

Dataselektering en -manipulering vir Statistiese Engels-Afrikaanse Masjienvertaling

Me CA McKellar
20032609

*Skripsie voorgelê ter gedeeltelike nakoming van die vereistes vir die graad MA
in Algemene Taal en Literatuurwetenskap aan die Potchefstroomkampus van
die Noordwes-Universiteit.*

Studieleier: Mnr HJ Groenewald

Medestudieleier: Prof JC Roux

Februarie 2011

VOORWOORD

Graag word die volgende persone en instansies bedank:

- ∅ Die Sentrum vir Tekstegnologie, (CTexT®) en die Navorsingseenheid: Tale en Literatuur in die Suid-Afrikaanse konteks vir befondsing.
- ∅ My studieleier, Handré Groenewald, vir al die uitstekende advies. Dit was 'n voorreg om by iemand met soveel navorsingsondervinding te kon leer.
- ∅ My mede-studieleier, Prof JC Roux, vir die goeie raad en leiding.
- ∅ Die voormalige direkteur van die Navorsingseenheid: Tale en Literatuur in die Suid-Afrikaanse Konteks, Prof Hein Viljoen, vir volgehoue ondersteuning en belangstelling in dit wat ek doen.
- ∅ Die mense by CTexT, vir al die aanmoediging, hulp en ondersteuning.
- ∅ My familie, wat my elke tree van die pad moed ingepraat het.
- ∅ My vriende, wat geduldig na my praatjies geluister het, al het dit alles vir hulle soos Grieks geklink.

OPSOMMING

Dataselektering en -manipulering vir Statistiese Engels-Afrikaanse Masjienvertaling

Die sukses van enige masjienvertaalsisteem hang grootliks van die hoeveelheid en kwaliteit van die beskikbare afrigtingsdata af. 'n Sisteem wat met foutiewe of lae-kwaliteit data afgerig is, sal uiteraard swakker afvoer lewer as 'n sisteem wat met korrekte of hoë-kwaliteit data afgerig is. In die geval van hulpbronarm tale waar daar min data beskikbaar is en data dalk noodgedwonge vertaal moet word vir die skep van parallelle korpora wat as afrigtingsdata kan dien, is dit dus baie belangrik dat die data wat vir vertaling gekies word, so gekies word dat dit teksgedeeltes insluit wat die meeste waarde tot die masjienvertaalsisteem sal bydra. Dit is ook in so 'n geval uiters belangrik om die beskikbare data so goed moontlik aan te wend.

Hierdie studie stel ondersoek in na metodes om afrigtingsdata te selekteer met die doel om 'n optimale masjienvertaalsisteem met beperkte hulpbronne af te rig. Daar word ook aandag gegee aan die moontlikheid om die gewigte van sekere gedeeltes van die afrigtingsdata te verhoog om sodoende die data wat die meeste waarde tot die masjienvertaalsisteem bydra te beklemtoon. Alhoewel hierdie studie spesifiek gerig is op metodes vir dataselektering en -manipulering vir die taalpaar Engels-Afrikaans, sou die metodes ook vir toepassing op ander taalpare gebruik kon word.

Die evaluasieproses dui aan dat beide die dataselekteringsmetodes, asook die aanpassing van datagewigte, 'n positiewe impak op die kwaliteit van die resulterende masjienvertaalsisteem het. Die uiteindelijke sisteem, afgerig deur 'n kombinasie van verskillende metodes, toon 'n 2.0001 styging in die NIST-telling en 'n 0.2039 styging in die BLEU-telling.

Sleuteltermes: Statistiese masjienvertaling; Masjienvertaling; Engels; Afrikaans; Dataselektering; Afgrigtingsdata.

ABSTRACT

Data Selection and Manipulation for Statistical English-Afrikaans Machine Translation

The success of any machine translation system largely depends on the amount and quality of the available training data. A system that is trained with faulty or low quality data, will necessarily yield poorer output than a system that was trained with correct, high quality data. In the case of resource scarce languages where there is very little data available, it becomes necessary for data to be translated in order to create parallel corpora for use as training data. In such cases it is critically important to choose the data that adds the most information to the machine translation system. Under those circumstances it is also important to make the best possible use of the available training data.

This study therefore researches methods of training data selection in order to be able to train the optimal machine translation system with limited resources. The possibility of adapting the weights of certain parts of the training data, thereby emphasising the data that is the most valuable to the machine translation system, were also examined. Although this study was specifically conducted for the language pair English-Afrikaans, the methods investigated in this study could also be adapted to be used for other language pairs.

The evaluation process indicates that both the data selection methods as well as the adaptation of the data weights have a positive impact on the quality of the resulting machine translation system. The final machine translation system showed a 2.0001 increase in the NIST score and a 0.2039 increase in the BLEU score.

Key Terms: Statistical machine translation; Machine translation; English; Afrikaans; Data selection; Training data.

INHOUDSOPGAWE

Voorwoord	i
Opsomming	ii
Abstract.....	iii
Inhoudsopgawe.....	iv
Tabelle	vi
Figure	vi
1. Inleiding	1
1.1. Kontekstualisering	1
1.2. Toepaslikheid van navorsing	1
1.3. Probleemstelling.....	2
1.4. Navorsingsdoelstellings.....	4
1.5. Basiese hipotese stelling	4
1.6. Metode van ondersoek.....	4
1.7. Dokumentatoorsig	6
2. Literatuurstudie	7
2.1. Inleiding	7
2.2. Masjienvertaling	7
2.2.1. Oorsig	7
2.2.2. Statistiese masjienvertaling	9
2.2.3. Moses	10
2.3. Verwante studies.....	12
2.3.1. Masjienvertaling tussen Engels en Afrikaans.....	12
2.3.2. Manipulering van afrigtingsdata.....	13
2.3.3. Selekering van afrigtingsdata	15
2.4. Samevatting	20
3. Dataselektering- en manipulasie-eksperimente.....	22
3.1. Inleiding	22
3.2. Metode van ondersoek.....	22
3.3. Data	24
3.4. Programmatuur	25
3.5. Berekening van resultate.....	26

3.6. Standaardtaalmodel	28
3.7. Standaardsisteem	28
3.8. Basislyn	29
3.9. Selekering van bykomende afrigtingsdata	32
3.9.1. Selekering van data met vertrouetellings (Eksperimente 1 en 2)	33
3.9.2. Selekering van data met frekwensietellings (Eksperiment 3 en 4)	35
3.9.3. Evaluasie van Eksperimente 1-4	40
3.10. Aanpassing van gewigte van beskikbare data (Eksperiment 5)	43
3.11. Aanpassing en selekering van taalmodeldata	47
3.11.1. Aanpassing van taalmodeldata (Eksperiment 6)	47
3.11.2. Selekering van taalmodeldata (Eksperiment 7)	50
3.12. Kombinerings van metodes (Eksperiment 8)	51
3.13. Samevatting	53
4. Gevolgtrekkings en verdere navorsingsmoontlikhede	55
4.1. Gevolgtrekkings	55
4.2. Verdere navorsingsmoontlikhede	57
4.3. Slot	60
Bronnelys	61

TABELLE

Tabel 1: Samestelling van die eentalige data	25
Tabel 2: Samestelling van die tweetalige data.....	25
Tabel 3: NIST- en BLEU-tellings van die standaardsisteem.....	29
Tabel 4: NIST-tellings van die basislynsisteme	30
Tabel 5: BLEU-tellings van die basislynsisteme	31
Tabel 6: Evaluasieresultate (NIST) van die dataselekteringseksperimente (1-4).....	40
Tabel 7: Evaluasieresultate (BLEU) van die dataselekteringseksperimente (1-4)	41
Tabel 8: Evaluasieresultate van die aanpassing van datagewigte (Eksperiment 5).....	44
Tabel 9: Evaluasieresultate van die aanpassing van die gewigte van taalmodeldata (Eksperiment 6).....	48
Tabel 10: Evaluasieresultate van die selektering van taalmodeldata (Eksperiment 7)	50
Tabel 11: Evaluasieresultate van die kombinasie van verskeie metodes (Eksperiment 8)	52

FIGURE

Figuur 1: Vloeidiagram van die afrigting van 'n vertaalmodel met Moses	11
Figuur 2: Leerkurwes van die basislynsisteme (volgens NIST-tellings)	31
Figuur 3: Leerkurwes van die basislynsisteme (volgens BLEU-tellings)	32
Figuur 4: Vloeidiagram van Eksperimente 1 en 2.....	34
Figuur 5: Vloeidiagram van Eksperiment 3	40
Figuur 6: Vloeidiagram van Eksperiment 4	41
Figuur 7: Evaluasieresultate (NIST) van die dataselekteringseksperimente (1-4)	41
Figuur 8: Evaluasieresultate (BLEU) van die dataselekteringseksperimente (1-4)	42
Figuur 9: Evaluasieresultate (NIST) van die aanpassing van datagewigte (Eksperiment 5).....	45
Figuur 10: Evaluasieresultate (BLEU) van die aanpassing van datagewigte (Eksperiment 5)	46
Figuur 11: Evaluasieresultate (NIST) van die aanpassing van die gewigte van taalmodeldata (Eksperiment 6)	50
Figuur 12: Evaluasieresultate (BLEU) van die aanpassing van die gewigte van taalmodeldata (Eksperiment 6)	51
Figuur 13: Evaluasieresultate (NIST) van die selektering van taalmodeldata (Eksperiment 7)	53
Figuur 14: Evaluasieresultate (BLEU) van die selektering van taalmodeldata (Eksperiment 7)	53

1. INLEIDING

1.1. KONTEKSTUALISERING

In Suid-Afrika bestaan daar 'n groot behoefte aan die beskikbaarstelling van inligting in al elf amptelike landstale. Om byvoorbeeld alle regeringsinligting deur professionele vertalers te laat vertaal kan 'n tydrowende en duur proses wees. Met behulp van masjienvertaalsisteme kan dokumente outomaties vertaal word en 'n professionele vertaler word dan slegs benodig om die afvoer van die masjienvertaalsisteem na te gaan en te korrigeer. Masjienvertaling kan dus die produktiwiteit van die professionele vertaler verhoog (Hutchins, 2001:5; Plitt & Masselot, 2010:10) en is daarom 'n baie belangrike hulpmiddel in die Suid-Afrikaanse konteks.

Die sukses van enige masjienvertaalsisteem hang grootliks van die hoeveelheid en kwaliteit van die beskikbare data af, aangesien sodanige sisteme afgerig word met twee tipes data, te wete eentalige data in die teikentaal, en parallelle data bestaande uit sinne in die brontaal (die taal waaruit daar vertaal word) wat belyn is met hulle vertalings in die teikentaal (die taal waarheen daar vertaal word). 'n Sisteem wat met swak of lae-kwaliteit data afgerig is, sal uiteraard swakker afvoer lewer as 'n sisteem wat met goeie of hoë-kwaliteit data afgerig is. Swak data kan sinne wat verkeerd belyn of vertaal is, of sinne wat slegs uit name en nommers bestaan, insluit. Dit is ook belangrik om soveel verteenwoordigende data (data wat die woordeskat en taalgebruik van die domein weerspieël) as moontlik in die afrigtingsproses te gebruik om te verseker dat ook lae-frekwensiewoorde deur die resulterende sisteem hanteer kan word. Parallelle data is moeiliker om te versamel en gewoonlik ook skaarser as eentalige data en daarom is dit belangrik om te verseker dat die beste moontlike data versamel of ontwikkel word en ook optimaal gebruik word.

1.2. TOEPASLIKHEID VAN NAVORSING

Volgens die Grondwet (1996) van Suid-Afrika moet die status van die land se elf tale verhoog en hulle gebruik bevorder word. Die regering is daarom betrokke by 'n aantal

projekte om veeltaligheid in Suid-Afrika te bevorder. Een van hierdie projekte, befonds deur die Departement van Kuns en Kultuur, is die Autshumato-projek (Groenewald & Fourie, 2009:190). Die Autshumato-projek behels die skep van verskeie hulpmiddels vir vertalers wat, onder andere, die ontwikkeling van masjienvertaalsisteme vir drie Suid-Afrikaanse taalpare insluit. Die drie masjienvertaalsisteme word ontwikkel met die spesifieke doel om dokumente in die regeringsdomein te vertaal. Hierdie studie vorm deel van die eksperimentele ondersoek wat poog om metodes te vind vir die ontwikkeling van goeie kwaliteit masjienvertalingstelsels vir hulpbronarm Suid-Afrikaanse tale. Die bevindinge van hierdie studie kan egter ook op ander hulpbronarm taalpare toegepas word.

1.3. PROBLEEMSTELLING

Statistiese masjienvertaling is 'n tegnologiese ontwikkeling wat 'n groot impak op Suid-Afrika kan hê. Statistiese masjienvertaling kan gebruik word om relatief vinnig 'n goeie masjienvertaalsisteem af te rig. Daarom kan dit, as gevolg van ons groot aantal amptelike tale, meer as 'n noodsaaklikheid as 'n luuksheid beskou word. Die probleem met die ontwikkeling van doeltreffende masjienvertaalsisteme vir Suid-Afrikaanse tale is die tekort aan hoë-kwaliteit parallelle data. Die parallelle data wat wel beskikbaar is, is ook dikwels nie ideaal vir gebruik as afrigtingsdata nie. 'n Voorbeeld hiervan is data wat meestal uit name of syfers bestaan, aangesien dit feitlik geen waarde tot die masjienvertaalsisteem bydra nie.

'n Moontlike oplossing vir die tekort aan bruikbare data is om data te genereer deur die vertaling van tekste, maar hierdie oplossing is baie duur en tydrowend. Die prys vir vertaling van data in Suid-Afrika wissel tans tussen R0.70 en R0.90 per woord. Dit beteken dat die koste om voldoende data vir die afrigting van 'n statistiese masjienvertaalsisteem te laat vertaal, betreklik hoog kan wees.

Dit is ook moeilik om te bepaal hoeveel data vir die afrigting van 'n suksesvolle statistiese masjienvertaalsisteem benodig word. Die algemene siening is dat meer data altyd beter is. Koehn gebruik die *Europarl*-korpus (2005:81) om masjienvertaalsisteme vir verskeie Europese taalpare af te rig. Die *Europarl*-korpus bevat ongeveer 700 000 tot 1 000 000 parallelle sinne per taalpaar. Koehn se resultate wissel tussen 0.1 en 0.4 BLEU-tellings

(Koehn, 2005:83) (sien Afdeling 3.5 vir 'n uitgebreide beskrywing van die BLEU-metriek vir die outomatiese evaluering van die prestasie van masjienvertaalsisteme).

In die geval van hulpbronarm tale waar daar noodgedwonge data vertaal moet word vir die skep van parallelle data, is dit dus baie belangrik dat die data wat vir vertaling gekies word, so gekies word dat dit teksgedeeltes insluit wat die meeste waarde tot die masjienvertaalsisteem bydra. Dit is ook belangrik om nie lae-kwaliteit data tot die afrigtingsdata by te voeg nie, aangesien dit kan lei tot 'n verswakking van die kwaliteit van die uiteindelijke masjienvertaling. Voor die vertaling van data kan begin, moet daar egter eers besluit word hoe hierdie data gekies gaan word. Hierdie vraag vorm die basis van die eerste navorsingsdoelstelling van hierdie studie, naamlik om 'n metode van dataselektering te ontwikkel wat tot die bes moontlike masjienvertaalsisteme, wat met die mins moontlike data afgerig is, sal lei.

Dit kan ook moontlik gebeur dat daar vir 'n taalpaar met min beskikbare bronne van parallelle data ook nie fondse vir die vertaling van bykomende data beskikbaar is nie. Dit is in so 'n geval dus uiters belangrik om die beskikbare data so goed moontlik aan te wend. Een so 'n metode is om die gewigte van sommige "goeie" parallelle sinspare (hoë-kwaliteit sinspare wat wesenlike bydraes tot die kwaliteit van die masjienvertaalsisteem lewer) in die afrigtingskorpus te verhoog om sodoende op die "goeie" sinne klem te lê. Die gewigte kan vermeerder word deur bykomende kopieë van die "goeie" sinne in die afrigtingskorpus tot die afrigtingskorpus te voeg. Die nuwe korpus sal geen bykomende inligting tot die masjienvertaalsisteem bydra in terme van woordeskat nie, maar die bykomende geleenthede om die vertaling van sekere frases te observeer kan wel tot 'n verbetering in die vertaling lei (Finch & Sumita, 2008:113).

Die afrigting van 'n masjienvertaalsisteem vereis ook 'n eentalige korpus van teks in die teikentaal. Daar is altyd meer eentalige data as parallelle data beskikbaar en eentalige data is oor die algemeen ook eenvoudiger om te versamel as parallelle data; dus kan goeie kwaliteit eentalige data in 'n korter tyd bekom word. Indien die verandering van die taalmodeldata dus tot 'n verbetering in die kwaliteit van die masjienvertaling sou lei, kan dit 'n vinniger en goedkoper oplossing wees as die versamel van bykomende parallelle data of die vertaling van data om parallelle data te skep.

1.4. NAVORSINGSDOELSTELLINGS

Gegewe die kontekstualisering hierbo, is die doelstellings van hierdie studie om:

1. die effek van dataselektering (van beide parallelle en eentalige data) op die prestasie van statistiese masjienvertaling vir Engels-Afrikaans te ondersoek;
2. die effek van die aanpassing van die gewigte van beskikbare data (beide parallelle en eentalige data) op die prestasie van statistiese masjienvertaling vir Engels-Afrikaans te ondersoek; en
3. deur middel van 'n noukeurige kombinasie van die bogenoemde metodes 'n verbetering in die masjienvertaling van Engels na Afrikaans teweeg bring.

1.5. BASIESE HIPOTESE STELLING

Die noukeurige selektering van bykomende afrigtingsdata sal tot 'n verbetering in die kwaliteit van die masjienvertaling lei. Die aanpassing van beskikbare afrigtingsdata deur die gewigte te verander kan ook tot 'n verbetering in die kwaliteit van masjienvertaling lei.

1.6. METODE VAN ONDERSOEK

Die eerste stap in hierdie studie behels 'n literatuuroorsig waarin daar 'n inleiding tot masjienvertaling gegee word, met 'n spesifieke fokus op statistiese masjienvertaling en die werking daarvan. Daar word ook gekyk na ander studies wat oor die selektering van data vir masjienvertaling handel, asook studies wat op die aanpassing van die beskikbare afrigtingsdata vir verhoogde prestasie fokus.

Die eerste stap in die praktiese komponent van die studie is die samestelling van korpora. Beide tweetalige parallelle data en eentalige taalmodeldata word benodig. Die data wat in die studie gebruik word, word in Afdeling 3.3 bespreek. Nadat die data ingesamel is, moet 'n standaardmasjienvertaalsisteem en 'n standaardtaalmodel afgerig word. Die standaardmasjienvertaalsisteem vorm die beginpunt van alle dataselekterings-eksperimente. Die standaardtaalmodel word vir alle masjienvertaalsisteme gebruik, buiten vir dié wat deel vorm van die eksperimente wat handel oor die verbetering van die

masjienvertaalsisteme deur die selektering van taalmodeldata en die aanpassing van die datagewigte in die taalmodeldata. Moses, 'n oopbron statistiese masjienvertaling-gereedskapstel (Koehn *et al.*, 2007:177) is gebruik om die masjienvertaalsisteme af te rig. Die taalmodelle is met die *SRI Language Modeling*-gereedskapstel (Stolcke, 2002:901) afgerig. Die standaardmasjienvertaalsisteem (asook alle ander masjienvertaalsisteme) is outomaties geëvalueer met die Bilingual Evaluation Understudy (BLEU) (Papineni *et al.*, 2002:311) en NIST (naam afgelei van die US National Institute of Standards and Technology) (Doddington, 2002:138) evaluasiemetrieke.

Die basislynsisteme vir dataselektering is afgerig deur ewekansig-geselekteerde tekseenhede by die standaardsisteem se data te voeg. Die standaardsisteem is op 100 000 woorde afgerig en vir die basislynsisteme is daar telkens 'n aantal ewekansig-geselekteerde tekseenhede (bestaande uit ongeveer 50 000 woorde) by die afrigtingsdata gevoeg. Daar was 10 sisteme in die basislyn; 'n totaal van 500 000 woorde is dus tot die standaardsisteem se afrigtingsdata gevoeg. Al die sisteme in die basislyn is ook met die standaardtaalmodelle afgerig en outomaties geëvalueer. Daarna is daar gepoog om deur dataselekteringsmetodes op die ewekansige basislyn te verbeter. Dit is gedoen deur verskeie dataselekteringsmetodes te implementeer en te evalueer. In hierdie eksperimente is daar ook telkens 50 000 woorde in tien inkremente tot die afrigtingsdata van die standaardsisteem gevoeg, maar die verskil lê in die metodes waarvolgens die 50 000 woorde telkens gekies is.

Vervolgens is daar toe gefokus op die tweede navorsingsdoelstelling en is daar dus bepaal of die aanpassing van die gewig wat verskeie dele van die afrigtingsdata dra, tot beter masjienvertaling kon lei. Hiervoor is gedeeltes van die data gekies en toe in die afrigtingskorpus verdubbel of verdriedubbel. Verskeie hoeveelhede data is gekies vir hierdie eksperiment. Die verskillende eksperimente is presies soos die basislynsisteem afgerig en geëvalueer.

Daar is ook ondersoek gedoen oor die effek van die verandering van die taalmodeldata op die kwaliteit van die uiteindelijke masjienvertaling. Dit behels, soos met die parallelle data, twee verskillende tipes ondersoeke. Eerstens is daar gekyk of die aanpassing van die gewigte van die taalmodeldata tot beter masjienvertaling lei. Dit is gedoen deur gedeeltes van die data te kies en die gewig wat daardie gedeeltes dra te vermeerder deur kopieë van die betrokke sinne tot die afrigtingskorpus te voeg. Daar is ook gekyk of die selektering van taalmodeldata 'n verbetering bied oor die ewekansige

standaardtaalmodelle (wat as basislyn vir die taalmodel-eksperimente dien). Die taalmodel-eksperimente is telkens uitgevoer deur die nuwe taalmodelle met die standaardmasjienvertaalsisteem te gebruik. Die masjienvertaalsisteem is toe outomaties geëvalueer met die NIST- en BLEU-evaluasiemetrieke om te bepaal of daar enige verbetering in die kwaliteit van die masjienvertaling was.

Die finale stap in die praktiese eksperimente was die afrigting van 'n masjienvertaalsisteem wat die metodes wat in die vorige toetse die beste resultate behaal het te kombineer. Hierdie sisteem is met die beste taalmodel en beste datastel afgerig. Die gewigte van hierdie sisteem se afrigtingsdata is ook aangepas volgens die metode wat die beste resultate in vorige eksperimente behaal het. Die gekombineerde masjienvertaalsisteem is ook met die NIST- en BLEU-evaluasiemetrieke geëvalueer.

1.7. DOKUMENTOORSIG

In hierdie hoofstuk word die konteks en toepaslikheid van hierdie studie gemotiveer. Daar word gekyk na die probleme wat aanleiding gegee het tot hierdie studie, asook die navorsingsdoelstellings en basiese hipotese van die studie. Laastens word 'n kort oorsig oor die metode van ondersoek gegee. In die volgende hoofstuk word toepaslike agtergrondkennis verskaf. Daar word na masjienvertaling in die algemeen en ook spesifiek na statistiese masjienvertaling gekyk. 'n Oorsig oor bestaande navorsing betreffende dataselektering en -manipulering vir masjienvertaling word ook verskaf.

Alle praktiese eksperimente word in die derde hoofstuk bespreek. Daar word gefokus op die standaard- en basislynsisteme, die dataselekteringseksperimente, asook die eksperimente wat oor die aanpassing van die datagewigte handel. Die eksperimente oor die verbetering van masjienvertaling deur die aanpassing van die data wat vir die afrigting van taalmodelle gebruik word, word ook bespreek. Laastens word die eksperiment bespreek wat handel oor die kombinasie van verskeie metodes wat goeie resultate gelewer het. Alle evaluasieresultate en die afleidings wat daaruit gemaak kan word, word ook in Hoofstuk 3 saamgevat. Die laaste hoofstuk bevat 'n samevatting van die studie. Die gevolgtrekkings wat uit die verskeie praktiese eksperimente gemaak kan word, word ook in Hoofstuk 4 bespreek. Laastens word daar gekyk na die impak van hierdie studie op die terrein van masjienvertaling vir Suid-Afrikaanse tale, asook moontlike toekomstige werk.

2. LITERATUURSTUDIE

2.1. INLEIDING

In die vorige hoofstuk word die konteks en toepaslikheid van hierdie studie gemotiveer. Die probleemstelling word bespreek en die navorsingsdoelstellings en basiese hipotese-stelling word uiteengesit. Die metode van ondersoek word ook uiteengesit en bespreek.

Die doel van hierdie hoofstuk is om agtergrondkennis te verskaf en om 'n oorsig te bied oor toepaslike navorsing wat as basis vir die res van die studie kan dien. Eerstens word 'n oorsig gegee oor masjienvertaling in die algemeen en daarna word statistiese masjienvertaling meer breedvoerig bespreek. Daar word gekyk na navorsing oor masjienvertaling vir die taalpaar Engels-Afrikaans (Engelbrecht & Schultz, 2005:3; Snyman *et al.*, 2007:225; Kato & Barnard, 2007:2) en ook na ondersoeke oor die verbetering van masjienvertaling (Finch & Sumita, 2008:113; Lu *et al.*, 2007:343) deur die aanpassing van die afrigtingsdata. Laastens word daar gekyk na ondersoeke (Haffari *et al.*, 2009:415; Kato & Barnard, 2007:1; Mandal *et al.*, 2008:261) wat handel oor die verbetering van masjienvertaling deur spesifieke dataselekeringsmetodes.

2.2. MASJIENVERTALING

2.2.1. OORSIG

Warren Weaver en Andrew Booth het in 1947, kort na die ontwikkeling van die eerste rekenaars, reeds die moontlikhede van outomatiese vertaling deur 'n rekenaar bespreek. In 1954 is die eerste openbare demonstrasie van 'n masjienvertaalsisteem, geskep deur Leon Dosterd in samewerking met IBM, gehou (Hutchins, 1995:433). In die jare wat daarop gevolg het, is daar baie navorsing oor masjienvertaling gedoen, maar dit het al hoe duideliker geword dat outomatiese vertaling 'n baie meer komplekse taak is as wat aanvanklik vermoed is. In 1966 het die *Automatic Language Processing Advisory Committee* (ALPAC) 'n verslag geskryf wat gelei het tot die einde van feitlik alle navorsing op die gebied van masjienvertaling in die VSA (Hutchins, 1995:435). Die meeste navorsing oor masjienvertaling het verskuif na Kanada en Europa, waar daar 'n groot behoefte was

aan outomatiese vertaling ten einde met die verskillende bedryfstale te werk (Hutchins, 1995:435). In die 1980's het navorsing oor masjienvertaling vinnig begin vooruitgaan, deels as gevolg daarvan dat verskeie werkende sisteme die publiek se aandag op dié gebied gevestig het (Hutchins, 1995:436). Vandag word masjienvertaling as een van die grootste uitdagings op die gebied van natuurliketaalprosessering beskou.

Masjienvertaling is 'n proses waartydens teks, met of sonder menslike hulp, deur gerekenariseerde sisteme vertaal word (Hutchins & Somers, 1992:3). Masjienvertaling word oor die algemeen op twee maniere benader, naamlik reëlgebaseerde metodes en datagedrewe metodes.

Reëlgebaseerde masjienvertaling (Izuha *et al.*, 2008:430; Tyers, 2010b:1) maak gebruik van reëls wat deur taalkenners geskryf word, asook tweetalige woordelyste om 'n vertaling van 'n teks te skep. Die reëls word byvoorbeeld gebruik om die struktuur van die brontaalsin na dié van die teikentaalsin om te skakel. Indien daar in die brontaal byvoorbeeld adjektiewe voor naamwoorde geplaas word, maar in die teikentaal word adjektiewe na naamwoorde geplaas, word 'n reël benodig wat bepaal dat die twee woordtipes moet omruil (Jurafsky & Martin, 2009:906). Die woordelyst word tesame met morfologiese analise gebruik om die vertalings van brontaalwoorde te verkry.

Datagedrewe metodes sluit voorbeeldgebaseerde masjienvertaling (Kit *et al.*, 2000:57; Somers, 1999:113) en statistiese masjienvertaling (Lopez, 2008:1; Zens *et al.*, 2002:18) in. Met voorbeeldgebaseerde vertaling word die afrigtingsdata as voorbeelde gestoor van hoe sekere tekseenhede vertaal moet word en gedurende vertaling soek die sisteem dan deur die gestoorde data vir 'n teksgedeelte wat soortgelyk is aan die teksgedeelte wat vertaal moet word. Statistiese masjienvertaling is gebaseer op die waarskynlikhede dat sekere tekseenhede in die brontaal, vertalings van tekseenhede in die teikentaal is. Hierdie waarskynlikhede word dan gedurende die vertalingsproses gebruik om die mees waarskynlike vertaling van 'n sin te bepaal.

Van al bogenoemde metodes behaal statistiese masjienvertaling gewoonlik die beste resultate (Zens *et al.*, 2002:18) en van die onderskeie statistiese masjienvertalingsmetodes lewer frasegebaseerde statistiese masjienvertaling oor die algemeen die beste prestasie (Zens *et al.*, 2002:30). Dit is daarom dan ook die benadering waaroor die meeste navorsing binne die veld van masjienvertaling tans gedoen word (Lopez, 2008:2; Zens *et al.*, 2002:18) en die masjienvertalingsmetode wat in hierdie studie ondersoek word.

2.2.2. STATISTIESE MASJIENTVERTALING

Alle vertalings moet aan twee belangrike vereistes voldoen, naamlik die vertaalde sin in die teikentaal moet getrou wees aan die oorspronklike sin (dit moet dieselfde betekenis hê) en moet ook 'n natuurlike sin ('n grammaties korrekte, vloeiende sin) in die teikentaal wees. Die doel van masjienvertaling kan dus beskou word as die maksimering van 'n funksie wat die belangrikheid van beide hierdie faktore beklemtoon (Jurafsky & Martin, 2009:911). Indien daar vanaf Engels (E) na Afrikaans (A) vertaal word kan die funksie deur die volgende formule (Formule 1) voorgestel word:

$$\hat{A} = \operatorname{argmax} P(E|A) P(A) \quad [1]$$

Waar $\hat{A}$ die beste Afrikaanse vertaling van die Engelse sin verteenwoordig (dus die een met die hoogste waarskynlikheid), $P(E|A)$ verteenwoordig die waarskynlikheid dat 'n spesifieke Afrikaanse sin die vertaling van die Engelse sin is en $P(A)$ die waarskynlikheid dat die spesifieke Afrikaanse sin voorkom. Statistiese masjienvertaling benodig dus twee komponente, naamlik 'n vertaalmodel wat die waarskynlikheid aandui dat 'n woord (of reeks woorde) in die brontaal na 'n bepaalde woord (of reeks woorde) in die teikentaal vertaal sal word ($P(E|A)$) en 'n taalmodel wat 'n aanduiding bied van die waarskynlikheid dat die woord (of reeks woorde) in die teikentaal voorkom ($P(A)$) (Jurafsky & Martin, 2009:912).

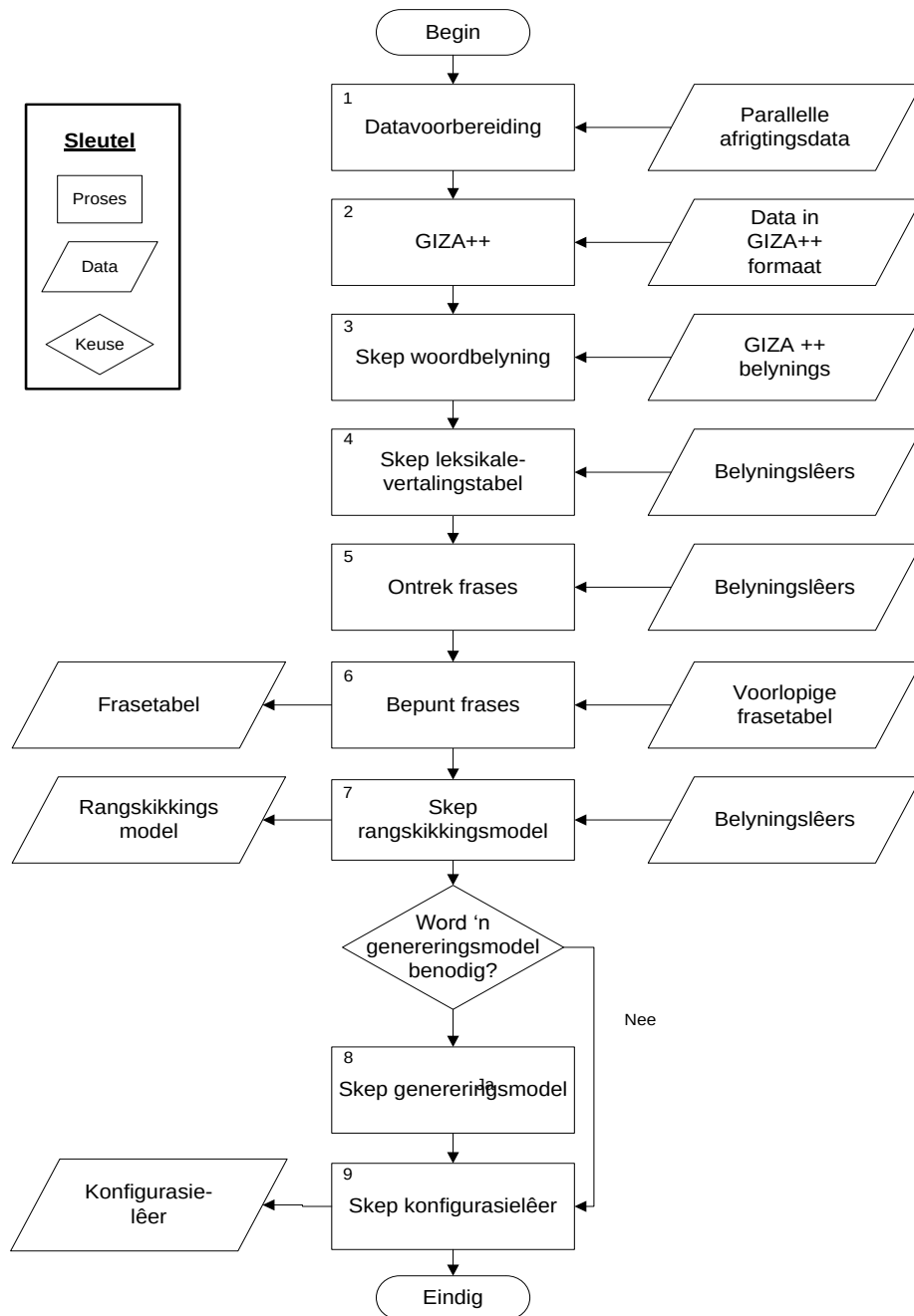
Die vertaalmodel wat in 'n statistiese masjienvertaalsisteem gebruik word, is die komponent wat, gegewe 'n brontaalsin en 'n teikentaalsin, die waarskynlikheid toeken dat die spesifieke brontaalsin die teikentaalsin sal genereer (Jurafsky & Martin, 2009:913). Een moontlike benadering tot die generering van die teikentaalsin vanuit die brontaalsin is om te kyk na hoe elkeen van die individuele woorde in die sin vertaal sal kan word. Vir frasegebaseerde vertaling word beide frases en enkelwoorde beskou wanneer die vertaalmodel (Jurafsky & Martin, 2009:913) ontwikkel word. 'n Frasegebaseerde vertaalmodel gebruik vertalingswaarskynlikhede (die waarskynlikheid dat 'n woord of frase in die brontaal met 'n bepaalde woord of frase in die teikentaal vertaal word) om die vertaling van bronteks te bepaal en afwykingswaarskynlikhede ('n maatstaf van die afstand tussen die posisies van frases in die twee tale) om die teikentaaltekste te herorden (Jurafsky & Martin, 2009:914). Beide hierdie tipes waarskynlikhede kan uit 'n belynde korpus afgelei word.

In statistiese masjienvertaling word taalmodelle gebruik om te bepaal of die voorgestelde vertaling 'n grammatikaal korrekte sin in die teikentaal is. Statistiese masjienvertaalsisteme maak gebruik van dieselfde n -gram taalmodelle as ander natuurliketaalprosessering toepassings soos spraakherkenning (Jurafsky & Martin, 2009:912). Die taalmodel word gebruik om die waarskynlikheid dat 'n spesifieke woord (of reeks woorde) in die teikentaal voorkom te bepaal. Die n -gram taalmodelle is probabilistiese modelle van reekse woorde, waar die reeks telkens n woorde lank is (Jurafsky & Martin, 2009:117). Die statistiese modelle kan gebruik word om die volgende woord in 'n string te voorspel of om te bepaal watter woord of woorde die mees waarskynlik op 'n voorafgaande reeks woorde sal volg.

2.2.3. MOSES

Moses, 'n oopbron, statistiese masjienvertalingsgereedskapstel (Koehn *et al.*, 2007:177) gebruik die frasegebaseerde benadering tot masjienvertaling. Dit is dan ook die program wat in saam *GIZA++* (Och & Ney, 2003:19) gebruik word vir die afrigting van masjienvertaalsisteme. Die twee programme bevat saam alle komponente wat benodig word om 'n ten volle funksionele masjienvertaalsisteem te ontwikkel. Dit sluit dus komponente in vir die verwerking van data, die afrigting van vertaalmodelle, die generering van vertalings, ens. (Koehn *et al.*, 2007:178). Die gereedskapstel kan ongeannoteerde afrigtingsdata, of data wat met faktore (lemmas, woordsoortetikette, ens.) geannoteer is, hanteer. Die afrigtingsdata, wat uit parallelle tekseenhede bestaan, word gebruik om die vertaalmodel mee af te rig.

Voordat die afrigting van die vertaalmodel kan begin moet die data in die korrekte formaat wees, naamlik daar moet twee lêers wees (een vir elke taal) wat die belynde afrigtingsdata bevat. Die tekseenhede wat vertalings van mekaar is, moet in ooreenstemmende lyntjies staan. Alle data moet in woordeenhede verdeel wees (leestekens moet dus apart van woorde geskryf word) en in kleinletters wees (Koehn, 2010:78). Dit is belangrik om te verseker dat die algoritme optimaal funksioneer. Die rede hiervoor is dat, indien 'n woord slegs een keer in die afrigtingsdata voorkom met 'n hoofletter of met 'n punt agterna, en dieselfde woord kom in die toetsdata voor sonder die hoofletter of punt, sal die masjienvertaalsisteem die woord nie herken nie en dus ook nie kan vertaal nie.



Figuur 1: Vloediagram van die afrigting van 'n vertaalmodel met Moses

Figuur 1 beeld die afrigting van 'n vertaalmodel uit, slegs die afvoerlêers wat deel uitmaak van die uiteindelijke vertaalmodel word as afvoer by die betrokke prosesse aangetoon (dus nie die tydelike lêers nie). In die eerste stap van die afrigtingsproses word die data voorberei deurdat dit omgeskakel word na die formaat wat deur *GIZA++* (Och & Ney, 2003:19) vereis word. *GIZA++* is 'n program wat gebruik word om die parallelle

tekseenhede op woordvlak te belyn (Koehn, 2010:83). *GIZA++* word gedurende die afrigtingsproses gebruik om die data in beide rigtings (van Engels na Afrikaans en andersom) op woordvlak te belyn. Die twee stelling belynings wat deur *GIZA++* geskep is vorm die basis van die vertaalmodel. Die twee belynings word dan deur een van 'n aantal heuristiese metodes gekombineer (Koehn, 2010:84) om 'n finale woordbelyning te vorm. Die kombineringsmetode word as deel van die afrigtingsinstruksie deur die gebruiker gespesifiseer.

Die vierde stap in die afrigtingsproses is die skep van 'n leksikale vertalingstabel. Die tabel bevat stelling moontlike vertalings op woordvlak. Elkeen van die vertalings het ook twee waarskynlikhede; die eerste toon aan hoe waarskynlik die spesifieke brontaalwoord as die spesifieke teikentaalwoord vertaal sal word. Die tweede waarskynlikheid toon aan hoe waarskynlik die teikentaalwoord as die betrokke brontaalwoord vertaal sou word (indien die rigting van vertaling omgeruil word) (Koehn, 2010:87). In die volgende stap word al die frases onttrek en in een lêer gekombineer om 'n enkele frasetabel te vorm. Die frases word onttrek deur woorde in die brontaal te gebruik wat langs mekaar staan en met teikentaalwoorde belyn is wat ook langs mekaar staan as frases wat vertalings van mekaar is. Gedurende die sesde stap word die waarskynlikhede dat elkeen van die brontaalfrases na die betrokke teikentaalfrases vertaal sal word, bereken (Koehn, 2010:88). Die volgende stap in die afrigtingsproses is die afrigting van 'n rangskikkingsmodel. Die rangskikkingsmodel leer verskillende herrangskikkings-waarskynlikhede vir elke frase en word gebruik om die teikentaalsin se woorde in die korrekte volgorde te plaas (Koehn, 2010:89). Indien die afrigtingsdata faktore (bv. lemmas) bevat, word 'n genereringsmodel daarna geskep. Laastens word 'n konfigurasielêer gemaak wat aandui waar die vertaalmodel en taalmodel op die bediener gestoor word en watter instellings en gewigte vir die gebruik daarvan gebruik moet word (Koehn, 2010:90).

2.3. VERWANTE STUDIES

2.3.1. MASJIENTVERTALING TUSSEN ENGELS EN AFRIKAANS

Alhoewel daar in die afgelope jare baie navorsing op die gebied van masjientvertaling gedoen is, is daar nog, te wete, redelik min navorsing gedoen vir vertaling tussen die amptelike landstale van Suid-Afrika. Engelbrecht en Schultz (2005:1) het 'n prototipe van 'n

Engels-Afrikaanse spraak-tot-spraak vertaalsisteem ontwikkel met statistiese masjienvertaling as onderdeel, die masjienvertaalsisteem is afgerig op Hansards data, bestaande uit 687 154 woorde, afkomstig vanaf Parlementêre sittings. Engelbrecht en Schultz (2005:5) het masjienvertaalsisteme van beide Engels na Afrikaans en van Afrikaans na Engels afgerig, beide met leestekens en daarsonder. Die sisteme is met die NIST- en BLEU-evaluasiemetrieke (sien Afdeling 3.5) geëvalueer. NIST-tellings wat wissel tussen 7.65 en 7.93, en BLEU-tellings wat wissel tussen 0.3413 en 0.3611, is gerapporteer (Engelbrecht & Schultz, 2005:6).

Die *EtsaTrans* voorbeeldgebaseerde masjienvertaalsisteem is sedert 2004 aan die Universiteit van die Vrystaat ontwikkel met die doel om vertalers met die vertaling van notules tot hulp te wees (Snyman *et al.*, 2007:225). Die masjienvertaalsisteem is ontwikkel met die oog op integrasie met die *Parnassus*-stelsel sodat die vertaling van inligting meer doeltreffend kan plaasvind en inligtingsvloei sodoende verbeter kan word (Snyman *et al.*, 2007:225). *Parnassus* is 'n stelsel wat ontwikkel is om die doeltreffendheid van vergaderings te verbeter en die samestelling van notules te fasiliteer (Snyman *et al.*, 2007:235). Die *EtsaTrans*-masjienvertaalsisteem maak gebruik van 'n soortgelykheidsfunksie om telkens die brontaalvoorbeeld wat die meeste met die bronteks ooreenstem uit 'n tweetalige databasis te selekteer (Snyman *et al.*, 2007:232).

Verder het Kato en Barnard (2007:1) 'n studie uitgevoer oor die ontwikkeling van statistiese masjienvertaalsisteme in die Suid-Afrikaanse konteks waar daar beperkte hoeveelhede afrigtingsdata beskikbaar is. Die studie het gefokus op vertaling tussen die taalpare Engels-isiXhosa, Engels-isiZulu, Engels-Setswana en Engels-Afrikaans. Die masjienvertaalsisteme is met data afkomstig vanaf regeringswebblaaie afgerig (Kato & Barnard, 2007:2). 'n BLEU-telling van 0.1999 is vir die Engels-Afrikaanse masjienvertaalsisteem gerapporteer. Francis Tyers werk ook aan 'n Engels-Afrikaanse, reëlgebaseerde masjienvertaalsisteem (Tyers, 2010a:1) alhoewel daar, te wete, nog geen amptelike vrystelling beskikbaar is nie.

2.3.2. MANIPULERING VAN AFRIGTINGSDATA

Dit is moontlik om die data wat die meeste waarde tot 'n akkurate vertaling sal bydra te beklemtoon deur die gewig wat aan daardie spesifieke gedeelte van die data toegeken is te vermeerder. Alhoewel hierdie benadering geen nuwe woordeskat aan die

masjienvertaalsisteem kan verskaf nie, kan dit wel lei tot die skep van beter belynings, wat weer sal lei tot beter vertalings van dokumente, aangesien die afrigtingsalgoritme meer geleenthede gebied word om sekere frases te waar te neem.

Finch en Sumita (2008:113) het verskeie masjienvertaalsisteme gebruik om sinne wat as “goeie data” beskou word uit te kies. Finch en Sumita (2008:113) definieer “goeie data” as die sinne wat korrekte en letterlike vertalings van mekaar is. Die gewig van die “goeie data” is dan vermeerder deur bykomende kopieë van die sinne tot die afrigtingskorpus by te voeg. Die frasegebaseerde statistiese masjienvertaalprogram wat gebruik is om die “goeie data” uit te wys, is aangepas om die teikentaalsin wat met die brontaalsin belyn is, te genereer deur van die frasetabel gebruik te maak. Indien dit nie moontlik is om die presiese teikentaalsin te produseer nie, is die vertaling leeg gelaat. Volgens Finch en Sumita (2008:115) sal die sinne wat baie “goeie” vertalings (korrek en letterlik vertaal) van mekaar is op hierdie wyse uitgelig word. Hierdie proses is tien keer herhaal (tienvoudige kruisvalidasie), elke keer is 90% van die data gebruik om ’n masjienvertaalsisteem mee af te rig en die ander 10% is vertaal. Die gewigte van die data wat wel op hierdie wyse vertaalbaar is, is dan aangepas deur twee bykomende kopieë van die betrokke sinne by die afrigtingsdata in te sluit. ’n Nuwe masjienvertaalsisteem is dan met die aangepaste data afgerig. ’n Verbetering van 0.012 in die BLEU-telling van ’n Japannees-Engelse masjienvertaalstelsel word met hierdie metode gerapporteer (Finch & Sumita, 2008:118).

’n Ander studie het gepoog om masjienvertaling meer toetsdata- of domeinspesifiek te maak om sodoende beter vertalings te bewerkstellig. Lu *et al.* (2007:343) bespreek twee verskillende metodes; die eerste vir bekende toetsdata of ’n bekende domein en die ander vir ’n onbekende domein. In die geval van ’n bekende domein of reeds bekende toetsdata word inligtingsherwinningmetodes gebruik om tekseenhede uit die afrigtingsdata te kies wat uit dieselfde domein as die toetsdata afkomstig is (Lu *et al.*, 2007:344). Hierdie tekseenhede word dan óf afsonderlik gebruik om ’n masjienvertaalsisteem mee af te rig, óf die gewigte word aangepas om sodoende die meer soortgelyke tekseenhede groter gewig te gee (Lu *et al.*, 2007:345). Die idee is dat die resulterende masjienvertaalsisteem goeie vertalings binne die betrokke domein sal lewer, terwyl swakker vertalings verwag kan word vir tekste wat buite die betrokke domein val.

In die geval waar die domein onbekend is of waar vertalings in ’n verskeidenheid domeine gelewer moet word, word verskillende modelle afgerig met subgedeeltes van die beskikbare afrigtingsdata. Daar word ook ’n algemene model afgerig met al die beskikbare

data. Indien die afrigtingsdata vanuit verskillende bronne afkomstig is, kan dit volgens die verskillende tipes bronne verdeel word (bv. nuusartikels, wette, ens.). Andersins kan 'n klein korpus van die brontaal vir elkeen van die moontlike domeine saamgestel word en dan gebruik word om deur middel van inligtingsherwinningmetodes die data in subkorpusse te sorteer (Lu *et al.*, 2007:345). Elkeen van die korpusse word dan gebruik om 'n aparte vertalingsmodel mee af te rig. Tydens die vertaling van 'n sin word daar dan eerstens van 'n soortgelykheidsfunksie gebruik gemaak om gewigte aan die verskillende vertalingsmodelle toe te ken, afhangende van hoe soortgelyk die afrigtingsdata van die model is aan die sin wat vertaal moet word (Lu *et al.*, 2007:346). Verskillende metodes word gebruik om gewigte aan die vertalingsmodelle toe te ken; onder andere word die vertalingsmodel met die hoogste soortgelykheidsfunksiewaarde gebruik om die teks te vertaal of die modelle word saam gebruik waar elkeen 'n ander gedeelte van die gewig bydra. Volgens Lu *et al.* (2007:347) toon beide hierdie metodes beter resultate as 'n masjienvertaalsisteem wat eenvoudig op alle beskikbare data afgerig is.

2.3.3. SELEKTERING VAN AFRIGTINGSDATA

Aangesien statistiese masjienvertaling so 'n sterk datagedrewe metode is, word die gehalte van die masjienvertaling gewoonlik verbeter deur die toevoeging van bykomende afrigtingsdata. Dit is te wyte aan die feit dat hoe meer afrigtingsdata beskikbaar is, hoe meer geleenthede het die algoritme om woorde en frases waar te neem en hul korrekte vertalings af te lei. Die grootte van die masjienvertaalsisteem se woordeskaf is dus eweredig aan die grootte van die afrigtingsdatastel. Afrigtingsdata kan uit tweetalige dokumente versamel word en dan outomaties of handmatig belyn word. Daar is egter nie altyd data beskikbaar nie, veral nie vir hulpbronnarm tale, soos die oorgrote meerderheid van die amptelike Suid-Afrikaanse tale, nie. 'n Gebrek aan parallelle data kan voorkom word deur sinne uit eentalige brontaaltekste te selekteer en dit deur professionele vertalers te laat vertaal. Aangesien sodanige vertaling 'n duur en tydrowende proses kan wees, is dit belangrik om te poog om dié sinne wat tot die grootste verbetering in die kwaliteit van die masjienvertaling sal lei te selekteer en te laat vertaal. Verskeie studies (Haffari *et al.*, 2009:415; Kato & Barnard, 2007:1; Mandal *et al.*, 2008:261) het reeds getoon dat dataselekteringsmetodes wat op aktiewe leermetodes gebaseer is tot beter masjienvertaalsisteme lei as die willekeurige kies van data.

Die eerste stap in enige masjienleerproses is die versameling van data. As die afrigtingsdata ewekansig geselekteer word, staan dit bekend as passiewe leer (Tong, 2001:3). Daar is dus geen poging om “goeie” data te kies nie. Dit is egter moontlik om die selekteringsproses op een of ander wyse te lei of te beïnvloed om sodoende die mees inligtinggewende data te kies (Tong, 2001:3). Hierdie keuses hoef nie vooraf bepaal te word nie, maar kan telkens op die kennis gebaseer word wat op daardie oomblik beskikbaar is. Aktiewe leer is die proses waar die keuse van bykomende data gebaseer word op die data wat reeds in gebruik is (Tong, 2001:4), met die idee dat meer inligtinggewende data gekies kan word as wat die geval is wanneer data ewekansig gekies word. Aktiewe leermetodes behaal gewoonlik beter resultate as passiewe leermetodes (Tong, 2001:5).

Haffari *et al.* (2009:415) het 'n aktiewe leerbenadering gebruik om die afvoer van 'n frasegebaseerde statistiese masjienvertaalsisteem te verbeter. Haffari *et al.* (2009:416) het met twee verskillende tipes data (buiten die toetsdata) gewerk, naamlik 'n klein tweetalige korpus en 'n groter eentalige korpus in die brontaal. Die tweetalige data is gebruik om 'n masjienvertaalsisteem mee af te rig wat op sy beurt gebruik is om die eentalige data mee te vertaal. Beide die oorspronklike tweetalige data en die vertaalde eentalige data is dan gebruik om 'n nuwe masjienvertaalsisteem mee af te rig wat gebruik word om die toetsdata vir evaluasiedoeleindes te vertaal. Die mees inligtinggewende sinne word dan uit die eentalige data gekies en vir 'n persoon gegee om te vertaal. Die vertaalde data word dan by die tweetalige data gevoeg (Haffari *et al.*, 2009:416). Hierdie proses word iteratief herhaal totdat 'n sekere vlak van vertalingskwaliteit (gemeet volgens die BLEU-metriek) bereik word.

Verskeie metodes om die inligtinggewende sinne vir vertaling uit te kies is getoets en met die ewekansige basislyn vergelyk om te bepaal watter metodes tot beter resultate lei (Haffari *et al.*, 2009:416). Die eerste stel dataselekteringsmetodes is gebaseer op die frekwensie van frases of n -gramme in beide die parallelle data, sowel as die eentalige data (Haffari *et al.*, 2009:417). Die beginsel is dat hoe meer dikwels 'n frase in die eentalige data voorkom, hoe belangriker is die insluiting daarvan in die afrigtingsdata; hoe meer dikwels 'n frase in die tweetalige data voorkom, hoe minder belangrik is die frase, aangesien dit met sy vertaling reeds beskikbaar is. Die gemiddelde waarskynlikheidsverhouding word vir die frases in 'n sin bereken en dan as aanduiding gebruik van watter sinne vir vertaling gekies moet word (Haffari *et al.*, 2009:417). Beide 'n

geometriese (Formule 2) en wiskundige (Formule 3) gemiddelde waarskynlikheidsverhouding is bereken.

$$\phi_g^p(s) = \left[\prod_{x \in X_s^p} \frac{P(x|U)}{P(x|L)} \right]^{\frac{1}{|X_s^p|}} \quad [2]$$

$$\phi_g^p(s) = \frac{1}{|X_s^p|} \sum_{x \in X_s^p} \frac{P(x|U)}{P(x|L)} \quad [3]$$

L verteenwoordig die tweetalige data en U verteenwoordig die eentalige data. Al die frases wat uit sin s onttrek kan word, word voorgestel deur X_s^p , en $P(x|U)$ of $P(x|L)$ is die waarskynlikheid dat frase x in die betrokke korpus voorkom (Haffari *et al.*, 2009:417). Die frases in die tweetalige data word outomaties gedurende die afrigtingsproses van die masjienvertaalsisteem onttrek; dit is egter moeiliker om die frases in die eentalige data te identifiseer. Daarom het Haffari *et al.* (2009:417) die formules aangepas om met n -gramme in plaas van frases te werk. Formules 4 en 5 gee die aangepaste formules vir die geweege kombinasie van die n -gramme se geometriese en wiskundige gemiddelde waarskynlikheidsverhoudings.

$$\phi_g^N(s) = \sum_{n=1}^N \frac{w_n}{|X_s^n|} \sum_{x \in X_s^n} \log \frac{P(x|U,n)}{P(x|L,n)} \quad [4]$$

$$\phi_a^N(s) = \sum_{n=1}^N \frac{w_n}{|X_s^n|} \sum_{x \in X_s^n} \frac{P(x|U,n)}{P(x|L,n)} \quad [5]$$

In die bostaande formules dui X_s^n die n -gramme in die sin aan en $P(x|L,n)$ of $P(x|U,n)$ dui die waarskynlikheid aan dat x in die stel n -gramme van die gegewe data voorkom. Die gewigte w_n pas die belangrikheid van die punte aan na aanleiding van die lengte van die onderskeie n -gramme (Haffari *et al.*, 2009:417).

Haffari *et al.* (2009:417) het ook sinne uit die eentalige korpus gekies vir vertaling volgens die mate waartoe dit met die tweetalige afrigtingsdata ooreenstem. Die soortgelykheid van 'n sin aan die tweetalige data is bepaal deur die geweege n -gram dekking te bereken (Haffari *et al.*, 2009:417). Hoe groter die verskil tussen die betrokke sin en die tweetalige data, hoe waarskynliker is dit dat die spesifieke sin vir vertaling gekies word. Daar is ook toetse gedoen met die kies van sinne deur van soortgelykheidsfunksies en die vertaalprogram se vertrouettelings gebruik te maak (Haffari *et al.*, 2009:417). Die verskeie dataselektieringsmetodes is ook gebruik as faktore in 'n lineêre rangskikkingsmodel wat

dan gebruik kan word om sinne in 'n orde te rangskik wat aandui watter sinne die meeste waarde tot die vertaalmodel kan bydra (Haffari *et al.*, 2009:418).

Verder het Haffari *et al.* (2009:418) ook die *Hierarchical Adaptive Sampling*-algoritme aangepas sodat dit op statistiese masjienvertaling toegepas kan word. Hierdie metode is gewaarborg om nie swakker as die ewekansige basislyn te presteer nie (Haffari *et al.*, 2009:418). Die algoritme begin deur 'n masjienvertaalsisteem met alle beskikbare tweetalige sinne af te rig en alle eentalige sinne in 'n enkele groep te plaas. Alle eentalige sinne word dan met die masjienvertaalsisteem vertaal waarna al die eentalige sinne dan in twee groepe verdeel word volgens hul ooreenkoms met die tweetalige data. Die α mees soortgelyke sinne word in die een groep (H_1) geplaas en die res in die ander groep (H_2) (Haffari *et al.*, 2009:418). Om K sinne te kies wat deur 'n persoon vertaal moet word, word β K sinne uit groep H_1 gekies en $(1-\beta)K$ sinne uit groep H_2 (Haffari *et al.*, 2009:418). Die waardes van beide α en β kan gekies word afhangende van die toepassing en aantal sinne wat op 'n slag vertaal kan word. Die vertaalde sinne word dan by die tweetalige data gevoeg, 'n nuwe vertaalsisteem afgerig en al die oorblywende eentalige sinne weer vertaal. Die oorblywende eentalige data is dus nou in twee groepe, H_1 en H_2 , verdeel. Albei groepe word vertaal en die groep met die laagste gemiddelde vertrouetsing word volgende verdeel (Haffari *et al.*, 2009:418). Hierdie proses word iteratief herhaal.

Die laaste metode wat deur Haffari *et al.* (2009:418) getoets is, is die omkeer van die rigting van vertaling. Elkeen van die eentalige sinne is na die teikentaal vertaal en die vertaling is dan met 'n tweede masjienvertaalsisteem vanaf die teikentaal terug na die brontaal vertaal. Die oorspronklike brontaalsin word dan met die vertaalde brontaalsin vergelyk (deur middel van die BLEU-telling) om te bepaal hoeveel inligting gedurende die vertaling verlore gegaan het (Haffari *et al.*, 2009:419). Die sinne waar die meeste inligting verloor is, word aan 'n persoon gegee om te vertaal, waarna die vertaalde sinne dan weer by die afrigtingsdata van die masjienvertaalsisteem gevoeg word.

Haffari *et al.* (2009:420) het gevind dat die selektering van sinne gebaseer op die geometriese gemiddeld van die waarskynlikheidsverhouding vir frases oor die algemeen beter resultate as die ander metodes lewer. Die gebruik van vertrouetsings wat deur die vertaalproses gegenereer is, het egter swakker resultate as die ander metodes gelewer (Haffari *et al.*, 2009:420).

Kato en Barnard (2007:1) het 'n kombinasie van beide aktiewe leer en semi-gekontroleerde leer gebruik om bykomende afrigtingsdata vir statistiese masjienvertaling te selekteer. Die benadering is ontwikkel met die aanname dat daar baie min tweetalige data beskikbaar is, maar dat daar wel bykomende eentalige data in die brontaal beskikbaar is. Eerstens is 'n voorlopige vertaalmodel afgerig met die beskikbare tweetalige data en dan gebruik om die eentalige data vanaf die brontaal na die teikentaal te vertaal (Kato & Barnard, 2007:2). Die vertaalde data word dan in twee stelle verdeel: die een stel bevat die data met die hoë vertrouetellings en die ander stel bevat die data met die lae vertrouetellings. Die datastel met die lae vertrouetellings word aan 'n vertaler gegee om te vertaal, waarna dit by die ander tweetalige data gevoeg word. Die nuwe afrigtingsdata word dan gebruik om 'n nuwe masjienvertaalsisteem mee af te rig wat weer gebruik word om die ander stel data (met die hoë vertrouetellings) mee te vertaal (Kato & Barnard, 2007:2). Die nuwe vertalings word by die afrigtingsdata gevoeg en 'n nuwe masjienvertaalsisteem word met hierdie data afgerig. Die proses kan iteratief herhaal word solank as wat die resultate 'n verbetering toon (Kato & Barnard, 2007:2). Die eksperiment is uitgevoer op vier verskillende Suid-Afrikaanse taalpare naamlik Engels-Afrikaans, Engels-isiZulu, Engels-Setswana en Engels-isiXhosa (Kato & Barnard, 2007:2). Die toepassing van aktiewe en semi-gekontroleerde leer op dataselektering het vir al die taalpare tot beter resultate as die ewekansige basislyn gelei (Kato & Barnard, 2007:4). Die grootste verbetering was by die taalpaar Engels-Setswana waar die BLEU-telling met 0.12 verhoog het (Kato & Barnard, 2007:3).

'n Ander studie (Mandal *et al.*, 2008:261) het twee verskillende dataselekteringsmetodes getoets en ook die twee metodes gekombineer om die uitwerking daarvan op die resulterende masjienvertaalsisteem (Mandaryn-Engels) te bepaal. Met die eerste metode is daar gepoog om die mees inligtinggewende sinne te kies deur te kyk na die mate waartoe bepaalde vertalings tussen verskillende masjienvertaalsisteme verskil (Mandal *et al.*, 2008:261). Dit is gedoen deur drie verskillende masjienvertaalsisteme (een gebaseer op hiërargiese frasegebaseerde vertaling en die ander twee op frasegebaseerde masjienvertaalsisteme) met dieselfde afrigtingsdata af te rig. Die drie sisteme is toe gebruik om die oorblywende data (brontaal) te vertaal. Die mate waartoe die drie vertalings verskil, word bepaal deur telkens een van die vertalings met die ander twee as verwysingsvertalings te vergelyk deur van 'n outomatiese evaluasiemetriek gebruik te maak (Mandal *et al.*, 2008:262). 'n Lineêre regressiefunksie word dan afgerig om die ware vertalingsprestasie op nuwe, onvertaalde brontaaldata te voorspel (Mandal *et al.*,

2008:262). Die TER-evaluasiemetriek (*Translation Error | Edit Rate*) (Snover *et al.*, 2006:223) word deur die funksie voorspel (Mandal *et al.*, 2008:263). Die sinne met die hoogste voorspelde TER word vir vertaling en byvoeging tot die afrigtingskorpus gekies (Mandal *et al.*, 2008:263).

Met die tweede metode is daar gepoog om sinne te kies wat selde in die afrigtingsdata voorkom, maar wel baie in die res van die data voorkom (en dus nie 'n buitengewone sin is nie). Hiervoor word twee verskillende taalmodelle gebruik, een wat afgerig is op die brontaalsinne in die tweetalige afrigtingskorpus en die ander wat afgerig is op alle beskikbare brontaalsinne (Mandal *et al.*, 2008:262). Die sinne word gekies deur van 'n perpleksiteitverhouding gebruik te maak wat laag sal wees indien die sin in beide korpusse (eentalig en tweetalig) skaars is. Indien die perpleksiteitverhouding dus laag is, word die sin as buitengewoon beskou en sal dit nie gekies word nie (Mandal *et al.*, 2008:262). Indien die verhouding egter hoog is (die sin is skaars in die tweetalige data en frekvent in die eentalige data) word die sin gekies om handmatig vertaal te word vir toevoeging tot die afrigtingsdata (Mandal *et al.*, 2008:262).

Die hibriede sisteem (kombinasie van die twee metodes) is geskep deur ewe veel sinne deur elkeen van die twee metodes te selekteer vir toevoeging tot die afrigtingsdata (Mandal *et al.*, 2008:262). Daar is met elkeen van die drie metodes verskillende masjienvertaalsisteme afgerig, waar daar telkens 4%, 10% en 20% van die moontlike bykomende data by die oorspronklike tweetalige afrigtingsdata gevoeg is. Daar is ook vir elkeen van die drie datahoeveelhede wat toegevoeg kan word 'n ewekansige basislyn afgerig en 'n sisteem is met alle beskikbare data afgerig. Mandal *et al.* (2008:262) het gevind dat die dataselekteringsmetodes oor die algemeen met die toevoeging van slegs 20% bykomende data beter resultate toon as met die toevoeging van al die beskikbare data.

2.4. SAMEVATTING

In hierdie hoofstuk word daar eerstens 'n oorsig gegee oor die algemene teorie en geskiedenis van masjienvertaling. Verskillende tipes masjienvertaling word kortliks genoem en statistiese masjienvertaling (insluitende frasegebaseerde statistiese masjienvertaling) word in meer diepte bespreek. In die res van die hoofstuk word verskeie studies bespreek waarin daar met datamanipulering en -selektering geëksperimenteer is.

Daar is nog redelik min navorsing oor masjienvertaling vir die taalpaar Engels-Afrikaans gepubliseer. Slegs drie studies wat oor hierdie onderwerp handel is in die literatuur gevind. Engelbrecht en Schultz (2005:1) het 'n Engels-Afrikaanse spraak-tot-spraak vertaalsisteem gemaak met statistiese masjienvertaling as die vertaalkomponent. Die *EtsaTrans*-sisteem is ontwikkel vir integrasie by die *Parnassus*-stelsel om die vertaling van notules te fasiliteer (Snyman *et al.*, 2007:225), terwyl Kato en Barnard (2007:1) ook 'n Engels-Afrikaanse masjienvertaalsisteem ontwikkel het as deel van hulle navorsing oor dataselektering vir hulpbronarm tale.

Daar word ook verwys na navorsing wat gedoen is oor die aanpassing van beskikbare data met die oog op die verbetering van die gehalte van masjienvertaling. Finch en Sumita (2008:113) het die afrigtingsdata aangepas deur kopieë van sekere sinne tot die data te voeg. Lu *et al.* (2007:343) het navorsing gedoen oor die aanpassing van afrigtingsdata om die resulterende masjienvertaalsisteem meer domeinspesifiek te maak en sodoende die gehalte van die masjienvertaalsisteem te verbeter. Dit het eksperimente vir beide bekende toetstekse/domein en onbekende toetstekse/domein ingesluit.

In die laaste gedeelte van hierdie hoofstuk word daar gekyk na navorsing wat handel oor die selektering van bykomende data om tot die afrigtingskorpus by te voeg. Haffari *et al.* (2009:415) het verskeie dataselekteringsmetodes, gebaseer op die beginsel van aktiewe leer, gebruik om bykomende data te kies om deur mense te laat vertaal vir toevoeging tot die afrigtingsdata. Kato en Barnard (2007:1) het 'n kombinasie van beide aktiewe leer en semi-gekontroleerde leer gebruik om bykomende afrigtingsdata vir statistiese masjienvertaling te selekteer. Hulle studie is uitgevoer op Suid-Afrikaanse taalpare en sluit ook 'n sisteem vir Engels-Afrikaans in. Mandal *et al.* (2008:261) het die onenigheid tussen verskillende masjienvertaalsisteme en die perpleksiteitverhouding tussen data gebruik om bykomende afrigtingsdata te selekteer. Hierdie twee metodes is ook gekombineer om 'n hibriede sisteem te vorm.

3. DATASELEKTERING- EN MANIPULASIE-EKSPERIMENTE

3.1. INLEIDING

In die vorige hoofstuk word daar eerstens 'n oorsig oor die algemene teorie en geskiedenis van masjienvertaling gegee. Verskillende tipes masjienvertaling word kortliks genoem en statistiese masjienvertaling word in diepte bespreek. Daar word spesifiek gekyk na frasegebaseerde statistiese masjienvertaling en die *Moses-gereedskapstel* (Koehn *et al.*, 2007:177). Daar is ook 'n oorsig oor studies wat handel oor masjienvertaling (Engelbrecht & Schultz, 2005:3; Snyman *et al.*, 2007:225; Kato & Barnard, 2007:2) vir die taalpaar Engels-Afrikaans gegee. Laastens word ondersoeke (Finch & Sumita, 2008:113; Lu *et al.*, 2007:343) oor die verbetering van masjienvertaling deur die aanpassing van die afrigtingsdata en ondersoeke (Haffari *et al.*, 2009:415; Kato & Barnard, 2007:1; Mandal *et al.*, 2008:261) bespreek wat handel oor die verbetering van masjienvertaling deur spesifieke dataselekteringsmetodes.

Hierdie hoofstuk bied 'n oorsig oor die praktiese eksperimente wat in hierdie studie uitgevoer is en 'n beskrywing van die data en sagteware wat in die verskeie eksperimente gebruik is, word gegee. Die besonderhede van die skep van 'n standaardmasjienvertaalsisteem en basislynmasjienvertaalsisteem, asook die afsonderlike eksperimente om op die prestasie van hierdie twee sisteme te verbeter, word verskaf. Die resultate van die outomatiese evaluasie van die verskillende eksperimente word ook voorgelê.

3.2. METODE VAN ONDERSOEK

Die eerste navorsingsdoelstelling van hierdie studie is om te bepaal of die toevoeging van sorgvuldig gekose data tot vinniger en groter verbeteringe in die kwaliteit van 'n statistiese masjienvertaalsisteem sal lei as die byvoeging van ewekansig geselekteerde data. Die voorvereistes hiervoor is 'n standaardsisteem wat verbeter moet word deur die toevoeging van data, asook 'n korpus wat bestaan uit 'n hoeveelheid belynde tekseenhede waaruit bepaalde tekseenhede geselekteer kan word vir byvoeging tot die standaardsisteem in 'n

poging om die standaardsisteem se prestasie te verbeter. Vir die taalpaar Engels-Afrikaans is daar gedurende die Autshumato-projek reeds 'n tweetalige, parallelle korpus versamel wat vir hierdie doel gebruik kan word; in die geval van ander taalpare sou daar moontlik slegs eentalige data beskikbaar wees wat dan vertaal moet word vir byvoeging tot die afrigtingsdata. In hierdie geval bestaan die basislynsisteem uit die standaardsisteem waarby 'n aantal ewekansig geselekteerde tekseenhede gevoeg is. Die basislyn vir dataselektering poog dus om die standaardsisteem se prestasie te verbeter deur die toevoeging van ewekansige data, terwyl die res van die dataselekteringseksperimente oor meer sorgvuldige selektering van afrigtingsdata handel.

Ten einde die resultate vergelykbaar te hou, is die hoeveelheid data wat tot die standaardsisteem bygevoeg word telkens konstant gehou. Die standaardsisteem waarop verbeter moet word, is op afrigtingsdata van ongeveer 100 000 woorde afgerig. In die eksperimente wat oor die selektering van bykomende afrigtingsdata handel word daar in totaal ongeveer 500 000 bykomende woorde afrigtingsdata tot die standaardsisteem bygevoeg. Dit word in tien inkremente van ongeveer 50 000 woorde elk gedoen sodat 'n leerkurwe vir elke metode bepaal kan word. Die woordtellings word telkens op die Engelse gedeelte van die data gedoen en alle leestekens word ingesluit.

Die tweede navorsingsdoelstelling van hierdie studie is om die effek van die aanpassing van die gewigte van sekere parallelle tekseenhede in die afrigtingsdata te ondersoek. Hiervoor is die standaardsisteem se datastel gebruik en aangepas. Verskillende toetse is gedoen om te bepaal wat die effek van die gewigsaanpassing van sekere gedeeltes van die afrigtingsdata is. Daar is ook gekyk na die verskille tussen die effek van die verdubbeling en verdriedubbeling van die spesifieke tekseenhede in die afrigtingsdata, wat gedoen word om die gewigte van die betrokke gedeeltes te verhoog. Hierdie benadering is getoets op beide die parallelle afrigtingsdata en die taalmodeldata.

Die finale doelstelling van die studie is om te bepaal of die kombinerings van verskillende tegnieke (naamlik dataselekteringsmetodes en die aanpassing van gewigte) en die toepassing daarvan op beide taalmodeldata en afrigtingsdata tot 'n groter verbetering in masjienvertaling lei as enige van die afsonderlike metodes. Ten einde hierdie ondersoek te doen is die resultate van al die verskillende eksperimente vergelyk en die eksperimente wat die heel beste resultate behaal het, is gekombineer.

Dit is baie belangrik om die verskillende masjienvertaalsisteme te kan evalueer ten einde te bepaal wat die effek van elkeen van die eksperimente op die kwaliteit van masjienvertaling is. Die masjienvertaalsisteme is outomaties met die NIST- en BLEU-metrieke geëvalueer. Alhoewel menslike evaluasie waarskynlik 'n beter aanduiding sou gee van die presiese verbeteringe wat deur die eksperimente veroorsaak is, sou dit in die geval van hierdie studie waar daar baie eksperimente (meer as 100 afgerigte masjienvertaalsisteme) uitgevoer is, onprakties wees. Die evaluasiemetrieke word in Afdeling 3.5 bespreek.

3.3. DATA

Soos reeds genoem, is een van die belangrikste vereistes vir die afrigting van 'n statistiese masjienvertaalsisteem die beskikbaarheid van data. Die prestasie van 'n statistiese masjienvertaalsisteem verbeter gewoonlik met die toevoeging van data, omdat daar dan meer data beskikbaar is vir die outomatiese onttrekking van parallelle frases. Daar is dus 'n groter kans dat skaarser woorde (en hul vertalings) in 'n baie groot korpus sal voorkom en daarom kan daar dan aangeneem word dat die herroeping (*recall*) van 'n masjienvertaalsisteem sal toeneem soos wat die afrigtingsdata vermeerder. In hierdie geval verwys herroeping na die hoeveelheid woorde in die brontaal wat na die teikentaal vertaal is. Daar is vir die taalpaar Engels-Afrikaans 'n redelike groot korpus met parallelle data beskikbaar (in vergelyking met wat vir ander Suid-Afrikaanse taalpare beskikbaar is). Aangesien hierdie eksperimente juis oor die selektering van data handel, en die resultate ook vir taalpare met minder data toepasbaar moet wees, word slegs 'n gedeelte van die totale beskikbare datastel vir elke eksperiment gebruik.

Vir die afrigting van 'n statistiese masjienvertaalsisteem word twee tipes data benodig, naamlik eentalige data in die teikentaal en tweetalige data wat op sinsvlak belyn is. Die eentalige data in die tekentaal (in hierdie geval Afrikaans) word benodig vir die afrigting van die taalmodelle (Afdeling 2.2.2). Die parallelle data word gebruik vir die afrigting van die vertaalmodel (Afdeling 2.2.2). Die parallelle data wat vir hierdie studie gebruik is, is versamel as deel van die Autshumato-projek (Groenewald & Fourie, 2009:190).

Die parallelle data wat vir hierdie studie gebruik is, is afkomstig van 'n verskeidenheid bronne. 'n Groot hoeveelheid van die data is regeringsdokumente en ook data wat vanaf

die regering se webwerf (www.gov.za) verkry is. Daar is ook data vanaf vryskutvertalers en vertalingsmaatskappye in die korpus. Die data bestaan uit volledige sinne, frases en losstaande woorde en daarom word daar telkens in hierdie studie verwys na “parallele tekseenhede” i.p.v. “parallele sinne”. Sommige gedeeltes van die data is handmatig belyn, terwyl die res van die data outomaties met behulp van Moore (2002:135) se outomatiese belyningsalgoritme belyn is. Vir die eentalige Afrikaanse data is die Media24-korpus (Pharos Dictionaries, 2006) gebruik. Die Media24-korpus is ’n baie groot korpus (sien Tabel 1 vir besonderhede) wat hoofsaaklik uit nuusberigte bestaan.

Tabelle 1 en 2 gee ’n oorsig oor die samestelling van die afrigtingsdata (beide eentalige data en tweetalige data wat uit parallele tekseenhede bestaan). Vir die eentalige data word die aantal sinne, sowel as die aantal woorde en aantal unieke woorde gegee. Vir die tweetalige data word die aantal parallele tekseenhede sowel as die aantal woorde en aantal unieke woorde per taal gegee. Alle tellings sluit leestekens in.

	Sinne	Woorde	Unieke Woorde
Media24-korpus	5 572 462	130 517 117	1 597 561

Tabel 1: Samestelling van die eentalige data

Taal	Parallele Tekseenhede	Woorde	Unieke Woorde
Afrikaans	956 107	7 663 818	374 238
Engels	956 107	7 851 722	218 704

Tabel 2: Samestelling van die tweetalige data

3.4. PROGRAMMATUUR

Vir die doel van hierdie studie is ’n groot aantal masjienvertaalsisteme afgerig. Hiervoor is *Moses*, ’n oopbron, statistiese masjienvertalinggereedskapstel (Koehn *et al.*, 2007:177) gebruik. *Moses* neem tweetalige afrigtingsdata en ’n aantal taalmodelle as toevoer en gebruik dit om ’n frasegebaseerde statistiese masjienvertaalsisteem te skep.

Die *SRI Language Modeling Toolkit* (Stolcke, 2002:901) is gebruik vir die afrigting van die taalmodelle. Die program gebruik 'n eentalige korpus in die teikentaal en skep n -gram taalmodelle daarmee. Vir die doel van die studie is drie verskillende taalmodelle gebruik. Al drie die taalmodelle is met dieselfde data afgerig, maar die n -gram tellings in die taalmodelle verskil, aangesien daar met 3-, 4- en 5-gram taalmodelle gewerk is.

Om te verseker dat die resultate van die eksperimente slegs 'n aanduiding van die doeltreffendheid van die verskeie metodes van dataselektering gee, is al die ander instellings wat moontlik die prestasie van die masjienvertaalsisteme kon beïnvloed konstant gehou. Voorbeelde van hierdie instellings is die keuse van die gladstrykingsmetode (Stolcke, 2002:902) wat gebruik word met die afrigting van die taalmodelle en die aantal n -gramme (Stolcke, 2002:902) waarmee die taalmodelle afgerig word. Dieselfde taalmodelle word vir al die eksperimente gebruik, behalwe waar die eksperimente juis oor die verandering van die taalmodeldata handel (sien Afdeling 3.11). Die instellings wat ter sprake is by die afrigting van die masjienvertaalsisteme word ook konstant gehou. Hierdie instellings sluit die keuse van die woordbelyningheuristiek (Koehn, 2010:158) en tipe herrangskikkingsmodel (Koehn, 2010:90) in.

3.5. BEREKENING VAN RESULTATE

Die uitvoering van die verskillende eksperimente wat in hierdie hoofstuk aangebied word, het meegebring dat 'n groot aantal verskillende masjienvertaalsisteme afgerig is. Elkeen van hierdie masjienvertaalsisteme se afvoer moes natuurlik ook geëvalueer word voordat enige sinvolle gevolgtrekkings oor die resultate van die eksperimente gemaak kon word. Omdat die menslike evaluasie van masjienvertalingsafvoer 'n tydrowende en duur proses kan wees, was menslike evaluering nie 'n moontlikheid in hierdie studie nie. Daarom is daar besluit om twee outomatiese evaluasiemetrieke te gebruik, naamlik die *Bilingual Evaluation Understudy* (BLEU) (Papineni *et al.*, 2002:311) en *NIST* (Doddington, 2002:138). Die NIST-metrik se naam is afgelei van die US National Institute of Standards and Technology.

BLEU is 'n outomatiese evaluasiemetrik wat die afvoer van 'n masjienvertaalsisteem met 'n aantal verwysingsvertalings (vertaling deur mense gedoen) van dieselfde teks vergelyk (Papineni *et al.*, 2002:311). Die basiese beginsel van BLEU is dat 'n "goeie"

masjienvertaling tot 'n groot mate met 'n verwysingsvertaling (deur 'n persoon vertaal) sal ooreenstem. Die mate van ooreenstemming kan gemeet word deur te kyk na die aantal oorvleuelende n -gramme. Die oorvleuelende 1-gramme tussen die masjienvertaling en die verwysingsvertaling gee 'n aanduiding van hoe goed die inhoud van die masjienvertaling die inhoud van die oorspronklike teks weerspieël. Die langer n -gramme gee 'n aanduiding van die vloeiendheid van die teks. BLEU ken 'n telling tussen nul en een toe as aanduiding van die kwaliteit van die masjienvertaling. 'n Telling van 0 is 'n aanduiding van 'n totaal onaanvaarbare vertaling en 1 'n aanduiding van 'n perfekte vertaling ('n masjienvertaling wat 100% ooreenstem met 'n menslike vertaling). Aangesien daar so baie verskillende, maar tog korrekte woordkeuses bestaan, behaal selfs menslike vertalings baie selde 'n BLEU-telling van een. Dit is ook die rede dat daar meer as een verwysingsvertaling gebruik moet word; hoe meer verskillende verwysingsvertalings daar beskikbaar is, hoe groter is die kans dat alle moontlike "korrekte" vertalings in die verwysingsvertalings sal voorkom. Indien daar slegs een verwysingsvertaling is, bestaan die moontlikheid dat korrekte vertalings lae tellings kan verkry omdat die verwysingsvertaling ander woordkeuses bevat.

Die NIST outomatiese evaluasiemetriek is op dieselfde beginsels as BLEU gebaseer (Doddington, 2002:138). Die evaluasiemetriek gebruik ook 'n vertaalde toetsteks met 'n aantal verwysingsvertalings en bepaal die kwaliteit van die vertaling deur te kyk na die oorvleuelende n -gramme. Daar is egter ook 'n aantal verskille tussen die twee evaluasiemetrieke. Waar BLEU 'n geometriese gemiddeld (sien tellings as afhanklik van mekaar) van die n -gram tellings gebruik, word 'n rekenkundige gemiddeld (sien tellings as onafhanklik van mekaar) vir die NIST-telling gebruik. Die NIST-formule ken ook meer gewig toe aan die n -gramme wat meer inligting in die sin dra (die n -gramme wat minder frekvent in die verwysingsvertalings voorkom). Laastens word die impak van klein variasies in die lengte van die vertalings op die NIST-telling geminimaliseer deur die bondigheidsboete (*brevity penalty*) te verander. Die doel van die bondigheidsboete is om te verseker dat vertalings nie te kort is nie aangesien dit moontlik sou wees om 'n vertaling te skep met net 'n paar woorde wat almal korrek is en dan die res van die sin weg te laat. Volgens Doddington (2002:141) toon die NIST-telling in verskeie gevalle 'n hoër korrelasie met menslike oordeel as die BLEU-telling. Die NIST-telling het geen absolute skaal nie, maar hoe hoër die telling, hoe beter is die vertaling.

Beide die evaluasiemetrieke vereis 'n toetstek in die brontaal, 'n vertaling van die toetstek deur die sisteem wat geëvalueer word en 'n aantal verwysingsvertalings. Vir hierdie studie is 'n toetstek van 200 sinne gebruik. Hierdie teks vorm deel van die METIS-II projek (Dirix *et al*, 2005:43) se evaluasietekste. Vier verwysingsvertalings deur professionele geakkrediteerde vertalers is gebruik.

Elkeen van die masjienvertaalsisteme wat gedurende hierdie studie afgerig is, is gebruik om die toetstek te vertaal. Daar is geen verandering aan die afvoer aangebring voor die BLEU- en NIST-tellings bereken is nie. Die evaluasieresultate word telkens saam met die betrokke eksperimente gegee.

3.6. STANDAARDTAALMODEL

Deel van hierdie studie is om te bepaal watter invloed die selektering van taalmodeldata op die kwaliteit van die uiteindelijke masjienvertaalsisteem het. Alle eksperimente, behalwe dié wat spesifiek die effek van geselekteerde taalmodeldata ondersoek, gebruik egter dieselfde standaardtaalmodelle. Aangesien daar in 'n latere eksperiment gepoog word om die kwaliteit van die masjienvertaling deur die selektering van taalmodeldata te verbeter, word slegs helfte van die beskikbare data gebruik om sodoende genoeg data te gebruik maar ook ruimte vir keuse in verdere eksperimente te laat. Vir die skep van die standaardtaalmodelle is die helfte van die sinne in die Media24-korpus ewekansig gekies. In die eksperiment wat handel oor die selektering van taalmodeldata word dieselfde hoeveelheid data op ander wyses uit die korpus gekies word. Daar is geen poging aangewend om die data te sorteer of skoon te maak deur die verwydering van vreemde karakters of teksgedeeltes in ander tale nie. Die ewekansig gekose afrigtingsdata is gebruik om 3-, 4- en 5-gram taalmodelle mee af te rig. Die drie taalmodelle word saam vir die skep van vertaalmodelle en vir die vertaling van tekste gebruik.

3.7. STANDAARDSISTEEM

Die standaardsisteem is die masjienvertaalsisteem waarop alle ander sisteme (insluitende die basislyn) probeer verbeter deur die toevoeging van bykomende afrigtingsdata, die verandering van datagewigte of die verandering van taalmodeldata. Die standaardsisteem

is afgerig op 'n redelik klein korpus, bestaande uit 100 000 Engelse woorde met vertalings, wat ewekansig geselekteer is. Die geselekteerde data is toe ook uit die afrigtingskorpus verwyder om te verseker dat dit nie weer tydens die dataselekteringseksperimente by die standaardsisteem se afrigtingsdata gevoeg kan word nie.

Daar is geen annotasie tot die data toegevoeg nie en ook geen poging aangewend om die data handmatig deur te gaan of skoon te maak deur die verwydering van vreemde karakters of teksgedeeltes in ander tale nie. Die standaardmasjienvertaalsisteem vorm die basis van al die eksperimente wat volg. Vir elkeen van die eksperimente is bykomende data gekies en tot die afrigtingsdata van die standaardmasjienvertaalsisteem toegevoeg, waarna 'n nuwe masjienvertaalsisteem met die vergrote datastel afgerig is. Die drie standaardtaalmodelle wat in Afdeling 3.6 bespreek word, is tydens die afrigting van die standaardsisteem gebruik.

Omdat al die eksperimente wat in hierdie studie uitgevoer word die doel het om op die prestasie van die standaardsisteem te verbeter, vorm die standaardsisteem dus 'n beginpunt vir al die ander eksperimente. Elkeen van die grafieke in die res van die hoofstuk bevat ook die NIST- of BLEU-telling (afhangende van die grafiek) van die standaardsisteem as die eerste punt op die leerkurwe, of as 'n vergelykende punt. Die resultate vir die standaardsisteem word in Tabel 3 gegee.

	NIST	BLEU
Standaardsisteem	6.0100	0.2496

Tabel 3: NIST- en BLEU-tellings van die standaardsisteem

3.8. BASISLYN

Die vorige afdeling bespreek die samestelling van die afrigtingsdata van die standaardmasjienvertaalsisteem wat as beginpunt vir al die dataselekteringseksperimente dien. Die volgende stap was om 'n basislynsisteem te skep wat 'n aanduiding bied van die verbetering in resultate wat verwag kan word indien die bykomende data ewekansig gekies en tot die afrigtingsdata van die standaardsisteem gevoeg word. Die resultate van hierdie basislynsisteem is toe vergelyk met dié van ander dataselekteringseksperimente om te bepaal of die sorgvuldige selektering van bykomende afrigtingsdata werklik 'n vinniger verbetering in masjienvertalingkwaliteit kan lewer. Die vergelyking is gedoen deur

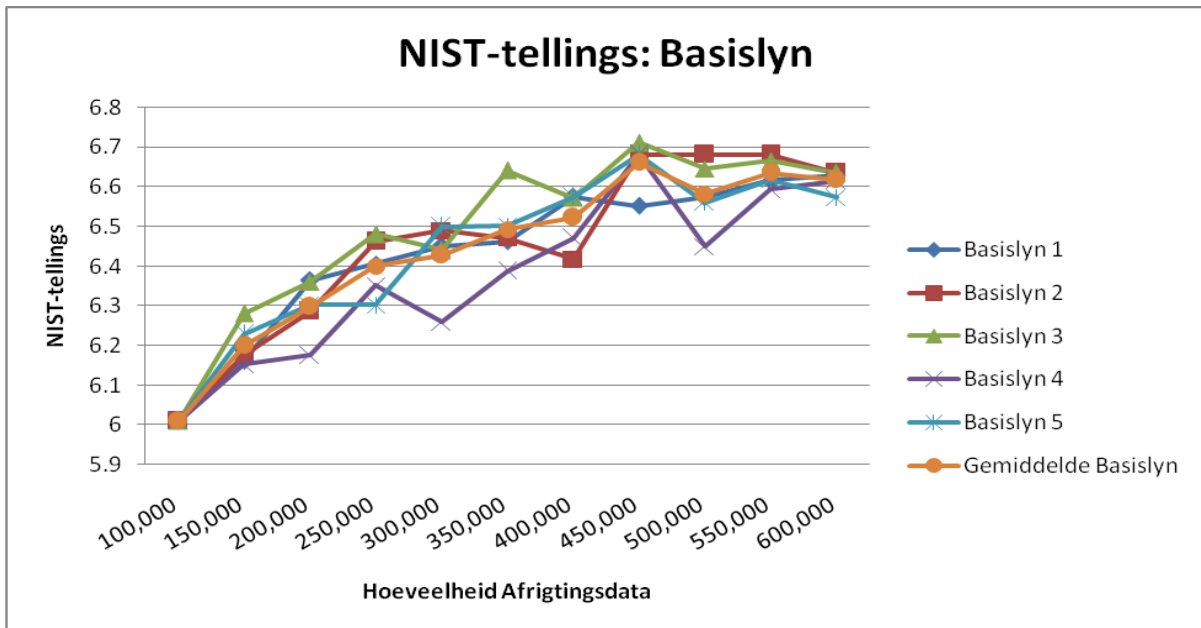
die basislyn saam met die ander sisteme grafies voor te stel en te kyk na die tempo waarteen die kwaliteit van die masjienvertaling verbeter (aangedui deur die helling van die leerkurwe).

Die basislynsisteem vir dataselektering is afgerig met bykomende afrigtingsdata met 'n totaal van 500 000 woorde. Soos met die standaardsisteem wat in Afdeling 3.7 bespreek word, is die woordtellings op die Engelse deel van die afrigtingsdata bereken en alle leestekens is by die telling ingesluit. Die bykomende data is ook, soos by die standaardsisteem, ewekansig geselekteer. Die toevoeging van die bykomende data is in tien stappe van ongeveer 50 000 woorde elk gedoen sodat 'n leerkurwe geteken kon word.

Om te verseker dat die ewekansige gekose data nie toevallig besondere “goeie” data is nie, is die proses van die ewekansige dataselektering vyf keer herhaal. Vir elkeen van die vyf eksperimente is die bykomende data ewekansig en onafhanklik van die ander vier eksperimente geselekteer. Die gemiddeld van die vyf basislynsisteme se resultate is geneem as die basislyn vir dataselektering. Tabel 4 toon die NIST-tellings vir die vyf basislynsisteme asook die gemiddelde NIST-telling, terwyl Tabel 5 presies dieselfde inligting in terme van BLEU-tellings weergee. Die leerkurwes van die sisteme word in Figuur 2 (volgens NIST) en Figuur 3 (volgens BLEU) grafies voorgestel.

	Iterasie 1	Iterasie 2	Iterasie 3	Iterasie 4	Iterasie 5	Iterasie 6	Iterasie 7	Iterasie 8	Iterasie 9	Iterasie 10
Basislyn 1	6.1614	6.3636	6.4057	6.4505	6.4628	6.5747	6.5519	6.5732	6.6169	6.6276
Basislyn 2	6.1761	6.2871	6.4617	6.4894	6.4712	6.4182	6.6811	6.6811	6.6811	6.6352
Basislyn 3	6.2810	6.3602	6.4805	6.4401	6.6408	6.5720	6.7112	6.6458	6.6651	6.6352
Basislyn 4	6.1530	6.1747	6.3516	6.2589	6.3868	6.4716	6.6811	6.4499	6.5933	6.6157
Basislyn 5	6.2286	6.3023	6.3021	6.4992	6.5008	6.5745	6.6811	6.5611	6.6177	6.5736
Gemiddeld	6.2000	6.2976	6.4003	6.4276	6.4925	6.5222	6.6613	6.5822	6.6348	6.6175

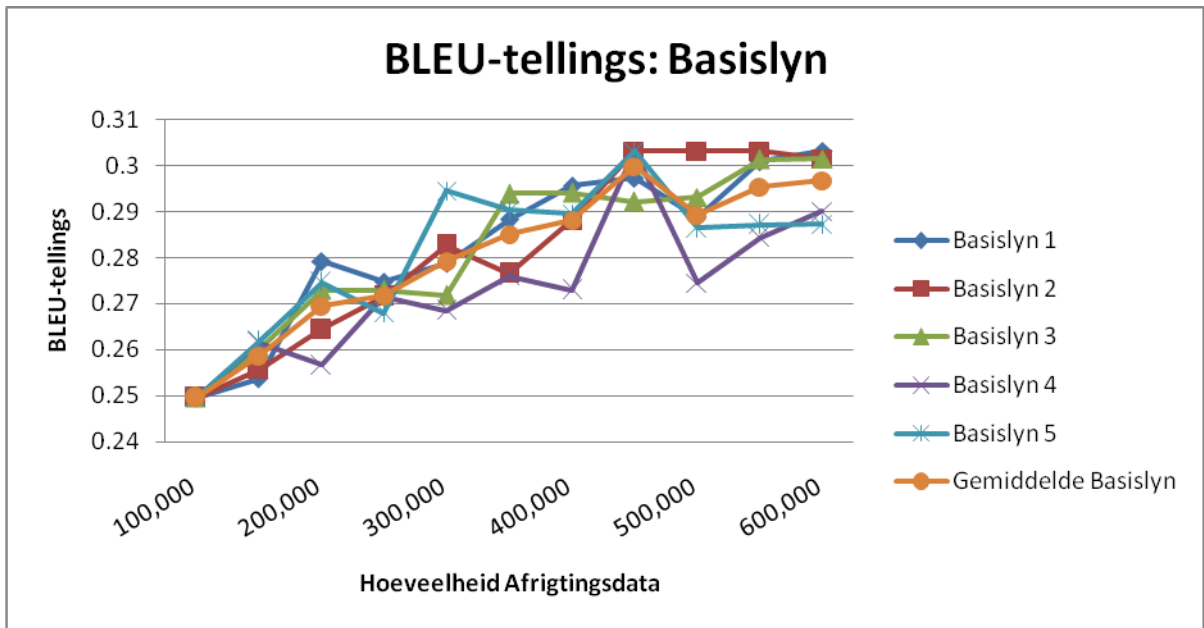
Tabel 4: NIST-tellings van die basislynsisteme



Figuur 2: Leerkurwes van die basislynstelsime (volgens NIST-tellings)

	Iterasie 1	Iterasie 2	Iterasie 3	Iterasie 4	Iterasie 5	Iterasie 6	Iterasie 7	Iterasie 8	Iterasie 9	Iterasie 10
Basislyn 1	0.2537	0.2792	0.2748	0.2788	0.2883	0.2957	0.2973	0.2888	0.3009	0.3032
Basislyn 2	0.2557	0.2644	0.2715	0.2827	0.2766	0.2881	0.3032	0.3032	0.3032	0.3015
Basislyn 3	0.2601	0.2729	0.2729	0.2718	0.2939	0.2941	0.2922	0.2931	0.3013	0.3015
Basislyn 4	0.2616	0.2566	0.2716	0.2685	0.2760	0.2730	0.3032	0.2745	0.2844	0.2901
Basislyn 5	0.2618	0.2746	0.2680	0.2945	0.2904	0.2895	0.3032	0.2864	0.2871	0.2873
Gemiddeld	0.2586	0.2695	0.2718	0.2793	0.2850	0.2881	0.2998	0.2892	0.2954	0.2967

Tabel 5: BLEU-tellings van die basislynstelsime



Figuur 3: Leerkurwes van die basislynsisteme (volgens BLEU-tellings)

3.9. SELEKTERING VAN BYKOMENDE AFRIGTINGSDATA

Die eerste vier eksperimente in hierdie afdeling handel oor die selektering van bykomende tweetalige afrigtingsdata. Die data is in tien inkremente van 50 000 woorde geselekteer sodat 'n leerkurwe vir elke dataselekteringsmetode bepaal kon word. Die uiteindelijke sisteme is dus telkens afgerig met 'n korpus van ongeveer 600 000 woorde (bereken op die Engelse helfte van die data).

In Eksperimente 1 en 2 is die vertrouettellings wat deur Moses gedurende die vertaalproses toegeken word om bykomende data te kies gebruik. Die verskil tussen die twee eksperimente is dat Eksperiment 1 slegs met langer tekseenhede (langer as vyf woorde) uitgevoer is en Eksperiment 2 met alle beskikbare data. Die eksperimente word volledig in Afdeling 3.9.1 bespreek.

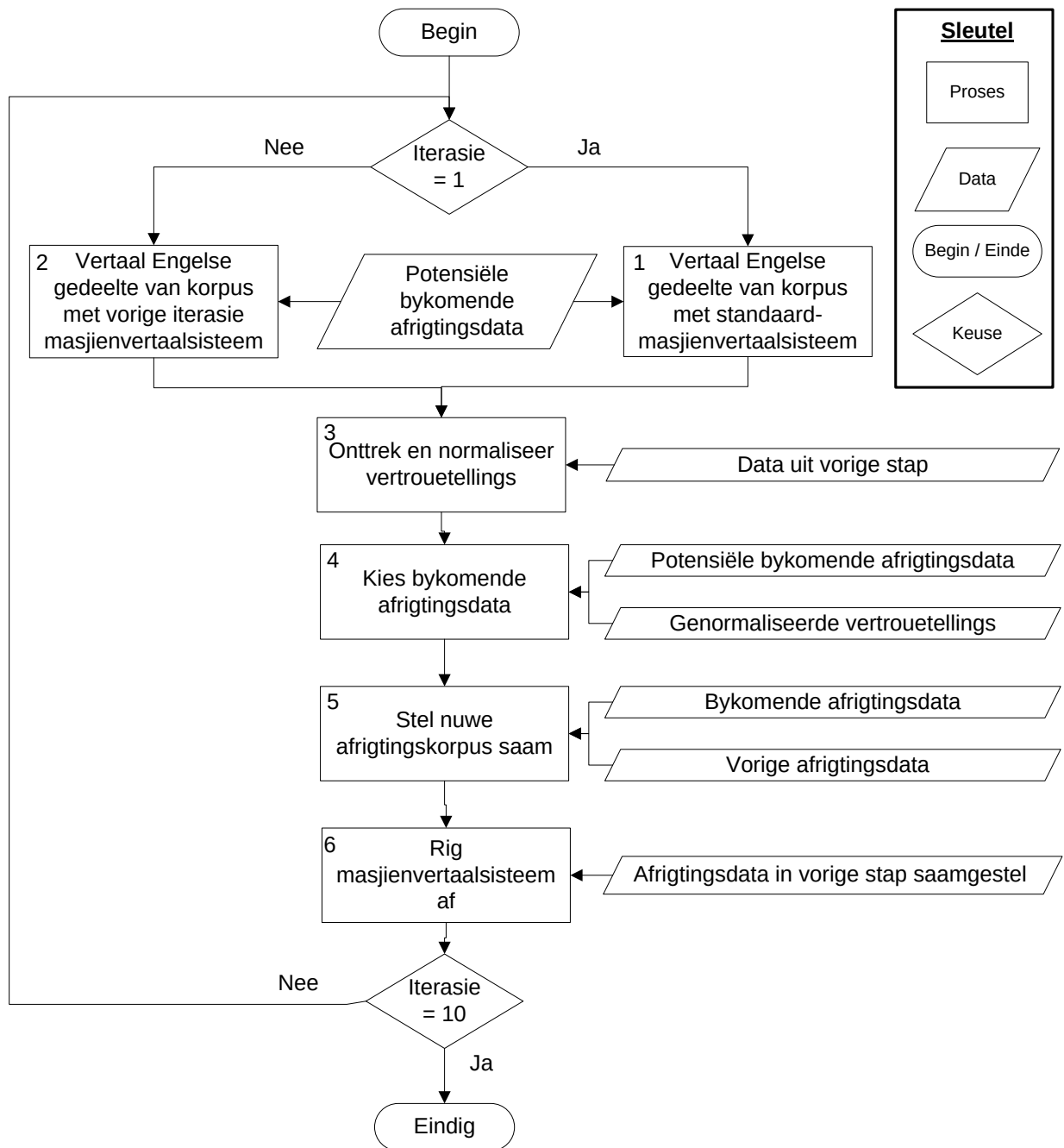
In Eksperimente 3 en 4 is bykomende data op grond van die frekwensietellings van die n -gramme in die Engelse tekseenhede geselekteer. Dit is bepaal deur 'n formule wat lyste van 1, 2, 3, 4 en 5-gramme gebruik, waar die n -gramme volgens frekwensie gesorteer is. Die frekwensies is met behulp van die *Europarl*-korpus (Koehn, 2005:79) bepaal. Die tekseenhede met die hoogste frekwensies, en dus die mees frekwente inhoud, is telkens

as bykomende data gekies. In Eksperiment 4 het slegs “ongesiene” n -gramme (n -gramme wat nie reeds in die afrigtingsdata voorkom nie) tot die frekwensietellings bygedra. Die twee eksperimente word volledig in Afdeling 3.9.2 bespreek.

3.9.1. SELEKTERING VAN DATA MET VERTROUETELLINGS (EKSPERIMENTE 1 EN 2)

Die *Moses* statistiese masjienvertalingsgereedskapstel (Koehn *et al.*, 2007:178) kan 'n vertrouetelling (*confidence score*) aan elke vertaalde tekseenheid toeken (Koehn, 2010:22) wat aandui hoe seker (in statistiese terme) die sisteem is dat die betrokke vertaling korrek is. Hoe moeiliker dit dus is om 'n tekseenheid te vertaal, hoe laer is die vertrouetelling. In 'n geval waar die tekseenheid uit frases bestaan wat redelik frekvent in die afrigtingsdata voorkom, sal die vertrouetelling hoog wees, want al die frases sal in die frasetabel voorkom en daarom sal dit dan “maklik” wees om akkuraat te vertaal. In 'n geval waar 'n tekseenheid baie ongesiene data bevat (data wat glad nie in die afrigtingsdata voorgekom het nie), sal *Moses* se vertrouetelling baie laag wees.

Daar kan dus onderskei word tussen data wat maklik vertaal kan word (en waaraan daar 'n hoë vertrouetelling toegeken word) en data wat moeiliker is om te vertaal (en wat dus 'n lae vertrouetelling het). Die data wat moeiliker is om te vertaal is dan juis die data waarvan bykomende voorbeelde in die afrigtingsdata benodig word. Dit maak dus sin om tekseenhede met 'n lae vertrouetelling tot die afrigtingsdata by te voeg, want dit veroorsaak dat die sisteem afgerig word om “problematiese” data beter te kan vertaal. In die geval van enkelwoorde en baie kort frases sal die woord/frase óf bekend, óf onbekend wees. Dit kan daartoe lei dat baie van die enkelwoorde/kort frases juis óf baie hoë, óf baie lae vertrouetellings het. Indien vertrouetellings gebruik word om nuwe afrigtingsdata te kies, kan dit moontlik gebeur dat 'n nuwe iterasie se bykomende data so te sê heeltemal uit woorde of kort frases bestaan. Om die effek hiervan te ondersoek is hierdie eksperiment twee keer herhaal. Die eerste keer is alle kort frases en losstaande woorde uit die potensiële bykomende afrigtingsdata verwyder (Eksperiment 1) en die tweede keer is alle beskikbare data gebruik (Eksperiment 2).



Figuur 4: Vloediagram van Eksperimente 1 en 2

In beide Eksperiment 1 (langer tekseenhede) en Eksperiment 2 (alle data) is dieselfde prosedure toegepas. Die vloediagram in Figuur 3 toon die stappe wat in die twee eksperimente gevolg is. Die standaardsisteem is as beginpunt geneem en toe gebruik om die Engelse gedeelte van die potensiële afrigtingsdata mee te vertaal (Stap 1) om die vertrouettelings vir elkeen van die tekseenhede te verkry. Die tellings is genormaliseer

(Stap 3) deur dit met die aantal woorde in die tekseenheid te deel om te verseker dat die lengte van die tekseenheid nie 'n invloed op die grootte van die vertrouelling het nie. Die 50 000 tekseenhede met die heel laagste tellings is toe uit die korpus onttrek (Stap 4) en saam met hul oorspronklike vertalings by die standaardsisteem se afrigtingsdata gevoeg (Stap 5). Hierdie vergrote afrigtingskorpus is toe gebruik om 'n nuwe masjienvertaalsisteem af te rig (Stap 6) wat toe weer gebruik is om al die oorblywende data te vertaal (Stap 2). Hierdie proses is 10 keer herhaal totdat 'n totaal van ongeveer 500 000 woorde tot die afrigtingsdata van die standaardsisteem bygevoeg is.

3.9.2. SELEKTERING VAN DATA MET FREKWENSIE TELLINGS (EKSPERIMENT 3 EN 4)

Studies toon dat 'n baie klein gedeelte van 'n taal se woordeskat 'n baie groot persentasie van die algemene taalgebruik dek (Nation, 2001:13). Die oorgrote meerderheid van die woordeskat van 'n taal word dus selde gebruik. Volgens Nation (2001:15) kan so min as 5 000 woorde reeds 89.4% van Engelse taalgebruik dek. Om dus die kwaliteit van masjienvertaling so vinnig as moontlik te verhoog moet die tekseenhede wat die meeste frekwente woorde bevat, eerste tot die afrigtingsdata toegevoeg word. So 'n benadering verseker dat die masjienvertaalsisteem eers leer hoe om die mees frekwente woorde te vertaal en op 'n later stadium die skaarser woorde aanleer.

In die twee eksperimente wat in die res van hierdie afdeling bespreek word, word hierdie beginsel op die selektering van bykomende afrigtingsdata toegepas om te bepaal of die kies van die mees frekwente tekseenhede wel vinniger tot beter kwaliteit masjienvertaling kan lei. Aangesien die tipe masjienvertaling wat in hierdie studie gebruik is op frases gebaseer is, is nie net die frekwensietellings van losstaande woorde in berekening gebring nie, maar ook die frekwensietellings van die frases waaruit die tekseenhede bestaan. Vir beide eksperimente is een- tot vyf-gramme gebruik sodat enkelwoorde, woordpare, korter en langer frases almal gedek is.

Voordat enige selektering van data kan plaasvind, word 'n lys van frekwente Engelse *n*-gramme benodig. Hierdie lys is outomaties onttrek uit die Engelse gedeelte van die *Europarl*-korpus (Koehn, 2005:79). Alhoewel dit eerstens belangrik is vir 'n masjienvertaalsisteem om die mees frekwente woorde of frases akkuraat te kan vertaal, is

dit egter ook belangrik om skaarser woorde te herken en korrek te vertaal. Vir hierdie rede is daar geen spesifieke afsnypunt vir die aantal woorde in die lys bepaal nie, wat beteken dat n -gramme met lae frekwensies ook ingesluit is. Daar is egter wel van 'n afsnypunt van 10 voorkomste gebruik gemaak om n -gramme uit te sny wat, onder andere, spelfoute of tikfoute bevat. Daar is dus met vyf lyste, elkeen met 'n ander n -gram telling, gewerk waar elk van die lyste alle n -gramme bevat wat meer as tien keer in die Engelse gedeelte van *Europarl* korpus voorkom. Elke lys is volgens frekwensie van hoog na laag gesorteer.

Daar was dus ook 'n metode benodig waarvolgens aan elkeen van die Engelse gedeeltes van die potensiële bykomende afrigtingsdata 'n punt gekoppel kon word wat 'n aanduiding gee van hoe frekwent die data in die tekseenheid voorkom. Hoe hoër hierdie punt was, hoe groter die kans dat die betrokke tekseenheid tot die afrigtingsdata toegevoeg moes word. Hierdie punte moes toe ook op een of ander wyse genormaliseer word om te verseker dat sinslengte nie die waarde van die punt beïnvloed nie. Daar moes ook aandag geskenk word aan die feit dat hoe meer frekwent 'n n -gram was, hoe hoër die tekseenheid se punt opgestoot moes word, terwyl die minder frekwente n -gramme ook tot die punt moes bydra. Om hierdie rede is daar besluit om vir elkeen van die n -gramme in die frekwensielys 'n punt toe te ken wat aandui hoe frekwent die n -gram was. Hierdie punt is toegeken deur die n -gram se posisie in die lys af te trek van die aantal n -gramme in die lys plus een en die resultaat deur die aantal n -gramme in die lys te deel.

Indien die n -gram se posisie in die lys as p voorgestel word en die aantal n -gramme in die lys as n , kan die punt wat aan elkeen van die n -gramme toegeken word as volg voorgestel word:

$$\text{N-gram punt} = \frac{n-p+1}{n} \quad [6]$$

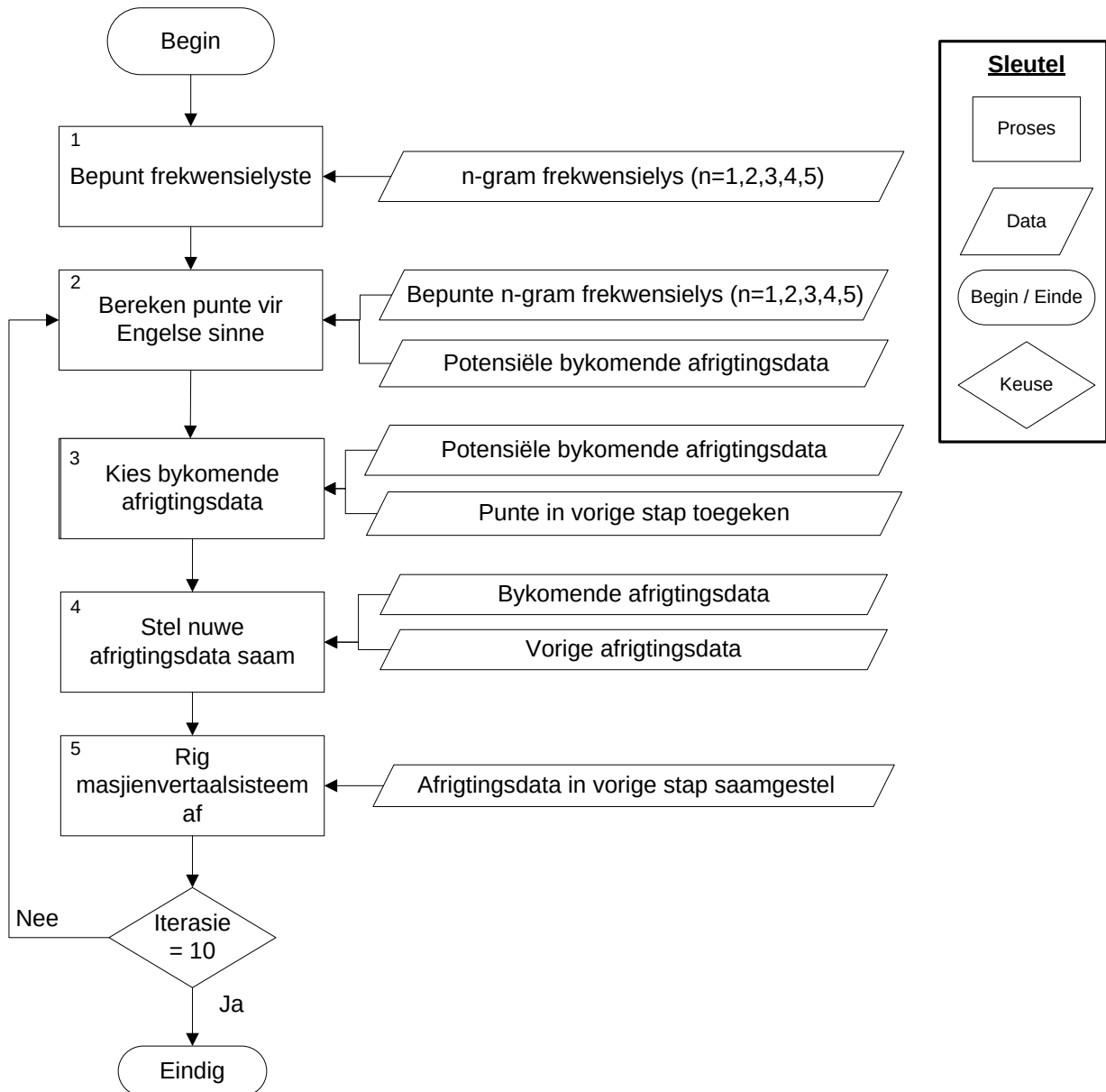
In die geval van 'n woordelys wat uit tien woorde bestaan, sal die mees frekwente woord dus 'n punt van 1 kry, die tweede woord 'n punt van 0.9 en die laaste woord 'n punt van 0.1.

Die uiteindelijke punt wat aan 'n tekseenheid toegeken is, was toe 'n aanduiding van hoe frekvent al die n -gramme in die betrokke tekseenheid was. Hierdie punt is bereken deur eenvoudig alle individuele n -gramme se punte op te tel en die totaal deur die aantal n -gramme in die tekseenheid te deel. Enige n -gram wat nie in die frekwensielys voorkom nie, het 'n punt van nul gekry. Die finale punt vir 'n tekseenheid is dus volgens Formule 7 bereken.

$$\text{Frekwensiepunt vir sin} = \frac{\sum \text{NgramPunte}}{\text{Aantal Ngramme}} \quad [7]$$

In Eksperiment 3 is die tekseenhede georden volgens die frekwensiepunte wat met behulp van Formule 7 bereken is. Die tekseenhede met die heel hoogste punte, en dus die mees frekwente inhoud, is heel eerste as bykomende afrigtingsdata bygevoeg. Weereens is hierdie data in tien iterasies van 50 000 woorde bygevoeg om 'n totaal van ongeveer 500 000 woorde tot die Engelse gedeelte van die afrigtingsdata te voeg. Vir elke iterasie is 'n masjienvertaalsisteem afgerig om die leerkurwe te bepaal.

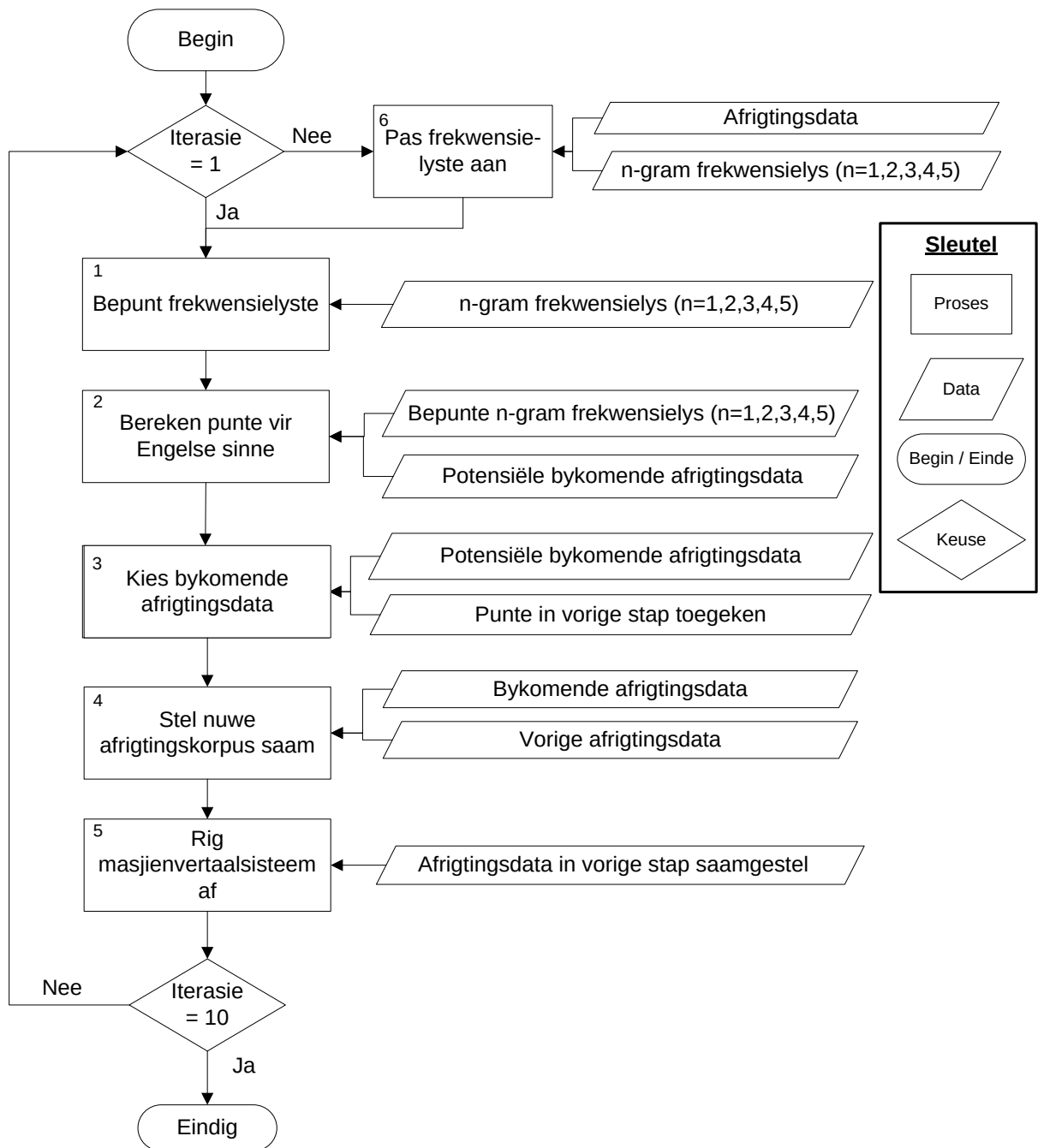
Die vloediagram in Figuur 5 toon die stappe wat in Eksperiment 3 gevolg is. Eerstens is punte aan die n -gram frekwensielyste toegeken op grond van die frekwensie van elke betrokke n -gram deur van Formule 6 (Stap 1) gebruik te maak. Die tweede stap het behels dat punte vir al die Engelse tekseenhede in die bykomende afrigtingsdata bereken word (Stap 2) deur van die frekwensielyste en Formule 7 gebruik te maak. Die tekseenhede met die hoogste punte is uit die korpus gehaal (Stap 3) en, saam met hul vertalings, by die vorige afrigtingsdata gevoeg (Stap 4) om 'n nuwe afrigtingskorpus te vorm. Hierdie nuwe data is dan gebruik om 'n nuwe masjienvertaalsisteem mee af te rig (Stap 5) en die oorblywende data is weer bepunt (Stap 2). Hierdie proses is 10 keer herhaal totdat 'n totaal van ongeveer 500 000 woorde tot die afrigtingsdata van die standaardsisteem bygevoeg is.



Figuur 5: Vloeiagram van Eksperiment 3

'n Vraag wat gedurende die beplanning en uitvoering van Eksperiment 3 ontstaan het, het ook aanleiding gegee tot 'n verdere eksperiment (Eksperiment 4). In Eksperiment 4 is daar spesifiek gefokus op frekwente woorde wat nog nie in die afrigtingsdata opgeneem is nie. Daar is dus nie net gekyk na watter tekseenhede die mees frekwente n -gramme bevat nie, maar ook na watter tekseenhede die meeste ongesiene frekwente n -gramme bevat. In Eksperiment 4 is 'n soortgelyke benadering gevolg as by Eksperiment 3. Die verskil was egter dat die punte wat aan die n -gramme toegeken is vir elke iterasie van vooraf bereken

is. Nadat die bykomende afrigtingsdata vir iterasie 1 gekies is, is alle n -gramme wat in die afrigtingsdata bestaan uit die frekwensielyste verwyder en die punte vir iterasie 2 van nuuts af bereken en toegeken. Die gevolg van hierdie benadering is dat enige n -gramme wat reeds in die afrigtingsdata voorkom geensins 'n bydrae tot die totale punt van die betrokke tekseenheid gelewer het nie.



Figuur 6: Vloeidiagram van Eksperiment 4

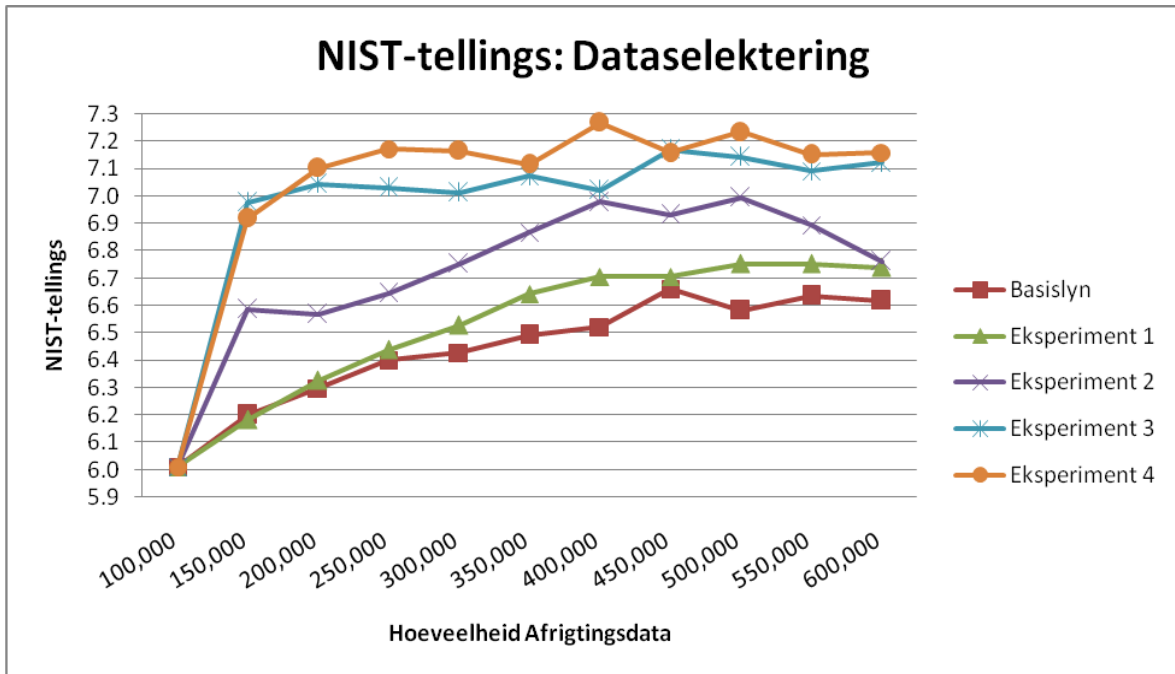
Die vloedigram in Figuur 6 toon die stappe wat in Eksperiment 4 gevolg is. Eerstens is punte aan die n -gram frekwensielyste toegeken (Stap 1); die potensiële bykomende afrigtingsdata is bepunt (Stap 2) en die tekseenhede met die hoogste punte is uit die korpus gehaal (Stap 3). Die nuwe afrigtingsdata is saamgestel deur die nuwe data by die vorige afrigtingsdata te voeg (Stap 4). 'n Masjienvertaalsisteem is met die nuwe data afgerig (Stap 5). Die frekwensielyste is toe aangepas deur al die n -gramme wat in die nuwe afrigtingsdata voorkom te verwyder (Stap 6). Die frekwensielyste is toe van voor af bepunt (Stap 1). Hierdie proses is 10 keer herhaal totdat 'n die totaal van ongeveer 500 000 woorde tot die afrigtingsdata van die standaardsisteem bygevoeg is.

3.9.3. EVALUASIE VAN EKSPERIMENTE 1-4

Die vorige twee afdelings handel oor die ontwerp van vier verskillende dataselekteringseksperimente waarvan die resultate in hierdie afdeling weergegee word. Die NIST- en BLEU-tellings van die vier eksperimente word in Tabelle 6 en 7 verskaf. Figure 7 en 8 bevat grafieke wat die leerkurwes van die eksperimente grafies voorstel. Die ewekansige basislyn, soos in Afdeling 3.8 bespreek, is ook in die leerkurwes vir vergelykingsdoeleindes ingesluit.

	Iterasie 1	Iterasie 2	Iterasie 3	Iterasie 4	Iterasie 5	Iterasie 6	Iterasie 7	Iterasie 8	Iterasie 9	Iterasie 10
Eksperi- ment 1	6.1818	6.3260	6.4381	6.5272	6.6428	6.7062	6.7047	6.7517	6.7519	6.7378
Eksperi- ment 2	6.5864	6.5673	6.6436	6.7516	6.8675	6.9779	6.9320	6.9941	6.8931	6.7616
Eksperi- ment 3	6.9765	7.0447	7.0312	7.0127	7.0737	7.0215	7.1694	7.1416	7.0916	7.1231
Eksperi- ment 4	6.9190	7.1028	7.1717	7.1657	7.1153	7.2689	7.1572	7.2335	7.1526	7.1586

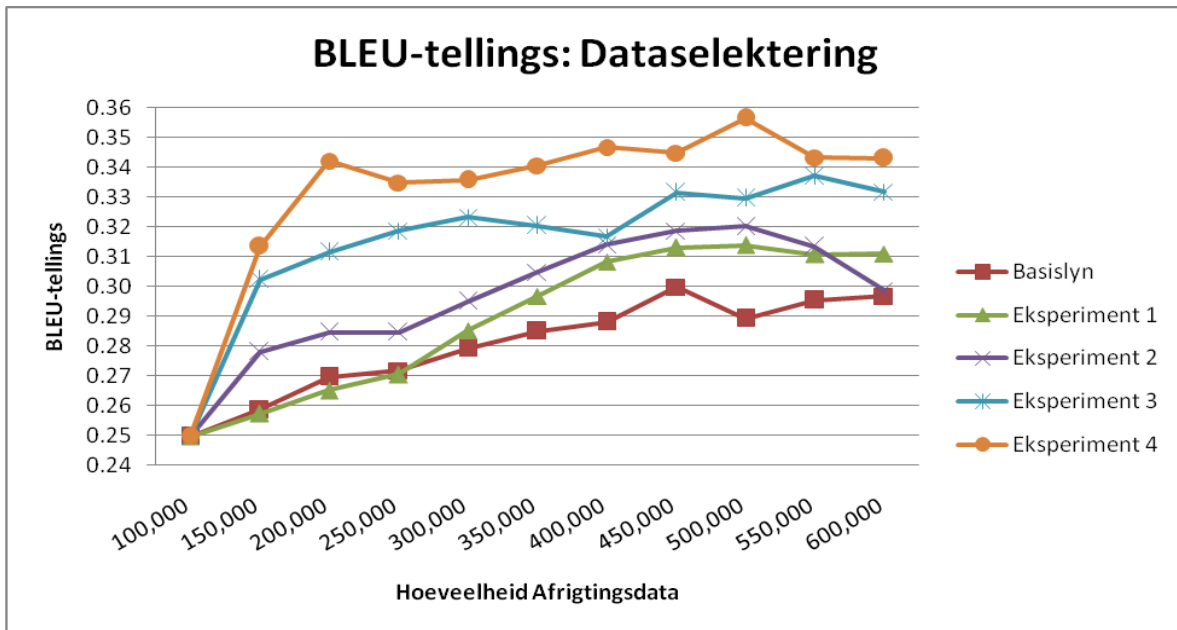
Tabel 6: Evaluasieresultate (NIST) van die dataselekteringseksperimente (1-4)



Figuur 7: Evaluasieresultate (NIST) van die dataselekteringseksperimente (1-4)

	Iterasie 1	Iterasie 2	Iterasie 3	Iterasie 4	Iterasie 5	Iterasie 6	Iterasie 7	Iterasie 8	Iterasie 9	Iterasie 10
Eksperi-ment 1	0.2573	0.2651	0.2705	0.2852	0.2968	0.3082	0.3130	0.3139	0.3107	0.3110
Eksperi-ment 2	0.2781	0.2846	0.2846	0.2950	0.3048	0.3141	0.3186	0.3201	0.3135	0.2987
Eksperi-ment 3	0.3022	0.3115	0.3187	0.3233	0.3203	0.3168	0.3315	0.3296	0.3370	0.3318
Eksperi-ment 4	0.3134	0.3421	0.3348	0.3357	0.3406	0.3467	0.3448	0.3565	0.3432	0.3431

Tabel 7: Evaluasieresultate (BLEU) van die dataselekteringseksperimente (1-4)



Figuur 8: Evaluasieresultate (BLEU) van die dataselekteringseksperimente (1-4)

Uit die resultate blyk dit dat elk van die dataselekteringsmetodes wat geëvalueer is beter resultate as die ewekansige basislyn lewer. Daar is egter baie variasie tussen die verskillende eksperimente. Die tempo waarteen die kwaliteit van die masjienvertaling in die eerste paar iterasies styg (duidelik sigbaar in beide Figure 7 en 8 as die helling van die leerkurwe) toon veral groot verskille tussen die verskeie eksperimente.

Eksperiment 1 lewer in die eerste paar iterasies swakker resultate as die basislyn, maar daarna lewer dit tog beter resultate as die basislyn. Vanaf die vierde iterasie toon al die eksperimente matige, maar tog konstante verbetering tot by die negende iterasie waar die kurwe begin afplat. Eksperiment 2 daarenteen wys 'n steiler styging aan die begin van die iterasies maar begin by iterasie agt sterk daal. Beide Eksperimente 1 en 2 lewer dus 'n verbetering bo die ewekansige basislyn. Uit die resultate vir die tien iterasies van die twee eksperimente is dit nie heeltemal duidelik watter een van die benaderings tot die beste resultate lei nie. Alhoewel Eksperiment 1 by die tiende iterasie hoër punte behaal het, het Eksperiment 2 vinniger gestyg en het dus vinniger tot goeie resultate gelei. Indien Eksperiment 2 in verdere iterasies aanhou daal, sou Eksperiment 1 'n beter dataselekteringsmetode wees. Indien die daling egter slegs tydelik is, is dit duidelik dat Eksperiment 2 die beste een van die twee metodes is.

Beide Eksperimente 3 en 4 toon 'n groot verbetering bo die basislyn. Albei toon in die heel eerste iterasie reeds 'n skerp styging in beide die NIST- en BLEU-tellings. Eksperiment 4

behaal beter resultate as Eksperiment 3. Albei eksperimente plat na slegs 'n paar iterasies af, vanwaar daar slegs kleiner veranderinge in die tellings is. Ten spyte hiervan toon beide eksperimente 'n groot verbetering bo die basislyn asook 'n verbetering oor die eerste twee dataselekteringsmetodes wat getoets is. Die dataselekteringsmetodes in Eksperimente 3 en 4 kan tot baie groei in die kwaliteit van masjienvertaling lei met die toevoeging van redelik min bykomende data. Dit sal veral nuttig wees in 'n situasie waar daar geen bykomende parallelle korpora is nie en Engelse data gekies moet word vir vertaling deur professionele vertalers.

3.10. AANPASSING VAN GEWIGTE VAN BESKIKBARE DATA (EKSPERIMENT 5)

In die vorige afdeling is die effek van die toevoeging van bykomende afrigtingsdata op grond van bepaalde dataselekteringsmetodes ondersoek. Dit is egter ook moontlik dat daar vir 'n taalpaar net geen bykomende parallelle data bestaan nie en dat daar ook geen hulpbronne is vir die verkryging of vertaling van bykomende data nie. Vir hierdie rede is daar in hierdie studie ook gekyk na die aanpassing van die gewigte van die beskikbare data in 'n poging om die kwaliteit van die masjienvertaling te verbeter sonder die toevoeging van bykomende data.

Die gewigte van bepaalde gedeeltes van die afrigtingsdata word verhoog deur veelvuldige kopieë van daardie bepaalde gedeeltes by die afrigtingsdata te voeg. 'n Gedeelte van die data word dus verdubbel of verdriedubbel om dan dubbel of driedubbel soveel gewig as die ander gedeeltes in die afrigtingsdata te dra. Daar word geen nuwe woordeskat bygevoeg nie, maar daar word verwag dat dit sal veroorsaak dat frekwente woorde met groter akkuraatheid vertaal sal word.

Vir hierdie eksperiment is daar weereens eerstens op die tekseenhede gekonsentreer wat frekwente n -gramme bevat. Die rede hiervoor is dat dit juis die frekwente data is wat beklemtoon moet word deur die aanpassing van die gewigte. Die formule en frekwensielyste wat vir Eksperiment 3 opgestel is, is ook hier gebruik om die gedeeltes van die afrigtingsdata waarvan die gewigte aangepas moet word te selekteer. Om 'n idee te kry van die optimale hoeveelheid data om te vermeerder is die selektering op verskeie vlakke gedoen. Die korpus wat vir die afrigting van die standaardsisteem gebruik is, is vir

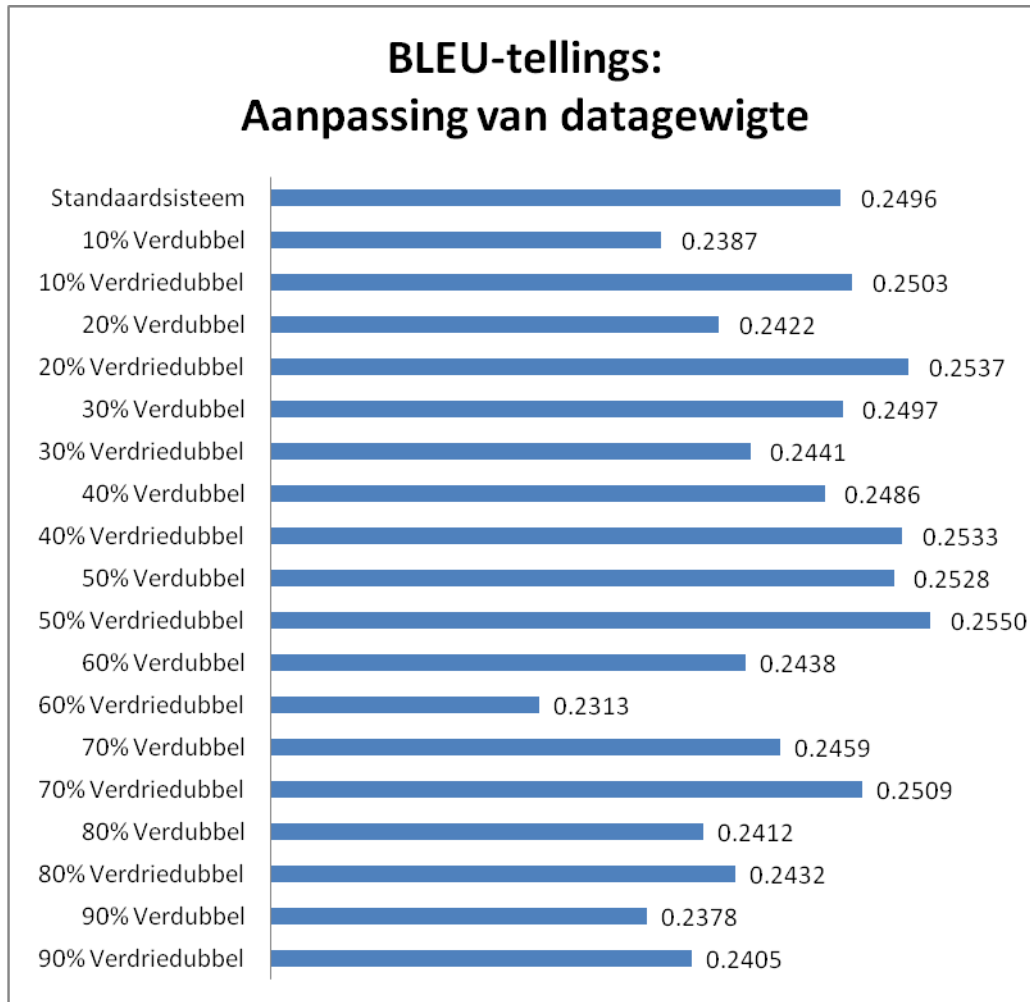
hierdie eksperiment gebruik. Eerstens is die 10% van die afrigtingsdata met die hoogste frekwensiepunte gekies en die gewig daarvan verdubbel deur 'n ekstra kopie van elkeen van die tekseenhede tot die korpus te voeg. Daarna is die gewig verdriedubbel deur twee kopieë toe te voeg. Die proses is toe herhaal met 20% tot 90% (in inkremente van 10%) van die tekseenhede met die hoogste frekwensietellings in die afrigtingsdata om te bepaal wat die effek van die hoeveelhede is en ook om te sien of die toevoeging van twee kopieë van die tekseenhede tot beter resultate sou lei as die toevoeging van slegs een kopie. Die resultate van Eksperiment 5 word in Tabel 8 gegee. Figure 9 en 10 bied 'n grafiese voorstelling van die onderskeie NIST- en BLEU-tellings.

Data Aangepas	NIST	BLEU	Data Aangepas	NIST	BLEU
10% Verdubbel	5.9487	0.2387	10% Verdriedubbel	6.1018	0.2503
20% Verdubbel	5.9901	0.2422	20% Verdriedubbel	6.0688	0.2537
30% Verdubbel	6.0118	0.2497	30% Verdriedubbel	5.9499	0.2441
40% Verdubbel	6.0413	0.2486	40% Verdriedubbel	5.9876	0.2533
50% Verdubbel	6.0997	0.2528	50% Verdriedubbel	6.1150	0.2550
60% Verdubbel	5.9562	0.2438	60% Verdriedubbel	5.8137	0.2313
70% Verdubbel	5.9574	0.2459	70% Verdriedubbel	5.9728	0.2509
80% Verdubbel	5.9287	0.2412	80% Verdriedubbel	5.9666	0.2432
90% Verdubbel	5.9167	0.2378	90% Verdriedubbel	5.8985	0.2405

Tabel 8: Evaluasieresultate van die aanpassing van datagewigte (Eksperiment 5)



Figuur 9: Evaluasieresultate (NIST) van die aanpassing van datagewigte (Eksperiment 5)



Figuur 10: Evaluasieresultate (BLEU) van die aanpassing van datagewigte (Eksperiment 5)

Soos uit die grafieke blyk, wissel die resultate van die verskeie toetse. Oor die algemeen toon die gevalle waar die geselekteerde data verdriedubbel word beter resultate as die gevalle waar die data slegs verdubbel word. Van al die verskillende toetse is die beste resultate behaal deur helfte (50%) van die afrigtingsdata te verdriedubbel.

Alhoewel die aanpassing van die data se gewigte geen nuwe woordeskat tot die masjienvertaalsisteem bygedra het nie, is dit duidelik dat dit tog in sekere gevalle tot 'n verbetering in die kwaliteit van die uiteindelijke vertaling gelei het. Die implikasie hiervan is dat dit, selfs in gevalle waar geen bykomende data beskikbaar is nie, moontlik is om deur die aanpassing van die data 'n verbetering in die masjienvertaling mee te bring. Hierdie bevinding is veral belangrik vir hulpbronarm tale wat baie min data beskikbaar het.

3.11. AANPASSING EN SELEKTERING VAN TAALMODELDATA

Alle vorige eksperimente het ten doel gehad om die effek van die selektering of aanpassing van die tweetalige afrigtingsdata op die kwaliteit van die masjienvertaalsisteem se afvoer te ondersoek. Daar is egter ook 'n ander moontlikheid vir die aanpassing van data en dit behels die selektering of aanpassing van die eentalige taalmodeldata. Hierdie afdeling bespreek dus die eksperimente wat poog om die masjienvertaling te verbeter deur slegs aan die eentalige data in die teikentaal te verander. Aangesien eentalige data baie makliker is om te bekom as parallelle data, sal enige aanduiding van verbetering in kwaliteit van groot nut wees vir enige hulpbronnarm taal.

Daar word twee verskillende ondersoeke gedoen om die effek van die taalmodeldata op die kwaliteit van masjienvertaling te ondersoek. Vir hierdie eksperimente word die standaardsisteem se parallelle afrigtingsdata gebruik en alle programinstellings en spesifikasies word dieselfde gehou as met die afrigting van die oorspronklike taalmodelle, daar word slegs aan die taalmodeldata verander. Die eerste ondersoek is om te kyk of die aanpassing van die gewigte van die beskikbare data tot moontlike verbeterings kan lei. Daar word ook gekyk of die noukeurige selektering van afrigtingsdata vir die taalmodelle tot beter resultate lei as die ewekansige selektering van data wat gevolg is by die skep van die oorspronklike taalmodelle.

3.11.1. AANPASSING VAN TAALMODELDATA (EKSPERIMENT 6)

Die eerste poging om die taalmodel te verbeter (Eksperiment 6) deur aanpassings aan die taalmodeldata is gedoen met die beperkende aanname dat alle beskikbare data reeds gebruik is. Die data wat in gebruik is, word dus op die een of ander wyse aangepas. Die metode wat ondersoek is, is soortgelyk aan die aanpassing van datagewigte wat in Eksperiment 5 ondersoek is. Die doel van die aanpassing van die data se gewigte is om die data wat meer frekwente taalgebruik verteenwoordig te beklemtoon.

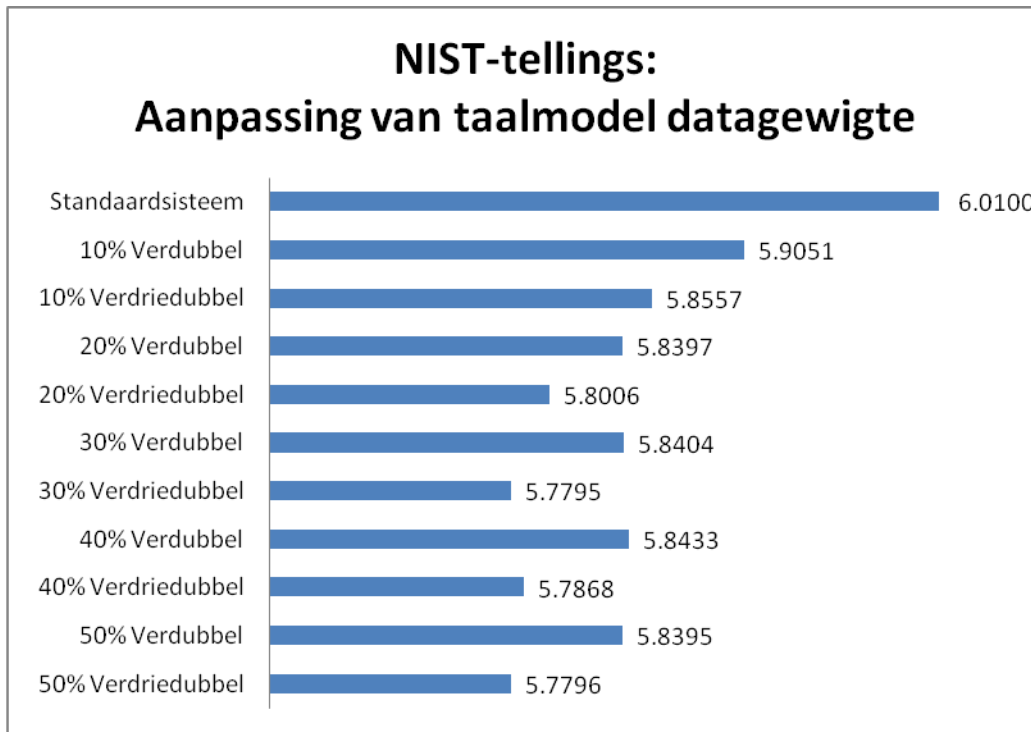
Vir die toekenning van frekwensiepunte aan die sinne word frekwensielyste van die verskeie n -gramme vir die teikentaal (Afrikaans) benodig. Hierdie lyste is uit die volledige Media24-korpus onttrek. Soos met die Engelse lyste (Eksperimente 3 en 4) is die n -

gramme wat minder as tien keer voorkom uit die lys verwyder en die lyste is volgens frekwensie (van hoog na laag) rangskik.

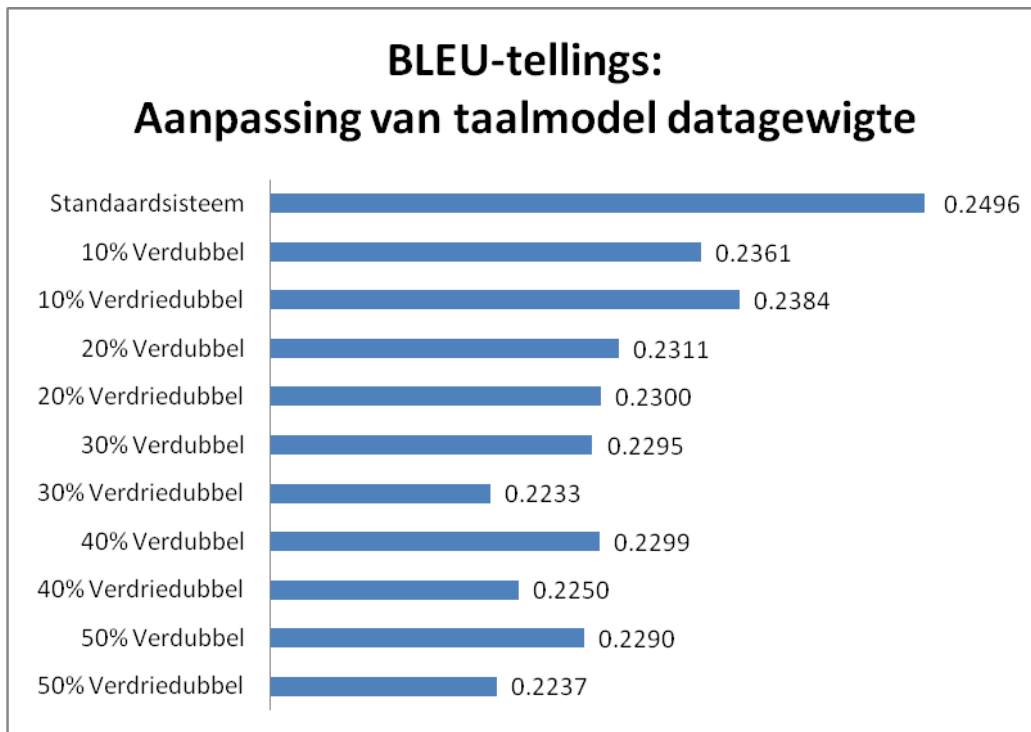
Die Afrikaanse frekwensielyste is gebruik om aan al die sinne in die oorspronklike taalmodeldata punte toe te ken. Die data met die hoogste punte is vir duplisering gekies. Daar is 10%, 20%, 30%, 40% en 50% van die data met die hoogste punte gekies en hierdie data is verdubbel en verdriedubbel. Die aangepaste datastelle is gebruik om nuwe taalmodelle af te rig en vir elkeen van die aangepaste stel taalmodelle is daar weer 'n masjienvertaalsisteem afgerig met die standaardsisteem se afrigtingsdata en die aangepaste taalmodelle. Die resultate van Eksperiment 6 word in Tabel 9 gegee. Figure 11 en 12 bied 'n grafiese voorstelling van die NIST- en BLEU-tellings.

Data Aangepas	NIST	BLEU	Data Aangepas	NIST	BLEU
10% Verdubbel	5.9051	0.2361	10% Verdriedubbel	5.8557	0.2384
20% Verdubbel	5.8397	0.2311	20% Verdriedubbel	5.8006	0.2300
30% Verdubbel	5.8404	0.2295	30% Verdriedubbel	5.7795	0.2233
40% Verdubbel	5.8433	0.2299	40% Verdriedubbel	5.7868	0.2250
50% Verdubbel	5.8395	0.2290	50% Verdriedubbel	5.7796	0.2237

Tabel 9: Evaluasieresultate van die aanpassing van die gewigte van taalmodeldata (Eksperiment 6)



Figuur 11: Evaluasieresultate (NIST) van die aanpassing van die gewigte van taalmodeldata (Eksperiment 6)



Figuur 12: Evaluasieresultate (BLEU) van die aanpassing van die gewigte van taalmodeldata (Eksperiment 6)

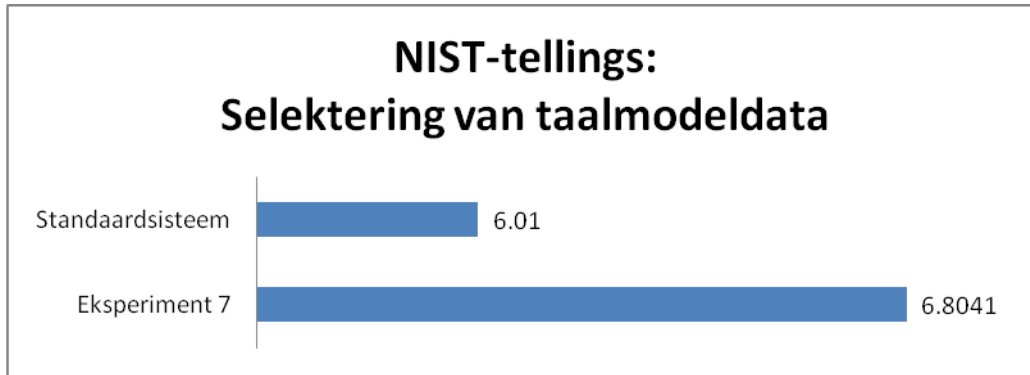
Soos uit Figure 11 en 12 blyk, het die aanpassing van die gewigte van die taalmodeldata nie tot enige verbetering in die kwaliteit van die masjienvertaling gelei nie. Dit kan moontlik te wyte wees aan die feit dat die aanpassing van die datagewigte veroorsaak dat die resulterende taalmodel nie meer 'n "getroue" beeld van die taal weerspieël nie.

3.11.2. SELEKTERING VAN TAALMODELDATA (EKSPERIMENT 7)

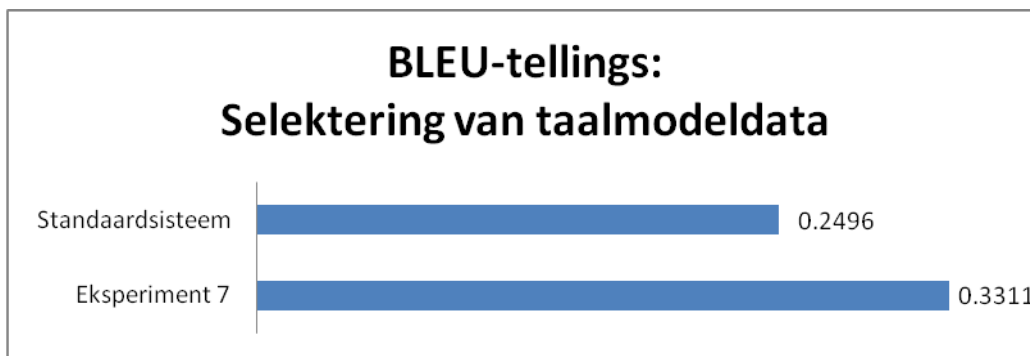
Daar is ook ondersoek ingestel tot die effek van die selektering van taalmodeldata. In hierdie eksperiment is daar van nuuts af begin met die selektering van taalmodeldata. Die volle Media24-korpus is as potensiële data geneem en dieselfde hoeveelheid data as wat gebruik is vir die oorspronklike taalmodelle is geselekteer. Die dataselektering vir hierdie eksperiment is weereens op frekwensie gebaseer maar aangesien die volle afrigtingskorpus op een slag geselekteer word en daar geen iterasies betrokke is nie, is die frekwensielyste nie, soos met Eksperiment 4, aangepas nie. Daar is vir elkeen van die sinne in die korpus 'n punt toegeken op grond van die frekwensie van die n -gramme in die sin en die sinne met die hoogste punte is geselekteer. Die geselekteerde data is gebruik om nuwe taalmodelle af te rig en die nuwe modelle, saam met die standaardsisteem se afrigtingsdata is gebruik om 'n nuwe masjienvertaalsisteem af te rig. Die resultate van Eksperiment 7 word in Tabel 10 gegee. Figure 13 en 14 bied grafiese voorstellings van die NIST- en BLEU-tellings.

	NIST	BLEU
Eksperiment 7	6.8041	0.3311

Tabel 10: Evaluasieresultate van die selektering van taalmodeldata (Eksperiment 7)



Figuur 13: Evaluasieresultate (NIST) van die selekering van taalmodeldata (Eksperiment 7)



Figuur 14: Evaluasieresultate (BLEU) van die selekering van taalmodeldata (Eksperiment 7)

Die resultate van Eksperiment 7 toon dat die selekering van taalmodeldata deur middel van frekwensietellings tot 'n merkbare verbetering in die kwaliteit van masjienvertaling kan lei. Die implikasie hiervan is dat dit moontlik is om verbeteringe in die kwaliteit van die masjienvertaling teweeg te bring sonder om bykomende tweetalige data beskikbaar te hê. Hierdie bevinding kan veral van groot nut wees vir hulpbronarm tale waar dit moeilik is om tweetalige data op te spoor maar waar eentalige data wel bekombaar is.

3.12. KOMBINERING VAN METODES (EKSPERIMENT 8)

In die voorafgaande afdelings is verskeie metodes vir die selekering en aanpassing van data geëvalueer om die effek op die kwaliteit van die masjienvertaalsisteem se afvoer te ondersoek. In hierdie afdeling word die metodes wat in die ander eksperimente tot die heel beste resultate gelei het gekombineer. Die doel hiervan was om te sien of 'n kombinasie van metodes dalk 'n beter resultaat as die individuele eksperimente sou lewer.

Die eerste keuse wat gemaak moes word, was die keuse van die taalmodelle vir hierdie eksperiment. Daar word in hierdie studie twee verskillende taalmodelleksperimente beskryf waarvan een verskeie toetse behels het. Die taalmodel wat die beste resultate met die standaardsisteem se datastel getoon het, is dié wat in Eksperiment 7 ontwikkel is, waar die data van die begin af deur middel van frekwensietellings gekies is.

Buiten die taalmodelle moes die tweetalige afrigtingsdata vir hierdie eksperiment (Eksperiment 8) ook gekies en 'n besluit geneem word oor die aanpassing van die gewigte van die tekseenhede in die afrigtingsdata. Hierdie eksperiment is nie in iterasies gedoen nie; slegs die finale afrigtingsdata (afkomstig uit iterasie 10) van die gekose eksperiment is gebruik. Die eksperiment wat die heel beste dataselekeringsmetode vir afrigtingsdata gelewer het, is Eksperiment 4. Die data in dié eksperiment is gekies deur te kyk na die tekseenhede wat die meeste ongesiene frekwente n -gramme bevat. Aangesien die aanpassing van dele van die data se gewig in die korpus ook tot verbeteringe in die masjienvertaalsisteem gelei het, is daar besluit om ook die effek te toets wanneer dit met die beste afrigtingsdata en taalmodelle gekombineer word. Die verdriedubbeling van die top 50% van die data, soos bepaal met frekwensiepunte, het in Eksperiment 5 die beste resultate gelewer. Dit is dus die metode wat in hierdie eksperiment gebruik is.

Eksperiment 8 behels dus twee toetse. In die eerste toets word 'n nuwe masjienvertaalsisteem afgerig met die finale datastel van Eksperiment 4 en die taalmodelle van Eksperiment 7. Daarna word 'n tweede sisteem afgerig waar dieselfde taalmodelle gebruik is maar die parallelle afrigtingsdata se gewigte aangepas is soos hierbo genoem. Die resultate van beide eksperimente word in Tabel 11 gegee.

	NIST	BLEU
Eksperiment 8: Sonder aanpassing van datagewigte	8.6176	0.5006
Eksperiment 8: Met aanpassing van datagewigte	8.5238	0.4919

Tabel 11: Evaluasieresultate van die kombinasie van verskeie metodes (Eksperiment 8)

Tabel 11 toon aan dat, indien die beste metodes vir die selektering van parallelle afrigtingsdata en taalmodeldata gekombineer word, die beste resultate (volgens beide evaluasiemetrieke) tot dusver behaal is. Om optimale groei in die kwaliteit van masjienvertaling te verseker, is dit dus baie belangrik om 'n kombinasie van metodes te oorweeg, aangesien dit beter resultate lewer as wanneer die verskillende metodes afsonderlik toegepas word.

Die aanpassing van die gewigte van die tweetalige afrigtingsdata het, anders as in Eksperiment 5, nie tot 'n verbetering gelei nie. Beide die NIST- en BLEU-punte het vir hierdie toets gedaal. Hierdie metode is dus minder doeltreffend wanneer dit op geselekteerde data toegepas word. Eksperiment 5 is ook op baie minder data (slegs 100 000 woorde) uitgevoer in vergelyking met Eksperiment 8 (600 000 woorde) wat ook tot die verswakking aanleiding kon gegee het.

3.13. SAMEVATTING

In hierdie hoofstuk word die verskillende praktiese eksperimente bespreek wat deel van die studie uitmaak. Verskillende dataselekteringsmetodes is toegepas en die resulterende masjienvertaalsisteme is telkens geëvalueer. Die besonderhede van die eksperimente en die resultate van die evaluasie word in die hoofstuk gegee.

In die eerste gedeelte van die hoofstuk word die data, programme en evaluasiemetode beskryf wat in die studie gebruik is. Die ontwikkeling van die standaardtaalmodelle, standaardmasjienvertaalsisteme en basislynsisteme vir die dataselekteringseksperimente, almal afgerig met ewekansig gekose data, word beskryf.

Die vier eksperimente wat handel oor die selektering van bykomende tweetalige afrigtingsdata word uiteengesit. Eksperimente 1 en 2 het die vertrouettellings, deur *Moses* gedurende die vertaalproses toegeken, gebruik om bykomende afrigtingsdata mee te kies. Die verskil tussen die twee eksperimente is dat Eksperiment 1 slegs met langer tekseenhede uitgevoer is en Eksperiment 2 met alle data. Eksperimente 3 en 4 het bykomende data geselekteer op grond van die frekwensietellings van die n -gramme in die Engelse tekseenhede. Die verskil tussen Eksperiment 3 en Eksperiment 4 is dat daar in Eksperiment 4 spesifiek gefokus is op frekwente woorde wat nog nie in die afrigtingsdata opgeneem is nie. Die eerste vier eksperimente se evaluasieresultate word in Afdeling 3.9.3 weergegee. Die evaluasieresultate het getoon dat alle dataselekteringsmetodes tot beter kwaliteit masjienvertaling lei as die ewekansige basislyn. Van die vier verskillende dataselekteringsmetodes het die metode van Eksperiment 4 die heel beste resultate getoon.

Die vyfde eksperiment wat in die hoofstuk beskryf word, handel oor die aanpassing van die bestaande data. In die eksperiment is die data verander deur telkens 'n hoeveelheid

daarvan te verdubbel of verdriedubbel om sodoende die gewig van daardie gedeelte van die data in die afrigtingsdata te vermeerder. Hierdie eksperiment het egter minder positiewe resultate getoon as die dataselekteringseksperimente. Die toets waar 50% van die data verdriedubbel is, het die grootste verbetering getoon.

Die sesde en sewende eksperimente wat in die hoofstuk beskryf word, het nie aan die tweetalige afrigtingsdata verander nie, maar aan die eentalige taalmodeldata. In Eksperiment 6 is die taalmodeldata aangepas deur 'n soortgelyke metode as die een wat vir Eksperiment 5 gebruik is. In Eksperiment 7 is die taalmodeldata heeltemal van nuuts af gekies. Die aanpassing van die gewigte van die taalmodeldata het in alle toetse tot swakker punte gelei. Die sorgvuldige selektering van taalmodeldata het egter tot 'n baie groot verbetering in die masjienvertaling gelei.

Die laaste afdeling in die hoofstuk beskryf die opstelling en resultate van Eksperiment 8. Die doel van die eksperiment was om te bepaal of die kombinerings van vorige eksperimente wat goeie resultate behaal het tot 'n groter verbetering in die kwaliteit van die masjienvertaling kan lei as wat met die benaderings afsonderlik behaal is. Om dit te toets, was die beste dataselekteringstrategie saam met die beste taalmodelle en die beste aanpassing van datagewigte gebruik. Die resultate toon dat die kombinerings van die beste dataselekteringstrategie en die beste taalmodel tot 'n verbetering in resultate lei. 'n Poging om die gewigte van die data in die kombinasiesisteme aan te pas was egter vrugtelos.

Die resultate wat in hierdie hoofstuk aangebied word, toon dat die selektering van data 'n baie belangrike deel van masjienvertaling uitmaak. Goeie dataselektering vir die taalmodeldata en die parallelle afrigtingsdata kan tot beter gehalte masjienvertaling lei.

4. GEVOLGTREKKINGS EN VERDERE NAVORSINGS-MOONTLIKHEDE

4.1. GEVOLGTREKKINGS

Aan die begin van hierdie studie is drie doelstellings geformuleer, naamlik om:

1. die effek van dataselektering (van beide parallelle en eentalige data) op die prestasie van statistiese masjienvertaling vir Engels-Afrikaans te ondersoek;
2. die effek van die aanpassing van die gewigte van beskikbare data (beide parallelle en eentalige data) op die prestasie van statistiese masjienvertaling vir Engels-Afrikaans te ondersoek; en
3. deur middel van 'n noukeurige kombinasie van die bogenoemde metodes 'n verbetering in die masjienvertaling van Engels na Afrikaans teweeg te bring.

Die hipotese wat aan die begin van die studie gestel is, is dat die noukeurige selektering van bykomende afrigtingsdata tot 'n verbetering in die kwaliteit van masjienvertaling sal lei en dat die aanpassing van beskikbare afrigtingsdata deur die gewigte te verander ook tot 'n verbetering in die gehalte van masjienvertaling kan lei.

Gedurende die verloop van hierdie studie is verskeie eksperimente uitgevoer om die bogenoemde hipotese te toets en die drie doelstellings te behaal.

Eerste navorsingsdoelstelling: *Stel ondersoek in na die effek van dataselektering (van beide parallelle en eentalige data) op die prestasie van statistiese masjienvertaling vir Engels-Afrikaans.*

Vier verskillende eksperimente is uitgevoer oor die selektering van parallelle afrigtingsdata en een oor die selektering van die taalmodel se afrigtingsdata. Al die metodes wat gebruik is om parallelle afrigtingsdata te kies het tot beter resultate gelei as die ewekansige basislyn. Die dataselektering basislyn het 'n NIST-telling van 6.6175 en 'n BLEU-telling van 0.2967 behaal. Die eksperimente toon 'n verbetering van tussen 0.1203 en 0.5411 in die NIST-telling en tussen 0.002 en 0.0464 in die BLEU-tellings. Uit die evaluasieresultate kan dus gesien word dat die selektering van afrigtingsdata die gehalte

van die resulterende masjienvertaalsisteem sal verbeter. Uit die leerkurwes (soos in Figure 7 en 8 uitgebeeld) kan daar ook gesien word dat die dataselekteringsmetodes tot 'n vinniger styging in die kwaliteit van die masjienvertaling gelei het. In die geval van die selektering van die taalmodeldata is 'n verhoging van 0.7941 in die NIST-telling en 0.0815 in die BLEU-telling gesien met die toepassing van dataselektering vir die afrigtingsdata. Die sorgvuldige selektering van data het dus ook in die geval van die eentalige taalmodeldata tot verbeterde masjienvertaling gelei.

Tweede navorsingsdoelstelling: *Stel ondersoek in na die effek van die aanpassing van die gewigte van beskikbare data (beide parallelle en eentalige data) op die prestasie van statistiese masjienvertaling vir Engels-Afrikaans.*

Vir hierdie ondersoek is daar telkens 'n gedeelte van die data gekies en die gewig van daardie gedeelte is dan vermeerder deur bykomende kopieë van die tekseenhede tot die korpus te voeg. In die geval van die parallelle afrigtingsdata het die verdriedubbeling van 50% van die afrigtingsdata tot 'n 0.105 verbetering in die NIST-telling en 'n 0.0054 verbetering in die BLEU-telling gelei. Die verbetering is gemeet deur die aangepaste sisteem met die standaardsisteem, wat as basislyn vir hierdie eksperimente gedien het, te vergelyk. In die geval van die aanpassing van die eentalige taalmodeldata het dit egter in alle gevalle geblyk dat die aanpassing van die data nie tot beter masjienvertaling gelei het nie. Die aanpassing van die parallelle afrigtingsdata kan dus wel tot beter kwaliteit masjienvertaling lei.

Derde navorsingsdoelstelling: *Probeer om deur middel van 'n noukeurige kombinasie van die eerste twee metodes 'n verbetering in die masjienvertaling van Engels na Afrikaans teweeg bring.*

Die metodes wat uit die eksperimente die beste resultate getoon het, is gekombineer met die verwagting dat dit tot 'n groter verbetering sou lei as wat met die verskillende metodes afsonderlik behaal is. Die kombinasiesisteem behaal tellings van 8.6176 (NIST) en 0.5006 (BLEU) in vergelyking met die basislynsisteem se tellings van 6.6175 (NIST) en 0.2967 (BLEU). Die kombinasiesisteem lewer ook beter resultate as wat elk van die metodes onderskeidelik gelewer het. Die aanpassing van die afrigtingsdata in hierdie sisteem het egter nie tot enige verbetering in die sisteem gelei nie. Daar kan dus afgelei word dat die selektering van afrigtingsdata die beste werk indien dit op beide die parallelle en eentalige data toegepas word.

Die hipotese wat aan die begin van die studie gestel is, is dus waar bewys. Die noukeurige selektering van bykomende afrigtingsdata het tot 'n verbetering in die kwaliteit van die masjienvertaling gelei; dit geld vir beide die eentalige taalmodeldata en die tweetalige afrigtingsdata. Die aanpassing van beskikbare afrigtingsdata deur die gewigte te verander het ook in die geval van die parallelle data tot 'n verbetering in die gehalte van masjienvertaling gelei.

4.2. VERDERE NAVORSINGSMOONTLIKHEDE

Hoewel daar 'n aantal interessante bevindinge gedurende hierdie studie gemaak is, het daar ook na aanleiding van hierdie studie 'n aantal nuwe navorsingsmoontlikhede ontstaan. In die res van hierdie afdeling word die vrae waartoe hierdie studie aanleiding gegee het genoem en die navorsingsmoontlikhede wat uit elk voortspruit, word bespreek.

- Sal menslike evaluasie dieselfde mate van verbetering toon met die toepassing van die ontwikkelde dataselektering en -manipuleringsmetodes as wat deur outomatiese evaluasie getoon is?

Ongeag die mate van ooreenstemming wat die NIST- en BLEU-tellings met menslike oordeel toon, sal die beste evaluasiemetode altyd die menslike wees. Daar is baie fyner verskille tussen potensiele vertalings wat, hoewel dit nie eintlik tot verskillende NIST- en BLEU-tellings sou lei nie, wel 'n groot verskil in die manier kan maak waarop die gebruiker die sin verstaan. Dit is daarom van die uiterse belang om ook die metodes op hierdie wyse te toets. Dit sou natuurlik onprakties wees om al die masjienvertaalsisteme deur mense te laat evalueer, maar 'n duidelike beeld sal wel verkry word deur slegs die finale iterasie van die sisteme met mekaar te kan vergelyk om te bepaal of die NIST- en BLEU-tellings wel 'n akkurate beeld van die verbeteringe weerspieël deur die toepassing van die dataselektering en -manipulering.

- Sou die ontwikkelde dataselektering en -manipuleringsmetodes dieselfde vlak van verbetering getoon het indien die vertaling vanaf Afrikaans (as brontaal) na Engels (as teikenstaal) gedoen is?

Die implikasie hiervan is dat die selektering en manipulering van parallelle data dan op die inligting in die Afrikaanse gedeelte van die data gebaseer word en die aanpassing van taalmodeldata op eentalige Engelse data uitgeoefen word. Die moontlikheid bestaan dat

dieselfde tegnieke tot 'n ewe groot verbetering in die kwaliteit van die Engelse afvoer sou lei as met dié van Afrikaans. Dit is egter ook moontlik dat die geneigdheid van Afrikaans om baie samestellings te bevat die doeltreffendheid van die metodes wat in hierdie studie bestudeer is, sal verlaag en dat die metodes aangepas moet word om die spesifieke morfologie van Afrikaans (as brontaal) in ag te neem.

- Kan die toevoeging van morfologiese inligting by die dataselekteringsproses tot beter vertalingskwaliteit lei?

In die geval van homonieme kan dit gebeur dat slegs een van die moontlike betekenisse van die woord in die afgrigtingsdata voorkom en dat die resulterende masjienvertaalsisteem dan telkens die spesifieke woord verkeerd vertaal deur nie die konteks waarin die woord gebruik word in ag te neem nie. Dit is dus belangrik dat die mees frekwente woorde in 'n taal nie net in die afgrigtingsdata voorkom nie, maar ook dat dit soveel keer voorkom as wat nodig is vir die afgrigtingsalgoritme om die korrekte vertaling van die woord in elke konteks te kan aflei. Die dataselekteringsmetodes kan aangepas word om morfologiese inligting, soos woordsoortetikette wat kan aandui wat die funksie van die woord in die sin is, in ag te neem vir die keuse van tekseenhede om tot die afgrigtingsdata te voeg.

- Kan die selektering van data vir vertaling ook gedoen word deur teikentaaldata te kies en na die brontaal te laat vertaal? Sou dit beter resultate lewer as die kies van brontaaldata of sou 'n kombinasie van die twee tegnieke dalk tot optimale resultate lei?

Dit sou moontlik kon gebeur dat sekere taalverskynsels en woorde in die teikentaal nie deur die masjienvertaalsisteem vasgevang word indien al die data juis met die brontaal as fokus geselekteer word nie. Dit sou interessant wees om te sien hoe die gehalte van die resulterende masjienvertaalsisteem verskil indien die dataselektering juis op die teikentaal uitgevoer sou word. Dit is ook moontlik dat die kombinasie van die twee benaderings juis alle belangrike inligting sal vasvang en daardeur tot goeie masjienvertaalsisteme sal lei.

- Moet die frekwensielyste uit domeinspesifieke data onttrek word?

Indien die afgrigtingsdata en die tekste wat met die uiteindelijke masjienvertaalsisteem vertaal gaan word uit dieselfde domein afkomstig is, sal die kwaliteit van die afvoer uiteraard hoër wees. Dit sou dus moontlik van nut wees om die frekwensielyste juis uit domeinspesifieke data te onttrek. Sodoende word die data gekies, of beklemtoon, wat

soortgelyk is aan die data wat uiteindelik deur die masjienvertaalsisteem vertaal moet word. Hierdie benadering kan egter slegs van nut wees indien die domein vasgestel is en daar genoeg domeinspesifieke data beskikbaar is.

- Behoort “swak” data uit die parallelle data verwyder te word?

’n Verdere moontlikheid vir verbetering van die afrigtingsdata is die verwydering van swak data uit die korpus. Dit kan data insluit wat verkeerd belyn of vertaal is, asook data wat buitengewone vertalings bevat soos, onder andere, idiome wat selde letterlik vertaal word. Die verwydering van sulke data sal moontlik tot beter frasetabelle (met minder foute) lei en sodoende ook ’n verbetering in die afvoer van masjienvertaling tot gevolg hê. ’n Aanpassing van die dataselekteringsmetode wat in Eksperiment 3 gebruik is, kan moontlik gebruik word om die tekseenhede uit te wys wat die laagste punte behaal en dus ook baie swak of buitengewone sinne in die betrokke taal is.

- Sal die metodes wat gedurende hierdie studie ondersoek en ontwikkel is soortgelyke verbeteringe toon vir ander taalpare of ander benaderings tot masjienvertaling?

Dit sou van enorme nut wees as die metodes wat in hierdie studie ontwikkel is ook gebruik kan word om die afrigting van masjienvertaalsisteme vir ander Suid-Afrikaanse taalpare, en ook ander hulpbronnarm tale, te bevorder. Dit sou dan hopelik moontlik wees om vinniger werkende sisteme vir die ander Suid-Afrikaanse taalpare te ontwikkel. Dit sou ook interessant wees om ondersoek in te stel na die werking van die metodes indien die dataselektering op die meer morfologies komplekse tale, soos isiZulu, toegepas sou word.

- Sal die metodes wat gedurende hierdie studie ondersoek en ontwikkel is soortgelyke verbeteringe toon vir ander benaderings tot masjienvertaling?

’n Ander aspek van aanpassing is die gebruik van dieselfde dataselekteringsmetodes vir ’n ander benadering tot datagedrewe masjienvertaling, soos voorbeeldgebaseerde masjienvertaling. Dit sou moontlik wees dat dieselfde tegnieke ook by ander benaderings tot masjienvertaling ewe goed sou werk, maar die moontlikheid bestaan ook dat ’n ander tipe benadering tot masjienvertaling ’n ander benadering tot dataselektering sal vereis.

4.3. SLOT

Soos in Hoofstuk 1 genoem, is masjienvertaling 'n tegnologie wat 'n baie groot impak op Suid-Afrika kan hê. Indien masjienvertaalsisteme vir al ons amptelike landstale beskikbaar was, sou inligting baie meer vrylik beskikbaar wees en sou dit aanleiding kon gee tot die verhoogde toegang tot inligting in die moedertale van die meerderheid Suid-Afrikaners. Die grootste struikelblok tot die realisering van so 'n droom is die tekort aan parallelle data vir die amptelike Suid-Afrikaanse tale. Met die tegnieke wat in hierdie studie ondersoek en ontwikkel is, kan die beskikbare data so goed moontlik benut word. Indien daar van professionele vertalers gebruik gemaak word om tekste uit Engels na die ander tale te vertaal in 'n poging om die parallelle korpora te vergroot, kan die metodes wat in hierdie studie ondersoek en ontwikkel is, gebruik word om data vir vertaling te selekteer. In hierdie studie is daar bewys gelever dat die sorgvuldige selektering van data wel baie vinniger tot 'n sekere vlak van masjienvertaling kan lei. Op hierdie wyse word daar verseker dat die fondse wat vir vertaling van data beskikbaar is so doeltreffend moontlik aangewend word en dat die uiteindelijke masjienvertaalsisteem goeie resultate kan behaal ten spyte van die tekort aan groot hoeveelhede parallelle korpora.

BRONNELYS

CALLISON-BURCH, C., OSBORNE, M. & KOEHN, P. 2006. Re-evaluating the role of bleu in machine translation research. (*In Proceedings of the 11th Annual Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. Trento, Italy. p. 249-256.)

DIRIX, P., SCHUURMAN, I. & VANDEGHINSTE, V. 2005. 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, Thailand. p. 43-50.)

DODDINGTON, G. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. (*In Marcus, M., ed. Proceedings of the 2nd International Conference on Human Language Technology Research*. San Diego, California. p. 138-145.)

ENGELBRECHT, H.A. & SCHULTZ, T. 2005. Rapid development of an Afrikaans-English speech-to-speech translator. (*In International Workshop on Spoken Language Translation. Conference Proceedings of the IWSLT*. Pittsburgh, Pennsylvania. p 1-8.)

FINCH, A. & SUMITA, E. 2008. Using statistical machine translation to grade training data. (*In Institute of Electrical and Electronics Engineers. Proceedings of the Second International Symposium on Universal Communication*. p. 113-119.)

GROENEWALD, H.J. & FOURIE, W. 2009. Introducing the Autshumato Integrated Translation Environment. (*In European Association for Machine Translation. Proceedings of the 13th Annual Conference of the EAMT*. Barcelona. p. 190-196.)

GRONDWET **see** SOUTH-AFRICA. 1996.

HAFFARI, G., ROY, M. & SARKAR, A. 2009. Active learning for statistical phrase-based machine translation. (*In Association of Computational Linguistics. Proceedings of Human Language Technologies. The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Boulder, Colorado. p. 415-423.)

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 Press. p. 431-445.)

HUTCHINS, J. 2001. Machine translation and human translation: in competition or in complementation?. *International journal of translation*, 13(1/2):5-20, January-December.

HUTCHINS, J.W. & SOMERS, H.L. 1992. *Introduction to machine translation*. London: Academic Press. 362 p.

IZUHA, T., KUMANO, A. & KURODA, Y. 2008. Toshiba rule-based machine translation system at NTCIR-7 PAT MT. (*In* *Proceedings of NTCIR-7 Workshop Meeting*. Tokyo. p. 430-434.)

JURAFSKY, D. & MARTIN, J.H. 2009. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*. 2nd ed. Upper Saddle River, N.J.: Prentice Hall. 988 p.

KATO, R.S.M. & BARNARD, E. 2007. Statistical translation with scarce resources: a South African case study. *SAIEE Africa research journal*, 98(4):26-30.

KIT, C., PAN, H. & WEBSTER, J.J. 2000. Example-based machine translation: a new paradigm. (*In* Chan, S.W., ed. *Translation and information technology*. Hong Kong: Chinese University of Honk Kong Press. p. 57-78.)

KOEHN, P. 2005. Europarl: a parallel corpus for statistical machine translation. (*In* *Proceedings of the Tenth Machine Translation Summit*. Phuket. p. 79-86.)

KOEHN, P. 2010. *Moses: statistical machine translation system: user manual and code guide*. <http://www.statmt.org/moses/manual/manual.pdf> Date of access: 19 April 2010.

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*

Association of Computational Linguistics, ed. Proceedings of the ACL 2007: demo and poster sessions. Prague. p. 177-180.)

LOPEZ, A. 2008. Statistical machine translation. *ACM computing surveys*, 40(3):1-47.

LU, Y., HUANG, J. & LIU, Q. 2007. Improving statistical machine translation performance by training data selection and optimization. (In Association of Computational Linguistics. Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Prague. p. 343-350.)

MANDAL, A., VERGYRI, D., WANG, W., ZHENG, J., STOLCKE, A., TUR, G., HAKKANITUR, D. & AYAN, N.F. 2008. Efficient data selection for machine translation. (In Institute of Electrical and Electronics Engineers. Proceedings of the 2008 IEEE workshop on Spoken Language Technology. Goa. p. 261-264.)

MOORE, R. 2002. Fast and accurate sentence alignment of bilingual corpora. (In Richardson, S.D., ed. Machine translation: from research to real users. Berlin: Springer. p. 135-144.) (Lecture notes in Computer Science, vol 2499.))

NATION, I.S.P. 2001. Learning vocabulary in another language. Cambridge: Cambridge University Press. 477 p. (Cambridge Applied Linguistics series.)

OCH, F.J. & NEY, H. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19-51, March.

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

PHAROS DICTIONARIES. 2006. Media24-Korpus. Kaapstad: Pharos.

PLITT, M. & MASSELOT, F. 2010. A productivity test of statistical machine translation post-editing in a typical localisation context. *Prague bulletin of mathematical linguistics*, 93(1):7-16, January.

SNOVER, M., DORR, B., SCHWARTZ, R., MICCIULLA, L. & MAKHOUL, J. 2006. A study of translation edit rate with targeted human annotation. (*In* Association for Machine Translation in the Americas. Proceedings of the 7th Conference of the Association for Machine Translation in the Americas. Cambridge, Massachusetts. p. 223-231.)

SNYMAN, C., EHLERS, L. & NAUDE, J. 2007. Development of the EtsaTrans translation system prototype and its integration into the Parnassus Meeting Administration System. *Southern African linguistics and applied language studies*, 25(2):225-238.

SOMERS, H. 1999. Review article: Example-based machine translation. *Machine translation*, 14(2):113-157.

SOUTH-AFRICA. 1996. Constitution of the Republic of South Africa. <http://www.info.gov.za/documents/constitution/index.htm>. Date of access: 16 November 2010.

STOLCKE, A. 2002. SRILM - an extensible language modelling toolkit. (*In* Hansen, J.H.L. & Pellom, B., eds. Proceedings of the 7th International Conference on Spoken Language Processing. Denver, Colorado. ISCA Archive. p. 901-904.)

TONG, S. 2001. Active learning: theory and applications. Palo Alto, Calif.: Stanford University. (Thesis - PhD.) 184 p.

TYERS, F. 2010a. Afrikaans and English. http://wiki.apertium.org/wiki/Afrikaans_and_English. Date of access: 08 February 2011.

TYERS, F.M. 2010b. Rule-based Brenton to French machine translation. (*In* European Association for Machine Translation. Proceedings of 14th Annual Conference of the EAMT. Saint-Raphael. p. 1-8.)

ZENS, R., OCH, F.J. & NEY, H. 2002. Phrase-based statistical machine translation. (*In* Jarke, M., Koehler, J. & Lakemeyer, G., eds. KI 2002: Advances in artificial intelligence. Berlin: Springer. p. 18-32.)