Sintaktiese Herrangskikking as Voorprosessering in die Ontwikkeling van ’n Engels na Afrikaanse Statistiese Masjienvertaalsisteem

Marissa Griesel

(née Van Rooyen)

13017527

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

Studieleier: Me. S. Pilon

Medestudieleier: Prof. J.C. Roux

September 2011

Opsomming

Sintaktiese Herrangskikking as Voorprosessering in die Ontwikkeling van ’n Engels na Afrikaanse Statistiese Masjienvertaalsisteem

Deur Marissa Griesel

Statistiese masjienvertaling na enige van die hulpbronskaars Suid-Afrikaanse tale, lewer oor die algemeen swak kwaliteit afvoer. Groot hoeveelhede afrigtingsdata is nodig om afvoer te genereer wat sinvol in ’n vertaalomgewing ingesluit kan word om menslike vertalers se werk te vergemaklik. Aangesien hierdie groot hoeveelhede data nie altyd beskikbaar is nie, moet ander tegnieke ondersoek word om die afvoer van die sisteme te verbeter. Een van die metodes in die internasionale literatuur wat goeie verbeteringe in die afvoer meebring, is om sintaktiese herrangskikking as voorprosessering toe te pas. Voorprosessering het ten doel om die dekoderingsproses te vereenvoudig aangesien minder verandering in hierdie stadium nodig sal wees. Afrigting word ook vergemaklik aangesien outomatiese woordbelynings makliker gemaak kan word omdat die woordvolgorde in beide die brontaal en die teikentaal meer eenders is. Die voorprosessering word verrig op beide die teikentaalafrigtingsdata en die teks wat vertaal moet word. Dit is in die vorm van reëls wat patrone in die etikette herken en die struktuur dienooreenkomstig aanpas. Die etikette word deur ’n sintaktiese analiseerder aan die teikentaalkant van die tweetalige afrigtingsdata toegeken. In hierdie navorsingsprojek word die tegniek vir vertaling van Engels na Afrikaans aangepas en die reëls hanteer herrangskikking van werkwoorde, modale hulpwerkwoorde, die verledetydskonstruksie, konstruksies met “to” en negatiewe. Die doel daarvan is om die Engelse (brontaal) struktuur te verander om meer na die Afrikaanse (teikentaal) struktuur te lyk. ’n Deeglike analise van die afvoer van ’n basislynsisteem moet as beginpunt gemaak word. Die foute wat in die afvoer voorkom, word in kategorieë verdeel en elkeen van die onderliggende konstrukte word vir Engels en Afrikaans bestudeer. Hierdie analise van die afvoer en die literatuur oor sintaksis vir die twee tale word gekombineer om die linguisties gemotiveerde reëls te formuleer. Die module waarin die voorprosessering gedoen word, word in terme van presisie en herroeping geëvalueer en ’n F-telling word ook bereken wat hierdie twee metrieke saamvat in een syfer. Al drie hierdie metrieke lewer resultate wat goed met internasionale standaarde vergelyk. Verder word ’n vergelyking getref tussen die sisteem wat met die voorprosesseringsmodule verryk word en ’n basislynsisteem waarop geen ekstra prosessering toegepas word nie. Die vergelyking word aan die hand van twee metrieke (die BLEU- en NIST-tellings) wat outomaties bereken word, gedoen en toon baie positiewe resultate. Wanneer die dokument as geheel beoordeel word, het die BLEU-telling van 0,4968 na 0,5741 (7,7 %) gestyg en die NIST-telling van 8,4515 na 9,4905 (10,4 %).

Sleutelterme

Statistiese masjienvertaling; Afrikaans; Engels; sintaktiese herrangskikking; voorprosessering.

Abstract

Syntactic Reordering as Pre-processing in the Development of an English to Afrikaans Statistical Machine Translation System

By Marissa Griesel

Statistic machine translation to any of the resource scarce South African languages generally results in low quality output. Large amounts of training data are required to generate output of such a standard that it can ease the work of human translators when incorporated into a translation environment. Sufficiently large corpora often do not exist and other techniques must be researched to improve the quality of the output. One of the methods in international literature that yielded good improvements in the quality of the output applies syntactic reordering as pre-processing. This pre-processing aims at simplifying the decoding process as less changes will need to be made during translation in this stage. Training will also benefit since the automatic word alignments can be drawn more easily because the word orders in both the source and target languages are more similar. The pre-processing is applied to the source language training data as well as to the text that is to be translated. It is in the form of rules that recognise patterns in the tags and adapt the structure accordingly. These tags are assigned to the source language side of the aligned parallel corpus with a syntactic analyser. In this research project, the technique is adapted for translation from English to Afrikaans and deals with the reordering of verbs, modals, the past tense construct, constructions with "to" and negation. The goal of these rules is to change the English (source language) structure to better resemble the Afrikaans (target language) structure. A thorough analysis of the output of the baseline system serves as the starting point. The errors that occur in the output are divided into categories and each of the underlying constructs for English and Afrikaans are examined. This analysis of the output and the literature on syntax for the two languages are combined to formulate the linguistically motivated rules. The module that performs the pre-processing is evaluated in terms of the precision and the recall, and these two measures are then combined in the F-score that gives one number by which the module can be assessed. All three of these measures compare well to international standards. Furthermore, a comparison is made between the system that is enriched by the pre-processing module and a baseline system on which no extra processing is applied. This comparison is done by automatically calculating two metrics (BLEU and NIST scores) and it shows very positive results. When evaluating the entire document, an increase in the BLEU score from 0,4968 to 0,5741 (7,7 %) and in the NIST score from 8,4515 to 9,4905 (10,4 %) is reported.

Keywords

Statistical machine translation; Afrikaans; English; syntactic reordering; pre-processing.

Voorwoord

Ek wil graag die volgende mense en instansies bedank vir hul besondere bydrae:

- Die Navorsingseenheid: Tale en Literatuur in die Suid-Afrikaanse Konteks en die Sentrum vir Tekstegnologie (CTexT®) vir befondsing en ondersteuning. In die besonder prof. Hein Viljoen, Handré Groenewald en Ulrike Janke vir die tyd in Tilburg en die tyd terwyl ek aan hierdie verslag gewerk het.

- My studieleier, me. Suléne Pilon, en medestudieleier, prof. Justus Roux vir hulp, bystand en goeie raad, selfs as ek dit nie wou hoor nie.

- Prof. Menno van Zaanen en die dosente van HAIT by die Universiteit van Tilburg vir die onmisbare kennis wat ek daar kon inwin.

- Cindy McKellar vir tegniese hulp en geselskap. Dankie dat jy Moses getem het voordat ek moes probeer en gereeld sy grille en giere namens my opgelos het.

- Dirk, dankie dat jy al van Matriek af in my geglo het! Jou ondersteuning en liefde, selfs van duisende kilometers weg, het al die verskil in die wêreld gemaak. Maar regtig!

- Vir my ma en pa – dankie dat julle altyd gevra het hoe dit gaan, al was die antwoord altyd dieselfde. Dankie dat Ma vir my wag as ek alles wil lees en dat Pa my geleer het om dit te doen.

- Lené, dankie dat jy my weer aan die wonderwêreld van Anneli van Rooyen voorgestel het. Ek het al vergeet…

- Aan elke vriend wat êrens op ’n Saterdag vir my ’n vleisie gebraai het terwyl ek werk, of my omgepraat het om eerder saam te braai, dankie!

- Liewe Heer, baie dankie vir die vermoëns wat U aan my gegee het, en dat U nie toegelaat het dat ek een tree van hierdie pad alleen loop nie.

Inhoudsopgawe

Tabelle en figure

Hoofstuk 1: Inleiding

1.1 Inleiding en kontekstualisering

In ŉ veeltalige omgewing soos Suid-Afrika moet regeringsdokumente in soveel as moontlik van die 11 amptelike tale beskikbaar gestel word. Dit is duidelik dat menslike vertaling sonder veel rekenarisering nie die ideale oplossing vir hierdie situasie is nie, en daarom het die Nasionale Departement Kuns en Kultuur (DKK) in 2007 die *Autshumato*-projek van stapel gestuur.[1] Die doel van hierdie uitgebreide projek is om vertaalhulpmiddels vir al die amptelike landstale daar te stel, asook masjienvertaalsisteme (MV-sisteme) binne die openbareadministrasiedomein vir drie taalpare – Engels na isiZulu, Engels na Sesotho sa Leboa (Sepedi) en Engels na Afrikaans.

Masjienvertaling is ŉ proses waartydens spraak of teks in ŉ brontaal outomaties na spraak of teks in ŉ teikentaal vertaal word. In hierdie proses kan die rol van die rekenaar ŉ paar vorme aanneem. Die rekenaar kan bloot as vertaalhulpmiddel gebruik word om byvoorbeeld spelling na te gaan of deur van ŉ vertaalgeheue gebruik te maak. Hierdie proses word rekenaargesteunde menslike vertaling genoem. Die rekenaar kan ook ŉ groter rol speel soos om ŉ vertaling (gedeeltelik) te genereer waarna dit deur menslike vertalers nagegaan en gekorrigeer word, en dit word mensgesteunde rekenaarvertaling genoem (Hutchins, 1995:431-445). Die MV-sisteme wat deur die DKK aangevra is, kan in die eerste kategorie – rekenaargesteunde menslike vertaling – vervat word, aangesien die vertaling wat dit genereer, in ŉ vertaalomgewing (die sg. *Autshumato ITE*) as suggestie aangebied sal word. Dit bly dus die menslike vertaler se verantwoordelikheid om die suggestie aan te pas om ŉ aanvaarbare vertaling van die brontaalsin te wees. Die *Autshumato*-projekspan moet egter sorg dat die outomatiese vertaling van só ŉ aard is dat die menslike prosessering vinnig en effektief kan geskied.

ŉ Hele aantal metodes kan gebruik word om hierdie outomatiese vertaler te ontwikkel. Die gewildste metodes sluit reëlgebaseerde, voorbeeld- of geheuegebaseerde en statistiese MV in (Jurafsky & Martin, 2009: 895-944). Vir die eersgenoemde, word ŉ aantal reëls geskryf om linguistiese kennis na te boots en die vertaling word dan daarvolgens gedoen (Arnold *et al*., 1994:66-69). Die tweede metode behels die voorsiening van enorme hoeveelhede parallelle korpora wat as voorbeelde aan ŉ algoritme gegee word. Die voorbeelde word dan geënkodeer en opgeroep wanneer ŉ soortgelyke frase vertaal moet word (Somers, 2003:513). Die *Autshumato*-projekspan het die derde metode, te wete statistiese masjienvertaling (SMV), as basiese benadering tot die ontwikkeling van die sisteme gekies. Algoritmes wat hierdie benadering volg, genereer die mees waarskynlike vertaling van ŉ brontaalsin deur statistiese modelle wat van parallelle korpora afgelei is te gebruik (Somers, 2003:513). Die benadering lewer normaalweg goeie resultate en vaar dikwels beter as reëlgebaseerde metodes, maar verg groot hoeveelhede data (rondom 10 miljoen belynde sinspare) in die vorm van parallelle korpora (Arnold *et al*., 1994: 139-154). In Hoofstuk

[1] Sien http://autshumato.sourceforge.net/ vir meer besonderhede oor die projek. Die module wat in hierdie navorsingsprojek ontwikkel is, asook die nodige toetsdata, sal ook op hierdie webtuiste beskikbaar gemaak word.

2 word die keuse van die metode, asook die werking daarvan, in meer besonderhede bespreek. Die *Autshumato*-masjienvertaalsisteme is tans onder ontwikkeling en die Engels na Afrikaanse sisteem is nou in die finale ontwikkelingsfase. Die korpora waarmee hierdie sisteem afgerig word, sluit ongeveer 470 000 belynde tekseenhede in en bestaan uit belynde sinspare en tweetalige woordelyste. Geen verdere prosessering word nog gedoen om die sisteem te verbeter nie. Dit is egter belangrik om in gedagte te hou dat die sisteme wat in hierdie projek ontwikkel word, nie net as navorsingsmodelle gebruik sal word nie, maar dat dit bedoel is om die werkslading by regeringskantore in die praktyk te verlig. Die projekspan moet dus seker maak dat die hulpbronne wat ontwikkel word, hierdie behoefte bevredig. Die kwaliteit van die afvoer van die sisteme moet daarom hoog wees sodat dit die vertaling van regeringsdokumente kan vergemaklik en nie die proses onnodig belemmer nie.

1.2 Literatuurstudie

Tot dusver is daar nog weinig navorsing oor die verbetering van MV-sisteme vir vertaling na enige van die Suid-Afrikaanse tale gedoen. Hierdie studie sal dus op grond van navorsing vir ander taalpare gedoen moet word. Internasionale navorsing oor tegnieke om MV-sisteme te verbeter sluit onder andere reëlgebaseerde naprosessering, dataseleksie as voorprosessering en voorprosessering deur sintaktiese herrangskikking in.

Reëlgebaseerde naprosessering is ŉ gewilde veld waarin verskeie studies al positiewe resultate gelewer het. Volgens hierdie metode word die afvoer van ŉ basislynsisteem aan ŉ stel reëls gemeet en veranderinge word dan op grond daarvan aangebring. Hierdie veranderinge kan verbeterings in die gebruik van lees- en skryftekens insluit, maar ook meer komplekse probleme soos woordvolgorde en leesbaarheid oplos (Simard *et al*., 2007; Och, 2003; Krings & Koby, 2001).

In die enigste ander studie oor die verbetering van die kwaliteit van die *Autshumato*-SMV-sisteem se afvoer, word die korpora in voorprosessering gemanipuleer en sorgvuldig gekies om die meeste inligting moontlik te bevat (McKellar, 2011). Die intuïsie agter hierdie studie is dat die afrigting van die SMV-sisteem geoptimeer word om die meeste inligting uit die klein hoeveelheid parallelle data wat beskikbaar is te ontgin. Die studie lewer goeie resultate en ŉ verhoging van 20% in die BLEU-telling word gerapporteer.

ŉ Verdere metode wat in die literatuur kompeterende resultate lewer met betrekking tot die verbeterings wat dit te weeg bring, is voorprosessering deur middel van reëlgebaseerde sintaktiese herrangskikking. Die metode behels dat sekere sistematiese verskille in die sintaksis van die bron- en teikentale gebruik word om die twee tale struktureel nader aan mekaar te bring (Parlikar, 2008). Die brontaalsinne word herrangskik sodat die sinstruktuur daarvan meer na die teikentaalstruktuur lyk voordat afrigting van ŉ MV-sisteem selfs begin (Badr *et al*., 2009). Dit lewer goeie resultate en is geskik vir ŉ sisteem waar die teikentaal ŉ hulpbronskaars taal is, aangesien geen gespesialiseerde natuurliketaalprosesseringshulpmiddels daarvoor benodig word nie (vgl. Wang *et al*., 2007; Collins *et al*., 2005; Badr *et al*., 2009; Parlikar, 2008). Die enigste hulpbronne wat ekstra bygevoeg moet word, is ŉ sintaktiese analiseerder vir die brontaal (in

hierdie geval Engels) plus linguistiese kennis van die teikentaal (Afrikaans). Enkele studies wat van hierdie metode gebruik maak, sal vervolgens bespreek word. [2]

Badr *et al*., (2009) rapporteer 'n toename in die BLEU-telling van 'n SMV-sisteem met Engels as brontaal en Arabies as teikentaal wanneer van herrangskikking as voorprosessering gebruik gemaak word. In hierdie studie is die brontaalafrigtingsdata met die *Collins Parser* (Collins, 1997) geanaliseer en die herrangskikkingsreëls is daarna op die geanaliseerde data toegepas. Die herrangskikkingsreëls is deur menslike kenners van die twee tale ontwikkel en berus dus op linguistiese kennis. Die reëls kan in twee kategorieë verdeel word – een stel wat die Subjek-Werkwoord-volgorde manipuleer, en 'n tweede stel wat die struktuur van naamwoordstukke herrangskik. 'n Nuwe SMV-sisteem is met die herrangskikte data afgerig en met die METIS II-toetsteks geëvalueer. Die BLEU-telling wys 'n toename van 0,3217 (sonder herrangskikking) na 0,3246 (wanneer herrangskikkingsreëls gebruik word).

Sjinees na Engelse SMV baat ook by hierdie metode in die studie van Wang *et al*. (2007). Die volgorde van werkwoordstukke, naamwoordstukke en lokaliseringsfrases verskil sistematies in Sjinees en Engels en die reëls fokus daarop om die Sjinese struktuur te herrangskik sodat dit nader aan die Engelse struktuur is. Nadat die Sjinese data herrangskik is, is 'n nuwe MV-sisteem met die *Moses*-SMV-gereedskapstel (sien 2.2.1) ontwikkel en getoets. Die BLEU-tellings het in hierdie geval toegeneem van 0,2852 (sonder herrangskikking) na 0,3086 wanneer die herrangskikkingsreëls gebruik word. Die reëls is ook afsonderlik vir akkuraatheid getoets. 'n Toetsteks van 200 sinne is handmatig herrangskik en vergelyk met die afvoer van dieselfde teks wat outomaties herrangskik is. Lokaliseringsreëls is 77,6% akkuraat, naamwoordstukreëls 54,6% en werkwoordstukreëls 65,7%. Die voorprosesseringsmodule behaal dus 'n gemiddelde akkuraatheid van 62,1%. In die studie word aangetoon en beklemtoon dat die kwaliteit van die sintaktiese analise 'n groot rol in die uiteindelike resultate speel, omdat die patrone wat herken moet word, hierop berus. As daar tydens die sintaktiese analise verkeerde patrone toegeken word, sal die reëls ook verkeerdelik toegepas word.

In 'n derde studie vir vertaling van Engels na Duits, word dieselfde metode as hierbo gevolg om 'n verbetering van 0,2520 na 0,2680 in die BLEU-telling te kry (Collins *et al*., 2005). In hierdie studie word menslike evaluasie ook op die herrangskikte sisteem gedoen deur 100 sinne uit die toetsteks wat met die basislynsisteem en die nuwe sisteem vertaal is, vir twee beoordelaars te gee. Die beoordelaars moes eenvoudig aandui watter een van die twee vertalings hulle verkies. Die eerste beoordelaar het die afvoer van die herrangskikte sisteem in 40 sinne verkies, 40 as onveranderd geklassifiseer en 20 sinne van die oorspronklike sisteem verkies. Die tweede beoordelaar het die afvoer van die herrangskikte sisteem in 44 sinne verkies, 37 sinne as onveranderd gesien en 19 sinne van die oorspronklike sisteem verkies. Albei die beoordelaars het die herrangskikte sisteem in die meerderheid sinne bo die oorspronklike sisteem verkies.

[2] Sien ook 4.3.1 vir 'n volledige beskrywing van die BLEU- en NIST-tellings, asook die METIS II-toetsteks waarna gereeld in die literatuurstudie verwys word.

Herrangskikking van die brontaaldata lewer dus goeie resultate vir verskeie taalpare. Vir Engels na Arabies bring dit ŉ verbetering in die BLEU-telling van 0,3217 na 0,3246, vir Sjinees na Engels ŉ toename in die BLEU-telling van 0,2852 na 0,3086 en vir Engels na Duits ŉ verbetering van 0,2520 na 0,2680. In die Suid-Afrikaanse konteks waar min hulpbronne vir die Suid-Afrikaanse tale beskikbaar is, sou hierdie metode dus geskik kon wees, aangesien dit nie van duur kerntegnologieë afhanklik is nie.

1.3 Probleemstelling en navorsingsvrae

Aangesien daar nog voorheen geen Engels-Afrikaanse MV-sisteem ontwikkel is nie, is daar ook nog geen navorsing oor die verbetering van so ŉ sisteem gedoen nie. Uit die konteks wat in die vorige afdelings geskets is, kom ŉ behoefte aan akkurate MV-sisteme vir die Suid-Afrikaanse tale na vore en is dit daarom nodig om maniere te ondersoek om MV-sisteme vir hierdie tale te optimaliseer. Aangesien daar nie groot hoeveelhede korpora beskikbaar is vir hulpbronskaars tale soos Afrikaans, isiZulu en Sesotho sa Leboa nie, moet ander kreatiewe oplossings gevind word om die SMV-algoritme optimaal op kleiner datastelle te laat funksioneer.

Die literatuurstudie in 1.2 wys daarop dat sintaktiese herrangskikking in voorprosessering ŉ belowende roete kan wees. Zwarts en Dras (2007) is van mening dat so ŉ sintaktiese herrangskikking die brontaaldata meer toeganklik vir die meganismes van SMV maak en dat dit een van die redes is waarom hierdie metode oënskynlik so goed werk. ŉ Tweede rede vir die effektiwiteit van die benadering wat in die artikel uitgelig word, is dat die herrangskikking die brontaalsintaksis verander om ŉ nader voorstelling van die teikentaalsintaksis te weerspieël en daarom beter afvoer toon. Die omskakeling van brontaalsintaksis na teikentaalsintaksis word dus deur menslike kenners in voorprosessering nageboots en dit word nie aan die statistiese model oorgelaat nie. Verbeterings in die BLEU-telling, asook menslike evaluasie wys op die feit dat die kwaliteit van die afvoer van die verskillende SMV-sisteme verhoog het met die toepassing van herrangskikkingsreëls.

Dit is egter nodig om die invloed van so ŉ voorprosesseringsmodule op die kwaliteit van die afvoer van ŉ SMV-sisteem vir die taalpaar Engels-Afrikaans verder na te vors. Daar is nog geen navorsing vir hierdie tipe voorprosessering vir die spesifieke taalpaar gedoen nie. Die afvoer van die sisteem is ook nog nie geanaliseer om vas te stel wat die areas is waar verbeter kan word nie en die twee tale is ook nog nooit vergelyk met die ontwikkeling van ŉ SMV-sisteem in gedagte nie. Die volgende basiese navorsingsvrae kan dus onderskei word:

1. Wat is die vertalingsfoute wat in die afvoer van die *Autshumato*-SMV-sisteem voorkom en wat deur middel van reëlgebaseerde sintaktiese herrangskikking as ŉ voorprosesseringstap voorkom kan word?

2. (a) Wat is die relevante verskille tussen Engelse en Afrikaanse sintaksis wat moontlik aanleiding tot die foute in (1) kan gee, en

   (b) hoe kan hierdie verskille gebruik word om reëls te formuleer wat in sintaktiese herrangskikking gebruik sou kon word?

3. Tot watter mate sal reëlgebaseerde sintaktiese herrangskikking die huidige *Autshumato*-sisteem beïnvloed met betrekking tot die BLEU- en NIST-tellings?

In die lig van die navorsingsvrae wat bo uiteengesit is, kan die volgende doelstellings vir die voorgenome studie gestel word:

1. Om die afvoer van die *Autshumato* Engels na Afrikaanse SMV-sisteem te analiseer en vertalingsfoute wat moontlik deur voorprosessering d.m.v. reëlgebaseerde sintaktiese herrangskikking voorkom kan word te identifiseer.

2. (a) Om die verskille tussen Engelse en Afrikaanse sintaksis wat moontlik vir die foute verantwoordelik kan wees na te vors, en

   (b) om linguisties gemotiveerde reëls te formuleer wat in die voorprosesseringsmodule gebruik kan word. Hierdie reëls sal ook afsonderlik geëvalueer word om die effektiwiteit daarvan na te gaan.

3. Om die afvoer van die resulterende SMV-sisteem (hierna die afvoer van die *uitgebreide sisteem* genoem) te evalueer en krities met die huidige *Autshumato*-SMV-sisteem te vergelyk. Evaluasie behoort die internasionaal aanvaarde BLEU- en NIST-tellings in te sluit. Die m

1.4 Hipotese en metodologie

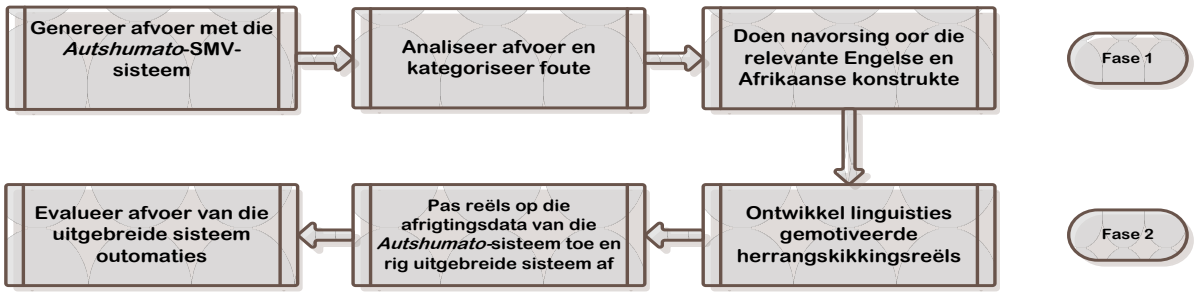
Hierdie navorsingsprojek veronderstel dat ŉ grondige analise van die afvoer van die *Autshumato*-SMV-sisteem tot die identifikasie van sekere probleemareas of -kategorieë sal lei. Dit word verder aangeneem dat sommige van hierdie foutkategorieë uit sistematiese verskille tussen die sintaksis van Engels en Afrikaans voortspruit en dat hierdie verskille in herrangskikkingsreëls wat vir voorprosessering gebruik kan word, geformaliseer kan word. Die manipulasie van die brontaalteks sal volgens die hipotese tot verbeteringe in die afvoer van die sisteem waarop hierdie voorprosessering toegepas word, lei.

Die navorsingsprojek kan as toegepaste navorsing beskryf word, aangesien dit hier gaan om die kennisbasis wat eers deur ŉ literatuurstudie opgebou moet word en daardie kennis wat dan uiteindelik op ŉ spesifieke probleem toegepas kan word (OECD, 2002:78). Die navorsing kan in twee fases verdeel word:

1. As ŉ eerste stap sal die vertalingsfoute in die afvoer van die *Autshumato*-SMV-sisteem saamgegroepeer word om sodoende onderliggende verskille tussen Engelse en Afrikaanse sintaktiese konstrukte te vind. Die relevante sintaktiese verskille kan dan in herrangskikkingsreëls geformaliseer word. Navorsingsvrae (1) en (2a) word dus in hierdie fase ondersoek en moontlike oplossings vir die probleem word voorgestel.

2. Die tweede fase neem ŉ aanvang met die ontwikkeling en implementering van die voorprosesseringsmodule (Navorsingsvraag (2b)). ŉ Uitgebreide sisteem sal ontwikkel word met die voorprosessering as eerste komponent en die afvoer van hierdie SMV-sisteem sal ook geëvalueer moet word om die bruikbaarheid te bepaal. In die tradisie van die internasionale navorsing wat reeds bespreek is, sal die BLEU- en NIST-tellings bereken word, en die afvoer van die uitgebreide sisteem sal ook deur menslike beoordelaars geëvalueer word.

Figuur 1 toon ŉ skematiese voorstelling van die fases in die navorsingsmetode.



Figuur 1: Fases in die navorsingsprojek

1.5 Samevatting

Mensetaaltegnologie en Natuurliketaalprosessering is relatiewe nuwe velde in Suid-Afrika en min navorsing binne die veld van masjienvertaling is tot dusver vir die hulpbronskaars tale gedoen. Hierdie studie poog dus om ŉ bydra te maak tot die sukses van die eerste SMV-sisteem vir Engelse na Afrikaanse vertaling en ontwikkel en toets nuwe tegnieke om die bruikbaarheid van die afvoer van so ŉ sisteem te verhoog. Sou die tegnieke suksesvol blyk te wees, kan dit ook op die ander hulpbronskaars tale, en in besonder isiZulu en Sesotho sa Leboa, van toepassing gemaak word om tot die sukses van dié sisteme, wat ook deel van die DKK se *Autshumato*-projek uitmaak, by te dra.

Hoofstuk 2 bevat ŉ grondige oorsig oor die terrein van statistiese masjienvertaling, asook oor die gereedskap wat beskikbaar is om dit te ontgin. Verder word die analise van die huidige *Autshumato* Engels na Afrikaanse SMV-sisteem ook gedoen om probleemareas te identifiseer. Die relevante sintaktiese konstrukte word ook in hierdie hoofstuk uiteengesit.

Hoofstuk 3 beskryf die ontwikkeling en implementering van die voorprosesseringsmodule met spesifieke aandag aan die linguistiese herrangskikkingsreëls wat in ŉ voorprosesseringsmodule vervat sal word. Hierdie module en die uitgebreide SMV-sisteem word dan in Hoofstuk 4 geëvalueer aan die hand van internasionaal erkende metrieke voordat Hoofstuk 5 die gevolgtrekkings wat uit die navorsingsprojek gemaak kan word gee en as samevatting van die projek dien.

Hoofstuk 2: Analise van die afvoer van die *Autshumato*-masjienvertaler

2.1 Inleiding

Die *Autshumato*-SMV-sisteem vorm die grondslag vir die navorsing wat hier gedoen word, aangesien dit in hierdie projek as die basislynsisteem (*baseline system*) vir outomatiese vertaling van Engels na Afrikaans dien. Enige sisteme wat spruit uit verdere ontwikkeling of prosessering sal dus met hierdie sisteem vergelyk word om die invloed van die veranderinge op die afvoer te evalueer. Dit is daarom belangrik om in dié hoofstuk die werking van die masjienvertaalalgoritme te beskryf, asook om ŉ deeglike analise van die kwaliteit van die afvoer wat tans gegenereer word te doen. Op die manier kan die gebreke in die afvoer van die standaardsisteem geïdentifiseer word, voordat daar in volgende hoofstukke na ŉ moontlike oplossing gesoek kan word.

Hoofstuk 2.2 gee ŉ oorsig oor statistiese masjienvertaling en in besonder die *Moses*-SMV-gereedskapstel (2.2.1). Die foute in die *Autshumato*-afvoer word in 2.3 beskryf en met voorbeelde toegelig voordat ŉ samevatting in 2.4 gemaak word.

2.2 Statistiese masjienvertaling

Masjienvertaling is in die 1960's as die heilige graal van natuurliketaalprosessering beskryf omdat dit so ŉ komplekse taak is (Bar-Hillel, 1960). Gesofistikeerde masjienvertaling behels nie net ŉ woord-vir-woord vertaling met ŉ tweetalige woordelys nie, maar vereis ook dat die vertaling natuurlik en getrou aan die oorspronklike moet wees (Jurafsky & Martin, 2009: 911). Statistiese masjienvertaling (SMV) modelleer die afrigtingsdata om juis aan hierdie twee vereistes te voldoen. Probabilistiese modelle word opgestel om die natuurlikste vertaling, wat so na moontlik aan die oorspronklike sin is te vind. Hierdie modelle bestaan uit frasetabelle wat belynde groepe woorde in beide die bron- en teikentale opsom. Vir elke item in so ŉ tabel is daar dus ŉ brontaalfrase, ŉ ooreenstemmende teikentaalfrase en ŉ waarskynlikheid dat dié twee frases vertalings van mekaar is. ŉ Teikentaalfrase mag dus meer as een maal in die tabel voorkom, maar nooit saam met dieselfde brontaalfrase nie. Die waarskynlikheidsaanduiding word bepaal deur die frekwensie van ŉ spesifieke belyning (d.i. die frekwensie van die spesifieke kombinasie) teen die totale hoeveelheid belynings te normaliseer.

Die resultate van die SMV-afvoer is afhanklik van die kwaliteit en kwantiteit van die parallelle tekskorpora wat tydens afrigting aan die sisteem beskikbaar is (Arnold *et al*., 1994: 139-154). Hoe meer data gebruik word om die modellering te doen, hoe vollediger sal die tweetalige frasetabel wees wat tydens hierdie stap onttrek word. Die waarskynlikheidsaanduiding van elke vertaling sal ook realistieser wees omdat meer voorbeelde van elke item in die frasetabel gebruik word om hierdie waarskynlikheid te bepaal. Die waarskynlikheid wat in die frasetabel vir ŉ belynde paar aangegee word, is dus ŉ beter voorstelling van die frekwensies waarin hierdie frases in die regte wêreld voorkom. Omdat die totale frasetabel beter saamgestel kan word met meer data, kan die sisteem vertalings van ŉ hoër gehalte genereer.

Die *Autshumato*-projekspan het statistiese masjienvertaling (SMV) as benadering vir die volgende redes gekies (Groenewald & Du Plooy, 2010):

- SMV is tans die benadering wat deur verskeie internasionale industriële en akademiese navorsingslaboratoria verkies word;
- moderne SMV-gereedskapstelle (*toolkits*) is vrylik as oopbronprogrammatuur beskikbaar; en
- minder ekspert-linguistiese kennis is nodig om ŉ werkende basislynsisteem (*baseline system*) met hierdie metode daar te stel as met ŉ reëlgebaseerde benadering.

Een van die bekendste gereedskapstelle wat vir masjienvertaling ontwikkel is, is die *Moses*-SMV-pakket (Koehn *et al*., 2007). Hierdie oopbronprogrammatuur laat die gebruiker toe om outomaties SMV-sisteme vir enige taalpaar af te rig en sal vervolgens bespreek word.

2.2.1 Die *Moses*- SMV-gereedskapstel

Koehn *et al*. (2007) noem dat een van die redes vir die ontwikkeling van ŉ oopbron-SMV-gereedskapstel was om die veld te help groei. Voordat hierdie stel hulpbronne beskikbaar gemaak is, was die meeste navorsing op die gebied tot interne projekte of duur inisiatiewe beperk (Koehn *et al*., 2007). Die *Moses*-SMV-gereedskapstel maak dit egter moontlik om relatief vinnige vordering te maak en die afvoer van die sisteme effektief met mekaar te kan vergelyk. Hierdie gereedskapstel vorm dan ook die raamwerk waarbinne die basislynsisteem en verdere ontwikkeling in die *Autshumato*-projek gedoen sal word. Die tipes data wat in die verskillende stappe nodig is, word vervolgens bespreek waarna die stappe wat nodig is vir die daarstel van die basislynsisteem uiteengesit sal word.

2.2.1.1 Data

Drie tipes data is nodig om ŉ SMV-sisteem met die *Moses*-gereedskapstel af te rig.

- Teks in die teikentaal (Afrikaans) is nodig om ŉ taalmodel te skep.
- ŉ Parallelle korpus wat op sinsvlak belyn is, word vir afrigting gebruik.
- ŉ Aparte datastel word ook gebruik om die sisteem te toets.

Tabel 1 gee die hoeveelheid data van elke tipe wat in die *Autshumato*-projek gebruik word. Elkeen van die tipes data word daarna bespreek.

Korpus	Aantal tekseenhede	Bronne
Eentalige korpus vir taalmodelle	5 572 462 sinne	*Media24*-korpus (Pharos Dictionaries, 2006)
Parallelle korpus vir afrigting	470 019 belynde pare	Saamgestel uit data van www.services.co.za, ander regeringsdata van die NLS en Hansards, asook korpora van privaatinstansies wat vertaalgeheues en tydskrifte soos *Pula Imvula* insluit.
Toetsdata	200 sinne	METIS II-toetsteks (Dirix *et al*., 2007)

Tabel 1: Aantal tekseenhede in elke korpus

Die taalmodelle wat met teikentaaldata afgerig word, gee linguistiese inligting aan die sisteem. Patrone in die woordvolgorde en ander taalspesifieke konvensies word daarin gemodelleer en met waarskynlikheidsaanduidings verbind (Stolcke, 2002). Die taalmodelle word in die dekodeerder gebruik om die gegenereerde vertaling meer na die patroon van die teikentaal te laat lyk. Die eentalige data in die teikentaal wat hiervoor gebruik word, kan ook met ekstra annotasies soos morfologiese analise en lemma-inligting verryk word, maar enige verdere inligting moet met taalspesifieke hulpbronne toegevoeg word. Vir die *Autshumato*-SMV-sisteem is geen ekstra inligting toegevoeg nie, omdat interne eksperimente gewys het dat verryking van die data met woordsoortetikette en inligting oor die lemmas swakker resultate lewer (die NIST-telling het van 8,3610 na 7,7655 gedaal en die BLEU-telling van 0,4811 na 0,4136). Hierdie resultate kan moontlik toegeskryf word aan die relatief klein hoeveelheid data wat gebruik word om die sisteem mee af te rig. Omdat daar nie baie voorbeelde van woorde in verskillende kontekste en daarom met verskillende stelle linguistiese inligting voorkom nie, is hierdie inligting te meerduidig om tot die kwaliteit van die afvoer by te dra. ŉ Verdere faktor wat hier ŉ rol speel, is dat die nodige tegnologieë vir Afrikaans nog nie op internasionale standaard is nie.

Die taalmodelle wat in hierdie navorsingsprojek en in die *Autshumato*-projek gebruik word, bevat dus net patrone wat uit die *Media24*-korpus (Pharos Dictionaries, 2006) onttrek is en geen ekstra annotasie word gedoen nie. Die *Media24*-korpus is ŉ versameling Afrikaanse nuusartikels wat die *Autshumato*-projekspan vir navorsingsdoeleindes bekom het en bevat meer as 5 miljoen sinne. ŉ Volledige beskrywing van die opstel van taalmodelle word in 2.2.1.2 gegee.

Die parallelle korpus is die belangrikste in statistiese masjienvertaling, aangesien dit uit hierdie korpus is wat die probabilistiese modelle en die frasetabelle onttrek word. Dit is dus belangrik om soveel data moontlik hierby in te sluit. Dit is ook belangrik om die sinsvlakbelyning so goed as moontlik te doen. Onakkurate belynings tussen brontaal- en teikentaalsinne kan die frasetabel en waarskynlikheidsaanduidings beïnvloed, aangesien woorde of frases verkeerdelik met mekaar verbind sal word. Vir die *Autshumato*-projek is data van die regeringsdomein van die internet onttrek, meestal van die webtuiste www.services.co.za. Hierdie data is ŉ versameling dokumente oor die dienste wat die Suid-Afrikaanse regering lewer en is verteenwoordigend van die tipes dokumente wat die vertalers by die Nasionale Taal-

diens (NLS)[3] op ’n daaglikse basis vertaal. Die dokumente bevat terminologie wat uniek aan die regeringsdomein is, en wys ook ’n skryfstyl wat deur die NLS gehandhaaf word. Dit is daarom gepaste data om vir die afrigting van ’n masjienvertaler wat in die NLS moet funksioneer te gebruik, aangesien die frasetabelle wat tydens afrigting onttrek word hierdie eienskappe sal weerspieël. Ander data, soos vertaalgeheues van privaatinstansies wat deur die *Autshumato*-span vir navorsingsdoeleindes ingesamel is, is ook gebruik.

Die toetsdata moet soortgelyk aan die afrigtingsdata wees en moet dus verkieslik uit dieselfde domein as die afrigtingsdata wees (Cieri, 2007: 229). Dieselfde voorprosessering wat op die afrigtingsdata toegepas is, moet ook op die toetsdata toegepas word. Die afrigtingsdata en toetsdata moet byvoorbeeld met dieselfde tekseenheididentifiseerder verdeel word om toe te sien dat die akkuraatheid van hierdie stap dieselfde vir beide datastelle is. Dit is egter ook belangrik dat die toetsdata nie in die afrigtingsdata vervat word nie, aangesien dit sal lei tot ’n wanvoorstelling van die kwaliteit van die afvoer (Jurafsky & Martin, 2009: 126). Sinne wat in die toetsdata en afrigtingsdata teenwoordig is, sal perfek vertaal word, en wys nie die vermoë van die masjienvertaler om ongesiene data te vertaal nie. Verwysingsvertalings word ook tydens evaluasie gebruik om die afvoer van die MV-sisteem te evalueer. Hierdie vertalings van die toetsdata moet deur linguiste of taalpraktisyns opgestel word. [4]

Hierdie drie tipes data word in verskillende stappe by die afrigting van ’n SMV-sisteem gebruik, maar die kwaliteit van elkeen dra tot ’n groot mate by tot die uiteindelike kwaliteit van die afvoer van die sisteem. Daar is ses stappe in die afrigtingsproses, te wete datavoorbereiding, taalmodellering, woordbelyning, frasetabelonttrekking, herrangskikkingsmodelonttrekking en genereringsmodelonttrekking (Koehn, 2010). Elkeen van hierdie stappe word vervolgens bespreek met verwysing na die data en hoe dit in die verskillende stappe aangewend word. Daar sal ook aangetoon word hoe die resulterende modelle in die dekodeerder gebruik word wanneer data vertaal word.

2.2.1.2 Stappe in die afrigting van ’n SMV-sisteem

Die *Moses*-gereedskapstel bevat ’n afrigtingsalgoritme en dekodeerder as sentrale modules. Die dekodeerder is die module wat uiteindelik vir die vertaling verantwoordelik is, maar dit moet afgerig word om korrekte waarskynlikheidsaanduidings en belynde frases te bevat. Die dekodeerder is dus grootliks afhanklik van die sukses en kwaliteit van die afvoer van elkeen van die stappe wat hieronder bespreek word.

Voorbereiding van die data

Die eerste stap is om die afrigtingsdata voor te berei deur beide kante van die parallelle korpus (Engels en Afrikaans in hierdie geval) op sinsvlak te belyn, alles na kleinletters om te skakel en sinne langer as 100 woorde te verwyder. Dit is nodig om die data na kleinletters om te skakel om

[3] Sien http://www.dac.gov.za/chief_directorates/language_services.htm vir meer besonderhede oor hierdie diens.

[4] Hoofstuk 4.3.1 brei verder oor die toetsdata en verwysingsvertalings uit.

sodoende die woordbelyningsproses te vergemaklik. Lang sinne (meer as 100 woorde) word vervolgens verwyder omdat die belynings wat in die volgende stap gedoen word, nie effektief daarmee kan funksioneer nie. Wanneer ŉ sin te lank word, verskil die posisies van woorde in die bron- en teikentaal te veel van mekaar en kan ŉ outomatiese belyning nie met sekerheid gedoen word nie (Koehn *et al*., 2007).[5]

Opstel van die taalmodelle

Alhoewel die algoritme wat hierdie deel van die proses hanteer as deel van die gereedskapstel versprei word, word dit ook apart beskikbaar gestel sodat taalmodelle vir ander toepassings opgestel kan word. Die *SRILM*-gereedskapstel (Stolcke, 2002) is oopbronprogrammatuur wat vrylik vir navorsingsdoeleindes gebruik kan word en het ten doel om taalmodellering so maklik en vinnig as moontlik te maak. Die gereedskapstel is dus ontwerp om al die elemente wat vir hierdie taak nodig is te bevat en ook om maklik in die *Moses*-omgewing in te skakel.

Statistiese taalmodellering behels die opstel van modelle wat die waarskynlikheid van sekere woordstringe weergee. Die algoritme stel eerstens ŉ lys *n*-gramme op. *N*-gramme is stringe woorde van ŉ bepaalde lengte (*n*) en word opgestel deur ŉ venster oor elke sin in die eentalige afrigtingsdata te skuif en al die *n*-gramme so te onttrek (Jurafsky & Martin, 2009: 117). In die frase "ons eet graag pasta", kan die volgende 3-gramme onttrek word (epsilon dui ŉ leë woord aan):

1. "*epsilon ons eet*"
2. "*ons eet graag*"
3. "*eet graag pasta*"
4. "*graag pasta epsilon*"

In die volgende stap word waarskynlikheidsaanduidings aan elkeen van hierdie stringe gekoppel om aan te dui hoe gereeld die spesifieke *n*-gram in die afrigtingsdata voorkom in verhouding tot die totale hoeveelheid *n*-gramme van dieselfde lengte. Aparte taalmodelle word vir verskillende waardes van *n* opgestel en die gebruiker kan self die verskillende waardes bepaal. Eksperimente in die *Autshumato*-projek het getoon dat 3-, 4- en 5-gramme die beste modellering van Afrikaanse struktuur lewer. Drie modelle, een vir elkeen van die verskillende lengtes *n*-gramme, word dus deurgaans in hierdie navorsingsprojek gebruik. Die drie modelle word dan saam gebruik om Afrikaanse strukture van verskillende lengtes te modelleer.

[5] Al die modules wat nodig is om hierdie voorbereiding te doen, word as deel van die *Moses*-gereedskapstel versprei en kan by http://www.statmt.org/ afgelaai word.

Woordbelynings

Die stappe wat hierna volg, gebruik almal die parallelle afrigtingskorpus wat op sinsvlak belyn moet wees en al die voorprosessering wat vroeër in hierdie afdeling genoem word, moet reeds gedoen wees. Hoofletters in die korpus moet dus reeds na kleinletters omgeskakel wees en geen sinne langer as 100 woorde mag in die korpus wees nie.

Verdere belynings word nou in twee stappe gedoen. In die eerste stap word woorde outomaties op ŉ growwer vlak belyn en in die tweede stap word die woordbelynings verfyn in die lig van soortgelyke woordpare. Die eerste stap belyn woorde wat op die oog af dieselfde spelling het, of wat gereeld in dieselfde konteks voorkom. As die sin "ek eet graag pasta" gereeld in die parallelle korpus met ŉ frase "*i like eating pasta*" belyn word, kan die algoritme aflei dat "pasta" in al twee sinne met mekaar belyn kan word omdat die spelling identies is. Die woorde "eet" en "*eating*", asook "ek" en "*i*" kan ook belyn word omdat hierdie twee pare woorde waarskynlik gereeld in ander sinne wat met mekaar belyn word, voorkom. Die oorblywende woorde "graag" en "*like*" kan dan ook belyn word omdat dit die enigste dele is wat nog nie gekoppel is nie.

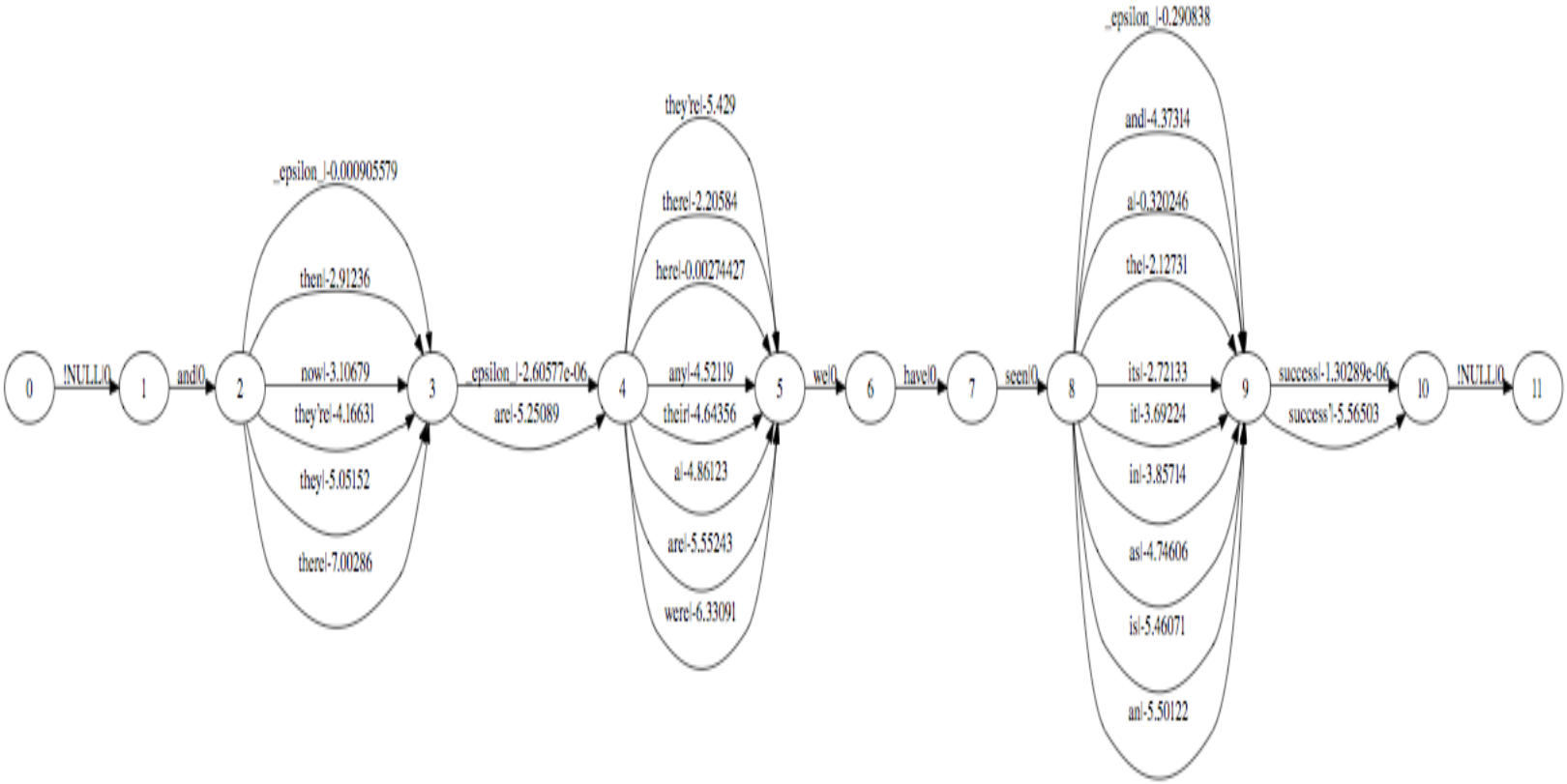
Die volgende stap in die woordbelyningsfase, verfyn hierdie growwe belynings deur teenstrydighede op te los en veralgemenings te maak. In die voorbeeld wat bo genoem word, kan "*like"* ook aan "hou van" in ŉ variant van die Afrikaanse sin gekoppel word. Die gevolglike teenstrydighede word opgelos deur na die frekwensies van die verskillende belynings te verwys. Indien albei die moontlikhede gereeld in die parallelle korpus voorkom, word albei as geldige belynings aanvaar. Indien ŉ belyning ŉ baie laer frekwensie het, word dit as ŉ fout gemerk en nie verder in berekening gebring nie. Veralgemenings mag ook insluit om groter frases met mekaar te belyn. Dit beteken dat vaste uitdrukkings, bv. "in verband met" in Afrikaans met hul ooreenstemmende woordgroep, "*with regard to"* in Engels belyn kan word.

Stel ŉ vertalingstabel op woord- en frasevlak op

Die derde stap in die afrigtingsproses gebruik die woord- en frasebelynings wat in die vorige stap opgestel is om ŉ vertalingstabel op woordvlak op te stel. Hierdie stap stel ŉ tweetalige leksikon op en verskaf verder ŉ waarskynlikheidsaanduiding vir elke vertaling. Die waarskynlikheidsaanduiding word later in die dekodeerder gebruik om keuses tussen verskillende moontlikhede te maak en word op die kante van die diffusienetwerk gebruik (sien Figuur 2). ŉ Soortgelyke tabel word ook vir frases opgestel sodat die dekodeerder nie woord-vir woord vertalings hoef te genereer nie, maar die langste moontlike string kan gebruik.

Stel die herrangskikkingsmodel op

In die volgende stap word ŉ herrangskikkingsmodel opgestel wat ŉ koste aan die volgorde van woorde toeken. Woorde in ŉ frasepaar wat oorkruis belyn word, sal swaarder beboet word as woorde wat in dieselfde volgorde in beide die brontaalfrase en teikentaalfrase is. Die kostes wat in hierdie stap toegeken word, speel ook ŉ rol in die uiteindelike waarskynlikheidsaanduidings wat in Figuur 2 aangegee word, aangesien dit ten doel het om woorde wat gereeld in ŉ spesifieke volgorde voorkom te bevoordeel omdat hierdie woordvolgorde beter vertalings behoort te lewer.



Figuur 2: Grafiese voorstelling van ’n diffusienetwerk[6]

[6] Beskikbaar by http://www.statmt.org/moses/ ?n=Moses.ConfusionNetworks.

Stel die genereringsmodel op

In die laaste stap word ŉ genereringsmodel opgestel en al die elemente word in ŉ logiese manier georden om die diffusienetwerk te vorm met al die inligting wat uit die vorige stappe onttrek word. ŉ Diffusienetwerk is ŉ geweegde, gerigte grafiek met die spesiale eienskap dat elke pad van die beginnode deur al die ander nodes loop tot by die eindnode (Bertoldi & Federico, 2005). Tussen elkeen van die nodes (die kante of "*edges*") word ŉ etiket met ŉ woord en waarskynlikheidsaanduiding aangegee. Die totale waarskynlikheid van ŉ pad van die begin na die einde, word bepaal deur die waarskynlikheidsaanduidings van die kante te vermenigvuldig.

Uit die diffusienetwerk wat in Figuur 2 voorgestel word, kan ons die frase "*and here we have seen the success*" genereer deur telkens die woord met die hoogste waarskynlikheid te kies (let op dat "*epsilon*" as ŉ leë woord vertaal word).

Die genereringsmodel en konfigurasies word uiteindelik deur die dekodeerder gebruik om, wanneer die masjienvertaler gebruik word, ŉ vertaling te lewer. Hierdie vertaling word uit die diffusienetwerk onttrek op dieselfde manier as wat bo beskryf is. Die vertalingstabel verskaf die moontlike parallelle woorde of frases en die dekodeerder vind dan die pad met die hoogste waarskynlikheidsaanduiding vir die spesifieke konteks in die diffusienetwerk. Die dekodeerder is dus die algoritme wat al die inligting wat deur die verskillende stappe in die afrigtingsalgoritme ontgin is, gebruik om sistematies van die brontaal na die teikentaal te vertaal (Koehn *et al*., 2007).

2.2.1.3 Samevatting

Die *Moses*-gereedskapstel kan moeiteloos gebruik word om SMV-sisteme af te rig en vertalings met die resulterende sisteem te genereer. Die afrigtingsfase verloop in ses stappe, te wete datavoorbereiding, taalmodellering, woordbelyning, frasetabelonttrekking, herrangskikkingsmodelonttrekking en genereringsmodelonttrekking. In die eerste stap word hoofletters met kleinletters vervang, lang sinne word verwyder en die korpus word op sinsvlak belyn om ŉ Engels-Afrikaanse parallelle korpus te vorm. Die volgende stap neem die eentalige Afrikaanse korpus en onttrek taalmodelle (3-, 4- en 5-gramme) om die struktuur van die teikentaal na te boots. Die derde stap onttrek woordbelynings uit die parallelle korpus en in die vierde stap word frases uit hierdie korpus onttrek om die verskillende vertalingsmoontlikhede in die twee tale te gee. Die dekodeerder sal uiteindelik hierdie tabelle gebruik om woorde en frases te vertaal. Elkeen van die elemente in hierdie tabelle word ook van waarskynlikheidsaanduidings verskaf om die frekwensie van die element in die afrigtingskorpus voor te stel. Die vyfde stap behels die opstel van ŉ herrangskikkingsmodel. Hierdie model bevoordeel uiteindelik woorde wat gereeld in ŉ spesifieke volgorde voorkom soos vaste uitdrukkings (bv. "ten spyte van" en "na gelang van"). In die laaste stap word ŉ diffusienetwerk opgestel wat inligting uit die vorige stappe orden en so voorstel dat die dekodeerder daaruit keuses kan maak. Die verskillende tabelle en modelle wat in die vorige vier stappe opgestel is, word nou saamgevoeg om een voorstelling te vorm.

Die belangrikste aspek van die afrigting is die onttrekking van die waarskynlikheidsaanduidings uit die parallelle afrigtingskorpus. As die verkeerde waarskynlikheidsaanduidings onttrek word, sal die diffusienetwerk in die dekodeerder ook nie die korrekte keuses kan maak nie. Die waarskynlikheidsaanduidings is ook kontekssensitief wat inhou dat ŉ waarskynlikheid vir ŉ gegewe kant afhanklik is van die waarskyn-

likhede van die kante daarvoor en daarna. Hoe meer afrigtingsdata die sisteem dus beskikbaar het, hoe beter kan dit hierdie kontekste modelleer (sien byvoorbeeld Lü *et al.*, 2007 en Mandal *et al*., 2008).

In die volgende afdeling sal die foute wat die *Autshumato*-sisteem maak, bespreek word. Hierdie foute sluit die woordvolgorde van vertaalde sinne (2.3.1), probleme met ontkenning (2.3.2) en verlede tyd (2.3.3), asook foute met betrekking tot die posisie van die werkwoorde in die Afrikaanse vertaling (2.3.4) in. ŉ Algemene kategorie word ook onderskei waarin probleme wat nie deur die voorgestelde voorprosesseringsmodule aangespreek kan word nie, bespreek word (2.3.5).

2.3 Foute in die *Autshumato*-vertaling

Alhoewel beide Engels en Afrikaans sogenaamde SVO-tale is (sinne word in die volgorde Subjek – Werkwoord – Objek georganiseer), is die toepassing van die interne norme van die twee tale baie verskillend. Interne norme "is gegrond op verskynsels wat eie is aan die taalkundige struktuur van ŉ besondere taal" (Carstens, 2004: 39). Soos vroeër bespreek, is dit nodig om die afvoer van die basislynsisteem te analiseer om sodoende areas te identifiseer waar die interne norme van die twee tale sistematies van mekaar verskil sodat hierdie verskille met behulp van herrangskikkingsreëls geminimaliseer kan word. Om hierdie analise te doen is ŉ teks van 300 sinne met die basislynsisteem vertaal. Hierdie sinne is willekeurig uit die totale afrigtingsdatastel gekies en uit die afrigtingskorpus verwyder. Hierdie aparte ontwikkelingsteks het ook geen ekstra voor- of naprosessering ondergaan nie. Die sinne uit die ontwikkelingsteks is daarna handmatig in ses kategorieë gegroepeer op grond van die soort foute wat daarin voorkom. Die kategorieë sluit woordvolgorde (2.3.1), ontkenning (2.3.2), verlede tyd (2.3.3), werkwoorde (2.3.4) en ŉ algemene kategorie (2.3.5) in. Hierdie foutgroepe word vervolgens bespreek deur eerstens na die spesifieke konstruksie in Afrikaans en daarna in Engels te verwys. ŉ Vergelykende opsomming met voorbeelde word daarna vir elke foutgroep gegee. Vir elke voorbeeld word die oorspronklike Engelse sin, die *Autshumato*-vertaling waarin die fout voorkom, en daarna die korrekte Afrikaanse sin gegee. Dit moet egter hier duidelik gestel word dat die sintaktiese verskille tussen Engels en Afrikaans aan die hand van kursoriese en eksemplariese hoëvlakveralgemenings gedoen word, aangesien dit op hierdie vlak is wat die outomatiese herrangskikking toegepas sal word.

2.3.1 Woordvolgorde

Carstens (2004: 40-46) en Du Plessis (1985: 19-28) lig die onderstaande belangrike punte rakende Afrikaanse woordvolgorde uit.

- Afrikaans toon ŉ volgorde van subjek-werkwoord-komplement (of -objek) in ongemerkte sinne (bv. "Ek stap daagliks").
- Ander elemente (bv. bywoorde) kan voor die werkwoord geplaas word en die subjek skuif dan na die posisie ná die werkwoord (bv. "Soggens **stap ek** graag").
- Die posisie van die werkwoord verskil in verskillende tipes sinne.
    - Die werkwoord skuif na die begin van ŉ vraagsin of na die posisie net ná die vraagwoord (bv. "**Stap** jy ook gereeld?").
    - Die werkwoord skuif na die begin van ŉ wenssin of seënwense (bv. "**Was** ek tog maar ŉ stapper!").

- Die werkwoord(stuk) in bysinne staan nader aan die einde van die sin en die hulpwerkwoord van tyd staan dan daarna (bv. "Ek stap graag soggens, maar geniet dit as ek in die aand **gerus het**").
- Om 'n subjek in die beginposisie te sit vir klemtoon (vooropstelling), is ook 'n aanvaarbare verskuiwing in Afrikaans (bv. "**Stap** is my lewe").

Volgens Biber *et al.* (2002: 400-405) vertoon die ongemerkte woordvolgorde in Engels die onderstaande eienskappe.

- Die subjek gaan gewoonlik die werkwoord vooraf en die werkwoord gaan gewoonlik die komplemente vooraf (bv. "*I sing every day*").
- Alle frase-elemente wat deur *wh*-woorde gerealiseer word, word gereeld in die beginposisie van 'n frase geplaas (bv. "*I don't know **what** to sing"*).
- Frases word gewoonlik nie deur ander elemente verbreek nie (bv. *"I **do not understand** this"* word nie verdeel in *"I **do** this **not understand**")*.
- Vooropstelling kom gereeld voor en dus word 'n konstituent wat normaalweg na die werkwoord voorkom, dikwels beklemtoon deur die konstituent in die beginposisie te plaas (bv. "***Every day** I sing*").
- Inversie (*inversion*) is ook algemeen in Engels. Die werkwoordstuk skuif dan tot voor die subjek (bv. "***Singing** is what I do*").

Foute in die woordvolgorde het heel dikwels saam met ander foutgroepe voorgekom. Die verledetydsmerkers sou byvoorbeeld in die verkeerde posisies geplaas word en Voorbeeld 1 (V1) wys dat die woordvolgorde van die negatiewe sin nie korrekte Afrikaanse volgorde óf negativering bevat nie. Die eerste ontkennende woord ("nie") moet na die werkwoord ("moet") kom en die werkwoord moet in hierdie geval voor die naamwoordstuk ("die administrasie") wees. Die tweede ontkenningswoord ontbreek ook.

> (V1.1) Engels: "*In future, the administration must not be able to react.*"
>
> (V1.2) MV-afvoer: "In die toekoms, die administrasie **nie moet in staat wees** om te reageer."
>
> (V1.3) Korrekte Afrikaans: "In die toekoms **moet** die administrasie **nie in staat wees** om te reageer **nie**."

Daar was ook heelwat sinne waar die woorde almal korrekte vertalings was, maar die volgorde nie aanvaarbaar was nie, aangesien die vertaalde sin steeds 'n Engelse woordvolgorde gehad het. In V1 en V2 word al die Engelse woorde na korrekte Afrikaanse woorde vertaal, maar die volgorde van die Afrikaanse sin is verkeerd.

> (V2.1) Engels: "*Without his strong support for peace, Europe would have looked different today."*
>
> (V2.2) MV-afvoer: *"*Sonder sy sterk steun vir vrede, **Europa sou** vandag anders gelyk het.*"*
>
> (V2.3) Korrekte Afrikaans: "Sonder sy sterk steun vir vrede **sou Europa** vandag anders gelyk het."

Uit hierdie voorbeelde word die raakpunte tussen Afrikaans en Engels wat woordvolgorde betref duidelik, maar daar is ook heelwat verskille wat die MV-sisteem se taak kan bemoeilik. Die oorgrote meerderheid van die sinne wat in hierdie foutkategorie ingedeel kan word, toon ŉ anglisistiese sinstruktuur wat nie aanvaarbare vertalings is nie en selfs soms tot verwarrende sinne kan lei. Hierdie foute kan toegeskryf word aan die onvermoë van die masjienvertaalalgoritme om Afrikaanse sinstruktuur effektief uit ŉ relatief klein hoeveelheid data te leer. Die volgorde van woorde in die teikentaalsin word bepaal deur die inligting wat in die taalmodelle vervat word. In 2.2 is reeds gestel dat die kwaliteit van dié modelle direk afhang van die vermoë van die algoritme om akkurate *n*-gramme en waarskynlikheidsaanduidings te onttrek waarvolgens die woorde in die teikentaalsin herrangskik kan word. As die taalmodelle dus nie met genoegsame data afgerig word nie, kan dit nie al die moontlike kontekste en strukture modelleer nie en sal die uiteindelike woordvolgorde van die afvoer nader aan die brontaalstruktuur wees omdat die vertaling eerder na ŉ woord-vir-woord vertaling neig. Die voorbeelde het ook gewys dat die werkwoord dikwels deel van die verwarde woordvolgorde is. Later in 2.3.4 word foute wat spesifiek met die posisie van die werkwoord in ŉ sin te make het, meer volledig bespreek.

2.3.2 Ontkenning

Carstens (2004: 57) sê van Afrikaanse negatiewe dat dit belangrik is om die reëls daarrondom noukeurig te volg, aangesien dit ŉ belangrike komponent van die struktuur van die taal is. In enkelvoudige negatiewe sinne kom die ontkenningswoord "nie" prototipies twee maal in ŉ sin voor. Die eerste ontkenningswoord volg gewoonlik direk na die werkwoord(stuk) wat die aksie wat genegativeer word uitbeeld en die tweede na aan die einde van die sin. Carstens wys egter op twee uitsonderings op hierdie basiese patroon.

- Die eerste uitsondering op die normale vorming van ŉ ontkennende sin is dat die tweede ontkenningswoord nie altyd gebruik word nie, veral wanneer die werkwoord nie ŉ kompliment neem nie. Carstens (2004: 57) gee die volgende voorbeelde:

  (V3) "Ek ken haar nie."

  (V4) "Aanstaande jaar kan ek nie, maar wel nou."

- Verder is daar ook ander ontkennende woorde wat in die plek van die eerste "nie" kan staan. Hierdie woorde sluit "niemand", "niks" en "nimmer" in (vergelyk V6).

Ontkenning in Engels, daarteenoor, is ŉ eenvoudiger konstruksie. Biber *et al.* (2002: 239-240) onderskei twee tipes ontkenning – *not-negation* waar "not" of "-n't" voor die element wat genegativeer word, ingevoeg word (sien V5 onder), en *no-negation* waar die negatief gevorm word deur nes vir Afrikaans, ander woorde soos "nothing", "no" en "none" in te voeg (V6 gee ŉ voorbeeld hiervan).

(V5.1) Engels: "*I have **not** signed yet."*

(V5.2) Korrekte Afrikaans: "Ek het nog **nie** geteken **nie**."

(V6.1) Engels: "*Europe has **nothing** to do with the African continent."*

(V6.2) Korrekte Afrikaans: "Europa het **niks** met die Afrika-kontinent te doen **nie**."

Die feit dat Afrikaans in die meeste gevalle twee ontkennende woorde neem en Engels slegs een, sorg vir heelwat foute in die *Autshumato*-afvoer. Die tweede "nie" ontbreek nie net soms nie, maar word ook verkeerdelik ingevoeg wanneer daar reeds ŉ ander negatief soos "nimmer" of "nooit" in die sin gebruik is, of wanneer die sin positief behoort te wees. Die afrigtingsdata in die *Autshumato*-projek bevat nie genoeg voorkomste van die verskillende afwykings van die normale patroon om realistiese waarskynlikheidsaanduidings vir die patrone te onttrek nie. Die patrone wat wel voorkom word dan soms verkeerdelik toegepas. Die onderstaande voorbeelde (V7 en V8) wys hierdie twee foute.

(V7.1) Engels: "*Firstly, as we see it, expansion is **not** only a responsibility*."

(V7.2) MV-afvoer: "Ten eerste, soos ons dit sien, uitbreiding is **nie** net ŉ verantwoordelikheid."

(V7.3) Korrekte Afrikaans: "Eerstens, soos ons dit sien, is uitbreiding **nie** net ŉ verantwoordelikheid **nie**."

(V8.1) Engels: "*We have **never** come to a conclusion.*"

(V8.2) MV-afvoer: "Ons het nog **nooit nie** tot ŉ gevolgtrekking gekom **nie**."

(V8.3) Korrekte Afrikaans: "Ons het nog **nooit** tot ŉ gevolgtrekking gekom **nie**."

In V7 word die tweede "nie" uitgelaat en lewer ŉ Afrikaanse vertaling wat nie aan die interne norme van die taal gehoor gee nie. In V8 word oorbodige ontkennende woorde gebruik en dit lei tot ŉ verwarrende Afrikaanse vertaling. Die Afrikaanse leser mag aflei dat die groep wat hier aan die woord is, altyd tot ŉ gevolgtrekking kom, terwyl die oorspronklike Engelse betekenis juis inhou dat die groep nog nie tot ŉ gevolgtrekking oor ŉ spesifieke saak gekom het nie.

2.3.3 Verlede tyd

Afrikaanse verlede tyd word normaalweg deur die invoeging van die "het ge-"-konstruksie gevorm. Alhoewel "het" redelik sistematies na die naamwoordstuk gebruik word en ge- gewoonlik as prefiks aan die werkwoord gevoeg word, is daar volgens Carstens (2004: 88-93) en die Afrikaanse Woordelys en Spelreëls (Taalkommissie, 2002: 188-192) die onderstaande vier uitsonderings op die toevoeging van "ge-" by die werkwoord.

- Die imperfekvorm word as aanduiding van die verlede tyd gebruik (bv. "dink – dog/dag").
- Woorde wat met "be-", "er-", "ge-", "her-", "mis-", "ont-", "ver-" en "weer-" begin, kry gewoonlik nie ŉ "ge-" vooraan nie (bv. "Hy **het** gister **be**gin werk").
- Woorde waarvan die klem op die tweede (of verdere) lettergreep val, word ook sonder "ge-" in die verlede tyd geskryf (bv. "Sy **het** haar pa behoorlik **aanbid**").
- Woorde wat op "-eer" eindig, word soms met of sonder ŉ "ge-" geskryf (bv. "Sy **het probeer** leer", maar "Sy **het** daaroor **geargumenteer**").

Biber *et al.* (2002: 116-117) identifiseer ses basiese patrone om die Engelse verlede tyd te vorm. Hierdie patrone word vervolgens genoem en beskryf.

- Daar word ŉ -t-suffiks aan die einde van die werkwoord geplaas en mag ŉ -d of -t aan die einde van die basisvorm vervang. Woorde wat hier as voorbeeld kan dien sluit *"send – sent*" en "*spoil – spoilt*" in.
- Die basisvorm kan ook ŉ -t of -d-suffiks aan die einde neem, plus die vokaal in die basisvorm verander (bv. "*sell – sold*" en "*think – thought*").
- Die werkwoord kan ŉ -ed-suffiks neem (bv. "*show – showed*" en "*laugh – laughed*").
- Die vokaal in die basisvorm verander (bv. "give – gave" en "know – knew").
- Daar is ook werkwoorde wat geen verandering ondergaan nie (bv. "*cut*" en "*hit*").
- Die verledetydsvorm kan ook heeltemal verskil van die basisvorm (bv. "*go – went*").

Verskeie foute uit hierdie kategorie het in die *Autshumato*-afvoer voorgekom. Sommige sinne het te veel verledetydsmerkers gehad (sien V9). Die "het" of "ge-" is ook meermale uitgelaat en soms op die verkeerde plek ingevoeg (vgl. V10).

(V9.1) Engels: "*It happened in 2003*."

(V9.2) MV-afvoer: "Dit **was** gebeur **het** in 2003."

(V9.3) Korrekte Afrikaans: "Dit het in 2003 gebeur."

(V10.1) Engels: "*He **was winning** at the Games*."

(V10.2) MV-afvoer: "Hy **was gewen** by die Spele."

(V10.3) Korrekte Afrikaans: "Hy **was besig om te wen** by die Spele."

Net soos in die vorige kategorie kan hierdie foute ook aan die grootte van die afrigtingskorpus en die onvermoë om al die verbuigings en kontekste te bevat, toegeskryf word. Omdat Engels geredelik van verbuigings van die werkwoorde gebruik maak om die verlede tyd aan te dui (bv. "*winn**ing***"), word die belyning bemoeilik omdat Engels nie altyd ekstra woorde invoeg om die verlede tyd aan te dui soos in Afrikaans nie. Dit beteken dat sogenaamde een-tot-baie belynings getrek word waar een Engelse woord na meer as een Afrikaanse woord vertaal word. In V10 moet "*winning"* byvoorbeeld met "besig om te wen" belyn word. Sulke belynings is moeilik om outomaties te maak.

2.3.4 Werkwoorde

Soos genoem onder 2.3.1, is die werkwoord ŉ element wat vir baie variasie in Afrikaanse woordvolgorde verantwoordelik is. Wat die literatuur betref, gee Du Plessis (1985: 19-28) en Ponelis *et al.* (1972: 122-127) die volgende beginsels vir hulpwerkwoorde in Afrikaans:

- Hulpwerkwoorde gaan gewoonlik die skakelwerkwoorde vooraf (bv. "Hulle **sal bly** luister").

- Wanner die hulpwerkwoord "het" saam met ŉ ander hulpwerkwoord gebruik word, skuif die "het" na die posisie na die hoofwerkwoord (bv. "Ons **kon** gewonder **het**").

- Modale hulpwerkwoorde vereis nie ge- vooraan die werkwoord nie.
- Saam met die modale hulpwerkwoorde "behoort" en "hoef" word "te" altyd ingevoeg (bv. "Jy **hoef** nie **te** gaan nie").
- Die negatief word vooruitgegaan deur "hulle" (bv. "Hulle **sal nie** gaan nie.").

Vir Engels gee Biber *et al.* (2002: 174-185) die volgende riglyne by die gebruik van modale hulpwerkwoorde. Dit is belangrik om hier te noem dat die skrywers van hierdie teks nie soos vir Afrikaans tussen hulpwerkwoord van tyd ("het") en modale hulpwerkwoorde (bv. "kon" en "sou") onderskei nie, maar eerder albei in een klas groepeer:

- Die vorm van die hulpwerkwoord of die hoofwerkwoord daarna, word nie aangepas om die verlede tyd of meervoude aan te dui nie (bv. "*He can go"*en *"They can all go"*).
- Nes vir Afrikaans gaan die woord "*they*" die negatief vooruit (bv. "*They **shall not** go*").
- Die hulpwerkwoord staan meestal voor die hoofwerkwoord en net in uitsonderlike gevalle daarna (bv. "*I **have to** get up*" en "*It **must have** fallen out*").

In die afvoer van die basislynsisteem kom veral foute voor wat te make het met die hulp- en koppelwerkwoorde. V11 wys een geval waar die hulpwerkwoorde in Engels direk na mekaar en direk voor die hoofwerkwoord voorkom, maar in Afrikaans (sien V11.2) skei die hoofwerkwoord die verskillende hulpwerkwoorde:

(V11.1) Engels: "*We **could have** wondered."*

(V11.2) Korrekte Afrikaans: "Ons **kon** gewonder **het**."

Hoofwerkwoorde ontbreek ook gereeld in die afvoer en dit is veral die werkwoorde wat aan die einde van ŉ vertaalde sin moet staan, wat ontbreek (vgl. V12). Die rede hiervoor kan moontlik aan die taalmodelle en die manier waarop dit die dekodeerder se keuses beïnvloed, toegeskryf word. Die taalmodelle ken ŉ hoër waarskynlikheid aan sinne van eenderse lengte toe. As die dekodeerder daarom ŉ sin uit die verskillende moontlike vertalings op grond van hierdie waarskynlikheidsaanduiding moet kies, word sinne met minder (hulp)werkwoorde na die einde van die sin ŉ nader ooreenstemming met die lengte van die Engelse sin hê en dus bo die (meer korrekte) langer vertaling gekies word.

(V12.1) Engels: "*This would enable Zimbabwe **to get** more European help*."

(V12.2) MV-afvoer: "Dit sou Zimbabwe in staat stel om meer Europese hulp."

(V12.3) Korrekte Afrikaans: "Dit sou Zimbabwe in staat stel om meer Europese hulp **te kry**."

Skeibare werkwoorde word ook moeilik vertaal vanweë die beperkings op die grootte van die korpus. Net soos in die geval van verbuigings van woorde om die verlede tyd aan te dui (sien 2.3.3), kan die frasetabelle wat met relatief min data onttrek is, nie al die verskillende vorme van die skeibare werkwoorde be-

vat nie. Die sisteem gebruik dus die vorm van die werkwoord wat die mees frekwente voorkom. Dit lei dikwels tot ongrammatikale sinskonstruksies soos wat in V13 gesien kan word.

(V13.1) Engels: "*The commision does not propose that this continues.*"

(V13.2) MV-afvoer: "Die kommissie voorstel nie dat dit aangaan."

(V13.3) Korrekte Afrikaans: "Die kommissie **stel nie voor** dat dit aangaan nie."

2.3.5 Ander foute

Benewens die foute wat bo genoem is, kom daar ook nog ander foute in die *Autshumato*-afvoer voor. Die foute in hierdie kategorie sal nie met reëls uitgeskakel kan word nie. Hierdie foute het eerder te doen met tekortkominge in die volledigheid van die frasetabelle en nie met reëlmatige verskille tussen die twee tale nie. Herrangskikkingsreëls sal dus nie soos in die vorige vier kategorieë gebruik kan word om die probleme op te los nie. Die voorprosesseringsmodule poog nie om hierdie foute te voorkom nie en dit word dus net kortliks hier bespreek.

Samestellings word in Afrikaans as een woord geskryf. Die Engelse "*coffee machine*" vertaal dus in Afrikaans na "koffiemasjien" (sien V14 en V15). Omdat samestelling so ŉ produktiewe morfologiese proses in Afrikaans is en nuwe samestellings gereeld voorkom (Pilon *et al*., 2008), is dit baie moeilik om voorbeelde van al die moontlikhede in die afrigtingsdata in te sluit. Dit is egter heel algemeen dat die verskillende woorde waaruit ŉ samestelling bestaan, wel vertaal kan word. So byvoorbeeld kom "*coffee*" en "*machine*" albei in ander kontekste in die afrigtingsdata voor en word dan as "koffie" en "masjien" vertaal. Afrikaanse spelreëls vereis egter dat dit as een woord geskryf word.

(V14.1) Engels: "*Our office will receive a new coffee machine today*."

(V14.2) MV-afvoer: "Ons kantoor kry vandag ŉ nuwe **koffie masjien**."

(V14.3) Korrekte Afrikaans: "Ons kantoor kry vandag ŉ nuwe **koffiemasjien**."

(V15.1) Engels: "*I have already talked about the price crisis*."

(V15.2) MV-afvoer: "Ek het reeds gepraat oor die **prys krisis**."

(V15.3) Korrekte Afrikaans: "Ek het reeds gepraat oor die **pryskrisis**."

Meervoude, verkleining, asook die attributiewe -e ontbreek gereeld of word verkeerd toegepas (vgl. V16 en V17). Hierdie foute het te make met die feit dat die afrigtingsdata nie voorbeelde van alle woorde en al hul fleksievorme bevat nie. Die SMV-sisteem kies dus ŉ woord uit die frasetabel wat die hoogste waarskynlikheid het om in die spesifieke konteks voor te kom. As die (korrekte) verbuiging van die woord nie in die frasetabel voorkom nie, kan dit nie gebruik word nie.

(V16.1) Engels: "*The commision's aims are unmoving*."

(V16.2) MV-afvoer: "Die kommissie se doelwit is **onbetwisbare**."

(V16.3) Korrekte Afrikaans: "Die kommissie se doelwit is **onbetwisbaar**."

(V17.1) Engels: "*All other issues may be discussed*."

(V17.2) MV-afvoer: "Alle ander **kwessie** kan bespreek word."

(V17.3) Korrekte Afrikaans: "Alle ander **kwessies** kan bespreek word."

Woorde wat nie vertaal kan word nie (weereens omdat dit nie in die frasetabel is nie) verswak ook die afvoer van die *Autshumato*-sisteem aansienlik (vgl. V18 en V19). Uit die sinne wat vertaal is, het 24% een of meer Engelse woorde bevat. Sommige idiomatiese uitdrukkings word ook direk (woord vir woord) vertaal en anglisismes kom gereeld voor (vgl. V20).

(V18.1) Engels: "*These debacles seem unneccesary.*"

(V18.2) MV-afvoer: "Hierdie **debacles** lyk sinneloos."

(V18.3) Korrekte Afrikaans: "Hierdie fiasko's lyk onnodig."

(V19.1) Engels: *"In West Africa it is an even bigger problem*."

(V19.2) MV-afvoer: "In **West Africa** is dit selfs 'n groter probleem."

(V19.3) Korrekte Afrikaans: "In **Wes-Afrika** is dit selfs 'n groter probleem."

(V20.1) Engels: "*Come on in if it suits you*."

(V20.2) MV-afvoer: "**Kom aan in** as dit jou pas."

(V20.3) Korrekte Afrikaans: "**Kom binne** as dit jou pas."

Die verskillende foute wat in 2.3 bespreek is, blyk almal dieselfde oorsprong te hê – 'n tekort aan hoë kwaliteit afrigtingsdata. Ongelukkig is data-insameling 'n duur en tydrowende proses wat verder vererger word wanneer een van die tale 'n hulpbronskaars taal is. In 2.3.1 tot 2.3.4 word vier foutkategorieë uitgelig wat wel patroonmatige afwykings toon en dus met herrangskikkingsreëls opgelos kan word. Die laaste foutkategorie wat in 2.3.5 bespreek word, bevat egter 'n aantal voorbeelde wat nie in hierdie studie hanteer kan word nie.

2.4 Samevatting

In hierdie hoofstuk is statistiese masjienvertaling eerstens bespreek en die proses wat met die afrigting van ŉ SMV-sisteem met die *Moses*-pakket gepaard gaan, is in fyner besonderhede uiteengesit. Hierdie proses bestaan uit ses fases. In die eerste fase moet die afrigtingsdata voorberei word deur leë lyne te verwyder en die sinne in die brontaalteks met die sinne in die teikentaalteks te belyn. Taalmodelle word volgende opgestel om die struktuur van Afrikaans met behulp van 3- 4- en 5-gramme te modelleer voordat die afrigtingsdata in die derde fase op woordvlak belyn word. In die vierde fase word frasetabelle opgestel wat ooreenstemmende frases in beide Engels en Afrikaans bevat. In die laaste twee fases word ŉ herrangskikkingsmodel en genereringsmodel opgestel wat uiteindelik gebruik word om ŉ moontlike vertaling te genereer.

Die effektiwiteit van die SMV-algoritme word deur die kwaliteit en kwantiteit van die afrigtingsdata bepaal. Die dekodeerder (die module wat uiteindelik al die modelle gebruik om die vertalings te lewer) gebruik net inligting uit die modelle en frasetabelle wat reeds tydens die voorbereiding van die algoritme onttrek word om die mees waarskynlike vertaling te genereer. Aangesien daar nie soveel afrigtingsdata vir die taalpaar Engels-Afrikaans beskikbaar is as vir ander suksesvolle internasionale projekte nie, kom foute in die struktuur van die vertaalde sinne meer gereeld voor. In 2.3 van hierdie hoofstuk is vyf foutkategorieë bespreek. Dit sluit foute in die woordvolgorde (2.3.1), ontkenning (2.3.2), verlede tyd (2.3.3), werkwoorde (2.3.4) en ŉ algemene kategorie (2.3.5) in. Veral die foute in (2.3.5) wys daarop dat die sisteem nog nie met genoegsame data afgerig is om goeie woordkeuses te maak nie. Sommige woorde word eenvoudig glad nie vertaal nie. Die strukture in die bron- en teikentaal wat konsekwent van mekaar verskil, kan tydens voorprosessering herrangskik word om die invloed daarvan op die kwaliteit van die vertaling te beperk.

In die volgende hoofstuk sal daar op die voorgestelde voorprosesseringsmodule gefokus word en sal aangetoon word hoe die analise van die data in 2.3 en die patroonmatighede wat hier uitgewys is, aanleiding tot herrangskikkingsreëls gegee het. Die reëls wat in die volgende hoofstuk bespreek word, is dus ŉ direkte poging om die verskille tussen die twee tale te minimaliseer om sodoende ŉ beter vertaling te lewer. Hoofstuk 3 sal die ontwerp en implementering van die voorprosesseringsmodule uiteensit. Die ontwikkeling van die herrangskikkingsreëls en voorbeelde van die toepassing daarvan sal bespreek word en daar sal ook aangetoon word hoe die voorprosesseringsmodule by die res van die afrigtingsprosedure inskakel.

Hoofstuk 3: Sintaktiese herrangskikking as voorprosesseringsmodule

3.1 Inleiding

In die vorige hoofstuk is die gebreke in die afvoer van die *Autshumato*-sisteem in verskillende kategorieë verdeel en geanaliseer om sodoende 'n beter begrip van die kwaliteit van die afvoer te kry. Hierdie analise het getoon dat die basislynsisteem nog heelwat verbeter behoort te word voordat dit vir die vertalers in die verskillende regeringskantore 'n onontbeerlike hulpmiddel sal wees. Die foutgroepe wat in 2.3 aangedui is, sluit woordvolgorde, ontkenning, verlede tyd en werkwoorde in. Die laaste kategorie wat in 2.3.5 onderskei is, noem 'n aantal foute wat met 'n tekort aan afrigtingsdata te make het en dus nie verder in hierdie navorsingsprojek aandag sal geniet nie.

Die doel van hierdie hoofstuk is om 'n moontlike oplossing in die vorm van 'n voorverwerkingsmodule vir die eerste vier foutkategorieë voor te stel. In die volgende afdeling sal die herrangskikkingsreëls wat in die module gebruik word, bespreek word. Daarna sal beskryf word hoe die module ontwikkel is en hoe dit in die *Moses*-omgewing inpas.

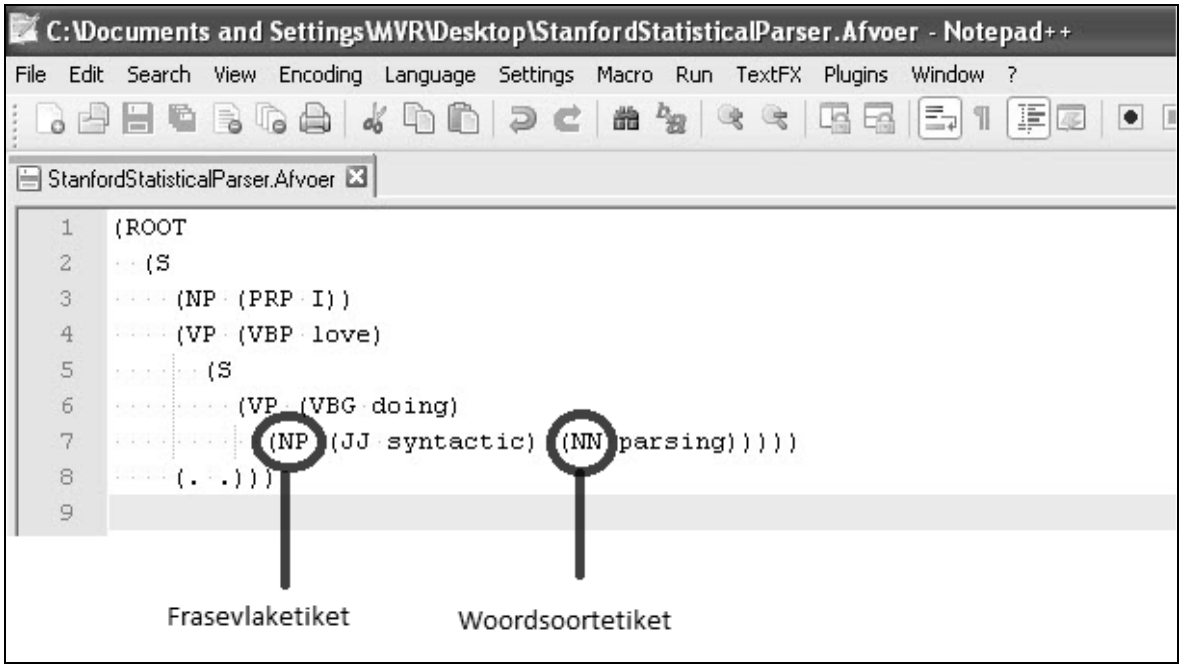
3.2 Linguisties gemotiveerde herrangskikkingsreëls

'n Analise van die afvoer van die basislynsisteem is in 2.3 beskryf en na aanleiding van die bevindinge van die genoemde analise is sintakties gemotiveerde reëls ontwikkel. Hierdie reëls is verkeie male op 'n kleiner ontwikkelingsteks van 50 sinne wat nie in die afrigtingsdata of METIS II-toetsteks voorkom nie, toegepas en geëvalueer. Die 300 sinne waaruit die foutkategorieë afgelei is, is ook gebruik om die reëls intyds te evalueer deurdat die sinne gereeld herrangskik is en die afvoer van die module is dan handmatig nagegaan om verfynings of veranderings in die reëls te maak. Eers nadat die reëls korrek op hierdie sinne toegepas kon word, is die groter afrigtingsdatastel daarmee herrangskik en 'n nuwe SMV-sisteem met die *Moses*-SMV-gereedskapstel afgerig.

Die reëls wat in hierdie module geïmplementeer is, kan beskou word as linguisties gemotiveerd omdat die verskuiwings wat aan die brontaalstruktuur gemaak word op die sintaktiese beginsels berus wat in 2.3 bespreek is. Die reëls is taalspesifiek, aangesien die Engelse struktuur na gelang van Afrikaanse struktuur herrangskik word om sodoende die verskille tussen die brontaal en teikentaal van die *Autshumato*-sisteem te minimaliseer.

Om die patrone wat in die vorige hoofstuk bespreek is (in besonder 2.3) op te spoor, is dit nodig om die data met sintaktiese inligting te annoteer. Die doel van die herrangskikkingsreëls is om frases in die Engelse brontaaldata te manipuleer, maar woordsoort-inligting is ook soms nodig om meer gedetailleerde patrone te kan beskryf (soos om byvoorbeeld 'n hulpwerkwoord in 'n ondergeskikte klous te herken). Die *Stanford Statistical Parser* (Klein & Manning, 2003) word gebruik om hierdie annotasie te doen en ken verskillende etikette aan die verskillende elemente toe (sien 3.4.1 vir meer besonderhede). Woordsoorte-

tikette word eerstens aan elkeen van die woorde in die sin toegeken en die verskillende woorde word dan in frases gegroepeer. Hierdie frases word dan met frasevlaketikette gemerk.[7] 'n Voorbeeld van die afvoer word in Figuur 3 gegee. Die oorspronklike sin ("*I love doing syntactic parsing.*") is met etikette op woordvlak en frasevlak geannoteer.



Figuur 3: Afvoer van die *Stanford Statistical Parser*

Die herrangskikkingsreëls kan in vyf kategorieë verdeel word na gelang van die hoofkomponent wat verskuif word: werkwoordherrangskikking (twee reëls), konstruksies met "to" (vier reëls), modale herrangskikking (twee reëls), asook reëls wat te doen het met ontkenning (een reël) en die verlede tyd (drie reëls). Elke reël in 'n kategorie is spesifiek ontwerp om 'n bepaalde tipe frase te hanteer en word vervolgens beskryf voordat voorbeelde van die toepassing daarvan gegee word.

3.2.1 Werkwoordherrangskikking

Hierdie twee reëls volg uit die feit dat die werkwoordstuk in Afrikaans na die einde van die bysin skuif (vergelyk 2.3.1 en 2.3.4) en poog dus om foute met betrekking tot die werkwoordvolgorde te verminder.

(R1.1) As {PP} {SBAR} {VP} {NP} dan {PP} {SBAR} {NP} {VP}

Die werkwoordstuk (VP) onmiddellik voor 'n naamwoordstuk (NP) skuif na die einde van die frase as die werkwoordstuk en naamwoordstuk deur 'n voorsetselstuk (PP) en relatiewe of ondergeskikte klous (SBAR) voorafgegaan word.

[7] 'n Lys van hierdie etikette kan in die *Moses*-handleiding (Koehn, 2010) gevind word.

(R1.2) As {PP} {SBAR} {VP} {PP} dan {PP} {SBAR} {PP} {VP}

Die werkwoordstuk (VP) onmiddellik voor 'n voorsetselstuk (PP) skuif na die einde van die frase as die werkwoordstuk en voorsetselstuk deur 'n ander voorsetselstuk (PP) en relatiewe of ondergeskikte klous (SBAR) voorafgegaan word.

3.2.2 Konstruksies met "to"

Die reëls in hierdie afdeling het ten doel om die "om te"-konstruksie in Afrikaans in die Engelse struktuur na te boots (vergelyk 2.3.1 en 2.3.4 in hierdie verband).

(R2.1) As {VP{TO to}} {VP{VB}} {ADJP} dan {VP{TO to}} {ADJP} {VP{VB}}

'n Adjektiefstuk (ADJP) skuif na die posisie tussen "to" (TO) en die hoofwerkwoord (VB). Hierdie verskuiwing plaas die adjektiefstuk binne die grense van die werkwoordstuk.

(R2.2) As {VP{TO to}}{VP{VB}} {ADVP} dan {VP{TO to}} {ADVP}{VP{VB}}

'n Bywoordstuk (ADVP) skuif na die posisie tussen "to" (TO) en die hoofwerkwoord (VB). Hierdie verskuiwing plaas die bywoordstuk binne die grense van die werkwoordstuk.

(R2.3) As {VP{TO to}}{VP{VB}} {NP} dan {VP{TO to}} {NP}{VP{VB}}

'n Naamwoordstuk (NP) skuif na die posisie tussen "to" (TO) en die hoofwerkwoord (VB). Hierdie verskuiwing plaas die naamwoordstuk binne die grense van die werkwoordstuk.

(R2.4) As {VP{TO to}}{VP{VB}} {PP} dan {VP{TO to}} {PP} {VP{VB}}

'n Voorsetselstuk (PP) skuif na die posisie tussen "to" (TO) en die hoofwerkwoord (VB). Hierdie verskuiwing mag die voorsetselstuk binne die grense van die werkwoordstuk plaas.

3.2.3 Modale herrangskikking

Modale hulpwerkwoorde word dikwels in Afrikaans van die hoofwerkwoord geskei. In 2.3.4 word die linguistiese onderbou vir hierdie groep reëls gegee.

(R3.1) As {VP {MD}} {VP{VB}{PP} dan {VP {MD}} {PP} {VP{VB}}

Die voorsetselstuk (PP) skuif na die posisie tussen die modale hulpwerkwoord (MD) en die werkwoordstuk met die hoofwerkwoord (VP{VB}).

(R3.2) As {VP {MD}} {VP{VB}{NP} dan {VP {MD}} {NP} {VP{VB}}

Die naamwoordstuk (NP) skuif na die posisie tussen die modale hulpwerkwoord (MD) en die werkwoordstuk met die hoofwerkwoord (VP{VB}).

3.2.4 Ontkenning

Die belangrikste verskil tussen Engels en Afrikaans (aangedui in 2.3.2) is dat die Afrikaanse konstruksie gereeld ŉ tweede negativeringselement neem.

(R4.1) As {RB not}{. .} dan {RB not} {NIE}{. .}

ŉ Merker wat nie deel is van die *Stanford Statistical Parser* se etiketstel nie ({NIE}), word aan die einde van die sin geplaas om die tweede "nie" in te voeg. Hierdie merker word nie saam met die ander etikette aan die einde van die voorprosessering verwyder nie, en dien as ŉ plekhouer vir die ontkenningswoord. Omdat die sinne ook nou meer ooreenstem wat lengte betref (gemeet in die aantal woorde in beide die Engelse en Afrikaanse sinne), behoort die belyningsproses ook vergemaklik te word.

3.2.5 Verlede tyd

Daar is verskillende maniere om die verlede tyd in Engels aan te dui, maar "het" word in die meeste gevalle in verledetydssinne in Afrikaans toegevoeg. Die verskillende aspekte van hierdie konstruksie word ook in 2.3.3 verder uiteengesit.

(R5.1) As {VB has/have/had} dan {VB has/have/had}{HET}{. .}

ŉ Merker wat nie deel is van die *Stanford Statistical Parser* se etiketstel nie ({HET}), word weereens ingevoeg om "het" in die sin te plaas as "*has*", "*have*" of "*had*" teenwoordig is. Engelse sintaksis vereis soms dat die vorm van die werkwoord verander om die verlede tyd aan te dui, byvoorbeeld "*sit*" word "*sat*" en "*preach*" word "*preached*" (sien 2.3.3). Die frase-etikette wat die *Stanford Statistical Parser* toeken, gee ongelukkig nie genoeg besonderhede om die ander vorme na te gaan nie en verdere morfologiese analise (in die vorm van ŉ lemma-identifiseerder) sou hier nodig wees om die Engelse woord weer na die teenwoordige tyd om te skakel. Verdere navorsing is nodig om die totale omvang van die probleem rakende die verlede tyd te kan oplos.

3.3 Voorbeelde van die toepassing van die reëls

In hierdie afdeling word voorbeelde uit die 300 sinne waarop die reëls getoets is, gegee. Dit het ten doel om die werking daarvan te illustreer. Vir elkeen van die voorbeelde word die oorspronklike sin met die sintaktiese annotasie eers gegee en daarna word die reëls wat op die sin toegepas is, gelys. Laastens word die herrangskikte sin met ŉ woord-vir-woord vertaling in Afrikaans gegee.

Voorbeeld 1: Werkwoordherrangskikking

Oorspronklike sin: *I visited two of these houses that I can recommend to anyone.*

Reëls toegepas: Werkwoordherrangskikking:

> Die werkwoordstuk (*can recommend*) onmiddellik na ŉ naamwoordstuk (*I*) skuif na die einde van die frase waar beide deur ŉ relatiewe of ondergeskikte klous (*that I can recommend to anyone*) gedomineer word.

Herrangskikte sin: *I visited two of these houses that I to anyone can recommend.*

Direkte Afrikaanse vertaling: Ek het twee van hierdie huise besoek wat ek vir enigeen kan aanbeveel.

Die werkwoord staan nou nader aan die einde van die Engelse sin en stem daarom ooreen met die posisie van die werkwoord in die voorgestelde Afrikaanse vertaling. Op hierdie manier behoort die dekodeerder nie meer 'n inkorrekte sin waarvan die werkwoord uitgelaat is, op grond van die woordbelynings te verkies nie (sien 2.3.4).

Voorbeeld 2: Konstruksies met "*to*"

Oorspronklike sin: *I call upon Bulgaria to comply with our request.*

Reëls toegepas: Konstruksies met "*to*":

'n Naamwoordstuk (*our request*) skuif na die posisie tussen "*to*" en die hoofwerkwoord (*comply*).

'n Voorsetselstuk (*with*) skuif na die posisie tussen "*to*" en die hoofwerkwoord (*comply*).

Herrangskikte sin: *I call upon Bulgaria to with our request comply.*

Direkte Afrikaanse vertaling: Ek doen 'n beroep op Bulgarye om aan ons versoek voldoen.

Die woordvolgorde in die Engelse sin is gemanipuleer om karakteristieke van die Afrikaanse woordvolgorde te toon. Die verskuiwing beteken dat die belyningsproses gladder sal verloop en dat die dekodeerder nie net op patrone wat in die taalmodelle teenwoordig is, staat maak nie (sien 2.3.1).

Voorbeeld 3: Modale herrangskikking

Oorspronklike sin: *Cyprus will be a sort of bridge with the countries in the area.*

Reëls toegepas: Modale herrangskikking

Die voorsetselstuk (*with the countries in the area*) skuif na die posisie tussen die modale hulpwerkwoord (*will*) en die werkwoordstuk met die hoofwerkwoord (*be*).

Die naamwoordstuk (*a sort of bridge)* skuif ook tussen die modale hulpwerkwoord en hoofwerkwoord in.

Herrangskikte sin: *Cyprus will a sort of bridge with the countries in the area be.*

Direkte Afrikaanse vertaling: Siprus sal 'n soort van brug met die lande in die area wees.

Die hulpwerkwoord se posisie is nou eenders in beide die Engelse en Afrikaanse sinne. Aangesien die werkwoord nou aan die einde van die sin staan, behoort dit ook nie verkeerdelik uitgelaat te word nie.

Voorbeeld 4: Ontkenning

Oorspronklike sin: *There are not, therefore, any amendments to the agenda for Friday.*

Reëls toegepas: Ontkenning

Herrangskikte sin: *There are not, therefore, any amendments to the agenda for Friday nie.*

Direkte Afrikaanse vertaling: Daar is nie, daarom, enige wysigings tot die agenda vir Vrydag nie.

'n Merker word ingevoeg om die ontkenning aan te dui. Hierdie merker dien as aanduiding dat 'n tweede ontkenningswoord tydens vertaling teenwoordig behoort te wees.

Voorbeeld 5: Verlede tyd

Oorspronklike sin: *We would not have any descendants.*

Reëls toegepas: Verlede tyd

Herrangskikte sin: *We would not any descendants have het nie.*

Direkte Afrikaanse vertaling: Ons sou nie enige afstammelinge gehad het nie.

'n Merker word ingevoeg om die verlede tyd aan te dui. Dit sorg dat die belynings makliker gemaak kan word en die sinslengtes meer ooreenstem.

3.4 Argitektuur van die voorprosesseringsmodule

Die module wat die linguisties gemotiveerde herrangskikkingsreëls op die data toepas, volg 'n aantal stappe om een sin op 'n slag te herrangskik totdat die totale teks gemanipuleer is. Dié teks word dan na die afrigtingsalgoritme of dekodeerder gestuur. In die res van hierdie afdeling sal die stappe wat op elke sin toegepas word, bespreek word, waarna 'n skematiese oorsig oor die uiteindelike proses in 3.4.4 gegee sal word.

3.4.1 Normalisering

Die eerste stap in die basiese voorprosesseringsmodule skakel enklitiese vorme soos *"don't"* en *"I'll"* weer na hul oorspronklike vorme (*"do not"* en *"I will"*) om. Hierdie normalisering word vir elke sin in die Engelse kant van die afrigtingsdata herhaal en ook vir elke sin in die Engelse teks wat vertaal moet word. Die stap is nodig om te verseker dat die frasetabel nie verkeerde belynings bevat nie, en dat twee woorde in Engels weer met twee ooreenstemmende woorde in Afrikaans belyn word. Sonder hierdie stap sou die frasetabel baie maklik "*don't"* met "moet" kon verbind, en nie met "moet nie". Tabel 2 bevat 'n lys van die mees frekwente woorde wat op hierdie manier genormaliseer is.

Oorspronklike woord	Genormaliseerde vorm
it's	*it is*
let's	*let us*
x'll	*x will*
won't	*will not*
don't	*do not*
I'm	*I am*
they're	*they are*
x'd	*x would*

Tabel 2: Enklitiese vorme wat genormaliseer word

3.4.2 Sintaktiese analise

Aangesien Engels 'n sogenaamde hulpbronryke taal is, bestaan daar heelwat goeie sintaktiese analiseerders vir die taal. Vir hierdie projek was akkuraatheid die grootste oorweging by die keuse van 'n sintaktiese analiseerder, aangesien die foute wat die analiseerder maak, noodwendig foute met die toepassing van die herrangskikkingsreëls tot gevolg sal hê (Collins *et al*., 2005).

Die *Stanford Statistical Parser* (Klein & Manning, 2003) lewer baie goeie resultate en 'n geweegde F-telling van 86 % word gerapporteer[8] (Klein & Manning, 2003). Dit is ook maklik om te implementeer en is redelik stabiel, aangesien min probleme met die toepassing van die sagteware voorkom (Wilcock, 2009: 51-52). Die *Stanford Statistical Parser* is met geannoteerde data uit die *Penn Treebank*-korpus afgerig (Mitchell *et al*., 1994). Hierdie korpus is handmatig deur linguiste geannoteer en dien dus as die goue standaard waarteen outomatiese analiseerders getoets kan word, of, soos in hierdie geval, waarmee analiseerders afgerig kan word. Die *Penn Treebank*-korpus bestaan uit geannoteerde weergawes van die *Wall Street Journal* (sake- en finansiële nuus met ongeveer 25 miljoen woorde) en die *Brown*-korpus ('n versameling aktuele Engelse tekste wat uit 1 miljoen woorde bestaan) (Mitchell *et al*., 1994).

Die *Stanford Statistical Parser* beskik oor modules wat 'n teks outomaties in sinne en tekseenhede verdeel (*sentenciser* en *tokeniser*) en geen voorprosessering in hierdie verband is dus nodig nie. Die gebruiker hoef slegs die data wat sintakties geannoteer moet word, in 'n tekslêer te gee. Die tweede stap in die voorprosesseringsmodule stuur die sin uit die afrigtingsdata of data wat vertaal moet word, na die sintaktiese analiseerder. Dié analise vorm die grondslag vir die herrangskikkingsreëls wat volgende toegepas word.

[8] Sien 4.2.1.3 vir 'n beskrywing van hierdie evaluasiemetriek.

3.4.3 Toepassing van die herrangskikkingsreëls

Die reëls wat in 3.2 beskryf is, bestaan uit patrone wat in die Engelse sintaktiese strukture herken word en dan verander word om meer na Afrikaanse sintaktiese strukture te lyk. Die voorprosesseringsmodule herken dus patrone in die etikette wat deur die *Stanford Statistical Parser* toegeken is en in die linkerkant van die herrangskikkingsreëls voorgestel word, en ruil dan die frases om sodat dit aan die patroon aan die regterkant van die reëls voldoen. Voorwaardestellings (sg. *if-then*-stellings) in die programmeringstaal Perl (ActiveState, 2005) word gebruik om die patroonherkenning te doen.

Nadat die sin herrangskik is, word al die ekstra annotasie wat deur die sintaktiese analise toegevoeg is, verwyder en die geprosesseerde sin word in die oorspronklike korpus teruggesit sodat dit steeds met die Afrikaanse ekwivalent belyn is (in die geval van die afrigtingsdata). Die proses word vir elkeen van die sinne in die Engelse afdeling van die afrigtingskorpus of teks wat vertaal moet word, herhaal totdat die hele afdeling herrangskik is.

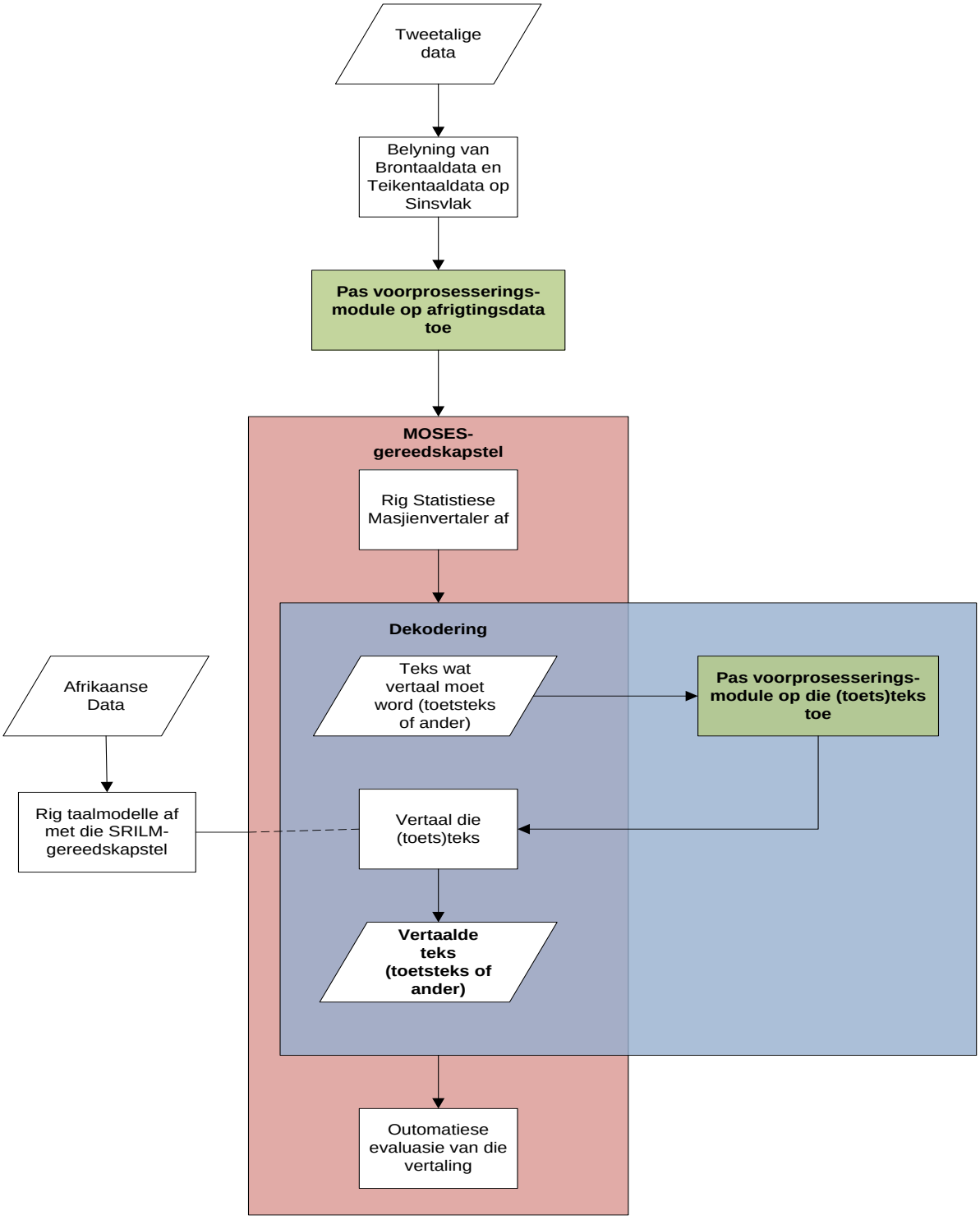
3.5 Skematiese oorsig oor die uitgebreide sisteem

Om die toepassing van die herrangskikkingsreëls so effektief moontlik te maak, is dit nodig om die ontwerp van die totale module waarin dit funksioneer goed te deurdink. Dit is belangrik dat die module moeiteloos in die *Moses*-omgewing inskakel. Twee ekstra stappe word tot die proses gevoeg, nl. ’n stap om die afrigtingsdata volgens die herrangskikkingsreëls te orden en ook om die data wat vertaal moet word met dieselfde module te prosesseer. Dieselfde module word dus twee maal toegepas – as voorprosessering van die afrigtingsdata en later as voorprosessering van die data wat vertaal moet word.

’n Skematiese voorstelling van die voorgestelde uitgebreide sisteem kan op die volgende bladsy in Figuur 4 gesien word[9]. Nadat die Engelse en Afrikaanse data met mekaar belyn is, word die voorprosesseringsmodule op die Engelse data toegepas. Die Engelse data word dus herrangskik en dit resulteer in ’n nuwe parallelle korpus. Hierdie nuwe korpus word dan saam met die taalmodelle wat met die *SRILM*-gereedskapstel afgerig is na die *Moses*-gereedskapstel gestuur. Die SMV-sisteem kan nou afgerig word. Die voorprosesseringsmodule word verder op die toetsteks gebruik om dit ook te manipuleer voordat dit met die SMV-sisteem vertaal en die kwaliteit van die afvoer outomaties geëvalueer word (sien Hoofstuk 4).

Die afrigtingsproses volg dus al die stappe wat in die diagram aangedui word en die stappe wat in die rooi blok is, word binne die *Moses*-gereedskapstel gedoen. Wanneer die resulterende sisteem gebruik word om ’n teks te vertaal, word slegs die deel van die proses wat met ’n blou blok gemerk is, gevolg. Die gebruiker voer dus ’n Engelse teks tot die sisteem toe waarna die voorprosesseringsmodule daarop toegepas word. Die gemanipuleerde teks word nou na die dekodeerder gestuur om ’n Afrikaanse vertaling te genereer wat uiteindelik as afvoer na die gebruiker gestuur word.

[9]Sien 2.2.1.2 vir ’n volledige beskrywing van al die komponente in die SMV-proses wat hier net kortliks genoem word.



Figuur 4: Afrigting van / vertaling met die uitgebreide SMV-sisteem

3.6 Samevatting

Hierdie hoofstuk bespreek die ontwikkeling van ŉ voorprosesseringsmodule wat op die afrigtingsdata van ŉ SMV-sisteem toegepas kan word. Die module het ten doel om die verskille tussen die brontaalsintaksis – in hierdie geval Engels – en die teikentaalsintaksis (Afrikaans) te minimaliseer ten einde die SMV-sisteem in staat te stel om beter vertalings te genereer. Die voorprosesseringsmodule bestaan uit twee hoofkomponente, nl. ŉ proses wat die Engelse sin met sintaktiese inligting annoteer en ŉ daaropvolgende stel reëls wat die sin herrangskik. Die sintaktiese annotasie word gedoen met behulp van die *Stanford Statistical Parser* en die herrangskikking word gedoen met tien linguisties gemotiveerde reëls.

Die herrangskikkingsreëls is pasgemaak vir die toepassing en is dus taalafhanklik. Kennis van beide die bron- en teikentaal moes vooraf ingewin word ten einde die reëls te konseptualiseer en ŉ iteratiewe toepassing daarvan op die 300 sinne wat deurentyd as toetsteks gebruik is, was ook nodig. Die reëls is dan na elke iterasie verfyn om die patrone wat deur die sintaktiese annotasie na vore gekom het, beter te beskryf.

Die voorprosesseringsmodule is gebruik om die Engelse kant van die parallelle afrigtingskorpus te herrangskik en ŉ SMV-sisteem is met die herrangskikte data afgerig. Om die sisteem te evalueer, is die voorprosesseringsmodule op die METIS II-teks toegepas waarna dit deur die herrangskikte SMV-sisteem vertaal is. ŉ Volledige beskrywing van hierdie evaluasie word in Hoofstuk 4 gegee.

Hoofstuk 4: Evaluasie

4.1 Inleiding

Om die doeltreffendheid van die voorprosesseringsmodule, asook die kwaliteit van die afvoer van die uitgebreide SMV-sisteem te bepaal, moet beide hierdie elemente met evaluasiemetodes wat aan internasionale standaarde voldoen, gemeet word. Die evaluasie van die elemente behoort nie net aan die einde van die navorsingsprojek gedoen te word nie, maar moet deurgaans geskied om vordering te meet (Jurafsky & Martin, 2009: 931). Om hierdie rede moet die metrieke maklik wees om te gebruik en ŉ goeie maatstaf vir die kwaliteit van die spesifieke element wees. In hierdie navorsingsprojek word die BLEU- en NIST-metrieke gebruik om die kwaliteit van die afvoer van die uitgebreide SMV-sisteem met die kwaliteit van die afvoer van die basislynsisteem te vergelyk. Hierdie metrieke is erkende evaluasiemetodes wat ook in internasionale studies aangewend word. Die gebruik van hierdie metrieke verseker dus dat die resultate van hierdie studie met dié van die studies wat in Hoofstuk 1 bespreek word, vergelyk kan word.

Om te verseker dat die kwaliteit van die afvoer van twee SMV-sisteme met mekaar vergelyk kan word, moet die opstelling van die evaluasie dieselfde wees. Daar moet byvoorbeeld nie ander veranderinge gemaak word aan die sisteem as die verandering waarvan die navorser die invloed wil toets nie (Hirschman & Mani, 2003: 418-419). Aangesien dit binne hierdie studie gaan om die invloed van die voorprosesseringsmodule, is dit belangrik dat die afrigtingsdata van die basislyn- en die uitgebreide sisteem dieselfde moet bly. Geen afrigtingsdata wat nie vir die basislynsisteem gebruik is, kan tot die afrigtingsdata van die uitgebreide sisteem toegevoeg word nie.

Albei sisteme moet ook op dieselfde toetsdatastel geëvalueer word. Die METIS II-toetsteks (Dirix *et al.*, 2007) word deurentyd vir evaluasie gebruik en bestaan uit 200 sinne uit die EUROPARL-korpus (Koehn, 2005). Hierdie korpus bevat ŉ versameling transkripsies van debatte in die Europese Parlement. Aangesien die afrigtingsdata en die toetsdata uit dieselfde domein afkomstig is, sal die skryfstyl en woordkeuse ooreenstem en dus die gepastheid van die sisteem vir ŉ spesifieke taak toets (Cieri, 2007: 229). ŉ Verdere rede vir die geskiktheid van die teks, is die feit dat dit heelwat gevalle bevat waarop elkeen van die reëls toegepas kan word. In Tabel 3 word die hoeveelheid relevante patrone gegee, d.i. gevalle waarop die herrangskikkingsreëls (sien 3.2) van toepassing is. Om die aantal patrone te bepaal, is die hoeveelheid voorkomste daarvan in die toetsteks getel. Die totale hoeveelheid patrone wat herrangskik kan word, word in die laaste ry van die tabel gegee. Hierdie toetsteks sal gebruik word om die voorprosesseringsmodule in isolasie te evalueer, asook om die kwaliteit van die afvoer van die basislyn- en die uitgebreide sisteme te vergelyk.

Herrangskikkingsreël	Hoeveelheid relevante patrone in die toetsteks
R1.1 en R1.2: Werkwoordherrangskikking	24
R2.1 tot R2.4: Konstruksies met "to"	32
R3.1 en R3.2: Modale herrangskikking	76
R4.1: Negatiewe	16
R5.1: Verlede tyd	32
Totale hoeveelheid patrone in die toetsteks:	**180**

Tabel 3: Samestelling van die METIS II-toetsteks

Die opstelling van die eksperimente moet verder konstant wees en dus moet die *Moses*-SMV-gereedskapstel en *SRILM*-gereedskapstel op dieselfde manier en met dieselfde parameters gebruik word. Vir hierdie navorsingsprojek word die onderstaande parameters deurgaans gebruik:

- Die taalmodelle word opgestel om 3-, 4- en 5-gramme in te sluit.
- Die outomatiese herrangskikking (*distortion limit*) wat deur die dekodeerder gedoen word, word tot ses beperk. Dit beteken dat die dekodeerder woorde tot ses posisies links of regs mag skuif om by die woordvolgorde in die taalmodelle aan te pas. Hierdie is die verstekopsie en word in die meeste gevalle aanbeveel (Koehn, 2010: 26-27).
- Die belyning in die verskillende tabelle word deurgaans met die verstekmetode gedoen, d.w.s. die snypunt van twee stelle belynings word eers ingesluit en addisionele belynings in die res van die korpus word dan op grond van verskillende heuristieke bygevoeg (Koehn, 2010: 84) (sien 2.2.1.2 vir ŉ volledige beskrywing).

In die vorige hoofstuk is die ontwerp van die voorprosesseringsmodule uiteengesit en die rol daarvan binne die uitgebreide SMV-sisteem verduidelik. In hierdie hoofstuk word die implementering daarvan geëvalueer en die sukses van die voorprosessering bepaal. In die eerste afdeling van hierdie hoofstuk (4.2), word die voorprosesseringsmodule in isolasie geëvalueer om te bepaal of die reëls korrek toegepas word. Herroeping, presisie en F-telling word gebruik om die effektiwiteit van die voorprosesseringsmodule te bepaal. Die volgende stap is om die impak van die voorprosesseringsmodule op die kwaliteit van die afvoer van ŉ SMV-sisteem te toets. Die afvoer van die basislynsisteem sal dus met die afvoer van die uitgebreide sisteem vergelyk word deur na die BLEU- en NIST-metrieke te verwys. Hierdie metrieke word in 4.3 in meer besonderhede bespreek waarna die evaluasieresultate gelys en bespreek word.

4.2 Evaluasie van die voorprosesseringsmodule

Sogenaamde "*embedded-component evaluation*" behels die evaluasie van 'n komponent wat binne 'n groter raamwerk moet funksioneer en behoort gereeld tydens die ontwikkeling van die komponent gedoen te word (Hirschman & Mani, 2003: 418-419). Benewens die evaluasie van die afvoer van die SMV-sisteem, is dit ook belangrik om te verseker dat die voorprosesseringsmodule optimaal funksioneer; veral met betrekking tot die identifikasie van sinne wat herrangskik behoort te word en die uiteindelike verwerking daarvan. In hierdie afdeling word twee aspekte getoets, nl. die herroeping (*recall*) en presisie (*presicion*) van die toepassing van die herrangskikkingsreëls in die voorprosesseringsmodule (sien 4.2.1.1 en 4.2.1.2 onderskeidelik). 'n Derde metriek, die F-telling, word ook bereken om die eerste twee metrieke in een syfer saam te vat (sien 4.2.1.3). Hierdie drie metrieke sal vervolgens bespreek word, voordat die resultate van die evaluasie op die METIS II-toetsteks in 4.2.2 bespreek word.

4.2.1 Metrieke

4.2.1.1 Herroeping

Die formule waarvolgens herroeping in inligtingonttrekking (*information retrieval*) bereken word, meet die vermoë van 'n sisteem om die relevante voorbeelde uit 'n versameling te onttrek (Manning *et al*., 2008: 155).

$$\text{“ Herroeping} = \frac{\text{\# (relevante items onttrek)}}{\text{\# (relevante items)}} \text{”}$$

Die gebruiker verskaf kriteria, bv. soekterme, en die sisteem soek dan dié voorbeelde in die versameling waarin hierdie soekterme voorkom. Die metriek is ook aangepas vir ander evaluasiedoeleindes, byvoorbeeld om die herroeping van speltoetsers te meet. Hier word die speltoetser se vermoë om spelfoute te identifiseer, getoets en meet dus hoeveel van die spelfoute wat in 'n teks voorkom, as sodanig deur 'n speltoetser gemerk is. Herroeping meet in hierdie studie die module se vermoë om patrone in die sintaktiese etikette wat aan die herrangskikkingskriteria voldoen (die "soekterme" wat in die dokument moet voorkom) te identifiseer. Die module moet dus die relevante patrone in die afrigtingsdata of teks wat vertaal moet word, effektief kan opspoor.

Dit is belangrik om te onthou dat 'n sin meer as een maal herrangskik kan word, afhangend van die patrone wat daarin voorkom (sien 3.2 vir voorbeelde). Dit is dus nie net die sinne wat herrangskik word wat in berekening gebring moet word nie, maar eerder die hoeveelheid patrone in die teks as geheel wat herrangskik word. Die aangepaste formule vir herroeping sien dus soos volg daaruit:

$$\text{Herroeping} = \frac{\text{\# (relevante patrone wat herken word)}}{\text{\#(relevante patrone in die goue standaard)}}$$

4.2.1.2 Presisie

Inligtingsonttrekking gebruik presisie om te meet hoeveel van die onttrekte voorbeelde relevant is (Manning *et al*., 2008: 155) en dit is dus 'n maatstaf van die gerigtheid van die sisteem. As 'n gebruiker dus na voorbeelde wat die soekterme "bank", "rivier" en "watersport" bevat, soek, moet die sisteem nie ook voorbeelde oor die finansiële instelling onttrek nie, aangesien die gebruiker dan deur voorbeelde moet

lees wat nie van belang is nie. In inligtingsonttrekking gaan dit hier dus oor die feit dat die gebruiker nie met onnodige voorbeelde belas moet word nie. In die geval waar speltoetsers geëvalueer word, word die sisteem se vermoë om tussen reg- en verkeerdgespelde woorde te onderskei getoets. Die metriek meet dus hoeveel van die woorde wat deur die speltoetser as foute gemerk is, is inderdaad verkeerdgespelde woorde.

$$\text{“ Presisie} = \frac{\#\text{ (relevante items onttrek)}}{\#\text{ (onttrekte items)}} \text{“}$$

Vir hierdie studie word presisie gebruik om te bepaal hoeveel van die reëls korrek toegepas is. ’n Toepassing word slegs as korrek beskou indien die herrangskikking die gewenste sintaktiese struktuur toon wat so na as moontlik aan die Afrikaanse struktuur is. Die formule kan soos volg aangepas word:

$$\text{Presisie} = \frac{\#\text{ (korrekte toepassings van die reëls)}}{\#\text{(toepassings van die reëls)}}$$

4.2.1.3 F-telling

Die twee metrieke wat bo bespreek is, het gesamentlik ’n groot invloed op die geskiktheid van die voorprosesseringsmodule. ’n Sisteem wat ’n lae herroeping het, gaan nie al die herrangskikkingsmoontlikhede onttrek nie en gaan dus steeds nie die gewenste resultate lewer nie. Aan die ander kant kan geen sisteem suksesvol wees as die presisie laag is nie. Die herrangskikking sal dan verkeerd gedoen word en steeds nie bydra tot die verbetering van die uiteindelike SMV-sisteem nie. Een maatstaf wat dié metrieke in kombinasie met mekaar weeg is dus nodig. Die harmoniese gemiddeld van herroeping en presisie, oftewel die F-telling (Manning et al., 2008: 156) word met die onderstaande formule bereken. In hierdie berekening is presisie en herroeping ewe belangrik vir die sukses van die sisteem.

$$\text{“ F-telling} = \frac{2*\text{Presisie}*\text{Herroeping}}{\text{Presisie} + \text{Herroeping}} \text{”}$$

4.2.2 Resultate

Vir die evaluasie van die voorprosesseringsmodule word die toetsteks daarmee herrangskik en dan vergelyk met ’n handmatige manipulasie van dieselfde teks. Die handmatige herrangskikking word in ’n sogenaamde goue standaard (*gold standard)* (Hirschman & Mani, 2003: 416) vervat. Om te verseker dat die reëls objektief in die goue standaard toegepas word, is die toetsteks eers met sintaktiese annotasie gemerk en daarna saam met die patrone wat ook in 3.2 beskryf word, aan ’n persoon met genoegsame taalkennis gegee om handmatig te herrangskik. Hierdie persoon kon dus slegs die reëls soos wat dit neergeskryf is, interpreteer en het nie ’n subjektiewe mening oor die doel van die reëls gehad nie.

Die verskille tussen die outomatiese herrangskikking en die goue standaard kan aan die hand van herroeping, presisie en F-telling uitgedruk word. Elkeen van die resultate word gegee as ’n syfer tussen 0 en 1 waar 1 ’n 100 % voorstel. Die drie metrieke soos bo bespreek, is telkens per reëlkategorie bereken om sodoende ’n oorsig oor die effektiwiteit van verskillende aspekte van die herrangskikkingsreëls te kry. Dit is maklik om sodoende te sien watter reëls vir die meeste herrangskikkings verantwoordelik was en hoe gepas elke reëlkategorie is. Om hierdie resultate te bereken, is die toetsteks met die reëls verryk en daarna handmatig nagegaan om te bepaal watter reëls korrek (volgens die kriteria wat in die reëls vervat word) toegepas is. Tabel 4 toon die resultate en word vervolgens bespreek.

Herrangskikkingsreël	Presisie	Herroeping	F-telling
R1.1 en R1.2: **Werkwoordherrangskikking**	1,0	1,0	1,0
R2.1 tot R2.4: **Konstruksies met "to"**	0,86	1,0	0,92
R3.1 en R3.2: **Modale herrangskikking**	0,74	1,0	0,85
R4.1: **Negatiewe**	1,0	1,0	1,0
R5.1: **Verlede tyd**	1,0	1,0	1,0
Gemiddeld:	**0,92**	**1,0**	**0,95**

Tabel 4: Evaluasiematriks per reëlkategorie

Uit Tabel 4 blyk dit dat die meerderheid van die reëls effektief toegepas word. Reëls wat die merkers invoeg om verlede tyd (R5.1) en negatiewe (R4.1) aan te dui, word perfek op hierdie toetsteks toegepas (het 'n herroeping en presisie van 100%). Dit kan toegeskryf word aan die feit dat die patrone waarvolgens hierdie twee konstrukte herken kan word, baie duidelik uiteengesit kan word. Die herrangskikkingsreëls kan dus maklik die relevante voorbeelde in die teks opspoor en die merkers invoeg. In V21 word die oorspronklike sin uit die METIS II-toetsteks eers gegee, waarna die herrangskikte sin (die weergawe wat deur die voorprosesseringsmodule gegenereer word) en die Afrikaanse vertaling gegee word. Hierdie voorbeeld wys dat die dekodeerder ook baat by die invoeging van merkers vir beide die verlede tyd en die negatief. Aangesien een-tot-een belynings nou moontlik is en omdat die werkwoordherrangskikking ook suksesvol toegepas is, kan 'n woord-vir-woord vertaling nou gemaak word.

(V21.1) Oorspronklike Engels: "She *should* ***not*** *have* ***received*** *any aid*."

(V21.2) Herrangskikte Engels: "She *should* ***not*** *any aid have* ***received het nie***."

(V21.3) Afrikaanse vertaling: "Sy behoort **nie** enige hulp te **ontvang het nie**."

Die reëls wat werkwoorde herrangskik toon ook 'n F-telling van 1,0. Hierdie twee reëls manipuleer baie spesifieke gevalle, nl. waar 'n voorsetselstuk en ondergeskikte klous die werkwoord voorafgaan, en laat baie werkwoorde onveranderd. Alhoewel die gevalle wat wel behandel word, krities tot die omskakeling van Engelse struktuur na Afrikaanse woordvolgorde is, behoort verdere studies hierdie groep reëls uit te brei om ook werkwoorde in hoofsinne na te gaan. Die posisie van die werkwoord in hierdie tipe sinne toon, soos reeds in 3.2.1 genoem, baie variasie en meer indringende sintaktiese en morfologiese annotasie is nodig om die probleem totaal op te los. In hierdie navorsingsprojek het die fokus egter op werkwoorde

aan die einde van langer sinne geval, aangesien hierdie woorde heel dikwels in die vertalings ontbreek het (vgl. 2.3.4). In V22 word ’n voorbeeld van korrekte werkwoordherrangskikking gewys. Aangesien die Engelse werkwoord nou ook aan die einde van die sin staan, behoort die dekodeerder ’n beter vertaling te lewer en nie die laaste woord uit te laat nie.

(V22.1) Oorspronklike Engels: “*We created a document that I can **send** to anyone*.”

(V22.2) Herrangskikte Engels: “*We created a document that I to anyone can **send***.”

(V22.3) Afrikaanse vertaling: “Ons het ’n dokument saamgestel wat ek vir enige een kan **stuur**.”

Herrangskikking van konstruksies met “*to*” en modale hulpwerkwoorde behaal ook 100%-herroeping. Al die patrone wat in die toetsteks teenwoordig is, word dus geïdentifiseer. Die presisie van hierdie twee reëlgroepe is egter nie perfek nie, maar lewer steeds goeie resultate (0,86 en 0,74 onderskeidelik). Die feit dat al die relevante patrone wel opgespoor word, maar nie altyd reg gemanipuleer word nie, wys dat daar moontlik uitsonderings bestaan op die reëls wat in 3.2.2 en 3.2.3 geformaliseer is. Die herrangskikkingsreëls is dus nog nie sensitief genoeg om al die voorkomste van hierdie twee patrone korrek om te skakel nie en meer reëls sal waarskynlik nodig wees. Die volgende voorbeeld (V23) van die toepassing van R2.1 wys dat twee bywoordstukke omgeruil moet word en nie net na die posisie tussen “*to*” en die hoofwerkwoord moet skuif nie.

(V23.1) Oorspronklike Engels: “*Then we will all be able to **move forward together***.”

(V23.2) Herrangskikte Engels: “*Then we will all be able to **forward together move***.”

(V23.3) Inkorrekte Afrikaanse vertaling: “Dan sal ons almal in staat wees om **vorentoe saam te beweeg**.”

(V23.4) Meer korrekte Afrikaanse vertaling: “Dan sal ons almal in staat wees om **saam vorentoe te beweeg**.”

Hierdie reël is egter ook in ander gevalle met groot sukses toegepas. In V24 word so ’n voorbeeld gegee waar die “*to*”-konstruksiereël goed toegepas word. In hierdie sin is die Engelse struktuur nou baie na aan die Afrikaanse struktuur en dit vereenvoudig die vertalingsproses.

(V24.1) Oorspronklike Engels: “*It is important **to fill** the vacuum between our two regions*.”

(V24.2) Herrangskikte Engels: “*It is important **to** the vacuum between our two regions **fill***.”

(V24.3) Afrikaanse vertaling: “Dit is belangrik **om** die vakuum tussen ons twee streke **te vul**.”

Aangesien die herrangskikkingsreëls effektief blyk te wees, kan die voorprosesseringsmodule nou gebruik word om die toets- en afrigtingsdata van die SMV-sisteem te manipuleer. In die volgende afdeling word die afvoer van die uitgebreide SMV-sisteem met die afvoer van die basislynsisteem vergelyk. Die module wat bo geëvalueer is, word dus nou toegepas en die afvoer van die resulterende sisteem word gebruik om BLEU- en NIST-tellings te bereken.

4.3 Evaluasie van ’n nuwe SMV-sisteem

Om vas te stel wat die invloed van die herrangskikkingsreëls op die kwaliteit van die Engels-Afrikaanse MV-sisteem is, is ’n SMV-sisteem, waarin die reëls geïmplementeer is, gebruik om die METIS II-toetsdatastel te vertaal. Dieselfde toetsdatastel is ook deur die basislynsisteem (sien 2.2) vertaal. Die BLEU- en NIST-tellings van die afvoer is outomaties bereken sodat die sisteme vergelyk kan word. Die outomatiese evaluasiemetrieke word in die volgende afdeling bespreek en daarna word die resultate van die outomatiese evaluasie gegee en bespreek.

4.3.1 Outomatiese evaluasiemetrieke

Outomatiese evaluasie van masjienvertaling verseker dat die kwaliteit daarvan gereeld gemonitor kan word. Die sukses van verskeie eksperimente soos die toevoeging van tweetalige woordelyste of parallelle sinspare uit ’n ander domein tot die afrigtingsdata, kan vinnig en koste-effektief bepaal word (sien o.a. Vilar *et al.*, 2007; Jurafsky & Martin, 2009: 931; Papineni *et al*., 2002; Nießen *et al.,* 2000). Die outomatiese evaluering van die SMV-sisteme moet volgens internasionaal erkende metrieke gedoen word om die relevansie daarvan te kan verseker en om die resultate vergelykbaar te maak met die resultate van ander soortgelyke sisteme. Die BLEU- en NIST-tellings word algemeen hiervoor gebruik en is ook die twee metrieke waarmee die *Autshumato*-sisteem deurentyd getoets is om die vordering te monitor.

Die BLEU-telling is in 2002 voorgestel as effektiewe evaluasiemetode vir grootskaalse projekte waar daar ’n behoefte aan vinnige en frekwente evaluasies is (Papineni *et al*., 2002). Die metriek vergelyk die afvoer van ’n outomatiese vertaler met een of meer vertalings wat deur menslike kenners gedoen is en gee dan ’n numeriese waarde aan die “nabyheid” van die outomatiese vertaling aan die verwysingsvertaling(s). *N*-gramme uit die outomatiese vertaling word met elkeen van die verwysingsvertalings vergelyk om die mate van ooreenstemming tussen die *n*-gramme te bepaal wat in die outomatiese vertaling gebruik word en dié wat in die menslike vertalings voorkom.

Hierdie korrelasie word met ’n aangepaste weergawe van die presisie-metriek wat vroeër bespreek is, bepaal. Elke woord wat in een van die verwysingsvertalings en in die outomatiese vertaling voorkom, tel as ’n “goeie” vertaling en dra by tot die mate van korrelasie of ooreenstemming vir daardie sin. Die metriek is egter aangepas om te voorkom dat oortollige woorde die telling verkeerdelik bevoordeel deur te verseker dat ’n woord nie meer tot die telling bydra sodra ’n geskikte verwysingsvertaling daarvoor gevind kan word nie. Wanneer twee identiese woorde derhalwe in ’n sin gebruik word waar net een nodig is (bv. *Ek het **die die** man gesien*), sal die telling nie twee maal verhoog word vir elkeen van die lidwoord nie, maar slegs een maal.

Die BLEU-telling word aangegee as ’n syfer tussen 0 en 1, waar 1 beteken dat die outomatiese vertaling identies is aan een van die verwysingsvertalings. Aangesien dit selfs vir menslike vertalers moeilik is om vertalings identies aan ’n verwysingsvertaling te lewer, is dit belangrik om meer as een verwysingsvertaling in die korpus in te sluit. Die aantal verwysingsvertalings beïnvloed ook die uiteindelike telling, aangesien meer frases met mekaar vergelyk kan word en daar dus ’n groter waarskynlikheid vir ’n ooreenstemmende frase bestaan. Hoe meer verwysingsvertalings ook gebruik word, hoe meer vertaalekwivalente sal in die verwysing voorkom (Papineni *et al*., 2002). Verskillende verwysings gebruik meermale verskillende sinonieme of omstellings en die BLEU-algoritme kan net ’n akkurate evaluasie lewer indien soveel moontlik variante in die verwysingsvertalings teenwoordig is. Menslike vertalers behaal ’n BLEU-telling

tussen 0,8 en 0,65, afhangend van die aantal verwysingsvertalings en die moeilikheidsgraad van die teks wat vertaal moet word (Papineni *et al*., 2002). Vir hierdie studie en in die *Autshumato*-projek, word ses verwysingsvertalings wat deur professionele taalpraktisyns gedoen is, gebruik.

Die tweede metriek wat vir outomatiese evaluasie gebruik word, is die NIST-telling (Doddington, 2002). Hierdie metriek berus ook op die ooreenstemmende *n*-gramme in die verwysingsvertaling en outomatiese vertaling, maar die algoritme weeg die tellings verder om *n*-gramme wat minder frekwent is en reg vertaal word te bevoordeel (Doddington, 2002). Frases wat nie algemeen voorkom nie, is moeiliker om te vertaal, omdat die waarskynlikheidsaanduiding daarvoor in die dekodeerder nie hoog is nie (sien 2.2.1.2). Die intuïsie van hierdie telling is dus om die sisteem te bevoordeel wanneer dit sulke minder frekwente frases reg vertaal en is dus 'n goeie aanduiding van die sisteem se vermoë om unieke frases of sinne te dekodeer. Menslike vertalers behaal 'n NIST-telling tussen 10 en 11 en dit word dus as die ideale resultaat gesien (Doddington, 2002).

4.3.2 Resultate van die outomatiese evaluasie

Die basislynsisteem en die waarop herrangskikking toegepas word, is gebruik om die toetsteks te vertaal waarna die BLEU- en NIST-tellings vir elkeen van die gegenereerde vertalings bereken word deur dit met ses verwysingsvertalings te vergelyk. In Tabel 5 word die NIST- en BLEU-tellings vir beide die basislynsisteem en die uitgebreide sisteem (waarop voorprosessering toegepas is) gelys. Die ooreenstemmingsresultate op verskillende *n*-gramvlakke (sg. "*individual n-gram scoring*" (Lin & Hovy, 2003)) word ook gegee. Hierdie ooreenstemmingsresultate gee 'n aanduiding van die impak van die toepassing van die voorprosesseringsmodule op kleiner skaal. Die laaste ry in Tabel 5 wys die resultate vir die dokument as geheel en neem alle *n*-gramvlakke in ag.

Die kwaliteit van die afvoer het op al die vlakke gebaat by die toepassing van die herrangskikkingsreëls en die BLEU- en NIST-tellings het in al die gevalle verbeter. 'n Totale verbetering van 10,4% kan in die NIST-telling op dokumentvlak gesien word (van 8,4515 na 9,4905) en die BLEU-telling het met 7,7 % van 0,4968 na 0,5741 verbeter. Die byvoeging van die voorprosesseringsmodule het ook op *n*-gramvlak 'n positiewe invloed gehad en die tellings toon op al nege vlakke (1- tot 9-gramme) 'n verbetering. Die grootste impak volgens die NIST-telling kan op die 1-gramvlak gesien word. Hierdie telling verbeter met 0,7 (oftewel 7%) en dui daarop dat die sisteem meer woorde korrek vertaal. Een van die tekortkominge in die afvoer van die basislynsisteem was die groot hoeveelheid woorde wat onvertaald gelaat is (sien 2.3.5) en die verbeterde telling op 1-gramvlak is daarom belangrik.

Die BLEU-telling wys die grootste verbetering op 3-gramvlak (8%) en dui daarop dat meer frases of woordgroepe korrek vertaal word. Heelwat vaste uitdrukkings in Afrikaans bestaan uit drie woorde (bv. "in terme van", "te wyte aan", "as gevolg van") en skeibare werkwoorde kom ook dikwels as drie woorde voor ("voorstel" as "voor te stel", "deurgee" as "deur te gee"). Die verbetering in die 3-gramme kan derhalwe toegeskryf word aan die vermoë van die herrangskikkingsreëls om vaste uitdrukkings en skeibare werkwoorde so te groepeer dat die vertaling makliker plaasvind.

Vlak	Metriek	Basislynsisteem	Uitgebreide sisteem
1-gramme	**NIST**	6,0998	6,8033*
	BLEU	0,8287	0,8696
2-gramme	**NIST**	1,7083	1,9610
	BLEU	0,5781	0,6529
3-gramme	**NIST**	0,4543	0,4995
	BLEU	0,4172	0,4995*
4-gramme	**NIST**	0,1258	0,1510
	BLEU	0,3048	0,3829
5-gramme	**NIST**	0,0634	0,0757
	BLEU	0,2342	0,3000
6-gramme	**NIST**	0,0303	0,0434
	BLEU	0,1712	0,2288
7-gramme	**NIST**	0,0164	0,0262
	BLEU	0,1216	0,1692
8-gramme	**NIST**	0,0038	0,0176
	BLEU	0,0792	0,1224
9-gramme	**NIST**	0,0019	0,0090
	BLEU	0,0534	0,0849
Dokument as geheel	**NIST**	**8,4515**	**9,4905**
	BLEU	**0,4968**	**0,5741**

Tabel 5: Outomatiese evaluasie

In V25 word ’n voorbeeld van verbeterde 1- en 3-gramme gewys. Die oorspronklike Engelse frase word eers gegee, daarna die herrangskikte Engelse sin, die basislynsisteem se vertaling en laastens die vertaling volgens die uitgebreide sisteem waarop voorprosessering toegepas is.

(V25.1) Oorspronklike Engels: “*I **nonetheless believe***”

(V25.2) Herrangskikte Engels: “*I **believe nonetheless***”

(V25.3) Basislynsisteem: “Ek **nonetheless glo**”

(V25.4) Uitgebreide sisteem: “Ek **glo nietemin**”

Omdat die volgorde in die herrangskikte Engelse sin nader aan die Afrikaanse sinstruktuur is en dus nader aan die patrone wat tydens afrigting uit die parallelle korpus onttrek is, kon die Engelse woord wat nog in V25.3 aanwesig is, vertaal word. Die woordvolgorde het ook verbeter.

Die tellings in Tabel 5 vir langer frases toon egter nie dieselfde goeie resultate nie en alhoewel die uitgebreide sisteem steeds beter vaar, is die impak van die voorprosesseringsmodule kleiner. Beide sisteme vertaal dus die woorde tot ŉ groot mate korrek, maar toon steeds nie perfekte woordvolgorde wanneer dit met die verwysingsvertalings vergelyk word nie. Die herrangskikkingsreëls wat in hierdie navorsingsprojek ontwikkel is, poog om verskuiwings tussen frases wat op mekaar volg, aan te bring. Die swakker tellings vir die 4- tot 9-gramme wys egter dat verskuiwings oor groter afstande dalk nodig is. Die reëls oor werkwoordherrangskikking moet nie tot die ondergeskikte klous beperk word nie, maar dit moet eerder uitgebrei word om die werkwoord(stuk) voor die ondergeskikte klous ook te herrangskik. Verdere studie in hierdie verband is egter nodig om die reëls sensitief genoeg te maak.

Nießen *et al.* (2000) noem dat daar vir enige brontaalsin ŉ aantal korrekte vertalings bestaan en selfs met ŉ paar verwysingsvertalings is die metriek nie buigsaam genoeg om die wye verskeidenheid opsies in ag te neem nie. Vir hierdie navorsingsprojek word ses verwysingsvertalings gebruik, maar in soortgelyke studies word tot agt en selfs twaalf aanbeveel (sien o.a. Nießen *et al.*, 2000; Lin & Hovy, 2003; Zwarts & Dras, 2007; Collins *et al*., 2005). Die rede hiervoor is dat vertalers ook unieke woordkeuses maak en ŉ eie skryfstyl het wat kan lei tot vertalings wat nie met dié van die SMV-sisteem ooreenstem nie. Die BLEU- en NIST-tellings ag net identiese woorde en konstruksies as korrek en daarom kan dit aanbeveel word dat die maksimum aantal verwysingsvertalings gebruik word. As ŉ illustrasie van die probleem gee voorbeeld V26 die oorspronklike Engelse sin gevolg deur een van die verwysingsvertalings en die vertaling wat deur die uitgebreide sisteem gegenereer is. Dit is duidelik dat die outomatiese vertaling (V26.3) ook die inligting in die oorspronklike sin korrek oordra, maar aangesien die woordkeuse nie identies aan die verwysingsvertaling is nie, word dit deur die evaluasiemetrieke as inkorrek bestempel.

(V26.1) Oorspronklike Engels: “*We are in the **middle** of the **negotiating process***.”

(V26.2) Verwysingsvertaling: “Ons **staan midde** in die **onderhandelingsproses**.”

(V26.3) Outomatiese vertaling: “Ons **is** in die **middel** van die **proses van onderhandeling**.”

Die outomatiese evaluasiemetrieke wat in hierdie afdeling bespreek is, gee net ŉ aanduiding van die kwaliteit van die vertaling en gee nog nie ŉ oortuigende aanduiding van die bruikbaarheid van die afvoer vir menslike vertalers of die werklike vertaalkwaliteit nie. Die feit dat die tellings egter toegeneem het met die byvoeging van ŉ voorprosesseringsmodule, wys dat verdere navorsing in sintaktiese herrangskikking ŉ positiewe uitkoms kan hê.

4.4 Samevatting

In hierdie hoofstuk is die voorprosesseringsmodule in isolasie geëvalueer en die effek daarvan op die SMV-sisteem getoets. Die evaluasie van die voorprosesseringsmodule in 4.2 wys dat die module goed funksioneer – sinne wat herrangskik moet word, kan geïdentifiseer word (herroeping) en daarna korrek volgens die herrangskikkingsreëls verwerk word (presisie). Die module behaal ŉ gemiddelde herroeping van 0,92 en ŉ presisie van 1,0. Die F-telling vir die module as geheel is 0,95. Uit die voorbeelde kan afgelei word dat meer reëls en ŉ verfyning van die bestaande reëls nodig is sodat al die kategorieë ŉ F-telling van 1.0 kan behaal.

Die resultate van hierdie evaluasie toon positiewe resultate en dit is duidelik dat die toepassing van die voorprosesseringsmodule ŉ impak op die kwaliteit van veral die 1-, 2- en 3-gramme in die gegenereerde vertaling het. Wanneer die dokument as geheel beoordeel word, het die BLEU-telling van 0,4968 na 0,5741 (7,7 %) gestyg en die NIST-telling van 8,4515 na 9,4905 (10,3 %). Die resultate het egter ook gewys dat die tien herrangskikkingsreëls in die vyf kategorieë wat in hierdie navorsingsprojek ontwikkel is, nie herrangskikking op ŉ groot genoeg skaal toepas nie. Frases behoort oor groter afstande in sinne gemanipuleer te word. Dit is egter belangrik om in ag te neem dat ŉ beduidende verhoging in die outomatiese metrieke reeds met hierdie klein stel reëls behaal word en dat ŉ uitbreiding van die reëls ook belowende vooruitsigte inhou. Meer afrigtingsdata sal ook ŉ positiewe effek op die effektiwiteit van die reëls hê omdat die frasetabelle dan meer volledig opgestel kan word. ŉ Laaste verbetering wat aan die einde van die vorige afdeling aanbeveel word, is om meer verwysingsvertalings te gebruik om die vergelyking te doen. Heelwat sinne toon ŉ ander (korrekte) woordkeuse as die ses verwysingsvertalings wat tans geïnkorporeer word, maar word dan deur die outomatiese metrieke as verkeerde vertalings bestempel.

In die volgende hoofstuk sal ŉ opsomming van die studie en die bydrae wat dit tot die veld van Taaltegnologie lewer, gegee word. Verdere navorsing en verbetering vir hierdie spesifieke taalpaar sal vervolgens aanbeveel word en die verwante werk wat reeds vir ander Afrikatale gedoen is, sal bespreek word.

Hoofstuk 5: Samevatting

5.1 Gevolgtrekkings en bydrae

Die doel van die *Autshumato*-projek is om vertalers wat regeringsdokumente in die amptelike landstale beskikbaar moet stel, se werkslading te verlig sodat inligting vinniger versprei kan word. Hierdie studie poog dus om ŉ bydrae te maak tot die sukses van die eerste SMV-sisteem vir Engelse na Afrikaanse vertaling en ontwikkel en toets nuwe tegnieke om die bruikbaarheid van die afvoer van so ŉ sisteem te verhoog. Om die projek te organiseer, is die volgende doelstellings gestel:

- Om die afvoer van die *Autshumato* Engels na Afrikaanse SMV-sisteem te analiseer en vertalingsfoute wat moontlik deur voorprosessering d.m.v. reëlgebaseerde sintaktiese herrangskikking voorkom kan word, te identifiseer.

- (a) Om die verskille tussen Engelse en Afrikaanse sintaksis wat moontlik vir die foute verantwoordelik kan wees, na te vors, en

  (b) om linguisties gemotiveerde reëls te formuleer wat in die voorprosesseringsmodule gebruik kan word en hierdie reëls te evalueer.

- Om die afvoer van die resulterende SMV-sisteem te evalueer en krities met die huidige *Autshumato*-SMV-sisteem te vergelyk. Evaluasie behoort die internasionaal aanvaarde BLEU- en NIST-tellings in te sluit.

Die analise van die afvoer van die huidige *Autshumato* Engels na Afrikaanse SMV-sisteem is in Hoofstuk 2 gedoen en vyf uiteenlopende probleemareas word geïdentifiseer, te wete woordvolgorde, ontkenning, verlede tyd, werkwoorde en ŉ algemene kategorie met ŉ verskeidenheid foute. Die strukture in die bron- en teikentaal wat konsekwent van mekaar verskil (en in die eerste vier foutkategorieë vervat is), kan tydens voorprosessering herrangskik word om die invloed daarvan op die kwaliteit van die vertaling te beperk. Die relevante sintaktiese konstrukte word ook in hierdie hoofstuk uiteengesit en vir Engels en Afrikaans vergelyk. Die vergelyking maak dit duidelik dat daar vir sommige taalverskynsels patroonmatige verskille bestaan wat die manipulasie van die Engelse struktuur om meer na die Afrikaanse struktuur te lyk, moontlik maak. Dit is dan juis hierdie verskille wat in die herrangskikkingsreëls gestalte kry.

Nadat die verskille in die sintaktiese strukture van Afrikaans en Engels uitgelig en geanaliseer is, is daar in die volgende hoofstuk (Hoofstuk 3) op die voorgestelde voorprosesseringsmodule gefokus. Die reëls wat in dié hoofstuk bespreek word, is dus ŉ poging om die verskille tussen die twee tale te minimaliseer om sodoende ŉ beter vertaling te lewer. Vyf groepe reëls is uiteindelik ontwikkel: Werkwoordherrangskikkingsreëls, modaleherrangskikkingsreëls, reëls vir die herrangskikking van konstruksies met "*to*", verledetydsherrangskikkingsreëls en reëls vir die manipulasie van ontkennende frases.

In Hoofstuk 4 is die voorprosesseringsmodule eerstens in isolasie geëvalueer en daarna is die effek daarvan op die SMV-sisteem getoets. Die geïsoleerde evaluasie is gedoen deur die data in die METIS II-toetsteks te herrangskik en daarna die afvoer van die module met ŉ handgemaakte goue standaard te vergelyk. Hierdie evaluasie toon aan dat die module goed funksioneer en dit behaal ŉ gemiddelde herroeping

van 0,92 en ’n presisie van 1,0. Die F-telling vir die module as geheel is 0,95. Uit die voorbeelde wat ook in Hoofstuk 4 bespreek word, blyk dit dat die module verbeter kan word deur meer reëls toe te voeg en ook deur die bestaande reëls te verfyn. Die strekking van die reëls moet ook vergroot word sodat manipulasie oor langer afstande in ’n sin kan plaasvind.

Om die effek van die voorprosesseringsmodule op die kwaliteit van die afvoer van die SMV-sisteem na te gaan, is die METIS II-teks met die basislynsisteem en die herrangskikkingsisteem vertaal en BLEU- en NIST-tellings is vir elkeen van die resulterende vertalings bereken. Hierdie metrieke meet op verskillende wyses die ooreenstemming van ’n gegenereerde vertaling met ’n aantal verwysingsvertalings. Hierdie evaluasie toon positiewe resultate en dit is duidelik dat die toepassing van die voorprosesseringsmodule ’n impak op die kwaliteit van veral die 1-, 2- en 3-gramme in die gegenereerde vertaling het. Wanneer die dokument as geheel beoordeel word, het die BLEU-telling van 0,4968 na 0,5741 (7,7 %) gestyg en die NIST-telling van 8,4515 na 9,4905 (10,3 %). Dit is belangrik om daarop te let dat ’n beduidende verhoging in die outomatiese metrieke reeds met hierdie klein stel reëls behaal word en dat ’n uitbreiding van die reëls ook belowende vooruitsigte inhou. Die resultate ondersteun ook die uitkomste van die internasionale studies wat in Hoofstuk 1 bespreek word deurdat die skaal waarop verbetering plaasvind in ooreenstemming met daardie studies is.

5.2 Aanbevelings

In die lig van die gevolgtrekkings wat in hierdie studie gemaak is, kan die volgende aanbevelings vir toekomstige werk gemaak word:

- Die herrangskikkingsreëls wat in hierdie studie ontwikkel is, moet vervolgens verder uitgebrei en verfyn word. Verdere navorsing is nodig om te bepaal of ander konstrukte ook patroonmatige verskille toon wat op dieselfde manier herrangskik kan word. Dit kan gedoen word deur addisionele sinne te evalueer met behulp van die proses wat in Hoofstuk 2 van hierdie studie beskryf word om sodoende addisionele foutkategorieë te identifiseer. Een kategorie wat moontlik ondersoek moet word, is die besitskonstruksie (bv. “*Europe’s policy*” teenoor “Europa se beleid”)

- Heelwat sinne toon ’n ander (korrekte) woordkeuse as die ses verwysingsvertalings wat tans geïnkorporeer word, maar word dan deur die outomatiese metrieke as verkeerde vertalings bestempel. Dit word daarom ook aanbeveel dat meer verwysingsvertalings by die evaluasieproses ingesluit word.

- ’n Verdere aanbeveling is dat ’n toetsteks uit ’n ander domein as die regeringsdomein, wat hier ter sprake is, gebruik word om die toepaslikheid van die reëls vir verskillende gebruike te bepaal.

- Data-insameling moet ook voortgaan, aangesien SMV afhanklik van die kwantiteit en kwaliteit van die parallelle korpora is. Net soos in die geval van die toetsteks, sou dit ook wenslik wees om data uit ander domeine by die bestaande korpus te voeg om sodoende ’n meer verteenwoordigende stelsel daar te stel.

- ’n Aantal van die foute wat in die basislynsisteem se afvoer voorkom (veral die onder 2.3.5 gelys), sou miskien beter tydens naprosessering hanteer kon word. Dit word dus hier aanbeveel dat eks-

perimente gedoen word om vas te stel wat die invloed van die toevoeging van ŉ spel- en grammatikatoetser[10] as naprosessering op die kwaliteit van die afvoer sal wees.

- Die herrangskikkingsreëls wat in hierdie studie ontwikkel is, is vir ŉ spesifieke taalpaar ontwikkel, maar dieselfde metode kan ook gebruik word om reëls vir ander taalpare te ontwikkel. Aangesien die *Autshumato*-projek twee ander taalpare, nl. Engels-isiZulu en Engels-Sesotho sa Leboa insluit, behoort die invloed van voorprosessering volgens die metodes wat in hierdie studie uiteengesit is, ook getoets te word op dié twee taalpare. ŉ Voorlopige eksperiment in hierdie verband is reeds gedoen vir die taalpaar Engels-Sesotho sa Leboa (Griesel *et al.*; 2010). In hierdie eksperiment is ŉ stel van vyf reëls deur ŉ taalkenner ontwikkel en op die bestaande *Autshumato* Engels-Sesotho sa Leboa SMV-sisteem toegepas. Die BLEU- en NIST-tellings het toegeneem van 0,2126 na 0,2530 en van 4,9893 na 5,5214 onderskeidelik. Die sisteem toon dus ŉ gemiddelde verbetering van 5%. Daar is egter nog heelwat konstrukte in hierdie taalpaar wat nie ondersoek is nie en verdere navorsing in hierdie verband, en ook vir die ontwikkeling van die reëls vir die taalpaar Engels-isiZulu, is dus nodig.

5.3 Slot

Hierdie projek het ten doel gehad om ŉ tegniek vir die verbetering van die kwaliteit van die afvoer van ŉ bestaande SMV-sisteem vir vertaling van Engels na Afrikaans te ondersoek. Die gekose metode, te wete voorprosessering d.m.v. linguisties gemotiveerde sintaktiese herrangskikkingsreëls, blyk suksesvol te wees wanneer die afvoer met outomatiese metrieke geëvalueer word. Die reëls wat in hierdie navorsingsprojek ontwikkel is, kan nou uitgebrei word om sodoende ŉ SMV-sisteem daar te stel wat afvoer van so ŉ gehalte lewer dat die werk van die NLS-vertalers verder vergemaklik kan word. Die tegniek wat hier beskryf is, is ook geskik vir die verbetering van SMV-sisteme vir ander hulpbronskaars teikentale, soos Sesotho sa Leboa, aangesien geen kerntegnologieë benewens die herrangskikkingsreëls vir die teikentaal nodig is nie. Op hierdie manier dra die studie nie net by tot die ontwikkeling van die kennisbasis oor masjienvertaling vir Afrikaans nie, maar hou dit ook voordele vir die ander Suid-Afrikaanse tale in.

[10] Afrikaanse SkryfGoed 2011 word aanbeveel vir hierdie doel. Sien www.spel.co.za vir meer inligting.

Bibliografie

ARNOLD, D., BALKAN, L., MEIJER, S., HUMPHREYS, R.L. & SADLER, L. 1994. Machine translation: an introductory guide. London: NCC Blackwell. 240 p.

ACTIVESTATE. 2005. *ActivePerl 5.8.7 Build 813*. [http://www.ActiveState.com/Products/ActivePerl/] Date of access: 9 March 2005. Software.

BADR, I., ZBIB, R. & GLASS, J. 2009. Syntactic phrase reordering for English-to-Arabic statistical machine translation. (*In* Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. Athens, Greece. p.86-93.)

BAR-HILLEL, Y. 1960. The present status of automatic translation of languages. (*In* Alt, F., *ed*. Advances in computers: vol. 1. San Diego, CA: Academic Press. p.91-163.)

BERTOLDI, N & FEDERICO, M. 2005. A new decoder for spoken language translation based on confusion networks. (*In* Proceedings of Automatic Speech Recognition and Understanding Workshop (ASRU). San Juan, Puerto Rico. p. 86-91.)

BIBER, D., CONRAD, S. & LEECH, G. 2002. Longman student grammar of spoken and written English. Essex: Pearson Education. 487 p.

CARSTENS, W.A.M. 2004. Norme vir Afrikaans: enkele riglyne by die gebruik van Afrikaans. 4de uitg. Pretoria: Van Schaik. 478 p.

CIERI, C. 2007. Linguistic resources, development, and evaluation of text and speech systems. (*In* Dybkjær, L., Hemsen, H. & Minker, W., *eds*. Evaluation of text and speech systems. Dordrecht: Springer. (Text, speech, and language technology, vol. 37.) p. 221-261.)

COLLINS, M. 1997. Three generative, lexicalised models for statistical parsing. (*In* Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL) and Eighth Conference of the European Chapter of the Association for Computational Linguistics (EACL). Stroudsburg, Pennsylvania. p. 16-23.)

COLLINS, M., KOEHN, P. & KUCEROVA, I. 2005. Clause restructuring for statistical machine translation. (*In* Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL). Ann Arbor, Michigan. p. 531-540.)

DIRIX, P., SCHUURMAN, I. & VANDEGHINSTE, V. 2007. Metis II: example-based machine translation using monolingual corpora: system description. (*In* Proceedings of the Example-Based Machine Translation Workshop held in conjunction with the 10th Machine Translation Summit. Phuket, Tailand. p. 43-50.)

DU PLESSIS, H. 1985. Sintaksis. 2de uitg. Pretoria: Human & Rousseau. 103 p.

DODDINGTON, G. 2002. Automatic evaluation of machine translation quality using *n*-gram co-occurence statistics. (*In* Proceedings of the 2nd International Conference on Human Language Technology Research. San Diego, California. p. 1-8.)

GRIESEL, M., McKELLAR, C.A. & PRINSLOO, D. 2010. Syntactic reordering as preprocessing step in statistical machine translation from English to Sesotho sa Leboa and Afrikaans. (*In* Proceedings of the 21st annual Symposium of the Pattern Recognition Association of South Africa (PRASA). Stellenbosch. p. 105-110.)

GROENEWALD, H.J. & DU PLOOY, L. 2010. Processing parallel text corpora for three South African language pairs in the *Autshumato* Project. (*In* Proceedings of the 2nd Workshop on African Language Technology. Valletta, Malta. p. 27-30.)

HUTCHINS, J. 1995. Machine translation: a brief history. (*In* Koerner, E.F.K. & Asher, R.E., *eds*. Concise history of the language sciences: from the sumerians to the cognitivists. Oxford: Pergamon. p. 431-445.)

HIRSCHMAN, L. & MANI, I. 2003. Evaluation. (*In* Mitkov, R., *ed*. The Oxford handbook of computational linguistics. Oxford: Oxford University Press. p. 414-429.)

JURAFSKY, D. & MARTIN, J.H. 2009. Speech and language processing: an introduction to natural language processing, computational linguistics and speech recognition. 2nd ed. New Jersey: Pearson Education. 1024 p.

KLEIN, D. & MANNING, C. 2003. Accurate unlexicalized parsing. (*In* Proceedings of the 41st Meeting of the Association for Computational Linguistics (ACL). Sapporo, Japan. p. 423-430.)

KOEHN, P. 2005. Europarl: a parallel corpus for statistical machine translation. (*In* Proceedings of the 10th MT Summit. Pukhet, Thailand. p. 79-86.)

KOEHN, P., HOANG, H., BIRCH, A., CALLISON-BURCH, C., FEDERICO, M., BERTOLDI, N., COWAN, B., SHEN, W., MORAN, C., ZENS, R., DYER, C., BOJAR, O., CONSTANTIN, A. & HERBST, E. 2007. Moses: open source toolkit for statistical machine translation. (*In* Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL). Prague, Czech Republic. p. 177-180.)

KOEHN, P. 2010. Moses statistical machine translation system: user manual and code guide. [http://www.statmt.org/moses/] Date of access: 13 Nov. 2010.

KRINGS, H.P. & KOBY, G.S. 2001. Repairing texts: empirical investigations of machine translation post-editing processes. Kent, OH: Kent State University Press. 635 p.

LIN, C.Y. & HOVY, E. 2003. Automatic evaluation of summaries using *n*-gram co-occurrence statistics. (*In* Proceedings of the Human Technology Conference (HLT-NAACL). Edmonton, Canada. p. 150-156.)

LÜ, Y., HUANG, J. & LIU, Q. 2007. Improving statistical machine translation performance by training data selection and optimization. (*In* Proceedings Of EMNLP-CoNLL. Prague, Czech Republic. p. 343-350.)

MANNING, C., RAGHAVAN, P. & SCHüTZE, H. 2008. Introduction to information retrieval. New York: Cambridge University Press. 482 p.

MITCHELL, M., KIM, G., MARCINKIEWICZ, M.A., MacINTYRE, R., BIES, A., FERGUSON, M., KATZ, K. & SCHASBERGER, B. 1994. The Penn Treebank: annotating predicate argument structure. (*In* Proceedings of the Human Language Technology Workshop. Boston, Mass. p. 110-115.)

MANDAL, A., VERGYRI, D., WANG, W., ZHENG, J., STOLCKE, A., TUR, G., HAKKANI-TUR, D., AYAN, N.F. 2008. Efficient data selection for machine translation. (*In* Proceedings of IEEE Spoken Language Technology (SLT) Workshop. Goa, India. p.261-264.)

McKELLAR, C.A. 2011. Dataselektering vir statistiese Engels-Afrikaanse masjienvertaling. Potchefstroom: NWU. (Skripsie – M.A.)

NIEßEN, S., OCH, F.J., LEUSCH, G. & NEY, H. 2000. An evaluation tool for machine translation: fast evaluation for MT research. (*In* Proceedings of the 2nd International Conference on Language Resources and Evaluation. Athens, Greece. p. 39-45.)

OCH, F.J. 2003. Statistical machine translation: from single-word models to alignment templates. Aachen: RWTH Aachen. (Verhandeling – PhD)

OECD. 2002. Proposed standard practice for surveys on research and experimental development (Frascati Manual). Eurostat. [http://www.oecd.org/document/6/0,3343,en_2649_34451_33828550_1_1_1_1,00.html] Date of access: 13 March 2010.

PAPINENI, K., ROUKOS, S., WARD, T. & ZHU, W.J. 2002. BLEU: a method for automatic evaluation of machine translation. (*In* Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL). Philadelphia, USA. p. 311-318.)

PARLIKAR, A. 2008. SMT: "reordering word problem" and solutions. Advanced MT Seminar. Pittsburgh, Pennsylvania: Carnegie Mellon University. [www.cs.cmu.edu/afs/cs.cmu.edu/.../cmt.../11-734_WhitePaper_aup.pdf] Date of access: 12 Jan. 2011.

PILON, S., PUTTKAMMER, M.J. & VAN HUYSSTEEN, G.B. 2008. Die ontwikkeling van 'n woordafbreker en kompositumanaliseerder vir Afrikaans. *Literator: Journal of Literary Criticism, comparative linguistics and literary studies*, 29(1): 65-91, Apr.

PONELIS, F.A., SENEKAL, H.E.J. & DE KLERK, W.J. 1972. Die patroon van Afrikaans. Johannesburg: Afrikaanse Pers. 395 p.

PHAROS DICTIONARIES. 2006. *Media24-Korpus*. Kaapstad, N.d.

SIMARD, M.,UEFFING, N., ISABELLE, P. & KUHN, R. 2007. Rule-based translation with statistical phrase-based post-editing. (*In* Proceedings of the 2nd Workshop on Statistical Machine Translation. Prague, Czech Republic. p. 203-206.)

SOMERS, H. 2003. Machine translation: latest developments. (*In* Mitkov, R. *ed*. The Oxford handbook of computational linguistics. Oxford: Oxford University Press. p. 512-528.)

STOLCKE, A. 2002. SRILM – an extensible language modelling toolkit. (*In* Proceedings of the 7th Internasional Conference on Spoken Language Processing (ICSLP). Denver, Colorado. p. 901-904.)

TAALKOMMISSIE VAN DIE SUID-AFRIKAANSE AKADEMIE VIR WETENSKAP EN KUNS. 2002. Afrikaanse woordelys en spelreëls. 9de uitg. Kaapstad: Pharos. 592 p.

VILAR, D., NEY, H., LEUSCH, G. & BANCHS, R.E. 2007. Human evaluation of machine translation through binary systems comparison. (*In* Proceedings of the Association for Computational Linguistics (ACL). Prague, Czech Republic. p. 96-103.)

WANG, C., COLLINS, M. & KOEHN, P. 2007. Chinese syntactic reordering for statistical machine translation. (*In* Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Praque, Czech Republic. p. 737-745.)

WILCOCK, G. 2009. Introduction to linguistic annotation and text analytics. San Rafael, CA: Morgan & Claypool. 159 p.

ZWARTS, S. & DRAS, M. 2007. Syntax-based word reordering in phase-based statistical machine translation: why does it work? (*In* Proceedings of the 11th MT SUMMIT. Copenhagen, Denmark. p. 559-566.)