 hieronder), maar wat nie deur preprosessering verander moet word nie. Voorbeelde hiervan sluit in:

- Plekname: Overijssel, Zoeterwoud, Wijchen, Beverwijk, Enschede, ensovoorts.
- Persoonname: Adelwijn, Neeltje, Gijs, Matthijs, De Bruijn, ensovoorts.
- Organisasiename: Koninklijke Boskalis Westminster, Verenigde Oost-Indische Compagnie, ensovoorts.

ŉ Probleem ontstaan wanneer benoemde entiteite identies aan selfstandige naamwoorde is, soos in die geval van die sanger “Koos Kombuis” waar *Kombuis* nie na *Keuken* verander moet word nie, of in die geval van die voormalige president van Suid-Afrika “F.W. de Klerk”, waar *Klerk* verkeerdelik na *Kantoorbediende* verander kan word.
In ŉ ander geval kan dit gebeur dat die benoemde entiteit van Nederlands afkomstig is, byvoorbeeld “van Wijk*”,* waar *Wijk* nie verander moet word na *Wyk* nie. ŉ Leksikon van moontlike benoemde entiteite, wat hierdie probleme veroorsaak, kan saamgestel word om hierdie probleem op te los. Elke entiteit moet dan eers in die leksikon opgesoek word voordat dit vernederlands word.

2.3.2. Nie-identiese kognate

Benoemde entiteite in hierdie kategorie vertoon sistematiese verskille tussen Afrikaans en Nederlands. Verskille kan impliseer dat ŉ letter of twee verander moet word, of selfs dat die entiteit

nie met ŉ hoofletter in Nederlands geskryf word nie. Voorbeelde van entiteite in hierdie kategorie sluit in:

- z-> s (zondag -> Sondag)
- sch -> sk (scheikunde -> skeikunde)
- tie -> sie (Deense-vakantiedag -> Deense-vakansiedag)
- c -> k (Congo -> Kongo)
- ch -> g (Biotechnologie -> Bio-tegnologie)
- c -> s (Centrum –> Sentrum)
- ij -> y (Argentijn -> Argentyn)

(Van Huyssteen & Pilon, 2009)

- Kleinletters -> Hoofletters (by dae van die week en maande van die jaar) (Ehlers & Van Beek, 2004).

2.3.3. Onverwante entiteite

Onverwante entiteite vertoon min of geen ooreenkoms met mekaar nie en kan moontlik die resultate van die Nederlandse BEH op Afrikaanse data negatief beïnvloed. ŉ Voorbeeld van ŉ onverwante entiteit is:

- Fryslân (ND) -> Friesland (AF).

Ten opsigte van bogenoemde verskille tussen Afrikaanse- en Nederlandse data, blyk dit dat daar wel op een of ander manier gepoog moet word om hierdie verskille te minimaliseer. Daar kan ook nie net op die benoemde entiteite in hierdie opsig gefokus word nie, maar dit is ook belangrik om te fokus op die konteks waarin die benoemde entiteite voorkom, aangesien dit bepaal of ŉ woord ŉ benoemde entiteit is, al dan nie. Weens hierdie rede is daar besluit om van A2DC (Van Huyssteen & Pilon, 2009) gebruik te maak om die Afrikaanse teks te vernederlands in ŉ poging om die resultate van die Nederlandse BEH op Afrikaanse data te verbeter. ŉ Volledige bespreking van A2DC word in Hoofstuk 3 gegee.

2.4. Samevatting

Aangesien die ontwikkeling van kerntegnologieë vir enige hulpbronskaars taal ŉ baie duur en tydsame proses kan wees, blyk dit uit die verskeie voorbeelde wat in hierdie literatuurstudie genoem word dat tegnologieherwinning gebruik kan word om die ontwikkeling van hierdie tegnologieë vir hulpbronskaarstale te bespoedig. Daar word dan verder aanvaar dat tegnologieherwinning ook vir Afrikaans gebruik kan word deur ŉ reeds ontwikkelde Nederlandse benoemde-entiteitherkenner aan te pas om Afrikaanse benoemde entiteite te identifiseer en ook te klassifi-

seer. Hierdie proses word verder ook deur die literatuur bevestig deur die klein semantiese en ortografiese verskille wat tussen Afrikaanse- en Nederlandse benoemde entiteite bestaan. Hierdie verskille is van so ’n aard dat dit moontlik sal wees om die Nederlandse BEH met Afrikaanse data te gebruik deur sekere pre- en/of postprosesseringstappe by te voeg (soos byvoorbeeld A2DC). Aangesien die tegnologieherwinning goeie resultate vir ander toepassings en ander taalpare gelewer het, gaan dit in hierdie studie gebruik word om ’n BEH vir Afrikaans te ontwikkel.

3. Die ontwikkeling van ’n BEH vir Afrikaans

3.1. Inleiding

In die vorige hoofstuk is verskeie voorbeelde van studies verskaf waar tegnologieherwinning suksesvol aangewend is om verskeie probleme in die praktyk op te los. Daar is ook in die vorige hoofstuk verwys na die verskille tussen Afrikaanse- en Nederlandse benoemde entiteite. In hierdie hoofstuk sal ’n Afrikaanse BEH ontwikkel word en in die ontwikkelingsproses sal ses verskillende eksperimente uitgevoer en bespreek word. ’n Uiteensetting van die verskillende eksperimente is soos volg:

- **Nederlandse Eksperiment:** In hierdie eksperiment is Nederlandse data deur *Frog*[2] (sien afdeling 3.2), ’n Nederlandse BEH, geannoteer en die resultate daarvan word geanaliseer.
- **Roudataeksperiment:** In hierdie eksperiment word *Frog* gebruik om rou Afrikaanse data te annoteer en die afvoer daarvan word geëvalueer.
- **A2DC-eksperiment:** In hierdie eksperiment word die Afrikaanse data deur middel van ’n pre-prosesseringsstap, A2DC (sien afdeling 3.4), vernederlands en *Frog* word weer gebruik om die data daarvan te annoteer. Die afvoer van *Frog* sal weereens geëvalueer word.
- **Veranderde A2DC eksperiment:** Gegewe die feit dat A2DC nie ontwikkel is om sekere entiteite te hanteer nie (Van Huyssteen & Pilon, 2009), was dit nodig om die module aan te pas om akronieme sowel as dubbelloopname wel te kan hanteer. Die Afrikaanse data word dan deur die veranderde A2DC vernederlands waarna die akkuraatheid van *Frog* op hierdie data geëvalueer word.
- **Roudataeksperimemt met gazetteers:** Hierdie eksperiment volg dieselfde proses as die roudataeksperiment, behalwe dat daar ’n ekstra pre-prosesseringsstap bygevoeg word waar entiteite in gazetteers opgesoek en dan geëtiketteer word.
- **Veranderde A2DC-eksperiment met gazetteers:** Hierdie eksperiment is in beginsel dieselfde as die vorige veranderde A2DC eksperiment, maar in hierdie eksperiment word ’n preprosesseringsstap (opsoek van entiteite in gazetteers) eers toegepas om die entiteite te identifiseer en te etiketteer. Daarna word die Afrikaanse data deur A2DC vernederlands en deur *Frog* geannoteer en die resultate daarvan geëvalueer.

Die etiketstel vir die benoemde-entiteitherkenning van *Frog* bestaan uit ses etikette wat elk IOB-notasies het (Tjong Kim Sang, 2002). Die IOB-notasies maak gebruik van drie soorte notasies I-

[2] Frog was voorheen bekend as TADPOLE (http://ilk.uvt.nl/downloads/pub/papers/tadpole-final.pdf)

,O- en B- wat onderskeidelik “insluitend”, “uitsluitend” en die “begin” van benoemde entiteite aandui. Die etikette stel die volgende kategorieë voor:

- ORG (Organisasiename soos “Suid-Afrikaanse Poliesiediens”);
- PER (Persoonsname en vanne soos “President Thabo Mbeki”);
- EVE (Gebeurtenisse soos “Vryheidsdag”);
- PRO (Produkname soos “The World koerant”);
- LOC (Plekname soos “Tshwane”); en
- MISC (Alle benoemde entiteite wat nie in bogenoemde kategorieë val nie soos “MIV”). (Chinchor & Robinson, 1998)

Daar is altesaam twaalf etikette omdat elke etiket beide ’n I- en ’n B-notasie bevat. Die “O”-notasie stel dat ’n entiteit of woord nie deel van ’n benoemde entiteit is nie en dus vorm dit nie deel van die etiketnotasies vir benoemde entiteite nie.

In afdeling 3.2. volg ’n beskrywing van *Frog* gevolg deur ’n bespreking van die BEH wat in *Frog* gebruik word. Daarna, in afdeling 3.3, volg ’n beskrywing van A2DC (“Afrikaans-to-Dutch Converter”). In afdeling 3.4 word die ses eksperimente beskryf en die resultate van elke eksperiment word bespreek. Afdeling 3.5 bevat ’n samevatting van die bevindinge van die eksperimente.

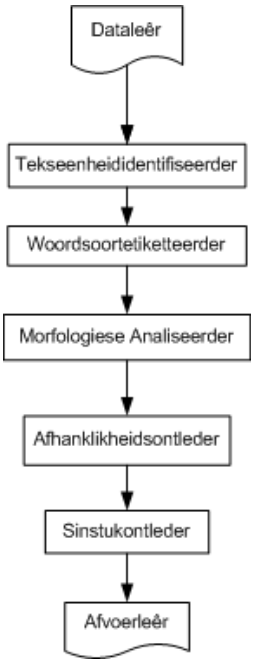
3.2. Frog

Volgens Van den Bosch *et al*. (2007a) is *Frog* ’n module-georiënteerde sintaktiese etiketteerder, analiseerder en sinstukontleder vir Nederlands. Die kern van die modules is gebaseer op geheue-gebaseerdeleer wat bestaan uit ’n *k*-Naastebuurpuntklassifiseerder en IGTREE (besluitnemingsboom). IGTREE is bekend daarvoor dat dit die klassifikasieproses van *k*-naastebuurpunte meervuldig bespoedig, maar nog steeds die akkuraatheid daarvan behou (Van den Bosch *et al*., 2007).

Die hooffunksie van *Frog* is om Nederlandse teks outomaties met morfo-sintaktiese inligting te annoteer en ook om die sintaktiese verhoudings tussen woorde op sinsvlak te bepaal. ’n Tekseenheididentifiseerder word as preprosesseringstap gebruik. Hierdie reëlgebaseerde tekseenheididentifiseerder verwyder leestekens met behulp van lyste van Nederlandse afkortings en verdeel ook sinne ten opsigte van heuristiese reëls (Rynaert, 2007).

Nadat die teks in tekseenhede verdeel is, word dit deur ’n woordsoortetiketteerder en morfologiese analiseerder geannoteer. Nadat die woordsoortetiketteerder die regte woordsoortetikette voorspel het, word dit na die morfologiese analiseerder gestuur, wat die woordsoortetikette gebruik om tussen dubbelsinnige woorde te onderskei. Die woordsoortetikette word ook gebruik as toevoer vir die afhanklikheidsontleder. ’n Ander struktuur, die sinstukontleder, gebruik ’n vasgestelde lys van multiwoordfrases en multiwoordeiename om die afhanklikheid van woorde in ’n gegewe sin te bepaal (Van den Bosch *et al*., 2007) (Sien Figuur 3).

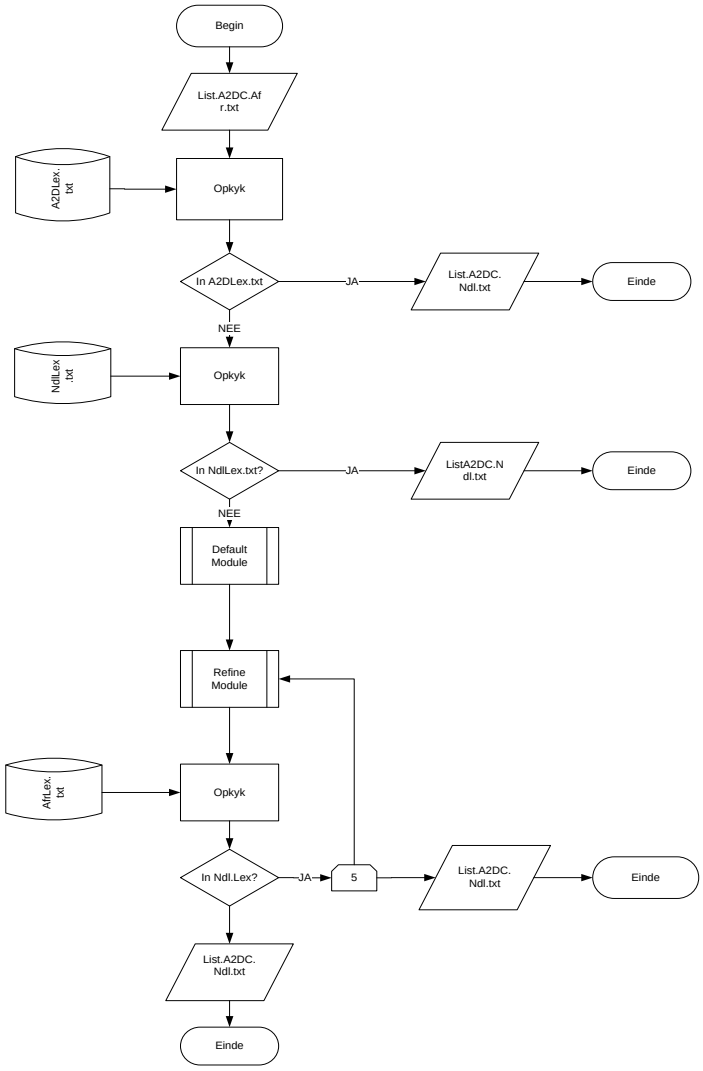
Volgens Van den Bosch (2012) werk *Frog* se BEH-module met ’n geheuegebaseerde-etiketeerder, genaamd die “memory-based tagger-generator” (MBT). Gedurende prosessering word al die opeenvolgende woorde geprosesseer ten opsigte van die konteks waarin hulle voorkom tesame met die vorige woord se klassifikasie. Op hierdie manier word die IOB-notasie voorspel wat aan die entiteit toegeken moet word. Die BEH in *Frog* bestaan uit twee modules, ’n IGTREE-gebaseerde module vir woorde wat bekend is en ’n TRIBL2-gebasseerde module vir onbekende woorde. TRIBL is ’n hibriede module wat beide IB1 en IGITREE kombineer (Dake, 2003). Die TRIBL2-algoritme begin op dieselfde manier as die IGTREE, maar skakel oor na IB1 wanneer ’n woord nie bekend is nie (Dake, 2003). Die bekende woorde word in ’n eienskapsvektor geplaas en vir die onbekende woorde word karaktereienskappe geskep wat dan help om hulle verder te klassifiseer (byvoorbeeld eerste letter, laaste letter, of dit ’n hoofletterwoord is, of dit ’n koppelteken of nommers bevat, ensovoorts).



Figuur 3: Vloeidiagram van prosesseringstappe van Frog

3.3. Afrikaans-na-Nederlands-omskakelaar (A2DC)

Om die resultate van *Frog* te verbeter gaan die data vernederlands word deur ’n Afrikaans-na-Nederlands-omskakelaar te gebruik. Van Huyssteen en Pilon (2009) het ’n omskakelaar, A2DC (“Afrikaans-to-Dutch converter”), ontwikkel wat Afrikaanse data na Nederlands kan verander. A2DC bestaan uit Perl-programmatuur en verskeie toevoer- en datalêers (sien Figuur 2). Van Huyssteen en Pilon (2009) word deurgaans as bronverwysing vir hierdie afdeling gebruik, behalwe waar anders gespesifiseer word.



Figuur 4: Vloeidiagram van die werking van A2DC

Enige teks wat deur A2DC gevoer word, moet eers vooraf in tekseenhede verdeel word. Die lys moet ook geen verkeerd gespelde woorde of akronieme bevat nie. Soos in Figuur 4 waargeneem kan word neem A2DC as toevoer ’n lys van Afrikaanse tekseenhede (List.A2DC.Afr.txt).

Vir die omskakeling van tekseenhede maak A2DC gebruik van twee leksikons, naamlik AfrDu.llex en AfrDu.tlex. AfrDu.llex word gebruik vir opsporing van identiese kognate (woorde

wat ortografies en semanties dieselfde is vir albei tale, en wat nie vertaal moet word nie) en AfrDu.tlex is ’n tweetalige lys van Afrikaanse tekseenhede en die Nederlandse vertalings van hierdie tekseenhede. Die doel van hierdie leksikon is om valse vriende (woorde wat ortografies dieselfde is, maar semanties verskil) en nie-kognate (woorde wat semanties dieselfde is, maar ortografies van mekaar verskil) te vertaal. Die grafeem-na-foneemomskakelingsreëls word outomaties deur die “Default and Refine”-algoritme (DR-algoritme) uit tweetalige data afgelei.

Die DR-algoritme is eintlik ontwikkel om outomaties grafeem-na-foneemomskakelingsreëls uit getranskribeerde data af te lei. Die algoritme maak gebruik van gulsige soek (“greedy search”) om die grafeem-na-foneem-reëls te soek wat die mees korrekte omskakelings in die afrigtingsdata sal veroorsaak (Davel & Barnard, 2004). Met gulsige soek word bedoel dat daar soveel gevalle as moontlik van ’n reël gesoek word en nie net te stop indien die reël vir die eerste keer opgespoor is nie. Elkeen van die grafeme, tesame met die kontekswoorde rondom dit, word met elke reël in die grafeem-na-foneem-stel vergelyk (Davel & Barnard, 2004). Die eerste reël wat pas, word dan op die grafeem toegepas. Na die toepassing van die reël word die grafeem met sy ooreenstemmende foneem belyn (Davel & Barnard, 2004).

Die afvoerleêr van A2DC (List.A2DC.Ndl.txt) bestaan uit ’n lys geëtiketteerde tekseenhede en toon die verandering aan wat elke woord ondergaan het in die omskakelingsproses. Die onderstaande etikette word vir klassifikasie gebruik.

- **<Translated>**: Hierdie etiket dui aan dat die tekseenheid volledig vernederlands is deur die DR-algoritme te gebruik.
- **<Untranslated>**: Hierdie etiket word gebruik wanneer daar geen veranderinge aan die tekseenheid aangebring is nie en dit ook nie in die teikentaal leksikon voorkom nie.
- **<NO WORD>**: Hierdie etiket dui aan dat die tekseenheid ’n getal, simbool of ander nie-woord karakters is, byvoorbeeld *2007*, *15*, #, ens.
- **<DrError>**: Hierdie etiket word toegeken aan tekseenhede wat karakters bevat wat nie deur die reëls hanteer kan word nie (DR-algoritme).
- **<LookupLex>**: Hierdie etiket dui aan dat die tekseenheid in die brontaalleksikon gevind is.
- **<TargetLex>**: Hierdie etiket dui aan dat die tekseenheid in die teikentaalleksikon gevind is.

3.4. Eksperimente

Vir hierdie studie is besluit om ’n eksperimentele benadering te ondersoek waar die toevoer data vir ses eksperimente verskillend is, maar elke eksperiment se prosedure min of meer dieselfde is. Vir die eerste eksperiment gaan Nederlandse data deur *Frog* geannoteer word. Hierdie Nederlandse data is soortgelyk aan die Afrikaanse data wat vir die volgende vyf eksperimente ge-

bruik gaan word met betrekking tot die totale aantal woorde, die totale aantal lyntjies en aantal benoemde entiteite wat dit bevat. Hierdie eerste eksperiment gaan ook dien as metingsmeganisme vir die eksperimente wat volg. Dit wil sê dat die Nederlandse eksperiment se resultate gebruik gaan word om aan te toon of die resultate van die ander eksperiment voldoende is, of nie.

In die tweede eksperiment word onveranderde (rou) Afrikaanse data deur *Frog* geannoteer en in die derde eksperiment word die Afrikaanse data met behulp van A2DC, vernederlands (Van Huyssteen & Pilon, 2009). Die vernederlandsde data word dan ook deur *Frog* geannoteer. In die vierde eksperiment is daar aan A2DC veranderinge aangebring om die probleme wat uit die afvoer van die vorige eksperiment opgetel is, reg te maak. Die Afrikaanse data is weereens deur die veranderde A2DC vernederlands en is daarna deur *Frog* geannoteer. In die laaste twee eksperimente word die tweede eksperiment (roudataeksperiment) en die vierde eksperiment (veranderde A2DC-eksperiment) weer ondersoek. By hierdie eksperimente word ’n ekstra preprosesseringstap ten opsigte van opsoek van entiteite in gazetteers bygevoeg.

Die teks vir die eerste eksperiment is vanaf die Europarl webtuiste (Khoen, 2005) verkry. Die Nederlandse gedeelte van die korpus bestaan uit 53,487,257 woorde. Die gedeelte wat wille keurig vir die eksperiment onttrek is, bevat 9 148 woorde en 373 benoemde entiteite wat ongeveer dieselde is as die Afrikaanse teks wat in die ander vyf eksperimente gebruik is.

Vir al vyf oorblywende eksperimente is dieselfde teks gebruik, naamlik “Die Staatsrede van Suid-Afrika” van 2007 (Suid-Afrikaanse Regering, 2012) . Die teks bestaan uit 9 124 woorde en 310 benoemde entiteite wat handmatig met IOB-annotasies (Tjong Kim Sang, 2002) geannoteer is om as goudstandaard in die evaluasies te dien. Die handmatige annotasies is dan verder gekontroleer deur ’n taalkundige om die korrekte standaarde te handhaaf. Die gazetteers wat vir die laaste twee eksperimente gebruik is, is saamgestel uit name, vanne, handelsname, dae van die week, maande van die jaar, titels en plekname wat vanaf die internet verkry is (Wikipedia, 2013a; 2013b; 2013c; 2013d; 2013e, 2013f).

3.4.1. Nederlandse eksperiment

Die teks vir die Nederlandse eksperiment is onttrek uit die Europarl-korpus (Koehn, 2002) en bevat 9 148 woorde en 373 benoemde entiteite. Hierdie data is gekies sodat dit ongeveer dieselfde aantal woorde en entiteite bevat as wat in die ander eksperimente gebruik is en is ook deur *Frog* geannoteer. Uit Tabel 6 kan waargeneem word dat 17.8% (76 uit 426) van die gevalle wat deur *Frog* as benoemde entiteite geïdentifiseer is, nie benoemde entiteite is nie (vals positiewes).

Totale aantal benoemde entiteite geïdentifiseer.	426
Aantal vals positiewes geïdentifiseer.	76
Aantal benoemde entiteite nie geïdentifiseer nie.	23

Tabel 6: Benoemde-entiteitidentifikasies van Nederlandse data

Vervolgens is 6,2% (23 uit 373) van die benoemde entiteite wat geïdentifiseer moes word, nie geïdentifiseer nie. Uit die analises van die data blyk dit dat die grootste probleem by titels (by-voorbeeld Meijnheer, Mevrouw, ens.) voorkom. Die analise toon ook dat titels soos Meijnheer en Mevrouw nie as benoemde entiteite herken word deur *Frog* nie.

Presisie	0.82
Herroeping	0.94
f-telling	0.88

Tabel 7: Resultate vir die benoemde-entiteitidentifikasie van Nederlandse data

Uit Tabel 7 blyk dat die identifikasie oor die algemeen goed is en vergelykbaar is met ander Nederlandse BEH's (Desmet & Hoste, 2010; Black & Vasilakopoulos, 2002; Burger *et al*., 2002; Cucerzan & Yarowsky, 2002; Malouf, 2002).

Etiket	**Herroeping**	**Presisie**	***f*-telling**
ORG	0.96	0.98	0.97
PER	0.68	0.84	0.75
EVE	0	0	0
PRO	0	0	0
LOC	0.97	0.52	0.68
MISC	0.83	0.68	0.75

Tabel 8: Resultate vir elke groep etikette van Nederlandse data

Die presisie, herroeping en ***f***-telling vir die verskeie kategorieë is op dieselfde manier bereken as wat dit vir die algehele eksperiment bereken is. Tabel 8 dui aan hoe akkuraat die klassifikasie proses vir duie Nederlandse data was. Uit Tabel 8 blyk dat (hoewel daar geen gebeurtenisse ("EVE") of produkte ("PRO") in die teks voorgekom het nie) die etikette goeie resultate toon. Die enigste probleem, is die feit dat die presisie vir LOC relatief laag is.

	ORG	EVE	PER	LOC	PRO	MISC
ORG	**156**	*	1	*	*	*
EVE	*	*	*	*	*	*
PER	1	*	**21**	*	1	1
LOC	4	*	8	**88**	2	14
PRO	*	*	*	*	*	1
MISC	2	*	1	*	1	**87**

Tabel 9: Verwarringsmatriks vir etiket-toekenning van Nederlandse data

Tabel 9 gee 'n aanduiding van hoeveel etikette korrek aan benoemde entiteite toegeken is[3].Uit Tabel 9 kan die rede vir die lae presisie van LOC opgemerk word. Volgens die tabel word 22,4% (28 van 116) van die LOC-entiteite as ander entiteite geklassifiseer. Nog meer is dat 12,0% (14 van 116) as MISC geklassifiseer is wat 53,8% uitmaak van die etikette wat verkeerdelik geklassifiseer is. Tabel 10 toon die presisie, herroepping en *f*-telling vir die etikettoekenning vir die Nederlandse data.

Presisie	0.73
Herroeping	0.89
f-telling	0.80

Tabel 10: Resultate vir etikettering van Nederlandse data

3.4.2. Roudataeksperiment

Vir hierdie eksperiment (asook die volgende 5 eksperimente) is "Die Staatsrede van Suid-Afrika" van 2007 gebruik wat vrylik op die regering se webtuiste (Suid-Afrikaanse Regering, 2012) beskikbaar is. Die teks bestaan uit 9 124 woorde en bevat 310 benoemde entiteite. Hierdie benoemde entiteite is handmatig met IOB-annotasies (Tjong Kim Sang, 2000) geannoteer om 'n goudstandaard te skep waarmee die afvoer van die BEH vergelyk kan word. Die resultate word in Tabel 11 uiteengesit.

Totale aantal benoemde entiteite geïdentifiseer	438
Aantal vals positiewes geïdentifiseer	83
Aantal benoemde entiteite nie geïdentifiseer nie	14

Tabel 11: Benoemde-entiteitidentifikasies van rou Afrikaanse data

Uit Tabel 11 kan waargeneem word dat 18.9% (83 uit 438) van die entiteite wat geïdentifiseer is, nie geïdentifiseer moes word nie (valse-positiewe). 'n Totaal van 14 entiteite wat as benoemde entiteite moes geïdentifiseer word, is glad nie geïdentifiseer nie. Uit die data blyk dit dat die

[3] Let Wel: Aangesien sommige van die benoemde entiteite uit meer as een woord bestaan en aangesien elke etikette afsonderlik geëvalueer is, is die aantal korrekte etikette nie noodwendig gelyk aan die aantal etikette wat geïdentifiseer is nie (sien1.5.4).

meeste misidentifikasies as gevolg van die name van maande (September), dae (Woensdag), afgeleide name uit ander tale (Khampepe) en ook titels (President) is. Van die redes waarom name van maande en dae nie korrek geïdentifiseer is nie, is omdat hierdie woorde met kleinletters in Nederlands geskryf word (Ehlers & Van Beek, 2004). Die rede vir die misidentifikasies is waarskynlik die feit dat dit nie algemeen in Nederlands voorkom nie.

Presisie	0.68
Herroeping	0.95
f-telling	0.79

Tabel 12: Resultate vir die benoemde-entiteitidentifikasie van rou Afrikaanse data

In vergelyking met die resultate van die Nederlandse data (Tabel 7) blyk dit uit Tabel 12 dat daar ŉ groot afname in presisie vir die rou Afrikaanse data is, maar die herroeping van hierdie eksperiment is wel beter. Die lae presisie dra daartoe by dat die *f*-telling vir die roudataeksperiment veel laer is as diè van die Nederlandse eksperiment. Hierdie verskynsel is nie onverwags nie, aangesien daar ortografiese verskille tussen Afrikaans en Nederlands voorkom.

Etiket	**Herroeping**	**Presisie**	***f*-telling**
ORG	0.59	0.50	0.54
PER	0.49	0.25	0.34
EVE	0.64	0.88	0.74
PRO	0.90	0.16	0.26
LOC	0.66	0.50	0.57
MISC	0.28	0.42	0.34

Tabel 13: Resultate vir elke groep etikette van rou Afrikaanse data

Uit Tabel 13 kan waargeneem word dat die etiketspesifieke herroeping, presisie en *f*-telling heelwat laer is as vir die Nederlandse data. Wat nog opmerkbaar is, is dat al die etiket kategorieë, behalwe EVE, se presisie baie laag is. Die moontlike rede vir die hoë presisie vir EVE is dat daar slegs vier gevalle vir EVE is en dat hierdie gevalle nie maklik verwar kan word met ander benoemde-entiteite (soos PER, ORG, ensovoorts) nie omdat dit nie in dieselfde konteks as PER en ORG gebruik word nie. ŉ Voorbeeld van ŉ EVE is “Vryheidsdag” wat dus dui op ŉ gebeurtenis in tyd.

	ORG	EVE	PER	LOC	PRO	MISC
ORG	**67**	*	20	12	12	12
EVE	*	**6**	*	5	*	1
PER	7	*	**25**	13	3	2
LOC	5	*	8	**60**	*	17
PRO	*	*	*	1	**8**	1
MISC	14	2	27	23	6	**33**

Tabel 14: Verwarringsmatriks vir etiket-toekenning van rou Afrikaanse data

Uit Tabel 14 kan die volgende afleidings gemaak word:

- ORG: 45.5% (56 van 123) word as ander etikette geklassifiseer (PER, LOC, PRO en MISC)
- MISC: Slegs 31.4% (33 van 105) is korrek as MISC geklassifiseer, maar 25% (27 van 105) daarvan is as PER geklassifiseer.

Sommige van die foute wat die BEH maak, is waarskynlik te wyte aan die feit dat die konteks waarin die Afrikaanse organisasiename in die data voorkom verskil van die konteks waarin Nederlandse entiteite voorkom. In die geval van die MISC-klassifikasies kan dit bloot wees dat die entiteite nie in Nederlands voorkom nie (byvoorbeeld “Mbeki”) wat ook verder kan bydra tot die verkeerdelike klassifikasie van MISC-entiteite as persoonsname (PER).

Tabel 15 toon die kombinasies van die verkeerd geëtiketteerde entiteite. Die tabel toon die korrekte etiket gevolg deur die verkeerde etiket (byvoorbeeld LOC (korrek) -> ORG (verkeerd)). Uit Tabel 15 kan waargeneem word dat die grootste probleem voorkom by entiteite wat as MISC geëtiketteer moes word, maar as LOC geëtiketteer is (23.08% van al die verkeerde etikette). Voorbeelde hiervan sluit in: “Februarie”, “Presidentlosie”, “Suid-Afrikaners” en “Hooggeregshof”. ’n Moontlike rede vir hierdie foute kan die konteks wees waarin hulle voorkom. Verder kan dit ook die feit wees dat sommige van hierdie entiteite ongewone woorde in Nederlands en daarom ook vir die Nederlandse BEH is. Die feit dat baie min entiteite van LOC->MISC geëtiketteer is, maar dat daar baie van MISC->LOC geëtiketteer is, toon aan dat daar wel nie ’n verwarring tussen die LOC en MISC plaasgevind het nie, aangesien die hoeveelhede nie ooreenstemmend is nie.

Ander etikettoekenningkombinasies wat hoofsaaklik bygedra het tot verkeerde etikettering is: MISC->PER (15.38%), ORG->PER (11.19%) en ORG->MISC (8.39%). Altesaam het hierdie vier etiketkombinasies bygedra dat 58.04% (83 uit 143) van die totale verkeerd geëtiketteerde entiteite. Die verkeerdelike toekenning van MISC->PER en ook ORG->MISC is reeds bespreek (sien bespreking by Tabel 14). Die rede vir die verkeerde klassifikasies van ORG->PER kan bloot wees dat *Frog* die organisasienaam verkeerdelik as ’n persoonsnaam identifiseer.

Tabel 15 gee ŉ aanduiding van die tipe kombinasies wat bygedra het tot etikettering van benoemde entiteite. Verder kan daar ook uit Tabel 15 afgelei word dat daar ŉ probleem bestaan ten opsigte van ORG en MISC. Aan die kombinasie ORG->MISC kan 8.57% van die verkeerde etikettoekenning toegeskryf word en MISC->ORG beslaan 7.14% van die verkeerde etikettoekenning. Daar kan dus aangeneem word dat dit ŉ probleem vir die sisteem is om tussen organisasiename en ander benoemde entiteite wat nie in die ander kategorieë val nie (MISC), te onderskei. Voorbeelde van hierdie tipe woorde is: “MIV” (MISC->ORG) en “EPWP” (ORG->MISC). Die grootste rede blyk dat die benoemde entiteite akronieme is wat nie noodwendig in Nederlands voorkom nie. Die presisie, herroeping en *f*-telling vir etikettering, word in Tabel 16 gegee.

Regte etiket –> Verkeerde etiket	Aantal	Persentasie
MISC->LOC	33	23.08
MISC->PER	22	15.38
ORG->PER	16	11.19
ORG->MISC	12	8.39
MISC->ORG	10	6.99
ORG->LOC	9	6.29
PER->LOC	9	6.29
LOC->PER	6	4.2
LOC->ORG	5	3.5
MISC->PRO	5	3.5
ORG->PRO	4	2.8
LOC->MISC	3	2.1
PER->ORG	2	1.4
EVE->LOC	2	1.4
PER->MISC	1	0.7
PRO->MISC	1	0.7
PRO->LOC	1	0.7
EVE->MISC	1	0.7
PER->PRO	1	0.7
TOTAAL	143	100

Tabel 15: Persentasie van verskillende etiketkombinasies van rou Afrikaanse data

Presisie	0.41
Herroeping	0.52
f-telling	0.46

Tabel 16: Resultate vir etikettering van rou Afrikaanse data

3.4.3. Vernederlandste data met behulp van A2DC

Vir hierdie eksperiment is die Afrikaanse data vernederlands deur A2DC (Van Huyssteen & Pilon, 2009) te gebruik. Daar word verwag dat die preprosesseringstap die akkuraatheid van die identifikasie en etikettering van die entiteite sal verbeter en dat dit resultate sal lewer wat vergelykbaar is met diè van die Nederlandse eksperiment. Van die 310 benoemde entiteite wat moes identifiseer word, is 16 nie geïdentifiseer nie. Die resultate word in Tabel 17 uiteengesit.

Totale aantal benoemde entiteite geïdentifiseer	412
Aantal vals positiewes geïdentifiseer	57
Aantal benoemde entiteite nie geïdentifiseer nie	16

Tabel 17: Benoemde-entiteitidentifikasies van A2DC data

In vergelyking met dieselfde resultate van die roudataeksperiment, kan daar in Tabel 17 waargeneem word dat die aantal vals positiewes van 18.9% afgeneem het na 13.8% (57 van 412). Alhoewel daar in totaal meer entiteite geklassifiseer is, is daar steeds entiteite wat nie as benoemde entiteite geklassifiseer is nie. In Tabel 17 kan waargeneem word dat daar steeds sestien (16) benoemde-entiteite is wat nie geïdentifiseer is nie. ’n Voorbeeld hiervan is UNDP. Uit die analise van die afvoer van *Frog* blyk dit dat UNDP as ’n selfstandige naamwoord eerder as ’n ORG benoemde entiteit geklassifiseer is, wat beteken dat *Frog* nie die woord as ’n akroniem herken het nie.

Presisie	0.71
Herroeping	0.95
f-telling	0.81

Tabel 18: Resultate vir die benoemde-entiteitidentifikasie van A2DC data

In vergelyking met die roudataeksperiment blyk dit uit Tabel 18 dat die herroeping konstant gebly het, maar die presisie van die sisteem het verbeter wat ook verder gelei het tot ’n beter *f*-telling (0.81). ’n Afleiding wat hieruit gemaak kan word, is dat die akkuraatheid van die sisteem wel verbeter het met behulp van die vernederlandsde data.

Etiket	Herroeping	Presisie	*f*-telling
ORG	0.66	0.60	0.63
PER	0.57	0.44	0.50
EVE	0.36	0.80	0.50
PRO	0.90	0.23	0.36
LOC	0.66	0.46	0.54
MISC	0.31	0.43	0.36

Tabel 19: Resultate vir elke groep etikette van A2DC data

In Tabel 19 kan ’n groot verbetering ten opsigte van die toekenning van etikette in die ORG-kategorie (0.54 na 0.63) waargeneem word, wat daarop dui dat die probleem met organisasie-name in die vorige eksperiment moontlik ’n taalverwante probleem was. Verder is daar verbetering vir al die ander etiketkategorieë, behalwe vir EVE, wat ’n afname getoon het (0.74 na 0.50). Uit die analise van die afvoer van *Frog* blyk dit dat sommige van die EVE benoemde entiteite eerder as MISC geklassifiseer is.

	ORG	EVE	PER	LOC	PRO	MISC
ORG	**75**	*	8	19	5	16
EVE	1	**3**	1	2	*	3
PER	2	*	**31**	10	5	1
LOC	8	*	7	**62**	*	16
PRO	*	*	*	2	**11**	*
MISC	15	2	13	31	6	**34**

Tabel 20: Verwarringsmatriks vir etiket-toekenning van A2DC data

Tabel 20, in vergelyking met die roudataeksperiment (Tabel 14), toon dat die helfte van die etiketkategorieë (ORG, PER en PRO) verbeter het ten opsigte van die korrekte klassifikasie. Die ander helfte van die etikette (EVE, LOC en MISC) wat ’n afname getoon het, se verskil was nie meer as drie of vier klassifikasies nie. Verder kan daar ook waargeneem word dat daar nog steeds ’n onreëlmatigheid ten opsigte van die klassifikasies vir die MISC- en ORG-kategorieë is. Ten opsigte van MISC word 30.6% as LOC geklassifiseer. Hierdie probleem het meestal te make met die vernederlandsing van plekname met koppeltekens soos byvoorbeeld “Zuid-Afrika”.

Ten opsigte van ORG word 39% as ander etikette geklassifiseer, waarvan LOC en MISC die grootste deel uitmaak. Dieselfde verskynsel kan ook by PER gesien word waar 36.7 % daarvan as ander etikette geklassifiseer word, meestal as LOC. ’n Verdere ondersoek lei tot die volgende gevolgtrekkings ten opsigte van die verkeerde geëtiketteerde entiteite (Tabel 21):

Regte etiket –> Verkeerde etiket	Aantal (Verskil)	Persentasie
MISC->LOC	29 (-4)	21.64
LOC->MISC	16 (+13)	11.94
ORG->LOC	16 (-7)	11.94
ORG->MISC	12 (-1)	8.96
MISC->ORG	11 (+2)	8.21
PER->LOC	9	6.72
MISC->PER	7 (-15)	5.22
LOC->ORG	7 (-2)	5.22
ORG->PER	5 (-11)	3.73
LOC->PER	5 (-1)	3.73
MISC->PRO	5	3.73
ORG->PRO	3 (-1)	2.24
PER->PRO	2 (+1)	1.49
PER->MISC	1	0.75
PER->ORG	1	0.75
EVE->LOC	1 (-1)	0.75
EVE->MISC	1	0.75
PRO->MISC	*	*
PRO->LOC	*	*
TOTAAL	134	100

Tabel 21: Persentasie van etikettoekenning van verskillende etiketkombinasies van A2DC data

Die eerste opmerkbare verandering uit Tabel 21 is dat beide die PRO->MISC- en PRO->LOC-kombinasies (vetdruk in Tabel 21) nie meer verkeerdelik geklassifiseer word in die vernederlandsde data nie. ’n Ander belangrike opmerking is dat die totaal verkeerd geklassifiseerde benoemde entiteite van 143 na 134 afgeneem het. Laasgenoemde waarneming gee die indruk dat dit tog voordelig kan wees om die Afrikaanse data te vernederlands.

Daar kan ook uit Tabel 21 waargeneem word dat daar wel veranderinge ten opsigte van die verskillende kombinasies van verkeerd geëtiketteerde entiteite plaasgevind het. Die drie mees opvallende veranderinge is by LOC->MISC wat met 13 gevalle vermeerder het, MISC->PER wat met 15 gevalle verminder het en ORG->PER wat met 11 gevalle verminder het.

Alhoewel daar baie voordele is in die vernederlandsing van Afrikaans wanneer dit met ’n Nederlandse BEH geannoteer word, is daar wel nog probleme wat voorkom. Die eerste hiervan is dat die kombinasie van MISC->LOC nog steeds bydra tot die meeste verkeerd geëtiketteerde entiteite (21.64%). MISC->LOC tesame met LOC->MISC, ORG->LOC, MISC->ORG en ORG-

>MISC dra by tot 62.68% (84 van 134) van die verkeerd geëtiketteerde entiteite. Voorbeelde van hierdie kombinasies is: “Millenniumontwikkelingsdoelwit” (MISC->LOC), “Suid-afrika”[4] (LOC->MISC), “Suid-afrikaanse Politiedienst”[4] (ORG->LOC), Ministerie van Veiligheid en Sekuriteit” (MISC->ORG) en “Reservebank” (ORG->MISC).

Presisie	0.48
Herroepping	0.56
f-telling	0.52

Tabel 22: Resultate vir etikettering van A2DC data

Tabel 22 toon aan dat die *f*-telling vir etiket klassifikasie oor die algemeen hoër is as vir die roudataeksperiment (0.46 na 0.52). Tabel 22 toon ook ’n verandering ten opsigte van die presisie en herroeping vir die etikettoekenning van *Frog.*

Alhoewel daar noemenswaardige verbeteringe in hierdie eksperiment teenoor die roudataeksperiment is, is die resultate nog nie vergelykbaar met die resultate wat vir die Nederlandse eksperiment verkry is nie.

Daar is wel ’n paar veranderinge wat aan A2DC aangebring kan word om voorsiening te maak vir die probleme (soos akronieme en dubbelloopname) wat tydens analise ontdek is. Soos Pilon & Van Huyssteen (2009) genoem het, is daar geen voorsienning gemaak vir akroniemidentifikasie en die vertaling daarvan nie. Hierdie is ’n probleem aangesien daar Afrikaanse akronieme is wat as benoemde entiteite voorkom wat dat nie volledig deur A2DC na hul Nederlandse vorm omgeskakel kan word nie.

Die probleem veroorsaak ook dat die Afrikaanse akronieme dan saam met die ander vertaalde tekste vir die Nederlandse BEH gegee word om te annoteer wat ’n laer akkuraatheid vir die sisteem tot gevolg kan hê.

Nog ’n probleem wat ook waargeneem is, kom voor by dubbelloopname (byvoorbeeld “Suid-Afrika”). Hierdie name word as een woord geïdentifiseer en word dan deur A2DC verander sodat die eerste letter ’n hoofletter is en die res almal kleinletters (byvoorbeeld “Suid-Afrika” word dus “Zuid-afrika” in plaas van “Zuid-Afrika”). Hierdie probleem dra daartoe by dat die woorde wat voor en na die koppelteken verskyn, nie in ag geneem word vir verandering of vertaling nie, wat dan ook die akkuraatheid van die sisteem benadeel.

[4] LET WEL: Hierdie woorde is direk ontrek uit die afvoer van *Frog* en die spelfoute was dus deel van die analise. Na verdere ondersoek blyk dit dat A2DC die oorsaak van die spelfout was en verandering is aangebring sodat nog ’n eksperiment uitgevoer kon word met die veranderde A2DC (sien 3.4.3).

3.4.4. Veranderde A2DC-eksperiment

Volgens die hipotese wat in hoofstuk 1 gestel is, behoort hierdie eksperiment met die veranderde A2DC se resultate die naaste te wees aan die Nederlandse eksperiment se resultate en ook die beste resultate te lewer vir die Afrikaanse data. Vir hierdie eksperiment gaan dieselfde Afrikaanse data wat vir die vorige eksperimente gebruik is, deur die veranderde A2DC gestuur word sodat dit vernederlands kan word. Daarna word die vernederlandsde data, soos in die geval van die vorige eksperimente, deur *Frog* gestuur om geannoteer te word. Die resultate word daarna geanaliseer.

Totale aantal benoemde entiteite geïdentifiseer	403
Aantal vals positiewes geïdentifiseer	56
Aantal benoemde entiteite nie geklassifiseer nie	16

Tabel 23: Benoemde-entiteitidentifikasies van veranderde A2DC data

Uit Tabel 23 kan waargeneem word dat daar minder entiteite geïdentifiseer is (412 na 403), alhoewel die aantal benoemde entiteite wat nie geïdentifiseer is nie, dieselfde gebly het. Een van die entiteite wat nie geïdentifiseer is nie is “President FW de Klerk”. Hierdie is ongewoon aangesien “President FW de Klerk” gedeeltelik in die vorige twee eksperimente geïdentifiseer is. Die aantal vals positiewes het wel van 57 na 56 afgeneem.

Presisie	0.73
Herroeping	0.95
f-telling	0.82

Tabel 24: Resultate vir die benoemde-entiteitidentifikasie van veranderde A2DC data

Vanuit die vorige tabel (Tabel 23) word daar verwag dat die presisie, herroeping en *f*-telling wel beter sal wees as in die vorige eksperiment en dit is inderdaad die geval. In Tabel 24 toon beide die presisie en *f*-telling verbeteringe en herroepping het onveranderd gebly. Die *f*-telling vir hierdie eksperiment is ook baie naby aan die *f*-telling vir die eksperiment met die Nederlandse data (0.82 in vergelyking met 0.88 wat vir die Nederlandse-eksperiment gekry is).

Etiket	**Herroeping**	**Presisie**	***f*-telling**
ORG	0.70	0.59	0.64
PER	0.51	0.46	0.48
EVE	0.36	0.80	0.50
PRO	0.90	0.32	0.47
LOC	0.66	0.46	0.54
MISC	0.27	0.51	0.35

Tabel 25: Resultate vir elke groep etikette van veranderde A2DC data

Alhoewel Tabel 25 'n klein afname in terme van die algemene presisie herroeping en *f*-telling toon blyk dit dat die veranderinge wat aan A2DC aangebring is nie veel van 'n verskil maak nie. Al etiketkategorieë wat wel 'n verbetering getoon het, is die LOC-groep (0.36 na 0.47). In vergelyking met die resultate vir die Nederlandse eksperiment is die klassifikasie van etikette nog ver van vergelykbaar af.

	ORG	EVE	PER	LOC	PRO	MISC
ORG	**79**	*	11	19	4	16
EVE	2	**3**	1	2	*	3
PER	3	*	**27**	10	3	1
LOC	8	*	7	**77**	*	1
PRO	*	*	*	1	**8**	1
MISC	13	2	8	39	5	**29**

Tabel 26: Verwarringsmatriks vir etiket-toekenning van Veranderde A2DC data

Uit Tabel 26 is dit duidelik dat die klassifikasie van etikette nog nie optimaal is nie. By ORG is 38.75% (50 van 129) van die entiteite foutiewelik as PER, LOC, PRO en MISC geklassifiseer wat aantoon dat die veranderinge aan A2DC nie veel van 'n bydra by hierdie kategorie gelewer het nie. 'n Ander voorbeeld is by MISC waar 40.63% (39 van 96) verkeerdelik as LOC geklassifiseer is.

Die eerste afleiding wat uit Tabel 27 gemaak kan word, is die twee ekstra kombinasies wat voorkom (gemerk met 'n *) wat nie in die vorige twee eksperimente verskyn het nie. Dit is ongewoon aangesien daar geen veranderinge gemaak is wat direk 'n invloed op die EVE-, ORG- of PER-kategorieë sou hê nie. Na verdere ondersoek is die volgende opgemerk. "Fifa-wêreldsokkerbeker" kom as EVE->ORG voor in hierdie eksperiment, maar was in die roudataeksperiment en in die A2DC-eksperiment as EVE->MISC geëtiketteer. Onder die ander kombinasie EVE->PER val die entiteit "Gesamentlike Sitting". Hierdie entiteit is in die roudataeksperiment as EVE->LOC en in die A2DC-eksperiment as EVE->PER geëtiketteer.

'n Groot probleem wat nog voorkom, is die MISC->LOC kombinasie wat weereens tot die grootste gedeelte van die verkeerd geëtiketteerde entiteite bydra met 30.56% of te wel 38 van die 124 verkeerd geëtiketteerde entiteite. Saam met die MISC->LOC kombinasie is daar drie ander kombinasies wat ook 'n groot aandeel het ten opsigte van verkeerd geëtiketteerde entiteite,
naamlik: ORG->MISC (10.48%), MISC->ORG (8.87%) en ORG->LOC (11.29%). Hierdie vier etiketkombinasies maak 61.20% (76 van 124) van die totale verkeerd geëtiketteerde entiteite uit. Vir hierdie eksperiment blyk dit weereens dat die BEH sukkel om te onderskei tussen ORG en onbekende entiteite (MISC), want vir ORG->MISC is daar 13 gevalle en vir MISC->ORG is daar 11 gevalle.

Regte etiket –> Verkeerde etiket	Aantal (Ver-skil)	Persentasie
MISC->LOC	38 (+9)	30.65
ORG -> LOC	14 (-2)	11.29
ORG -> MISC	13 (-1)	10.48
MISC -> ORG	11	8.87
LOC -> ORG	7	5.56
PER -> LOC	7 (-2)	5.56
MISC -> PER	5 (–2)	4.03
ORG -> PER	5	4.03
LOC -> PER	5	4.03
ORG -> PRO	5 (+2)	4.03
PER -> ORG	2 (-1)	1.61
*EVE -> ORG	2	1.61
LOC -> MISC	1 (-15)	0.81
PER -> MISC	1	0.81
EVE -> LOC	1	0.81
PER -> PRO	1	0.81
* EVE -> PER	1	0.81
PRO -> LOC	***1 (+1)***	***0.81***
PRO -> MISC	***1 (+1)***	***0.81***
MISC -> PRO	*****	*****
EVE -> MISC	*****	*****
TOTAAL	124	100

Tabel 27: Persentasie van etikettoekenning van verskillende etiketkombinasies van A2DC data

Oor die algemeen is daar ’n gedeeltelike afname ten opsigte van die etiketkombinasies met die grootste verskil by LOC->MISC wat ’n afname van 15 getoon het in vergelyking met die vorige eksperiment.

Presisie	0.52
Herroepping	0.58
f-telling	0.55

Tabel 28: Resultate vir etikettering van veranderde A2DC data

Uiteindelik toon Tabel 28 dat die presisie, herroeping en *f*-telling vir etikettoekenning wel bietjie verbeter het, maar dat die resultate nog nie vergelykbaar met die resultate van die Nederlandse eksperiment is nie. Aangesien die resultate uit die vorige drie eksperimente in terme van identi-

fikasie van benoemde entiteite aanvaarbaar was, maar die etikettoekenning van benoemde entiteite nie na wense was nie, is daar besluit om na ’n ander pre- en/of postprossesringstap te soek om die laasgenoemde probleem te probeer oplos. Die oplossing vir die probleem is gevind deur van gazetteers as pre-prosseseringstap aan te wend.

3.4.5. Roudataeksperiment met gazetteers

As gevolg van die onverwagte lae *f*-telling vir etikettoekenning by al drie die eksperimente met Afrikaanse data, is ’n oplossing gesoek om die telling te verbeter. ’n Moontlike oplossing blyk die gebruik van gazetteers (of naamlyste) as ’n pre-prosesseeringstap en dit is wat in hierdie eksperiment geëvalueer gaan word. Die Afrikaanse data word deur sagteware gevoer wat deur die gazetteers met name, vanne, titels, dae, maande, handelsmerke en plekname soek. Indien die benoemde entiteit waarna gesoek word in beide die Afrikaanse data en in een van die gazetteers voorkom, word dit in ’n lys geplaas met die korrekte etiket daarby. Die rou Afrikaanse data word dan deur *Frog* geanaliseer. Uit die afvoer van *Frog* word al die benoemde entiteite en hul etikette onttrek en vergelyk met die benoemde entiteite en hul etikette in die lys wat vroeër opgestel is. Indien daar enige verskille tussen die etikette van dieselfde benoemde entiteite voorkom, word die etiket aangepas soos dit in die lys geëtiketteer is. Die resultate word in Tabel 29 uiteengesit:

Totale aantal benoemde entiteite geïdentifiseer	491
Aantal vals positiewes geïdentifiseer	186
Aantal benoemde entiteite nie geklassifiseer nie	14

Tabel 29: Benoemde-entiteitidentifikasies van rou Afrikaanse data met gazetteers

Tabel 29 toon aan dat die totale entiteite wat geïdentifiseer is sowel as die aantal vals positiewes wat geïdentifiseer is, meer is as in die eerste roudataeksperiment (491 teenoor 438 vir die entiteite en 186 teenoor 83 vir die vals positiewes). Die aantal entiteite wat nie geïdentifiseer is nie het dieselfde gebly.

Presisie	0.61
Herroeping	0.97
f-telling	0.75

Tabel 30: Resultate vir die benoemde-entiteitidentifikasie van rou Afrikaanse data met gazetteers

Volgens Tabel 30 toon die eksperiment swakker resultate ten opsigte van presisie, herroeping en *f*-telling teenoor die eerste roudataeksperiment. Die swakker presisie (0.61 teenoor 0.68) wat ook gelei het tot ’n swakker *f*-telling (0.75 teenoor 0.79) is nie van veel belang nie aangesien die verskille baie klein is en die eintlike doel van hierdie eksperiment is om die *f*-telling van die etikettoekenning van benoemde entiteite te verbeter.

Etiket	Herroeping	Presisie	*f*-telling
ORG	0.46	0.76	0.58
PER	0.34	0.29	0.31
EVE	0.55	0.75	0.63
PRO	0.80	0.32	0.46
LOC	0.82	0.66	0.73
MISC	0.56	0.71	0.62

Tabel 31: Resultate vir elke groep etikette van rou Afrikaanse data met gazetteers

	ORG	EVE	PER	LOC	PRO	MISC
ORG	**65**	*	22	14	8	12
EVE	*	**6**	2	1	*	1
PER	4	*	**18**	10	2	7
LOC	6	*	5	**79**	1	*
PRO	*	*	*	1	**8**	*
MISC	10	1	16	16	4	**72**

Tabel 32: Verwarringsmatriks vir etiket-toekenning van rou Afrikaanse data met gazetteers

Tabel 31 dui daarop dat die f-telling van al die etikette 'n verbetering getoon het, in vergelyking met die eerste roudataeksperiment. Uit Tabel 31 kan afgelei word dat daar 'n minimale afname in die *f*-telling vir etikettoekenning vir EVE en PER is, in vergelyking met die eerste roudataeksperiment.

Uit Tabel 32 kan die rede vir die verlaging in *f*-telling vir PER duidelik gemerk word, aangesien 23 van die 41 (56.09%) PER-entiteite as MISC, ORG, PRO en LOC geklassifiseer is, waarvan LOC die meeste bydra (24.39%). Vir EVE is 4 van 10 (40%) van die entiteite verkeerdelik as PER, LOC en MISC geklassifiseer. Alhoewel MISC en ORG se *f*-telling toegeneem het wys Tabel 32 dat daar baie entiteite in hierdie kategorieë is wat verkeerd geëtiketteer word. Vir ORG word 56 van 121 (46.28%) verkeerdelik as PER, LOC, PRO en MISC geklassifiseer waarvan die meeste ten opsigte van PER (22 van 121, 18.18%) was.

Regte etiket –> Verkeerde etiket	Aantal (Verskil)	Persentasie
ORG->PER	22 (+6)	17.32
MISC->LOC	16 (-17)	12.6
MISC->PER	16 (-6)	12.6
ORG->LOC	14 (+5)	11.02
ORG->MISC	12	9.45
MISC->ORG	10	7.87
ORG->PRO	8 (+4)	6.3
LOC->ORG	6 (-1)	4.72
LOC->PER	5 (-1)	3.94
PER->ORG	4 (+2)	3.15
MISC->PRO	4 (-1)	3.15
PER->PRO	2 (+1)	1.57
*EVE->PER	2 (+2)	1.57
EVE->LOC	1 (-1)	0.79
PRO->LOC	1	0.79
*LOC->PRO	1 (+1)	0.79
*MISC->EVE	1 (+1)	0.79
EVE->MISC	1	0.79
TOTAAL	125	100

Tabel 33: Persentasie van verskillende etiketkombinasies van rou Afrikaanse data met gazetteers

In Tabel 33 kan waargeneem word dat ORG->PER, MISC->ORG, MISC->PER en ORG->LOC saam 53.54% (68 van die 125 etikettoekennings) van die totale foute in die eksperiment uitmaak.

Verder blyk dit dat daar ’n groot verskil is in die aantal foute wat tydens klassifikasie gemaak is in vergelyking met die eerste roudataeksperiment. Daar is ’n groot afname vir MISC->LOC (33 na 16, 51.51%) en MISC->PER (22 na 16, 27.27%). Daar is ook drie nuwe foutkombinasies wat te vore kom wat nie in die eerste roudataeksperimente was nie en dit word met ’n “*” in Tabel 33 aangedui. Dit is wel goed om te sien dat daar ’n afname in die totale aantal foute is (142 na 125, 12.58%) wat tot ’n toename in akkuraatheid vir etikettoekenning sal lui.

Presisie	0.62
Herroepping	0.56
f-telling	0.59

Tabel 34: Resultate vir etikettering van rou Afrikaanse data met gazetteers

In Tabel 34 kan gesien word dat daar ’n toename in presisie, herroeping en *f*-telling is vir hierdie eksperiment in vergelyking met die eerste roudataeksperiment. Presisie het van 0.41 na 0.62 toegeneem, herroepping van 0.52 na 0.56 en *f*-telling het verbeter van 0.46 tot 0.59.

3.4.6. Vernederlandsde-eksperiment met gazetteers

Na die belowende resultate wat in die vorige eksperiment verkry is, is daar besluit om die gazetteers saam met A2DC te gebruik in ’n poging om die resultate van die Nederlandse BEH te verbeter. Die resultate word in Tabel 35 gegee:

Totale aantal benoemde entiteite geïdentifiseer	400
Aantal vals positiewes geïdentifiseer	56
Aantal benoemde entiteite nie geïdentifiseer nie	14

Tabel 35: Benoemde-entiteitidentifikasies van A2DC met gazetteers

In vergelyking met die oorspronklike veranderde-A2DC-eksperiment is die resultate in Tabel 35 min of meer dieselfde. Al verskil is dat die aantal entiteite wat nie geïdentifiseer is nie van 16 na 14 afgeneem het en dat die totale aantal benoemde-entiteite wat geïdentifiseer is, van 403 na 400 afgeneem het. Tabel 36 toon ook geen groot verskil ten opsigte van presisie herroeping en *f*-telling nie.

Presisie	0.72
Herroeping	0.94
f-telling	0.82

Tabel 36: Resultate vir die benoemde-entiteitidentifikasie van A2DC met gazetteers

Etiket	**Herroeping**	**Presisie**	***f*-telling**
ORG	0.57	0.79	0.67
PER	0.51	0.59	0.55
EVE	0.27	0.60	0.38
PRO	0.80	0.38	0.52
LOC	0.81	0.54	0.65
MISC	0.50	0.78	0.61

Tabel 37: Resultate vir elke groep etikette van A2DC met gazetteers

Uit Tabel 37 kan daar wel verskille ten opsigte van die *f*-telling vir entiteitklassifikasie gesien word. ORG, PER, PRO, LOC en MISC het almal verbeteringe getoon ten opsigte van die oorspronklike veranderde-A2DC-eksperiment. Dit is net EVE wat se *f*-telling van 0.50 in die oorspronklike veranderde- A2DC-eksperiment verminder het na 0.38. Uit Tabel 37 kan daar wel waargeneem word dat die oorsaak van hierdie vermindering, die oorsaak is van ’n baie lae her-

roepping van EVE was (0.27) wat 'n aanduiding kan gee dat die identifisering van EVE-entiteit baie moeilik is om te bepaal. Tabel 38 gee ook meer duidelikheid rondom hierdie probleem.

	ORG	EVE	PER	LOC	PRO	MISC
ORG	**81**	*	7	20	7	15
EVE	2	**3**	1	2	*	3
PER	4	*	**27**	9	2	*
LOC	7	1	5	**78**	1	5
PRO	*	*	*	1	**8**	1
MISC	9	1	6	28	4	**65**

Tabel 38: Verwarringsmatriks vir etiket-toekenning van A2DC met gazetteers

Uit Tabel 38 kan daar duidelik 'n probleem waargeneem word ten opsigte van die toekenning van EVE-etikette. Slegs 3 uit die moontlike 11 (27.27%) EVE-etikettoekenning is korrek wat ook ooreenstem met die lae herroepping vir EVE in Tabel 38. Die ander EVE's is as ORG, PER, LOC en MISC geïdentifiseer. Verder blyk dit ook dat daar nog steeds probleme is by die toekenning van ORG en MISC etikette. By ORG is 49 van 130 (37.69%) etikette verkeerdelik as PER, LOC, PRO en MISC toegeken, waarvan LOC die meeste hieraan bygedra het (20 van 130, 15.38%). Vir MISC is 48 van 113 (42.47%) verkeerdelik as ORG, EVE, PER, LOC en PRO toegeken, waarvan die meeste ook aan LOC (28 van 113, 24.77%) toegeken word. Verder ondersoek sal ingestel word rondom die rede vir hierdie foute.

Uit Tabel 39 kan waargeneem word dat die aantal foute vir etikettoekenning vermeerder het van 124 na 141. Daar is ook drie nuwe foutkombinasies wat in Tabel 39 met 'n "*" gemerk is. Tabel 39 wys ook daarop dat die grootste bydra tot die foute afkomstig is van MISC->LOC, ORG->LOC en ORG->MISC wat 64 van die 142 (45.07%) foutkombinasies uitmaak. Hiermee word die kwessie verder versterk dat daar 'n probleem is ten opsigte van die toekenning van ORG en MISC-etikette.

Presisie	0.65
Herroepping	0.60
f-telling	0.62

Tabel 40: Resultate vir etikettering van A2DC met gazetteers

Volgens die resultate vir presisie herroepping en *f*-telling wat in Tabel 40 voorkom is dit duidelik dat die toevoeging van naamlyste 'n positiewe uitwerking op die akkuraatheid van etikettoekenning het. Alhoewel daar duidelike verbeteringe is ten opsigte van die etikettoekenning, toon Tabel 40 dat resultate baie nader is, maar tog nie vergelykbaar is met die resultate vir die Neder-

landse eksperiment nie. Verdere navorsing ten opsigte van groter naamlyste kan moontlik verbeterde resultate lewer.

Regte etiket –> Verkeerde etiket	Aantal (Verskil)	Persentasie
MISC->LOC	28 (-10)	19.72
ORG->LOC	20 (+6)	14.08
ORG->MISC	15 (+2)	10.56
MISC->ORG	9 (-2)	6.34
PER->LOC	9 (+2)	6.34
ORG->PER	7 (+2)	4.93
ORG->PRO	7 (+2)	4.93
LOC->ORG	7	4.93
MISC->PER	6 (-1)	4.23
LOC->PER	5	3.52
LOC->MISC	5 (+4)	3.52
PER->ORG	4 (+2)	2.82
MISC->PRO	4 (+4)	2.82
EVE->MISC	3 (+3)	2.11
EVE->ORG	3 (+1)	2.11
PER->PRO	2 (+1)	1.41
EVE->LOC	2 (+1)	1.41
*MISC->EVE	2	1.41
EVE->PER	1	0.7
PRO->LOC	1	0.7
*LOC->PRO	1	0.7
PRO->MISC	1	0.7
TOTAAL	141	100

Tabel 39: Persentasie van verskillende etiketkombinasies van A2DC met gazetteers

3.5. Samevatting

In hierdie hoofstuk is ’n eksperimentele benadering gevolg om te bepaal watter tipe pre- en/of post-prosesseringstappe aangewend moet word om ’n benoemde-entiteitherkenner vir Afrikaans te ontwikkel. Tegnologieherwinning is gebruik om ’n benoemde-entiteitherkenner, met behulp van ’n Nederlandse BEH, te ontwikkel wat vinniger, in terme van prosesseringspoed, en makliker implementeerbaar, ten opsigte van aantal eienskappe wat benodig word, is as die benoemde-entiteitherkenner wat Puttkammer (2006) ontwikkel het (Puttkammer, 2006:66). Oor die al-

gemeen blyk dit dat die eksperiment ŉ sukses was aangesien daar ŉ implementeerbare- en effektiewe BEH vir Afrikaans ontwikkel is. Ten opsigte van die eksperimente kan opgemerk word dat elke eksperiment wat met Afrikaanse data aangepak is, telkens beter resultate gelewer het in terme van die *f*-telling vir entiteitidentifikasie sowel as entiteitklassifikasie.

Die resultate van die laaste eksperiment met die Afrikaanse data was vergelykbaar met die resultate wat vir die Nederlandse data verkry is. In terme van die identifikasie van benoemde entiteite is ŉ *f*-telling van 0.82 vir die Afrikaanse data en ŉ *f*-telling van 0.88 vir die Nederlandse data verkry. In terme van die entiteitklassifikasie was die resultate nie so naby nie. Vir die klassifikasie van entiteite met Afrikaanse data was die *f*-telling 0.62 en vir die Nederlandse data was die *f*-telling 0.80. Hierdie resultate is nietemin belowend, aangesien daar verwag word dat die Nederlands BEH beter sou vaar op Nederlandse data as op Afrikaanse data. Dit blyk dus dat die konsep van tegnologieherwinning ŉ positiewe invloed op die ontwikkeling van ŉ BEH vir Afrikaans het en hopelik sal hierdie selfde metode in die toekoms gebruik kan word om meer kerntegnologieë vir hulpbronskaarstale te ontwikkel.

4. Slot

4.1. Inleiding

In hierdie studie is ŉ benoemde-entiteitherkenner (BEH) vir Afrikaans ontwikkel, deur tegnologieherwinning toe te pas op ŉ bestaande Nederlandse BEH (*Frog*). Alhoewel daar alreeds ŉ benoemde-entiteitherkenner vir Afrikaans bestaan (Puttkammer, 2006), is dit nie implementeerbaar nie weens die oormatige gebruik van eienskappe as afrigtingsdata, ten opsigte van verwerkingstyd vir klassifikasies en die geheue-kapasiteit wat nodig is vir afrigting van die sisteem (Puttkammer, 2006:66).

4.2. Opsomming

In Hoofstuk 1 word die noodsaaklikheid vir die ontwikkeling van ŉ implementeerbare benoemde-entiteitherkenner (BEH) vir Afrikaans bespreek deur eerstens te verwys na die Grondwet van Suid-Afrika (Republic of South Africa, 2003) se taalbeleid. Tweedens word die idee van ŉ BLARK (Krauwer, 2003) vir Suid-Afrikaanse tale bespreek wat gevolg word deur ŉ bespreking van ŉ oudit wat fokus op die aantal hulpbronne en verspreiding van mensliketaaltegnologie vir al elf Suid-Afrikaanse tale (Sharma Grover *et al*., 2010). Ten opsigte van die oudit is daar bevestig dat daar ŉ tekort aan teksgebaseerde-hulpmiddels vir Afrikaans bestaan.

Volgens die Grondwet van Suid-Afrika (Republic of South Africa, 2003) is dit die plig van die regering om alle inligting en kennis beskikbaar te maak vir al tien inheemse tale (uitsluitende Engels) van Suid-Afrika. Die regering het begin om reeds bestaande inligting van die inheemse tale aan die publiek beskikbaar te stel en het die hoeveelheid inligting wat beskikbaar is vir hierdie tale probeer vermeerder.

Volgens Krauwer (2003) bestaan daar ŉ inventaris vir die minimale aantal taal-verwante hulpbronne wat nodig word vir ŉ taal om kompeterend te wees op die vlak van navorsing en onderrig en staan bekend as ŉ BLARK (“Basic Language Resource Kit”). Krauwer (2003) stel verder dat BLARK se inhoud kan verskil ten opsigte van die behoefte van die gegewe taal, maar dat die BLARK van die taal aan ŉ infrastruktuur moet voldoen wat help om hulpbronne te bestuur, te onderhou en te versprei.. BLARK word onderverdeel in 3 kategorieë (Krauwer, 2003), naamlik: Standaarde, Data en Kerntegnologieë.

Ter voorbereiding vir die ontwikkeling van ’n BLARK vir al elf amptelike tale van Suid-Afrika, is ’n oudit in 2010 (Sharma Grover *et al*., 2010) begin om te bepaal watter inheemse tale ’n behoefte het aan watter tipe hulpbronne in terme van standaarde, data en kerntegnologieë. Om die hulpbronskaarsheid van tale te bepaal, is ’n mensliketaaltegnologie (MTT) taalindeks saamgestel. Ten opvolg hiervan is ’n MTT komponentindeks ontwikkel.

In hierdie studie is daar slegs op een van hierdie kategorieë, naamlik kerntegnologieë, gefokus met die spesifieke fokus op die ontwikkeling van ’n benoemde-entiteitherkenner (BEH) vir Afrikaans. Die doel van die taalindeks (Sharma Grover *et al*., 2010) is om aan navorsers ’n idee te gee oor die noodsaaklikheid vir ontwikkeling ten opsigte van hulpbronne vir die onderskeie Suid-Afrikaanse tale (Sharma Grover *et al*., 2010). Gegewe hierdie oudit blyk dit dat daar ’n groot aanvraag is vir die ontwikkeling van hulpmiddels vir teks in Afrikaans, wat dan meebring dat daar in hierdie studie gefokus word op die ontwikkeling van ’n benoemde-entiteitherkenner vir Afrikaans aangesien benoemde-entiteitherkenners belangrik is vir teksprosessering.

In Hoofstuk 2 is ’n beskrywing gegee van wat ’n entiteit en ’n benoemde entiteit is. Volgens Puttkammer (2006: 25) is 'n benoemde entiteit 'n “aansyn wat binne die konseptuele ruimte aan 'n enkele instansiëring veranker word deur middel van konvensie, 'n geïnstitusionaliseerde proses of 'n outoriteit en waarvan die skryfwyse of wetlik, of deur een of ander outoriteit bepaal word.” Verder in die hoofstuk word die proses van tegnologieherwinning verduidelik met behulp van ander studies waar die beginsel van tegnologieherwinning toegepas word.

Tegnologieherwinning behels die ontwikkeling van hulpbronne vir hulpbronskaarstale (L2) deur die herontwerp of verandering van kerntegnologieë van hulpbronryke tale (L1) (Rayner *et al*., 1997). Laastens is daar gekyk na die moontlike verskille wat tussen Afrikaanse- en Nederlandse benoemde entiteite mag voorkom. Hierdie verskille is vervolgens in drie kategorieë verdeel, naamlik: identiese kognate, nie-identiese kognate en onverwante entiteite. Om hierdie verskille tussen Afrikaanse-en Nederlandse benoemde entiteite te verminder, is ’n Afrikaans-na-Nederlands-omskakelaar (A2DC) (Van Huyssteen & Pilon, 2009) gebruik om die Afrikaanse data te vernederlands sodat dit moontlik beter resultate sal lewer wanneer die data deur ’n Nederlandse BEH geannoteer word.

Hoofstuk 3 begin met ’n beskrywing van *Frog* (Van den Bosch *et al*., 2007), die Nederlandse BEH wat in hierdie studie gebruik is, en die funksies en werking van die BEH-komponent van *Frog*. Daarna volg ’n beskrywing van die Afrikaans-na-Nederlands-omskakelaar (A2DC) (Van Huyssteen & Pilon, 2009) en laastens word die verskillende eksperimente wat gedoen is, uiteengesit.

Die studie bestaan uit ses eksperimente waarvan die eerste is om te sien wat die resultate van *Frog* op Nederlandse data is. Die tweede eksperiment het die effektiwiteit van *Frog* op onveranderde (rou) Afrikaanse data geëvalueer. Die volgende twee eksperimente het die resultate van *Frog* op vernederlandsde data geëvalueer. Die laaste twee eksperimente het weer die effektiwiteit op rou en vernederlandsde data geëvalueer, maar in hierdie eksperimente is gazetteers ook gebruik as deel van die preprosessering.

Volgens die resultate van die Nederlandse-eksperiment het *Frog*, soos verwag, goeie resultate gelewer met ŉ *f*-telling van 0.88 vir die identifikasie van benoemde entiteite en ŉ *f*-telling van 0.80 vir die klassifisering van hierdie benoemde entiteite. Die resultate vir hierdie eksperiment is vergelykbaar met ander Nederlandse BEH's (Desmet & Hoste, 2010; Black & Vasilakopoulos, 2002; Burger *et al*., 2002; Cucerzan & Yarowsky, 2002; Malouf, 2002).

In die roudataeksperiment was die resultate, soos verwag, swakker as vir die Nederlandse data met ŉ *f*-telling van 0.79 vir die identifisering van benoemde entiteite. Die resultate was baie swakker vir die klassifisering van benoemde entiteite met ŉ *f*-telling van 0.45. Die lae *f*-telling vir klassifikasies is te verwagte omdat daar wel ortografiese verskille tussen die twee tale voorkom.

Vir die A2DC-eksperiment is die Afrikaanse data vernederlands met behulp van A2DC (Van Huyssteen & Pilon, 2009). Hierdie eksperiment het beter resultate gelewer as in die roudataeksperiment met ŉ *f*-telling van 0.81 vir die identifisering van benoemde entiteite. Daar was wel nog ŉ probleem met die klassifisering van benoemde entiteite wat ŉ *f*-telling van 0.52 gelewer het. Alhoewel dit beter is as die resultate van die roudataeksperiment is dit nog nie vergelykbaar met die resultate van die Nederlandse-eksperiment nie. Na ŉ verdere ondersoek ten opsigte van die afvoer van A2DC (Van Huyssteen & Pilon, 2009) blyk dit dat die grootste probleem te make het met die omskakeling van dubbelloopname (soos "Suid-Afrika") en akronieme (soos "ABSA").

Nadat hierdie probleme opgelos is deur veranderinge aan die programmatuur aan te bring, is die eksperiment weer uitgevoer. Vir identifikasie van benoemde entiteite het die veranderde A2DC eksperiment ŉ *f*-telling van 0.82 gelewer. Die *f*-telling vir die klassifikasie van entiteite was 0.55. Hierdie resultate is steeds nie vergelykbaar met die resultate van die Nederlandse eksperiment nie en daar is ŉ oplossing gesoek om die resultate te verbeter.

Die oplossing het gekom in die vorm van gazetteers as preprosesseringsstap. Daar is nog twee eksperimente by gedoen. Die eerste was die herhaling van die roudataeksperiment, maar hierdie keer met die gebruik van gazetteers. In terme van identifikasie van entiteite was die *f*-telling 0.75 en vir klassifikasie van entiteite was die *f*-telling 0.59, wat 'n verbetering is in vergelyking met die oorspronklike roudataeksperiment. Die veranderde A2DC-eksperiment was ook herhaal met die gazetteers en het vir identifikasie van entiteite 'n *f*-telling van 0.82 en vir klassifikasie van entiteite 'n *f*-telling van 0.62 behaal, wat 'n verbetering was en die resultate was amper vergelykbaar met die Nederlandse eksperiment.

4.3. Gevolgtrekking

Die BEH wat in hierdie studie ontwikkel is toon beter resultate as die een wat deur Puttkammer (2006) ontwikkel is. In vergelyking met die BEH wat deur Puttkammer ontwikkel is, wat 'n *f*–telling van 0.81, presisie van 0.81 en herroeping van 0.97 gelewer het, het die BEH wat in hierdie studie ontwikkel is 'n *f*-telling van 0.82, presisie van 0.73 en herroeping van 0.94 gelewer (Puttkammer, 2006:99) [5].

Soos daar deur Puttkammer (2006) beweer is en soos dit ook in hierdie studie waargeneem is, "is dit duidelik dat dit 'n veel eenvoudiger taak is om benoemde entiteite te herken as wat dit is om te onderskei tussen verskillende tipes benoemde entiteite." (Puttkammer, 2006:86).
Uit die afleidings wat uit Puttkammer (2006) se studie gemaak is, wil dit voorkom of die BEH wat in hierdie studie ontwikkel is meer implementeerbaar is, ten opsigte van die aantal data sowel as die geheuekapasiteit wat benodig word vir die stoor van die data wat moontlik kan bydra tot 'n vertraging van verwerkingstyd.

Om die tekseenheididentifiseerder van Puttkammer af te rig is 224 eienskappe vir elke woord benodig, wat bestaan uit sewe tekseenhede vir elke fokuswoord (drie tekseenhede voor en drie tekseenhede na die woord) saam met 32 ander eienskappe wat die woord help identifiseer (Puttkammer. 2006:11). Daar word in totaal met 40 906 woorde afgerig waarvan 3 068 benoemde entiteite is (Puttkammer. 2006:11).

Weens die aard van *k*-Naastebuurpuntalgoritme moet al die woorde en hul eienskappe in die rekenaar se geheue gestoor word, wat 'n baie hoë geheuekapasiteit benodig en ook die spoed van prosessering kan vertraag.

[5] Daar word ook ander resultate vir Puttkammer se BEH gegee, maar daar is besluit om hierdie te gebruik aangesien dit opsommend is van sy studie.

In vergelyking met die BEH wat in hierdie studie ontwikkel is en gebruik maak van ’n reeds ontwikkelde Nederlandse BEH, *Frog*, ’n preprosesseringstap, A2DC, en agt gazetteers wat elk nie een meer as 250 items bevat het nie, kan die afleiding gemaak word dat die BEH wat in hierdie studie ontwikkel is, makliker is om te implementeer as die BEH wat deur Puttkammer ontwikkel is.

4.4. Toekomstige navorsing

Ten einde hierdie navorsing af te sluit is dit van groot belang om toekomstige navorsing moontlikhede ook te identifiseer.

- Indien daar wel verbeteringe aan die resultate aangebring wil word, is dit van groot belang om ’n ander sisteem te vind, wat Afrikaanse data kan vernederlands in terme van die konteks waarin die woorde (of benoemde entiteite) hulself vind.

- Aangesien *Frog* oopbronkodesagteware is, kan dit moontlik intern verander word deur byvoorbeeld ’n geannoteerde lys van Afrikaanse woorde by te voeg om beter resultate te lewer.

- Daar kan ook na ander alternatiewe Nederlandse BEH’s gekyk word. Aangesien *Frog* deel uitmaak van ’n module-georiënteerde sintaktiese etiketteerder kan beter resultate dalk verkry word met ’n Nederlandse BEH wat spesifiek op benoemde-entiteitherkenning fokus.

- ’n Ander benadering wat ook gevolg kan word, is om ’n taalafhanklike BEH te vind en dan die resultate van die sisteem op Afrikaanse data te bepaal.

- Die akkuraatheid van klassifikasie van benoemde entiteite behoort ook baie te verbeter deur slegs die aantal items van die gazetteers te vergroot.

- In toekomstige studies kan nuwe kerntegnologieë ontwikkel word vir die ander nege Suid-Afrikaanse tale deur dieselfde metodes wat in hierdie studie toegepas is te gebruik.

5. Bibliografie

ASAHARA, M. & MATSUMOTO, Y. 2003. Japanese named Entity Extraction with Redundant Morphological Analysis. (*In* Sha, F. & Pareira, F., *eds.* Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, Edmonton, Canada. Volume 1, p. 8-15).

BIKEL, D. M., MILLER, S., SCHWARTZ, R. & WEISCHEDEL, R. 1998. Nymble: A High Performance Learning Name-finder. (*In* The fifth conference on Applied Natural Language Processing, p. 194-201).

BLACK, W. J. & VASILAKOPOULOS, A. 2002. Language Independent Named Entity Classification by modified Transformation-based Learning and by Decision Tree Induction. (*In* Sang, T. K. & Erik, F., *eds*. Sixth conference on Natural language learning, Edmonton, Canada. Volume 20, p. 1-4).

BORTHWICK, A., STERLING, J., AGICHTEIN, E. & GRISHMAN, R. 1999. NYU: Description of MENE Named Entity Recognition Used in MUC-7. (*In*: The 40th Annual Meeting on Association for Computational Linguistics, Pennsylvania, USA, p. 473-480).

BURGER, J. D., HENDERSON, J. C. & MORGAN, W. T. 2002. Statistical Named Entity Recogniser Adaptation. (*In* Sang, T. K. & Erik, F., *eds*. Sixth conference on Natural language learning , Edmonton, Canada. Volume 20, p. 1-4).

CARRERAS, X., MARQUES, L. & PADRO, L. 2002. Named entity extraction using AdaBoost. (*In* Sang, T. K. & Erik, F., *eds*. Sixth conference on Natural language learning. Volume 20, p. 1-4, Edmonton, Canada).

CUCERZAN, S. & YAROWSKY, D. 2002. Language Independent NER using a Unified Model of Internal and Contextual Evidence. (*In* Sang, T. K. & Erik, F., *eds*. Sixth conference on Natural language learning, Edmonton, Canada. Volume 20, p. 1-4).

CHINCHOR, N. A. & ROBINSON P. 1998. Proceedings of the Seventh Message Understanding Conference (MUC-7): Named Entity Task Definition, (*In* Proceedings of the Seventh Message Understanding Conference (MUC-7), Fairfax, VA, p. 21).

DAKE, J. M. 2003. Explorations of the speed-accuracy trade-off in Memory Based Learning algorithms. Tegniese verslag ilk0302, ILK Navorsingsgroep. Tilburg University, Tilburg.

DAVEL, M. & BARNARD, E. 2004. A default-and-refinement approach to pronunciation prediction. (*In* The 15th Annual Symposium of the Pattern Recognition Association of South Africa, Computer Speech & Language, Grabouw, South Africa, 25 to 26 November 2004. Volume 22, p.374-393).

DESMET B. & HOSTE V. 2010. Dutch Named Entity Recognition using Classifier Ensembles. (*In* Westerhout, E. Markus, T. & Monachesi P. *eds.* Proceedings of the 20th Meeting of Computational Linguistics in the Netherlands, LOT Occasional Series 16, Utrecht, p. 29-41).

EHLERS, D. & VAN BEEK, P. 2004. Oranje boven. Nederlands voor Zuid-Afrika, Pretoria, Protea: Boekhuis, 2004.

EBART, 2010. Aktuelna arhiva. Medijska dokumentacija Ebart, http://www.arhiv.rs.

FELLBAUM, C. 1998. WordNet: An Electronical Lexical Database, Cambridge, MA. MIT Press, 2012.

GROENEWALD, H.J. & DU PLOOY, L. 2010. Processing Parallel Text Corpora for Three South African Language Pairs in the Autshumato Project. (*In* De Pauw, G., Groenewald, H. & De Schryver, G., *eds.* The Second Workshop on African Language Technology, AfLaT 2010, Valetta, Malta. p. 27-30).

INNOVISCOP, 2006. Experimental development Definition, http://www.innoviscop.com/en/definitions/experimental-development, 2013.

JURAFSKY, D. & MARTIN, J.H. 2010. Speech and Language Processing, An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, 2nd ed., Prentice-Hall.

KHOEN, P. 2005. Europarl: A Multilingual corpus for evaluation of machine translation. (*In* Nagao, M.,Takana H., Makino, T., Nomura, H., Uchinda, H. & Ishizaki, S. *eds.* Proceeding of the tenth Machine Translation Summit, AAMT, Phuket, Thailand. p.79-86).

KRAUWER, S. 2003. The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadmap. (*In*: International Workshop SPECOM, 2003, Moscow, Russia. p. 8-15).

MANNING, C. D. RAGHAVAN, P. & SCHÜTZE, H. 2008. Introducution to Information Retrieval. Cambridge University Press, NY, USA.

MARRERO, M. SÁNCHEZ-CUADRADO, S. LARA, J. M. ANDREADAKIS, G. 2009, Evaluation of Named Entity Extraction Systems. Research in Computing Science, 41:47-58.

MALOUF, R. 2002. Markov models for language-independent named entity recognition. (*In* Sang, T. K. & Erik, F. *eds*. Sixth conference on Natural language learning, Edmonton, Canada. Volume 20, p. 1-4).

MARTINOVIC, M. 2008. Transfer of Natural Language Processing Technology: Experiments, Possibilities and Limitations Case Study: English to Serbian. (*In* Krstev, C. *ed.*, Journal of Information and Library Science, College of New Jersey, NJ. No. 1-2, Volume 11, p. 11-21).

MACULLUM, A. & LI, W. 2003. Early results for named Entity Recognition with Conditional Fields, Feature Induction and Web-enhanced Lexicons. (*In* The seventh conference on Natural language learning at HLT-NAACL, 2003. Volume 4, p.188-191).

MBEKI, T. 2007. Staatsrede van Suid-Afrika. http://www.info.gov.za/speeches/son/sona_afrikaans.htm Datum van gebruik: 25 Sept. 2012.

NADEAU, D. & SEKINE, S. 2006. A survey of named entity recognition and classification. In: Sekine, S. and Ranchhod, E. Named Entities: Recognition, classification and use. Special issue of Lingvisticæ Investigationes. 30(1): p. 3-26).

Pilon, S. 2005. Outomatiese Afrikaanse Woordsoortetikettering.Ongepubliseerde MA-verhandeling. Potchefstroom: Noordwes-Universiteit.

PILON, S., VAN HUYSSTEEN, G.B. & AUGUSTINUS, L. 2010. Converting Afrikaans to Dutch for technology recycling. (*In* Herbts, B. & De Waal, A., *eds.* The 21st Annual Symposium of the Pattern Recognition Association of South Africa, Stellenbosch, Suid-Afrika. p. 219–224).

PUTTKAMER, M.J. 2006. Outomatiese Tekseenheididentifiseerder vir Afrikaans, NWU: Potchefstroom (Verhandeling-MA).

RAYNER, M. CARTER, D. BRETAN, I. EKLUND, R. WIREN, M. HANSEN, S. L. KIRCHMEIER-ANDERSEN, S. PHILP, C. SORENSEN, F. & THOMSEN, H.E. 1997. Recycling Lingware in a Multilingual MT System. (*In* Burstein, J., and Leacock, C. *eds.*, From research to commercial applications: making NLP work in practice, ACL, Somerset, 1997, p. 65-70).

REYNAERT, M. 2007. Sentence-splitting and tokenization in D-Coi, Tegniese verslag, ILK 07-03. ILK Research Group, 2012.

REPUBLIC OF SOUTH AFRICA, 2003. Language Policy Framework. http://www.dac.gov.za/policies/LPD_Language%20Policy%20Framework_English%20_2_.pdf Datum van gebruik: 28 Mrt. 2011.

SALTON, G. 1971. The SMART Retrieval System: Experiments in Automatic Document Processing. Prentice-Hall, Englewood Cliffs, NJ. p. 313-323, 2012.

SCANNELL, K. P. 2006. Machine translation for closely related language pairs. (*In* LERC 2006, p. 103-107).

SEKINE, S. 1998. NYU: Description of the Japanese NE system used for MET-2. (In Proceedings of the Seventh Message Understanding Conference (MUC-7), WA, USA).

SHARMA GROVER, A. 2009. A Technology Audit: The State of Human Language Technologies R&D in South Africa (Masters Research report). University of Pretoria: Graduate School of Technology Management.

SHARMA GROVER, A. VAN HUYSSTEEN, G. B. & PRETORIUS, M. W. 2011. An HLT profile of the official South African languages. (*In* De Pauw, G., Groenewald, H. & De Schryver, G., *eds.* The Second Workshop on African Language Technology, AfLaT 2010, Valetta, Malta. p.3-7).

SKUT, W. BRANDS, T. 2002. A maximum-entropy partial parser for unrestricted text. In Sixth Workshop on Very Large Corpora, p.143-151, 2011.
Suid-Afrikaanse Regering ***kyk*** Mbeki.

TJONG KIM SANG, E. F. 2002. Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. (*In* Sang, T. K. & Erik, F., *eds*. Sixth conference on Natural language learning, Edmonton, Canada. Volume 20, p. 1-4).

VANDEGHINSTE, V., SCHUURMAN, I., MARKANTONATOU, S., BADIA, T. & CARL, M. 2006. METIS-II: Machine Translation for Low Resource Languages. (*In* Nicoletta Calzolari, N., Choukri**,** K**.,** Gangemi, A., Maegaard, B., Mariani, J., Odijk, J., & Tapias, D., *eds.* The fifth international conference on language, resource and evaluation, Genoa, Italy. p. 1284-1289).

VAN DEN BOSCH, A. 2012a. Workings of Frog's named entity recognition, E-pos aan Antal.vdnBosch@uvt.nl Datum van gebruik: 15 Nov. 2012.

VAN DEN BOSCH, A. 2012b. Frog: A Morpho-syntactic analyser and dependency parser, http://ilk.uvt.nl/frog Datum van gebruik: 16 Nov. 2012.

VAN DEN BOSCH, A. BUSSER, B. CANISIUS, S. DAELEMANS, W. 2007. An efficient memory-based morphosyntactic tagger and parser for Dutch. (*In* Dirix P., Schuurman, I., Vandeghinste, V. & Van Eynde, F. *eds.* Proceedings of the 17th Meeting of Computational Linguistics in the Netherlands (CLIN), CCL Leuven, Leuven, The Netherlands. p. 191-206).

VAN HUYSTEEN, G.B. & PILON, S. 2009. Rule-based Conversion of Closely-related Languages: A Dutch-to-Afrikaans Converter. (*In* Nicolls, F., *ed.* 20th Annual Symposium of the Pattern Recognition Association of South Africa (PRASA). Stellenbosch, South Africa. p. 23-28).

VAN HUYSSTEEN, G.B. 2000. Die Reduplikasiekonstruksie in Afrikaans: Enkele Aspekte van 'n Kognitiewe Gebruiksgebaseerde Beskrywingsmodel vir Afrikaans. PU vir CHO*:* Potchefstroom (Proefskrif – PhD).

VILLAZÓN-TERRAZAS, B. SUÄREZ-FIGUEROA, S. C. & GÓMEZ-PERÉZ, A. 2008, Pattern-Based Method for Re-Engineering Non-Ontology Resources into Ontologies, (*In* Domingue, J. & Anutariya, C. *eds.* ASWC '08 Proceedings of the 3rd Asian Semantic Web Conference on The Semantic Web, p.167-181).

WIKIPEDIA, 2013a, List of organizations with international domain names, http://en.wikipedia.org/wiki/List_of_organizations_with_.int_domain_names Datum van gebruik: 14 Jan. 2013.

WIKIPEDIA, 2013b, List of people, http://en.wikipedia.org/wiki/Lists_of_people. Datum van gebruik: 15 Jan. 2013.

WIKIPEDIA, 2013c, List of international organization leaders in 2012, http://en.wikipedia.org/wiki/List_of_international_organization_leaders_in_2012, Datum van gebruik: 16 Jan. 2013.

WIKIPEDIA, 2013d, List of titles, http://en.wikipedia.org/wiki/List_of_titles, Datum van gebruik: 18 Jan. 2013.

WIKIPEDIA, 2013e, List of confectionery brands, http://en.wikipedia.org/wiki/List_of_confectionery_brands, Datum van gebruik: 21 Jan. 2013.

WIKIPEDIA, 2013f, List of countries by population in 2000, http://en.wikipedia.org/wiki/List_of_countries_by_population_in_2000, Datum van gebruik: 22 Jan. 2013.

YOURPARENTING, 2009. Baby names. http://www.yourparenting.co.za/tools/baby-names?name=&gender=&baby_name_origin_id=47&meaning= Datum van gebruik: 23 April 2012.