SKONER EN KLEINER VERTAALGEHEUES

deur

FRIEDEL WOLFF

voorgelê luidens die vereistes vir die graad

PHILOSOPHIAE DOCTOR

in die vak

REKENAARWETENSKAP

aan die

UNIVERSITEIT VAN SUID-AFRIKA

STUDIELEIER: PROF. LAURETTE PRETORIUS

MEDESTUDIELEIER: DR. PAUL BUITELAAR

Januarie 2018

Ek verklaar dat *Skoner en kleiner vertaalgeheues* my eie werk is en dat alle bronne wat ek gebruik of aangehaal het, erken is en deur middel van volledige verwysings aangedui is.

Verder verklaar ek dat ek nie vantevore hierdie werk, of 'n deel daarvan, vir eksaminering by Unisa of enige ander instelling van hoër onderrig ingedien het nie.

SAMEVATTING

Rekenaars kan ’n nuttige rol speel in vertaling. Twee benaderings is vertaalgeheuestelsels en masjienvertaalstelsels. By hierdie twee tegnologieë word ’n vertaalgeheue gebruik — ’n tweetalige versameling vorige vertalings. Hierdie proefskrif bied metodes aan om die kwaliteit van ’n vertaalgeheue te verbeter.

’n Masjienleerbenadering word gevolg om foutiewe inskrywings in ’n vertaalgeheue te identifiseer. ’n Verskeidenheid leerkenmerke in drie kategorieë word aangebied: kenmerke wat verband hou met tekslengte, kenmerke wat deur kwaliteittoetsers soos vertaaltoetsers, ’n speltoetser en ’n grammatikatoetser bereken word, asook statistiese kenmerke wat met behulp van eksterne data bereken word.

Die evaluasie van vertaalgeheuestelsels is nog nie gestandaardiseer nie. In hierdie proefskrif word ’n verskeidenheid probleme met bestaande evaluasiemetodes uitgewys, en ’n verbeterde evaluasiemetode word ontwikkel.

Deur die foutiewe inskrywings uit ’n vertaalgeheue te verwyder, is ’n kleiner, skoner vertaalgeheue beskikbaar vir toepassings. Eksperimente dui aan dat so ’n vertaalgeheue beter prestasie behaal in ’n vertaalgeheuestelsel. As ondersteunende bewys vir die waarde van ’n skoner vertaalgeheue word ’n verbetering ook aangedui by die opleiding van ’n masjienvertaalstelsel.

SLEUTELWOORDE

taaltegnologie; natuuurliketaalverwerking; vertaalgeheue; rekenaargesteunde vertaling; parallelle korpus; datakwaliteit; masjienleer; masjienvertaling; regressie; klassifikasie

CLEANER AND SMALLER TRANSLATION MEMORIES: ABSTRACT

Computers can play a useful role in translation. Two approaches are translation memory systems and machine translation systems. With these two technologies a translation memory is used — a bilingual collection of previous translations. This thesis presents methods to improve the quality of a translation memory.

A machine learning approach is followed to identify incorrect entries in a translation memory. A variety of learning features in three categories are presented: features associated with text length, features calculated by quality checkers such as translation checkers, a spell checker and a grammar checker, as well as statistical features computed with the help of external data.

The evaluation of translation memory systems is not yet standardised. This thesis points out a number of problems with existing evaluation methods, and an improved evaluation method is developed.

By removing the incorrect entries in a translation memory, a smaller, cleaner translation memory is available to applications. Experiments demonstrate that such a translation memory results in better performance in a translation memory system. As supporting evidence for the value of a cleaner translation memory, an improvement is also achieved in training a machine translation system.

KEYWORDS

language technology; natural language processing; translation memory; computer-assisted translation; parallel corpus; data quality; machine learning; machine translation; regression; classification

VOORWOORD

In die vroeë jare 2000 het ek per geleentheid kans gehad om industriekongresse by te woon soos die Localisation Research Centre in Limerick, Ierland se LRC-kongres. Ek het diep onder die indruk gekom van die opgang wat masjienvertaling (destyds veral statistiese masjienvertaling) in die kommersiële lokaliseringsindustrie maak. My kennismaking met statistiese masjienvertaling in daardie dae het 'n paar indrukke gelaat:

- Die stand van sake destyds was sodanig dat 'n redelike kwaliteit bereik kon word in sekere taalpare, veral as hulle taalkundig verwant is. Dié sukses was slegs haalbaar met reusehoeveelhede data van die regte kwaliteit en domein.
- Baie verwerkingskrag is nodig om stelsels op te lei met nietriviale datastelle.
- Die hulpbronne en moeite wat nodig is om die belofte van masjienvertaling te verwerklik, was buite bereik vir min of meer alle taalpare — selfs heelwat taalpare met Engels as brontaal. As mens dink aan die duisende tale van die wêreld, of selfs net dié met min of meer gestandaardiseerde ortografieë, sou mens tong in die kies kon sê dat die waarskynlikheid van statistiese masjienvertaling weglaatbaar gering is.
- Sekere taalkundige kompleksiteite het in daardie stadium geblyk nog buite die bestek van die ondersoek te val. Veral die hantering van tale met komplekse morfologie was problematies.
- 'n Realiteit wat sake verder kompliseer, is dat minderheidstale ook weens ekonomiese oorwegings soos 'n kleiner mark, nie ondersteuning van masjienvertaaldienste geniet nie, en dat vertaalgeheuestelsels juis deur vertalers in sulke tale benodig word [31].

Ek het gevolglik besluit om my daarop toe te spits om vertaalgeheuestelsels te verbeter, aangesien dit ’n eenvoudiger tegnologie is en baie makliker aan vertalers hulp kan verleen in enige taalpaar. Deur vertaling te vergemaklik, sal dit weldra bydra tot die korpus van parallelle tekste wat nodig is vir die opleiding van masjienvertaalstelsels.

Alhoewel daar baie jare verloop het sedert ek hierdie besluit geneem het, blyk die indrukke in ’n groot mate steeds waar te wees. Alhoewel die vereistes vir verwerkingskrag vir die bou van statistiese stelsels deesdae makliker bereikbaar is, eis die neurale benadering opnuut sy pond vleis, en is die bou van ’n goeie byderwetse stelsel weereens iets wat slegs vir goed toegeruste navorsingseenhede (kommersieel of akademies) toeganklik is.

Daarteenoor is selfs die goedkoopste persoonlike rekenaar van ’n dekade gelede voldoende om ’n vertaler met ’n vertaalgeheue by te staan. Dit is realisties vir ’n vertaler om vanaf ’n leë geheue te begin en geleidelik ’n hulpbron op te bou wat toenemend nuttiger word. Enige moontlike verbetering in vertaalgeheuestelsels sal van groot waarde wees vir veral die kleinste tale met die minste hulpbronne, waar ’n vertaalgeheuestelsel dalk die mees gevorderde vertaalsteun is waartoe ’n vertaler toegang het.

Met hierdie proefskrif hoop ek om ’n bydrae te lewer tot voordeel van vertalers en kwesbare tale wêreldwyd.

INHOUDSOPGAWE

1 INLEIDING 1
- 1.1 Agtergrond . . . 2
- 1.2 Probleemstelling . . . 10
- 1.3 Doel . . . 11
- 1.4 Metodologie . . . 12
- 1.5 Bydrae . . . 14
- 1.6 Oorsig oor die proefskrif . . . 15

2 LITERATUUROORSIG 17
- 2.1 Die waarde van parallelle korpusse . . . 18
- 2.2 Die impak van lae kwaliteit . . . 19
- 2.3 Bou en onderhoud van vertaalgeheues . . . 21
- 2.4 Skoonmaak . . . 24
- 2.5 Evaluasie . . . 26
- 2.6 Gevolgtrekking . . . 29

3 METODOLOGIESE SLAGGATE IN DIE EVALUASIE VAN VERTAALGEHEUES 31
- 3.1 Bestaande evaluasiemetodes . . . 33
- 3.2 Soortgelykheidsmate . . . 37
- 3.3 Implementasiebesonderhede . . . 45
- 3.4 Eksperimentele opset . . . 51
- 3.5 Bespreking . . . 62
- 3.6 Gevolgtrekking . . . 64

4 OBJEKTIEWE WAARDE VAN SOORTGELYKHEIDSMATE 67
- 4.1 Lineêre regressie . . . 68
- 4.2 Aanvangsdata . . . 72
- 4.3 Verrykte data . . . 74
- 4.4 Aanvangsmodel . . . 75
- 4.5 Verfynde model . . . 79
- 4.6 Resultate . . . 88
- 4.7 Gevolgtrekking . . . 90

5 IDENTIFISERING VAN VUIL INSKRYWINGS IN 'N VERTAALGEHEUE 93

5.1 Die probleem 95
5.2 Klassifikasie in masjienleer 97
5.3 Kenmerke 100
5.4 Klassifiseerders 110
5.5 Belangrikheid van kenmerke 112
5.6 Resultate 115
5.7 Gevolgtrekking 117
6 EVALUASIE VAN SKONER VERTAALGEHEUES 119
6.1 Gekontroleerde, semigekontroleerde en ongekontroleerde leer 121
6.2 Aanpassing vir ongekontroleerde leer 121
6.3 Intrinsieke evaluasie 125
6.4 Ekstrinsieke evaluasie: vertaalgeheuestelsel 128
6.5 Ekstrinsieke evaluasie: masjienvertaalstelsel 131
6.6 Gevolgtrekking 135
7 SLOT 137
7.1 Bydraes 137
7.2 Beperkinge en toekomstige werk 138
7.3 Breëre betekenis van die werk 139

LYS VAN FIGURE

Figuur 3.1 DGT en-fr onder edit4 56
Figuur 3.2 DGT en-fr onder edit4ngram 57
Figuur 3.3 DGT en-fr onder ngp 58
Figuur 4.1 Die residue teen edit3 78
Figuur 4.2 Histogram van tyd 80
Figuur 4.3 Histogram van $\log(\mathtt{tyd}+1)$, tyd in ms. . 80
Figuur 4.4 Histogram van $\log(\mathtt{tyd}+1)$, tyd in s. . . . 81
Figuur 4.5 Histogram van dist_edit4 85
Figuur 4.6 Histogram van $\log(\mathtt{dist\_edit4}+1)$ 86

LYS VAN TABELLE

Tabel 3.1 Gewigte by die sleuteldrukmaat 43
Tabel 3.2 Grense op lengtes 1 46
Tabel 3.3 Grense op lengtes 2 49
Tabel 3.4 Korpusstatistiek 52
Tabel 3.5 Resultate: DGT en-hu: F_1-tellings 60
Tabel 3.6 Resultate: DGT en-fr: F_1-tellings 61
Tabel 3.7 Resultate: GNOME en-hu: F_1-tellings . . . 62
Tabel 3.8 Resultate: GNOME en-fr: F_1-tellings . . . 63
Tabel 4.1 Korrelasies tussen afstandmate 82
Tabel 4.2 Gemiddelde R^2 per model 88
Tabel 5.1 Verdeling van klasetikette 96
Tabel 5.2 Reëls in LanguageTool 106
Tabel 5.3 Korrelasie voor en ná aanpassings 108
Tabel 5.4 Leksikale reëls wat in elke rigting onttrek is111
Tabel 5.5 Resultate vergeleke met ander stelsels . . 116
Tabel 6.1 Resultate: slegs segmente uit klas 1 126
Tabel 6.2 Resultate: foutiewe klas 1-aanname 127
Tabel 6.3 Resultate: oorspronklike DGT-subversameling 130

Tabel 6.4 Resultate: skoner DGT-subversameling . . 130
Tabel 6.5 Resultate: EU-boekwinkelkorpus 134
Tabel 6.6 Resultate: nuuskommentaarkorpus 134

1

INLEIDING

Vertaling is die proses waarmee teks in een taal oorgedra word in ’n ander taal. Baie vertalers gebruik rekenaars as deel van hulle vertaalwerk, en die gebruik van toegewyde sagteware daarvoor is nou algemeen. ’n Gewilde hulpmiddel is die *vertaalgeheue*[*] — ’n versameling vorige vertalings. ’n *Vertaalgeheuestelsel*[*] is sagteware wat die vertaler met relevante voorstelle uit ’n vertaalgeheue voorsien tydens vertaling.

(en) *translation memory, TM*

(en) *translation memory system*

Die bestuur van vertaalgeheues word gesien as ’n noodsaaklike taak vir ’n vertaler of agentskap wat ’n groot versameling vertaalgeheues het. Een aspek hiervan is kwaliteitsbeheer van die geheues. Die teenwoordigheid van foutiewe inskrywings in die geheue beteken dat voorstelle aan ’n gebruiker gewys kan word wat die vertaalproses belemmer. Deur sulke inskrywings te verwyder, kan die werking van die stelsel dus verbeter word. As so ’n projek wetenskaplik aangepak word, sal die vraag noodwendig ontstaan oor hoe só ’n verbetering geëvalueer kan word. Die evaluasie van vertaalgeheuestelsels is nie ’n uitgemaakte saak nie, en dit beperk navorsing in hierdie area.

Uit bogenoemde blyk dit dat rekenaarsteun vir die kwaliteitverbetering van ’n vertaalgeheue praktiese waarde het. Verder is die probleem van hoe om te evalueer tersaaklik vir alle navorsers in hierdie veld.

Alhoewel enkele belangrike terme pas kortliks verklaar is, sal daar volgende meer sorgvuldig gekyk word na sekere belangrike konsepte vir die verstaan van hierdie werk. Daarna kan ’n meer volledige agtergrond tot die studie verskaf word.

1.1 AGTERGROND

In hierdie afdeling word 'n bondige agtergrond verskaf wat nodig is vir die res van hierdie inleiding. Meer volledige agtergrond tot die res van die proefskrif word in **hoofstuk 2** aangebied. Alhoewel enkele terme reeds aan die begin kortweg gedefinieer is, sal hier in meer diepte na hierdie terme gekyk word, en ook na 'n paar ander terme wat binnekort gebruik word.

1.1.1 *Vertaling*

Vertaling is die eeue-oue proses waarmee teks in een taal oorgedra word in 'n ander taal om 'n soortgelyke boodskap oor te dra of 'n soortgelyke funksie te verrig. Die oorspronklike teks word die bron of bronteks* genoem, en die eindresultaat van die vertaling word die doel of doelteks* genoem.

(en) *source text*

(en) *target text*

Nie alle tipes vertaalwerk vind ewe veel baat by die gebruik van 'n vertaalgeheue nie. Die vermoë om 'n vorige vertaling weer te gebruik, is veral van waarde in regstekste, tegniese tekste (bv. handleidings, programkoppelvlakke), of enige teks met heelwat herhaling. Dit is ook van waarde in 'n geval waar 'n dokument vertaal word wat wesenlik dieselfde is as in 'n vorige vertaaltaak. Só hoef die vertaler meestal slegs aandag te skenk aan die teks wat nuut is in die dokument wat voorhande is.

1.1.2 *Rekenaargesteunde vertaling*

In die twintigste eeu is begin om van rekenaars gebruik te maak om allerlei aktiwiteite te ondersteun en/of te outomatiseer. Vertaling is nie 'n uitsondering nie. Reeds in die 1940's is volledig outomatiese vertaling met 'n rekenaar voorsien as 'n moontlikheid [40]. Nadat groot hoeveelhede befondsing vir navorsing oor sulke masjienvertaling onttrek is in die VSA weens die ALPAC-verslag [68] het die fokus in 'n mate verskuif in die rigting

van rekenaars as 'n hulpmiddel vir mense wat vertaal. Martin Kay het reeds in 1980 (ten spyte van sy skeptisisme oor volledig outomatiese vertaling) geargumenteer dat rekenaars tog 'n onontbeerlike rol kan speel in die vertaalproses omdat dit die druk op die vertaler vir die hantering van meganiese en roetine-aspekte kan verlig, en daardeur ook die proses vir die vertaler meer aangenaam kan maak [45].

Programme vir rekenaargesteunde vertaling* bevat dikwels 'n verskeidenheid funksionaliteite:

(en) *computer assisted translation, CAT*

- 'n vertaalgeheuestelsel (sien afdeling 1.1.4);
- konkordansiesoektogte in die vertaalgeheue;
- terminologiebestuur en terminologiehulp;
- hantering van plaasbare items* [3];

(en) *placeables*

- soekfunksionaliteit;
- omskakeling van lêerformate;*

(en) *file formats*

- statistiek — rapportering oor die grootte van die vertaaltaak en die verwagte waarde uit 'n vertaalgeheue;
- kwaliteitsbeheer — die bespeuring van kwaliteitsprobleme in die vertaling, bv. deur speltoetsers of ander fouttoetsers;
- belyningsfunksionaliteit* om 'n vertaalgeheue te bou vanaf 'n brondokument en sy vertaling.

(en) *alignment funtionality*

Hierdie proefskrif handel hoofsaaklik oor die werking van die vertaalgeheuestelsel, maar dit is waardevol om kennis te neem van sommige van die ander funksionaliteit. Die funksies vir kwaliteitsbeheer is relevant vir 'n poging om die kwaliteit van 'n vertaalgeheue te verbeter. Verder is dit ook goed om in gedagte te hou dat 'n vertaalgeheue gewoonlik gebruik word in 'n konteks van meerdere vertaalhulpmiddels. Die impak van 'n tekortkoming in die vertaalgeheuestelsel kan moontlik verminder word deur ander funksies in die sagteware.

1.1.3 *Vertaalgeheue*

(en) *translation memory, TM*

'n Vertaalgeheue* is 'n databasis van geassosieerde bron- en doeltekste. 'n Kerndoelwit in die opstel van 'n vertaalgeheue is spesifiek om hergebruik van die tekste te fasiliteer. Dit is algemeen dat segmentpare wat sinne verteenwoordig, gestoor word, maar dit is ook moontlik om te dink aan 'n vertaalgeheue as 'n versameling tekste op 'n fyner of growwer vlak van verdeling, bv. terme of paragrawe. In die geval waar paragraaf- of sinsegmentering gebruik word, is dit steeds moontlik dat opskrifte, lysinskrywings en ander korter fragmente ook in die geheue opgeneem word. Die term vertaalgeheue word soms gebruik om te verwys na die data asook die sagteware wat daarmee werk. Ter wille van duidelikheid word daar in hierdie proefskrif onderskei tussen hierdie twee konsepte. Die sagteware wat die navrae en bestuur van die vertaalgeheue behartig, word in hierdie werk 'n vertaalgeheuestelsel genoem, wat in die volgende afdeling bespreek word.

Die nut van 'n vertaalgeheue lê daarin dat 'n vertaler inspirasie kan put uit vorige vertalings en sodoende die proses versnel. 'n Vorige vertaling van 'n soortgelyke sin kan gebruik word as vertrekpunt, en benodig moontlik geen of min veranderinge, wat tyd kan spaar. Die gebruik van vertaalgeheues het ook waarde omdat dit kan help met konsekwentheid van terminologie en styl omdat voorstelle aan die vertaler voorgehou word vanuit vorige vertaalwerk. Waar 'n groot vertaaltaak deur meer as een vertaler behartig word, help die voorstelle van verwante vertaalwerk om makliker konsekwentheid tussen vertalers te weeg te bring.

Tydens die interaktiewe vertaalproses stoor 'n program vir rekenaargesteunde vertaling 'n vertaalgeheue as 'n byproduk. Die program stoor die geheue volgens die program se behoeftes, waarskynlik met 'n ontwerp wat geoptimeer is sodat navrae vinnig beantwoord kan word. Sulke databasisse is nie noodwendig ideaal vir die uitruil van vertaalgeheues nie, aangesien dit dalk nie oordraagbaar is tussen stelsels nie. Hiervoor is doelgemaakte lêerformate ontwikkel, waarvan die bekendste

TMX is — *Translation Memory eXchange*. As 'n gestandaardiseerde XML-formaat behoort uitruiling tussen programme wat die standaard korrek implementeer, te slaag.

Die uitruil van vertaalgeheues is aantreklik omdat dit 'n gebruiker in staat stel om voorstelle vanuit 'n groter, gekombineerde vertaalgeheue te ontvang. Om die meeste waarde uit bestaande vertaalgeheues te put, kan 'n vertaalagentskap al hulle vertaalwerk versamel en voortdurend poog om hulle prosesse te verbeter deur van vorige vertalings gebruik te maak en sodoende koste te verlaag. 'n Doelgerigte kombinering van geheues per kliënt, industrie of genre word dalk verlang, afhangend van die omstandighede. Die optimale bestuur van vertaalgeheues kan help om die meeste waarde daaruit te put.

Wanneer vertaalgeheues vanuit verskillende bronne gekombineer word, is dit natuurlik in 'n mate te verwagte dat daar 'n mengsel van vertaalstyle sal wees. Die vertaalwerk wat in die geheues vervat word, moes dalk vanweë kliëntevereistes aan verskillende stylgidse, voorskriftelike termlyste, ens. voldoen. Dit is ook moontlik dat nie alle vertaalwerk aan dieselfde mate van kwaliteitsbeheer onderwerp is nie. In so 'n geval kan 'n gekombineerde vertaalgeheue dus materiaal van varierende kwaliteit bevat tensy dit eers nagegaan en moontlik geredigeer word. Probleme in die proses van uitruiling (bv. weens gebrekkige TMX-implementasies) of kwaliteitsprobleme in die inhoud van gekombineerde geheues kan alles lei tot 'n vertaalgeheue wat problematiese inskrywings bevat. Só 'n vertaalgeheue word dan "vuil" genoem. Die skoonmaak van 'n vuil vertaalgeheue kan gesien word as deel van die bestuur van vertaalgeheues.

Ander aktiwiteite in natuurliketaalverwerking* gebruik ook parallelle tekste soos vertaalgeheues. 'n Belangrike toepassing is die opleiding van masjienvertaalstelsels. Ander toepassings sluit bv. tweetalige termonttrekking* [55] en kruistalige inligtingherwinning* [64] in.

(en) *natural language processing*

(en) *terminology extraction*

(en) *cross-lingual information retrieval, CLIR*

1.1.4 *Vertaalgeheuestelsel*

(en) *translation memory system*

'n Vertaalgeheuestelsel* is 'n program of programmodule wat voorstelle uit 'n vertaalgeheue onttrek. Die bestuur van die geheue kan gesien word as deel van die stelsel se funksies. Navrae kan interaktief op aanvraag hanteer word, of vooraf alles tesame [83]. In interaktiewe stelsels is dit belangrik dat die data en navrae effektief hanteer word sodat voorstelle vinnig gelewer kan word. In so 'n gebruiksopset sal die vertaalomgewing telkens outomaties die volgende segment van die bronteks as navraag vir die vertaalgeheue sien. Dit is ook moontlik om voorstelle vooraf te genereer voordat die vertaler die vertaaltaak aanpak. So 'n benadering kan dus gesien word as bondelverwerking.* In so 'n geval is die vermoë om 'n vertaalgeheuenavraag vinnig te kan beantwoord nie so belangrik nie.

(en) *batch procesesing*

Die bestuur van 'n vertaalgeheue kan aspekte insluit soos die byvoeging of verwydering van inskrywings. Byvoeging kan plaasvind as deel van die vertaalproses, of op aanvraag, bv. as 'n eksterne vertaalgeheue bygevoeg moet word.

'n Nuttige vertaalgeheuestelsel sal ook 'n manier hê om goeie voorstelle uit die databasis te onttrek, en nie die gebruiker te belas met voorstelle wat nie help nie. Ten einde 'n voorstel te onttrek wat vir die gebruiker van waarde sal wees, poog die stelsel om die rekord van bron- en doelteks uit die vertaalgeheue te onttrek waarvan die bronteks dieselfde of soortgelyk is aan die brontekssegment wat die gebruiker op daardie oomblik vertaal (die navraag aan die stelsel). Hoe hierdie soortgelykheid bepaal word of bepaal behoort te word, is 'n oop vraag. (Sien bv. die ondersoek in [82].) Hierdie soortgelykheidsmate* word in meer detail in hoofstuk 3 bespreek. Die bron- en doelteks wat uit die databasis onttrek is, word dan aan die gebruiker gewys. Die bronteks help om te sien wat die verwantskap is met die huidige segment wat vertaal word, en die doelteks kan gewoonlik maklik hergebruik word; 'n suksesvolle voorstel sal geen of slegs enkele veranderinge benodig.

(en) *similarity metrics*

Aangesien elke voorstel wat aan 'n gebruiker gewys word 'n kognitiewe las is [65], beperk die stelsel die aantal voorstelle.

Afgesien van 'n perk op die getal voorstelle per navraag, word voorstelle ook beperk tot dié wat hoogs waarskynlik vir die gebruiker nuttig sal wees. Só poog die stelsel om die aantal nuttelose voorstelle te beperk sodat hulle nie 'n onnodig negatiewe invloed op die produktiwiteit van die gebruiker sal hê nie. Hierdie besluit realiseer tipies as 'n vergelyking met 'n drempelwaarde* vir die soortgelykheidsmaat. Só 'n soortgelykheidsdrempel kan uitgebeeld word as bv. 70% of 0,7. 'n Inskrywing in die vertaalgeheue met 'n soortgelykheid laer as die drempel (bv. 45%) word nie aan die gebruiker gewys nie.

(en) *threshold value*

Sekere vertaalgeheuestelsels analiseer die teks op meer as een vlak. In 'n geval waar sinne as die eenheid van segmentasie hanteer word, kan die stelsel bv. sogenaamde subsegmentvoorstelle probeer genereer en stoor. Só kan frases (hetsy taalkundige frases of nie) ook voorgestel word. In hierdie werk word daar nie ondersoek ingestel na enige werking onder die segmentvlak nie.

1.1.5 *Masjienvertaling*

Aangesien vertaalgeheuestelsels in die konteks van taaltegnologie bestaan, is dit sinvol om 'n oomblik te neem om iets te sê oor die aanverwante veld van masjienvertaling.* 'n Masjienvertaalstelsel poog om 'n korrekte vertaling van teks volledig outomaties te genereer. Masjienvertaling saam met redigering agterna is al met groot sukses in vertaalprojekte gebruik. 'n Ander gebruik van masjienvertaling is om 'n leser 'n rowwe indruk te gee van 'n teks in 'n taal wat die leser nie goed verstaan nie.

(en) *machine translation, MT*

Verskillende benaderings tot masjienvertaling bestaan. Veral van belang vir hierdie studie, is die benaderings wat van 'n korpus parallelle teks gebruik maak. Statistiese masjienvertaling* is 'n benadering waarvolgens "vertaalreëls outomaties deur die sagteware ontdek word in 'n groot korpus van paral-

(en) *statistical machine translation, SMT*

lelle teks"[1] [50, p. xi]. Hierdie parallelle teks is in wese niks anders as 'n vertaalgeheue nie. Die jonger benadering van neurale masjienvertaling* gebruik ook parallelle teks as opleidingsdata en ontdek eweneens 'n stel vertaalreëls — weliswaar op 'n geheel ander manier [51].

(en) *neural machine translation*

Die evaluasie van masjienvertaling is 'n aktiewe navorsingsveld. By die jaarlikse werkswinkel vir masjienvertaling[2] is daar 'n toegewyde taak spesifiek vir die evaluasie van masjienvertaling. Juis omdat die stelsels gemene eienskappe het, is die evaluasie van masjienvertaalstelsels ook relevant tot hierdie studie. Veral die volledig outomatiese evaluasie is algemeen omdat dit vinnig, goedkoop en herhaaldelik gedoen kan word. Outomatiese evaluasie gebruik een of meer verwysingsvertalings tydens evaluasie. 'n Vergelyking van die afvoer van die stelsel met die verwysingsvertalings gee 'n aanduiding van die kwaliteit van die masjienvertaling. Die veld van kwaliteitskatting* in masjienvertaling het ontstaan as 'n poging om die kwaliteit van masjienvertaling te skat sonder verwysingsvertalings. Al twee hierdie aktiwiteite het raakpunte met werk wat in hierdie proefskrif vervat is.

(en) *quality estimation*

1.1.6 *Vertaalgeheuestelsels teenoor masjienvertaling*

Die bespreking hierbo oor masjienvertaling is geensins daarop gemik om 'n deeglike agtergrond tot die veld te gee nie. Dit is wel belangrik om masjienvertaling te bespreek om die navorsing beter te begrens deur masjienvertaling te kontrasteer met vertaalgeheuestelsels. Daar is heelwat oorvleueling tussen vertaalgeheuestelsels en masjienvertaalstelsels, veral statistiese masjienvertaalstelsels:

- Beide stelsels maak gebruik van 'n korpus van tweetalige materiaal.

1 "... that discovers the rules of translation automatically from a large corpus of translated text, ..."

2 Sien bv. http://www.statmt.org/wmt17/metrics-task.html

- Beide stelsels ontvang ’n toevoersegment in die bronteks as navraag.
- Beide stelsels poog om ’n bruikbare doelteks te lewer aan die gebruiker, alhoewel hier groot verskille is in hoe en wanneer hierdie voorstelle ontstaan.
- Beide stelsels kan gebruik word om ’n vertaalproses te versnel of te verbeter.

Hierdie twee tipes stelsels gebruik dus deels soortgelyke data, en die interaksie met ’n gebruiker is soortgelyk. Hulle kan op vergelykbare maniere in ’n vertaalproses geïntegreer word.

Daar is ’n paar verskille tussen masjienvertaalstelsels en vertaalgeheuestelsels:

- ’n Vertaalgeheuestelsel genereer geen taal nie, maar onttrek slegs bestaande teks uit ’n databasis — die vertaalgeheue. Masjienvertaling behels aansienlik meer kompleksiteit, tegnieke, ens. en vereis aansienlik meer verwerkingskrag.
- ’n Masjienvertaling word vir elke toevoersegment gegenereer. Daarteenoor sal ’n vertaalgeheuestelsel slegs voorstelle lewer waar iets gepas uit die geheue geïdentifiseer is. (Sien die bespreking in afdeling 1.1.4 oor die soortgelykheidsdrempel.)
- Die doelteks in ’n voorstel uit ’n vertaalgeheue is meestal ’n vertaling vir ’n ander bronteks as wat tans oorweeg word. Daarteenoor poog ’n masjienvertaalstelsel om ’n bruikbare vertaling te genereer vir die huidige bronteks (binne die beperkinge van die masjienvertaalstelsel). Derhalwe is dit sinvol vir ’n voorstel uit ’n vertaalgeheue om ’n bronteks as deel van die voorstel te wys, terwyl die bronteks van die masjienvertaling noodwendig die navraag self is.

Die evaluasiemetodes vir masjienvertaling is nie optimaal vir die evaluasie van vertaalgeheuestelsels nie. Hier is ’n paar van die belangrike verskille tussen die twee stelsels se evaluasie:

- Omdat 'n masjienvertaalstelsel vir elke segment 'n vertaling lewer, reken sommige van die evaluasietegnieke daarop. Aangesien 'n vertaalgeheuestelsel nie noodwendig vir elke segment 'n voorstel lewer nie, word aan daardie aanname nie voldoen nie.

- Die soortgelykheidsdrempel (sien afdeling 1.1.4) is 'n belangrike deel van die werking van 'n vertaalgeheuestelsel. Hierdie waarde moet dus geëvalueer word.

- 'n Vertaalgeheue mag dalk 'n perfekte vertaling bevat vir die navraag, maar sal meestal nie. Dus is dit nie sinvol om die voorstel te evalueer as 'n vertaling van die navraag nie, aangesien die vertaalgeheuestelsel slegs probeer om hulp te verleen, nie om 'n korrekte vertaling te lewer nie.

- By masjienvertaling word vloeiendheid gewoonlik geëvalueer om te bepaal in watter mate die stelsel teks kon genereer soos 'n moedertaalspreker. Aangesien 'n vertaalgeheuestelsel se voorstel uit die geheue kom, is dit waarskynlik iets wat deur 'n mens geskryf is, en is hierdie aspek waarskynlik nie sinvol in die evaluasie van die vertaalgeheuestelsel nie.

As gevolg van hierdie verskille kan evaluasietegnieke vir masjienvertaling nie sonder meer vir vertaalgeheuestelsels gebruik word nie.

Alhoewel daar heelwat raakpunte is tussen vertaalgeheuestelsels en masjienvertaalstelsels, is dit belangrik om kennis te neem van die klemverskille, veral wat evaluasie betref. Daar is wel heelwat te put uit die literatuur en metodes in die veld van masjienvertaling. Hierdie studie het dit egter ten doel om spesifiek in die veld van vertaalgeheues 'n bydrae te lewer.

1.2 PROBLEEMSTELLING

'n Eenvoudige, maar onpraktiese oplossing vir die probleem van 'n vuil vertaalgeheue is vir 'n professionele vertaler om elke

inskrywing na te gaan en waar nodig te redigeer. Dit is egter ondoeltreffend, veral aangesien dit nie moontlik is om vooraf te weet watter inskrywings enigsins hergebruik gaan word nie. Die tyd wat gebruik word om inskrywings te redigeer wat nooit gebruik gaan word nie, is gemors.

Soos in afdeling 1.1.4 genoem, is die bestuur van die vertaalgeheue die verantwoordelikheid van die vertaalgeheuestelsel, en die bestuur van die kwaliteit kan gesien word as deel hiervan. Deur slegs 'n handmatige proses vir die gebruiker aan te bied, blyk nie bevredigend te wees nie. Die navorsingsvraag wat hom hieruit voordoen is die volgende:

Hoe kan 'n vertaalgeheue outomaties skoongemaak word?

As daar 'n manier bestaan om 'n vertaalgeheue d.m.v. sagteware skoon te maak deur problematiese inskrywings te verwyder, kan hierdie aspek van die bestuur van vertaalgeheues as deel van vertaalgeheuestelsels geïmplementeer word.

Om hierdie navorsingsvraag te beantwoord, gaan die volgende drie subvrae beantwoord word:

- Hoe kan vertaalgeheues en vertaalgeheuestelsels geëvalueer word?
- Watter tegniek(e) kan gebruik word om foutiewe inskrywings in 'n vertaalgeheue te identifiseer?
- Is 'n skoongemaakte vertaalgeheue meer geskik vir 'n spesifieke taak as vantevore?

Die vraag oor evaluasie is belangrik omdat 'n poging om vertaalgeheues skoon te maak, geëvalueer moet kan word. Sonder 'n evaluasietegniek, sal die laaste vraag nie beantwoord kan word nie.

1.3 DOEL

Die doel van die studie is om die bestuur van vertaalgeheues te verbeter. Die spesifieke fokus is op die skoonmaak van vuil

vertaalgeheues deur die identifikasie van die vuil inskrywings (eerder as deur die korrigering van dié inskrywings).

'n Skoon vertaalgeheue behoort minder plek te beslaan sonder dat dit sy doel inperk. As foutiewe inskrywings nie meer as voorstelle deur die gebruiker gesien kan word nie, kan hierdie foutiewe inskrywings nie meer afbreuk doen aan die vertaalproses nie. Melby noem dat slegte vertalings in vertaalgeheues weens hergebruik kan propageer en voeg by dat "vertaalgeheues konsekwentheid bevorder sonder dat dit noodwendig enige effek het op akkuraatheid"[3] [61, p. 665]. Die skoonmaak van vertaalgeheues blyk dus van praktiese waarde te wees in die konteks van rekenaargesteunde vertaling. Vir ander toepassings soos termonttrekking of die opleiding van 'n masjienvertaalstelsel, kan 'n skoner vertaalgeheue moontlik beter resultate lewer.

In die aanvanklike verkennende werk vir hierdie proefskrif het dit duidelik geword dat die evaluasietegnieke nie gestandaardiseer is nie, en dat daar min ooreenstemming is oor hoe evaluasie behoort te geskied. Goeie evaluasie verhoog 'n mens se vertroue in die resultate van 'n eksperiment, en kan meer betroubare kennis help genereer.

Aangesien ons bewus is van die gebrekkige stand van sake wat betref die evaluasie van vertaalgeheuestelsels, is 'n noodwendige vraag wat hieroor gevra moet word *In watter mate kan die verbetering geëvalueer word?* 'n Verdere doel van hierdie studie is om die wetenskaplike geldigheid van die evaluasie van vertaalgeheuestelsels te ondersoek sodat hierdie vraag beantwoord kan word.

1.4 METODOLOGIE

Hierdie navorsing volg 'n kwantitatiewe navorsingsontwerp en steun veral op eksperimente. Die identifikasie van die vuil inskrywings in 'n vertaalgeheue word hier beskou as 'n klas-

3 "... [Translation Memory] promotes consistency without necessarily having any effect on accuracy"

sifikasieprobleem* — elke inskrywing moet geklassifiseer word as "skoon" of "vuil".

(en) *classification problem*

Die gebrekkige stand van sake wat evaluasie betref het aanleiding gegee tot die eerste fase van werk waar spesifiek ondersoek ingestel word na evaluasiemetodes en hulle parameters. Die tweede fase fokus op tegnieke vir die skoonmaak van vertaalgeheues. In die derde fase word die skoongemaakte geheues geëvalueer aan die hand van prestasie in toepassings. Verskillende tegnieke en metodologiese oorwegings bestaan vir die werk in elke afdeling. Daarom word hierdie sake in volledige detail in elke afdeling bespreek waar dit van belang is, en hier word slegs enkele oorsigtelike opmerkings gemaak.

In die *eerste fase* word daar ondersoek ingestel na bestaande evaluasiemetodes vir vertaalgeheuestelsels. Deur middel van 'n eksperiment word aangetoon dat daar gebreke is in die manier waarop hierdie metodes vantevore ingespan is. 'n Verfyning van 'n vorige evaluasiemetode word voorgestel wat sekere metodologiese probleme vermy. Aan die hand van data van interaktiewe vertaalsessies word daar deur middel van regressie ondersoek ingestel na geskikte parameters vir die evaluasiemetode.

In die *tweede fase* verskuif die aandag na die sentrale probleem — die skoonmaak van 'n vertaalgeheue. 'n Masjienleerbenadering* tot die skoonmaak van vuil vertaalgeheues word aangebied, en in 'n reeks eksperimente toegepas. In hierdie fase word van standaardevaluasiemetodes in die veld gebruik gemaak, nl. akkuraatheid en F-telling.*

(en) *machine learning approach*

(en) *accuarcy,* F *score*

In die *derde fase* word die waarde van 'n skoner vertaalgeheue ondersoek. Afgesien van 'n blote ondersoek na die intrinsieke klassifikasievermoëns in 'n klassifikasieprobleem is dit ook van waarde om na die ekstrinsieke waarde van die skoner datastel te kyk. Hiervoor word bestaande evaluasietegnieke gebruik in 'n reeks eksperimente. Hierdeur word die waarde van die skoner datastel in enkele toepassings vergelyk met die oorspronklike, "vuil" datastel. Die evaluasietegniek wat in die eerste fase ontwikkel is, word ook hier ingespan.

In alle gevalle word die stappe in detail beskryf om 'n herhaling van die eksperimente moontlik te maak. Waar moontlik, word geredelik beskikbare data gebruik om dit moontlik te maak om die eksperimente met dieselfde data te herhaal. Waar gepas, is tegnieke soos kruisvalidasie gebruik om die betroubaarheid van die resultate te verhoog. Die statistiese interpretasie word aangebied om die betroubaarheid van die uitslae te kwantifiseer. Die werk in die eerste fase is ook juis daarop gemik om 'n evaluasiemetode voor te stel wat die betroubaarheid van resultate in die derde fase kan verhoog. Hierdie proefskrif lewer dus ook 'n bydrae tot die metodologiese aspekte in hierdie navorsingsveld. Die bydrae van die proefskrif word vervolgens in meer detail bespreek.

1.5 BYDRAE

Die sentrale hipotese van hierdie werk kan opgesom word in hierdie stelling:

> *As dit moontlik is om die kwaliteit van die vertaalgeheue se inhoud te verhoog, sal die vertaalgeheue van meer waarde wees.*

In die ondersoek van hierdie hipotese maak hierdie proefskrif 'n bydrae van metodologiese en praktiese waarde.

In die *eerste fase* word 'n nuwe metode vir die evaluasie van vertaalgeheuestelsels aangebied. Die evaluasietegniek kan gebruik word om stelsels asook datastelle te evalueer. Buiten die gebruik van hierdie tegniek later in hierdie proefskrif, kan die tegniek ook help om die stand van volledig outomatiese evaluasie op die gebied van vertaalgeheues en vertaalgeheuestelsels te verbeter. Met die verbetering van evaluasiemetodes word die gaping tussen navorsingsuitsette en industrietoepassing meer suksesvol oorbrug aangesien navorsingsresultate beter met die werklikheid ooreenstem.

(en) *machine learning approach*

In die *tweede fase* word 'n masjienleerbenadering* tot die skoonmaak van vuil vertaalgeheues aangebied. 'n Stel leerkenmerke* wat suksesvol 'n verskeidenheid foute identifiseer, word

(en) *learning features*

aangebied en gekombineer in 'n gekontroleerde benadering.* (en) *supervised approach*

Die *derde fase* dui aan wat die impak van 'n skoner vertaalgeheue is in sekere toepassings. Hiermee is die resultate ook van waarde vir navorsers en praktisyns buite die veld van vertaalgeheuestelsels. Hiervoor word die aanvanklike gekontroleerde benadering uitgebrei sodat 'n ongekontroleerde benadering* ook aangebied kan word. Twee benaderings vir die skoonmaak van vertaalgeheues word dus bygedra tot die veld. (en) *unsupervised approach*

Vier gepubliseerde artikels het reeds uit die navorsing vir hierdie proefskrif voortgevloei [85–88].

1.6 OORSIG OOR DIE PROEFSKRIF

Die probleemdomein is nou uiteengesit met die nodige agtergrond om die navorsing se bydrae duidelik te maak. Die agtergrond wat vroeër in afdeling 1.1 verskaf is, was slegs inleidend met die oog op hierdie hoofstuk. Verdere agtergrond wat nodig is vir die materiaal in hierdie proefskrif word volgende in **hoofstuk 2** aangebied. Dit gee 'n volledige agtergrond van bestaande literatuur wat relevant is tot hierdie proefskrif. Daarna volg die hoofbydraes van die proefskrif.

Die eerste, meer introspektiewe fase van die studie begin in **hoofstuk 3** met 'n ondersoek na die gebrekkige stand van outomatiese evaluasie van vertaalgeheuestelsels. Dit dui aan dat evaluasie, soos dit algemeen tot nou toe gedoen is in navorsing oor vertaalgeheuestelsels, probleme het wat wetenskaplike geldigheid betref. 'n Verfynde metode vir die evaluasie van vertaalgeheuestelsels word voorgehou. In **hoofstuk 4** word parameters ondersoek wat meer betroubare evaluasieresultate lewer in die pas voorgestelde evaluasiemetode. Die uitkoms van die eerste fase is dus 'n manier om ingrepe in 'n vertaalgeheuestelsel te evalueer wat later in die proefskrif gebruik gaan word.

Die tweede fase handel oor die skoonmaak van 'n vuil vertaalgeheue. In **hoofstuk 5** word 'n masjienleermetode om 'n vertaalgeheue mee skoon te maak, beskryf. Die aanvanklike be-

nadering gebruik 'n volledig geannoteerde datastel, en kan dus gesien word as 'n gekontroleerde masjienleerbenadering. Die hoofstuk bespreek die leerkenmerke en leeralgoritmes wat gebruik word, en die klassifikasiekwaliteit van hierdie benadering word intrinsiek geëvalueer.

Die derde fase is afgestem op meer praktiese toepassing en evaluasie. 'n Meer prakties bruikbare benadering tot skoonmaak is waar geen (of min) geannoteerde data beskikbaar is nie. Hiervoor word die tegniek in **hoofstuk 6** uitgebrei om 'n soortgelyke taak uit te voer sonder dat 'n groot datastel eers geannoteer hoef te word. Daarna word die ekstrinsieke evaluasie aan die hand van die nuwe evaluasiemetode van hoofstuk 4, asook in die toepassing van masjienvertaling gedoen. Hierdie evaluasie dui op die geskiktheid van 'n skoongemaakte vertaalgeheue vir toepassings — ook buite die veld van rekenaargesteunde vertaling.

Die proefskrif word afgesluit in **hoofstuk 7** met 'n bespreking van die bydraes van die navorsing en 'n blik op toekomstige werk.

2

LITERATUUROORSIG

In die voorwoord is die navorser se persoonlike motivering vir hierdie studie genoem en 'n aanvanklike inleiding is in die vorige hoofstuk aangebied. In hierdie hoofstuk word 'n meer diepgaande agtergrond verskaf van die literatuur wat relevant tot hierdie studie is. Hiermee word die wetenskaplike motivering en konteks vir die studie beter uiteengesit.

Die gebruik van parallelle korpusse as vertaalgeheues is reeds in hoofstuk 1 genoem. Toepassings van parallelle korpusse buite die veld van vertaalgeheues word vervolgens bespreek asook die impak van datakwaliteit op hierdie toepassings. Hierdie dui ook aan watter toepassings moontlik die meeste kan baat by die kwaliteitverbetering wat vanaf hoofstuk 5 aangepak word. Gebrekkige datakwaliteit in vertaalgeheues kan insluip onder andere weens die manier waarop die geheues gebou word. Hierdie hoofstuk bespreek die bou van vertaalgeheues en vorige werk om datakwaliteit te verbeter — ook uit aanverwante velde soos masjienvertaling. Die dekking van hierdie onderwerpe beklemtoon die relevansie van die navorsing binne die vakgebied van natuurliketaalverwerking.

In die veld van masjienvertaling geniet evaluasie baie aandag, maar dit is nie die geval by vertaalgeheuestelsels nie. Die nou verband tussen hierdie twee tipes stelsels en die relatief minder volwasse stand van evaluasie in vertaalgeheuestelsels relatief tot masjienvertaalstelsels noop mens om ook ondersoek in te stel na die evaluasie van masjienvertaalstelsels. Laastens word 'n oorsig oor die evaluasie van vertaalgeheuestelsels aangebied.

2.1 DIE WAARDE VAN PARALLELLE KORPUSSE

(en) *linguistics*

Parallelle korpusse word vir velerlei take gebruik. Dit word gebruik oor 'n wye verskeidenheid velde vanaf taalkunde* tot natuurliketaalverwerking, en die gebruike strek vanaf teoretiese navorsing tot praktiese toepassings.

In die taalkunde is dit aantreklik om die gedrag van woorde en ander taalverskynsels in parallelle korpusse te ondersoek. Isabelle het in 1993 die volgende geskryf oor die waarde van parallelle korpusse [41]:

> ... bestaande vertalings bevat meer oplossings vir meer vertaalprobleme as enige ander beskikbare hulpbron.[1]

(en) *corpus linguistics*

(en) *comparable corpora*

Gevolglik word vertaalgeheues en vertaalgeheuestelsels gebruik tydens die opleiding van vertalers [28]. In korpustaalkunde* "word parallelle en vergelykbare korpusse* hoofsaaklik gebruik vir die bestudering van vertaalstudies en kontrasterende studies"[2] [60]. Dit kan ook as hulpmiddel dien in tweetalige leksikografie en terminologie [5, 46]. Toegewyde sagteware is al hiervoor ontwikkel, waaronder WordSmith,[3] Sketch Engine[4] [47] en ParaConc[5] [14] voorbeelde is. Nog 'n voorbeeld is Linguee[6] — 'n soekenjin vir parallelle tekste op die wêreldwye web.

Tweetalige korpusse is vroeg reeds gebruik as vertaalgeheues tydens rekenaargesteunde vertaling. Die tweetalige Hansard van Kanada was een van die eerste groot tweetalige hulpbronne wat elektronies wyd beskikbaar was. 'n Oorsig van vertaaltegnologie in Kanada [56] het vele toepassings beskryf wat moontlik gemaak is deur die beskikbaarheid van hierdie data.

1 "... existing translations contain more solutions to more translation problems than any other available resource."
2 "Parallel and comparable corpora are used primarily for translation and contrastive studies."
3 http://www.lexically.net/wordsmith/
4 https://www.sketchengine.co.uk/
5 https://www.paraconc.com/
6 https://www.linguee.com/

Jare nadat die ALPAC-verslag van 1966 'n demper op navorsing in masjienvertaling geplaas het [68], word daar aan die begin van die een-en-twintigste eeu weer met groot aandag teruggekeer na masjienvertaling, veral statistiese masjienvertaalstelsels wat met grootskaalse tweetalige korpusse opgelei word. By die werkswinkel vir masjienvertaling in 2017 se nuusvertaaltaak[7] is daar opleidingsdata in die ordegrootte van gigagrepe vir sommige taalpare aangebied. (Sien byvoorbeeld die Verenigde Nasies-korpus[8] [91].)

Daar is verskeie ander toepassings wat genoem kan word. 'n Tweetalige korpus verskaf inligting vir handmatige, semi-outomatiese of outomatiese tweetalige termonttrekking [53, 55]. Dit is ook gebruik as deel van werk aan eentalige parafrasering [10]. Met parallelle korpusse kan 'n verskeidenheid annotasies geprojekteer word, bv. woordsoortetikette*, benoemde entiteite*, ens. wat kan help met tegnologieoordrag na tale met minder volwasse taaltegnologie [89].

(en) *part of speech tags*
(en) *named entities*

Parallelle korpusse is ook gebruik om hibriede stelsels te bou wat aspekte van bogenoemde gebruike kombineer. So byvoorbeeld het Koehn en Senellart [52] 'n masjienvertaalstelsel gekombineer met 'n vertaalgeheuestelsel. Termonttrekking uit tweetalige data kan gebruik word om 'n masjienvertaalstelsel te verbeter [2]. Masjienvertaling het vele toepassings in ander velde van natuurliketaalverwerking, soos in kruistalige inligtingherwinning [64]. Woordbelynings kan gebruik word om die kwaliteit van sinsbelyning te verbeter in die samestelling van parallelle korpusse, wat weereens kan help om groter parallelle korpusse van hoër kwaliteit te bou vir al die bogenoemde toepassings.

2.2 DIE IMPAK VAN LAE KWALITEIT

Die impak van lae kwaliteit in 'n parallelle korpus, bv. belyningsfoute of die teenwoordigheid van foutiewe vertalings,

7 http://www.statmt.org/wmt17/translation-task.html
8 https://conferences.unite.un.org/UNCorpus

hang van die toepassing af. Die relatiewe waarde van meer data van swakker kwaliteit teenoor minder data van hoër kwaliteit moet in elke geval bepaal word. Hierdie soeke na balans in verskillende toepassings word in hierdie afdeling bespreek.

As 'n navorsingsvraag in vertaalstudies handel oor die foute wat vertalers maak, kan 'n parallelle korpus met foute gepas wees vir so 'n ondersoek. Indien 'n vergelykende studie egter oor 'n vuil datastel plaasvind, sal belyningsfoute, enkoderingsfoute, foutiewe taal, ens. nie bydra tot die ondersoek nie — tensy die navorsingsvraag spesifiek oor hierdie sake handel. Oor die algemeen vereis hierdie toepassings skoon datastelle met baie akkurate belynings [81, p. 5].

By 'n toepassing in kruistalige inligtingherwinning is daar ondersoek ingestel na drie faktore van die tweetalige korpusse wat die prestasie van sulke stelsels affekteer [78]. Uit die faktore (1) tematiese nabyheid aan die navrae, (2) belyningskwaliteit en (3) korpusgrootte is bevind dat die tematiese nabyheid 'n groter rol speel as die ander twee. Die gebruik van korpusse van laer kwaliteit (bv. vergelykbare korpusse) met tematiese relevansie blyk dus voordelig te wees ten spyte van swakker belyningskwaliteit.

In 'n vertaalgeheuestelsel kan 'n foutiewe inskrywing in die geheue beteken dat 'n gebruiker direk met 'n foutiewe inskrywing gekonfronteer sal word. Elke voorstel plaas 'n kognitiewe las op die vertaler [65], maar die foutiewe inskrywing bied minder waarde vergeleke met 'n inskrywing van hoë kwaliteit. Afgesien van die moontlikheid dat 'n wasige voorstel* geredigeer moet word, sal die foutiewe aspek(te) van die doeltaal ook nog verbeter moet word. Indien die gebruiker nie die fout gesien het nie, is dit moontlik dat die fout kan versprei — nie net na die vertaalde produk of artefak wat die vertaler op daardie stadium skep nie, maar ook weer verder in die vertaalgeheue [61, p. 665]. Daar is verder 'n aanduiding dat vertalers geneig is om te veel vertroue te plaas in die voorstelle uit die vertaalgeheue [1, 28], wat dus die negatiewe impak van kwaliteitsprobleme kan vergroot.

(en) *fuzzy match*

Frasegebaseerde statistiese masjienvertaling is redelik robuus teen opleidingsdata van lae kwaliteit [51], selfs teen belyningsfoute [32]. Met hierdie statistiese benadering word probeer om allerlei waarskynlikhede* te skat op grond van die frekwensie van verskynsels in die opleidingsdata. Dit blyk dat foute in die opleidingsdata wat eenmalig of selde voorkom, redelik maklik geïgnoreer word in hierdie benadering.

(en) *probabilities*

Daar is aanduidings dat data van lae kwaliteit 'n groter impak het in ander benaderings tot masjienvertaling. In 'n hiërargiese frasegebaseerde stelsel* het skoonmaak van die ontwikkelingsdatastel gehelp om kwaliteit te verbeter [59]. Die veld van neurale masjienvertaling is nog in sy kinderskoene, maar voorlopige resultate dui daarop dat sulke stelsels meer sensitief is vir foutiewe inskrywings in die opleidingsdata vergeleke met tradisionele statistiese stelsels [18, 20, 51].

(en) *hierarchical phrase-based system*

In elke geval sal hierdie toepassings in 'n mindere of meerdere mate baat vind by skoner datastelle. Alhoewel daar by baie van die bogenoemde toepassings gereeld beter resultate met groter datastelle behaal word, is die datahonger nie gratis nie. Afgesien van die moeite om groot datastelle te bekom en te bestuur, vereis meer data tipies meer stoorplek, en vereis datagedrewe stelsels meer geheue en meer verwerkertyd* tydens opleiding. Selfs al sou resultate nie noemenswaardig verswak waar vuil datastelle gebruik word nie, is dit aanloklik om datastelle kleiner te maak ten einde die stelselvereistes te beperk. Die gerief van vinniger eksperimente is aantreklik.

(en) *processor time*

2.3 BOU EN ONDERHOUD VAN VERTAALGEHEUES

'n Vertaler kan van niks af begin en 'n vertaalgeheue opbou tydens vertaling. Die program vir rekenaargesteunde vertaling stoor vertalings tydens interaktiewe vertaling. Dit staan bekend as die bou van 'n vertaalgeheue in *interaktiewe modus*. As dit die enigste manier is waarop 'n vertaalgeheue uitgebou word, word die groei dus beperk deur die vertaalspoed en die hoeveelheid vertaalwerk wat die vertaler verrig. Melby meen

"... interaktiewe modus vereis aansienlike tyd en moeite voordat 'n gewenste gelykbreekpunt bereik word"[9] [61, p. 668]. Twee moontlike oplossings hiervoor is die invoer van vertaalgeheues van elders, en die skep van vertaalgeheues d.m.v. belyning.

Alternatiewe bronne van vertaalgeheues bied 'n vertaler dus 'n manier om vinniger sy eie vertaalgeheue aan te vul. Die uitruil van vertaalgeheues d.m.v. TMX is op bladsy 4 bespreek en maak die oordrag van vertaalgeheues tussen vertalers moontlik, selfs in die geval van diverse programme vir rekenaargesteunde vertaling. Die uitvoer en invoer* van vertaalgeheues in TMX-formaat is nou 'n algemene funksie in sulke programme.

(en) *export and import*

'n Verdere moontlikheid is om vertaalde dokumente te belyn en die resultaat in die vertaalgeheue op te neem. Sommige programme vir rekenaargesteunde vertaling sluit belyningsfunksionaliteit in sodat die gebruiker die plaaslike vertaalgeheue daarmee kan uitbrei. Voorbeelde van sulke programme is SDL Trados Studio[10] en MemoQ.[11]

Die uitruil van vertaalgeheues is 'n moeilike probleem wat kopiereg betref [34, 73]. Onsekerheid oor eienaarskap kan met kontrakte gereël word, wat veral belangrik kan wees waar meer as een jurisdiksie betrokke is [34, p. 176].

In die afgelope dekade of wat het 'n verskeidenheid oplossings vir al die bogenoemde haakplekke verskyn. Outomatiese belyning het verbeter, groot hoeveelhede teks vanaf organisasies soos die Europese Unie en die Verenigde Nasies het algemeen beskikbaar geword, en belynde weergawes hiervan is vir algemene gebruik gepubliseer.

'n Eenvoudige tegniek vir sinsbelyning gebruik bloot die lengte van die sinne [30]. 'n Goeie byderwetse benadering sal leksikale inligting met lengte-inligting kombineer, en moontlik selfs tegnieke gebruik wat pasgemaak is vir die hulpbron [81, p. 53].

9 "... interactive mode requires considerable time and effort before reaching a desirable break-even point."
10 http://www.sdltrados.com/solutions/translation-alignment/
11 http://kilgray.com/memoq/2015-100/help-en/index.html?alignment.html

Die sukses van 'n outomatiese belyning hang van 'n verskeidenheid faktore af. Faktore sluit die volgende in:

- die kwaliteit van die twee tekste self — geskandeerde tekste of teks wat uit PDF-dokumente onttrek word, is dalk nie 100% getrou aan die oorspronklike dokument nie [72];
- of daar enige ander strukturele aanpassings aan een van die dokumente gemaak is wat nie in die ander een weerspieël word nie — 'n vertaling is nie noodwendig 100% getrou aan die bronteks nie [81, p. 34];
- die mate waarin die vertaling afwyk van die bronteks weens universele aspekte van vertaling* [6], soos eksplisering* en vereenvoudiging.*

(en) *translation universals*
(en) *explicitation*
(en) *simplification*

Met die gebruik van sulke outomatiese sinsbelyning, het dit moontlik geword om groot parallelle korpusse vry te stel, soos bv. Europarl[12] [49], die DGT-vertaalgeheue[13] [76] en verskeie datastelle uit die Opus-projek[14] [80]. Dit het 'n groot hupstoot gegee vir die navorsing in velde wat parallelle teksdata gebruik — veral masjienvertaling.

'n Afsonderlike ontwikkeling het te doen met die ontstaan van vertaalgeheuedienste aanlyn. Dienste soos MyMemory[15] laat gebruikers hulle vertaalgeheues aanlyn stoor. Alhoewel dit na 'n blote argitektuurverskil mag lyk, maak dit die deel van vertaalgeheues aansienlik makliker. Dit is gevolglik triviaal om data met vreemdelinge te deel as bydraes outomaties tot die gemeenskaplike vertaalgeheue bygedra word. MyMemory noem in 2017 dat hul databasis meer as 1,5 miljard bydraes bevat. So 'n diens kan natuurlik ook bogenoemde vertaalgeheues soos Europarl insluit en is dus nie volledig afhanklik van gebruikers se bydraes nie.

12 http://www.statmt.org/europarl/
13 https://ec.europa.eu/jrc/en/language-technologies/dgt-translation-memory
14 http://opus.lingfil.uu.se/
15 https://mymemory.translated.net/

Met interaktiewe modus en eie belynings is dit te verstane dat ’n vertaler self die gevolg sou dra van swak belynde data of swak vertalings in sy eie vertaalgeheue. ’n Gebruiker sou self die kwaliteit van die datastel in ’n mate kon bepaal deur hersiening daarvan, of deur met versigtigheid data in die vertaalgeheue in te sluit. Met die beskikbaarheid van groter hoeveelhede data, outomaties belynde datastelle, en reusedatastelle in dienste aanlyn is dit minder duidelik wie die verantwoordelikheid neem of behoort te neem. Dit is juis in die konteks van MyMemory wat met eksperimente begin is om die skoonmaak van vertaalgeheues te ondersoek [11].

2.4 SKOONMAAK

Heelwat tekshulpbronne in natuurliketaalverwerking is nie aanvanklik van optimale kwaliteit nie. Selfs vir eentalige hulpbronne is dit gereeld nodig om ’n verskeidenheid kwaliteitsprobleme die hoof te bied. ’n Voorbeeld van oplossings hiervoor is ’n versameling programme van Kenneth Heafield.[16] Onder hierdie programme is daar voorsiening gemaak vir die verwydering van segmente in ’n eentalige hulpbron met die volgende kwaliteitsprobleme:

- verkeerde karakterenkodering;
- ’n gebrek aan Unicode-normalisering;
- die segment is leeg (geen teks);
- gedupliseerde inhoud;
- onnodige spasies;
- inkonsekwente leestekengebruik;
- die segment is te lank.

Enige van hierdie sake kan ook problematies wees in tweetalige hulpbronne. Soortgelyke probleme is hanteer waar ’n tweetalige hulpbron gebou is uit eentalige tekste van lae kwaliteit [72].

16 https://github.com/kpu/preprocess

Selfs al sou al die bogenoemde in orde wees, kan 'n parallelle korpus steeds kwaliteitsprobleme hê. Probleme kan insluip weens foutiewe belyning of vertaalfoute — aspekte wat nie figureer in eentalige korpusse nie. Vervolgens word 'n oorsig gegee van relevante werk oor die skoonmaak van parallelle korpusse, ook buite die veld van vertaalgeheuestelsels.

Verskille tussen die bron- en doelteks by 'n vertaling is te verwagte — ook in goeie vertalings. Afwyking van die bronteks vanweë universele aspekte van vertaling soos eksplisering en vereenvoudiging dui nie noodwendig op swak of foutiewe vertaling nie. Dat daar egter iets soos foutiewe vertalings bestaan, is ongetwyfeld waar. Die grens tussen skoon en vuil inskrywings kan dus gesien word as 'n wasige gebied tussen twee uiterstes. Alhoewel die inskrywings op die uiterstes van skoon en vuil maklik van mekaar onderskei kan word, sal selfs kenners nie noodwendig saamstem oor inskrywings in die grensgebied nie. Vir 'n annotasietaak in hierdie verband was daar al karige ooreenstemming tussen annoteerders [13].

Dit is natuurlik te verstane dat handmatige skoonmaak van groot datastelle onhaalbaar is. Volledig outomatiese skoonmaakmetodes is dus nodig om datastelle te hanteer van die groottes wat nou algemeen is in natuurliketaalverwerking.

In die gewilde masjienvertaalprojek, Moses,[17] is daar 'n program `clean-corpus-n.perl` wat die opleidingsdata skoonmaak volgens enkele eenvoudige reëls. Dit identifiseer vuil inskrywings volgens hul lengte deur te toets of die getal woorde in die bron- en doelteks binne 'n sekere omvang is. Dit kan ook die verhouding van bronwoorde:doelwoorde en doelwoorde:bronwoorde toets teen 'n gegewe maksimum (by verstek* 9). Alhoewel hierdie program ontwikkel is vir toepassing op die opleidingsdata vir masjienvertaling, sou dit op enige soortgelyke datastel eweneens toegepas kon word.

(en) *default*

'n Studie oor die skoonmaak van data wat as vertaalgeheues versamel is, volg 'n masjienleerbenadering [11].[18] Hierdie werk

17 http://www.statmt.org/moses/

18 'n Oorsig oor relevante masjienleertegnieke word hier uitgelaat, maar is wel ingesluit in hoofstuk 5.

vind plaas in die konteks van MyMemory — 'n vertaalgeheue wat aanlyn deur gebruikers gebou word sonder sentrale beheer of toesig. Die skoonmaaktaak het in hierdie geval die opspoor van allerlei blatante foute ingesluit, soos inskrywings in die verkeerde taal en inskrywings waar gebruikers boodskappe aan mekaar intik in plaas van vertalings.

Hieruit het 'n gedeelde taak by 'n werkswinkel gespruit [13] waar deelnemers vertaalgeheues in drie taalpare moes skoonmaak. Hier is 'n onderskeid gemaak tussen ernstige foute en minder ernstige foute. Meer detail oor die opdrag in hierdie gedeelde taak is in hoofstuk 5, aangesien die raamwerk van die gedeelde taak ook ingespan word vir daardie hoofstuk. Aangesien geannoteerde data verskaf is vir die gedeelde taak, is die probleem deurgaans deur deelnemers aangepak met 'n gekontroleerde masjienleerbenadering soortgelyk aan [11].

Een van die inskrywings by die gedeelde taak is uitgebrei om bruikbaar te wees sonder 'n geannoteerde datastel [63]. So 'n benadering met ongekontroleerde masjienleer maak die stelsel bruikbaar in meer kontekste. Die aanpassing behels twee stappe: (1) 'n klassifikasie word afgelei vir segmente gebaseer op 'n kwaliteitskatting en (2) 'n ensemble van klassifiseerders word opgelei op grond van die afgeleide klassifikasie.

Kenmerke gebaseer op woordbelyning is gebruik in 'n masjienleerbenadering [63, 90]. In [18] was die fokus spesifiek op die bespeuring van segmente met semantiese verskille tussen die bron- en doeltaal. In die laaste twee gevalle is die verbetering in datakwaliteit geëvalueer in masjienvertaalstelsels.

2.5 EVALUASIE

Elkeen van die toepassings wat vroeër genoem is, regverdig 'n eie evaluasiemetode. Elke toepassing kan met vuil en skoon data geëvalueer word om die impak van skoner data te bepaal.

Alhoewel termonttrekking geëvalueer kan word deur die afvoer te vergelyk met 'n verwysingstel, is die beskikbaarheid van goudstandaarde 'n probleem, en kan die gebruikers se voor-

keure 'n groot rol speel in hulle indrukke van die termlys se kwaliteit [79, p. 258].

Die evaluasie van masjienvertaling geniet baie aandag. By die jaarlikse werkswinkel vir masjienvertaling[19] vorm evaluasie van die kompeterende stelsels 'n belangrike komponent. Deelnemers evalueer die ingediende masjienvertaalstelsels handmatig met 'n stelsel soos Appraise [29].[20] Daar is ook 'n gedeelde taak toegewy aan outomatiese evaluasiemate [16]. Die uitkomste van die handmatige evaluasie dien as 'n goudstandaard waarvolgens outomatiese evaluasiemetodes geëvalueer word. 'n Evaluasiemetode moet dus 'n evaluasie doen gebaseer op die gegenereerde vertalings en die verwysingsvertalings, en poog om die menslike deelnemers se oordeel so goed as moontlik na te boots. Vir outomatiese evaluasie is die mate BLEU [66], NIST [24], METEOR [9] en TER [74] gewild.

'n Ander faset van die evaluasie van masjienvertaalstelsels is die jonger veld van kwaliteitskatting [75]. Ook hiervoor is 'n gedeelde taak by die jaarlikse werkswinkel vir masjienvertaling.[21] Hiermee word gepoog om vertalings op segmentvlak te evalueer sonder verwysingsvertalings. 'n Tipiese benadering is gekontroleerde masjienleer van die redigeertyd wat nodig was om masjienvertalings te redigeer. 'n Bekende pakket hiervoor is QuEst++.[22]

In afdeling 1.1.6 is reeds 'n oorsig gegee van ooreenkomste en verskille tussen masjienvertaalstelsels en vertaalgeheuestelsels. Vorige werk op die gebied van die evaluasie van vertaalgeheuestelsels word volgende bespreek. Hierdie evaluasiemetodes is nog nie so volwasse as wat die geval is by masjienvertaling nie. Verskillende benaderings tot outomatiese en handmatige evaluasie is al gevolg.

Handmatige evaluasie met beoordeling deur mense is moontlik [15]. In daardie studie is medewerkers op Mechanical Turk gevra om 'n oordeel oor elke voorstel te vel op 'n 5-punt-Likert-

19 Sien bv. http://www.statmt.org/wmt17/
20 https://github.com/cfedermann/Appraise
21 http://www.statmt.org/wmt17/quality-estimation-task.html
22 https://github.com/ghpaetzold/questplusplus

skaal. In [33] is 'n evaluasiemetode ontwerp wat 'n vergelyking moontlik maak tussen stelsels wat verskillende benaderings gebruik vir soek en onttrekking, veral ten opsigte van segmentasie (of die gebrek daaraan). Aanvanklike werk aan evaluasie met oogvolgtoerusting* is aangebied in [65]. As vertalers voorstelle redigeer tot finale vertalings, kan die finale vertalings as verwysingsvertalings van hoë kwaliteit dien [74]. As hierdie verwysingsvertalings d.m.v. *translation error rate* (TER) met die doeltekste van die voorstelle vergelyk word, word daar verwys na HTER — *human-targeted translation error rate*. Vir 'n ondersoek na die gebruik van parafrases in 'n vertaalgeheuestelsel [35] is hierdie tegniek uit masjienvertaling oorgeneem, asook die HMETEOR variant van METEOR [23]. In dieselfde studie het deelnemers ook 'n subjektiewe evaluasie gedoen waar hulle by elke segment die beste tussen twee voorstelle gekies het.

(en) *eye tracking equipment*

Aangesien hierdie metodes op beoordeling of redigering deur mense staatmaak, is hulle nie geskik as outomatiese metodes vir die evaluasie van vertaalgeheues of vertaalgeheuestelsels nie. In teenstelling met die bogenoemde metodes is die fokus in hierdie proefskrif op volledig outomatiese evaluasiemetodes. Sulke metodes is van onskatbare waarde in vele velde, soos in masjienvertaling wat vroeër in hierdie afdeling bespreek is. Sulke outomatiese metodes is waardevol aangesien dit herhaalde eksperimente moontlik maak sonder dat beoordeling deur mense in elke eksperiment nodig is.

Vorige werk aan die outomatiese evaluasie van vertaalgeheuestelsels kan in drie groepe gesien word:

- Metodes gebaseer op die evaluasie van inligtingherwinning [7, 84]. Hierdie metodes is geskoei op die konsepte van presisie en herroeping.

- Metodes wat korrelasie met 'n verwysingsmaat toets [82]. Hier was dit TER wat as maatstaf vir evaluasie gedien het.

- Metodes gebaseer op die evaluasie van masjienvertaling [71]. Hier word die vertaalgeheuestelsel geëvalueer asof dit 'n masjienvertaalstelsel is.

Nie een van hierdie metodes geniet egter wye gebruik nie. Weens die sentrale rol wat vertaalgeheues in hierdie proefskrif speel, is 'n diepgaande bespreking oor hierdie evaluasiemetodes geregverdig. Hierdie metodes word in afdeling 3.1 in meer detail beskou, waar die kritiese bespreking ook sal dien as nodige agtergrond tot daardie hoofstuk. Om die oorsigtelike aard van hierdie hoofstuk te behou, word daar egter vir eers volstaan met die hoëvlakbeskrywing wat pas aangebied is.

2.6 GEVOLGTREKKING

Hierdie hoofstuk het die nuttigheid van parallelle korpusse vir verskeie toepassings aangetoon. Die kwaliteit van hierdie korpusse kan klaarblyklik 'n impak hê op sommige van hierdie toepassings. In hierdie proefskrif sal die impak van datakwaliteit in 'n vertaalgeheuestelsel en 'n masjienvertaalstelsel geëvalueer word. By masjienvertaaleksperimente is daar vir hierdie evaluasie 'n gevestigde praktyk, maar by vertaalgeheuestelsels is daar 'n gebrek aan standaardisering. Die volgende hoofstuk sal by hierdie punt begin deur eers 'n kritiese analise van bestaande outomatiese evaluasiemetodes vir vertaalgeheuestelsels te gee en dan een van hierdies te verfyn om die gebreke die hoof te bied.

3

METODOLOGIESE SLAGGATE IN DIE OUTOMATIESE EVALUASIE VAN VERTAALGEHEUES

Evaluasie is belangrik in natuurliketaalverwerking. Dit is onder andere essensieel om vooruitgang in die veld te meet en om kompeterende stelsels met mekaar te vergelyk [58].[1] 'n Belangrike beginsel van wetenskaplike vergelyking is die voorkoming van sydigheid.* Hierdie hoofstuk ondersoek moontlike bronne van sydigheid in die evaluasie van 'n vertaalgeheuestelsel.[2]

(en) *bias*

By vertaalgeheuestelsels kan 'n soortgelykheidsdrempel f dikwels deur die gebruiker gestel word sodat slegs voorstelle met 'n soortgelykheid gelyk aan of hoër as dié drempel voorgestel word. Hierdie drempel word dikwels uitgedruk as 'n persentasie soos 70% of 'n breuk 0,7. Of daar in enige spesifieke geval 'n voorstel gelewer word, hang natuurlik ook af van die inhoud van die vertaalgeheue. Anders as in die geval van masjienvertaalstelsels, gee 'n vertaalgeheuestelsel dus nie noodwendig voorstelle vir elke toevoersegment nie. Om elke vertaalgeheuevoorstel te oorweeg, dra by tot die kognitiewe las vir die vertaler [65], en die drempel is 'n meganisme om die getal nuttelose voorstelle wat vir oorweging aan die vertaler verskaf word, te beperk.

Verskeie aspekte van 'n vertaalgeheuestelsel kan geëvalueer word. Daar kan vrae wees oor hoe voorstelle onttrek word, die stelsel se parameters, en die datastel wat gebruik word. Ideaal gesproke behoort 'n evaluasiemetode van hulp te wees om vrae oor al hierdie aspekte te antwoord. Ons is dus op soek na 'n maat wat 'n plaasvervanger of aanduider* is vir die waarde van die voorstelle vir die vertaler. Hiermee kan die relatiewe waarde van datastelle vergelyk word in terme van die rela-

(en) *proxy*

1 Sien ook http://hlt-evaluation.org

2 Hierdie hoofstuk is deels gebaseer op [88] en [87].

tiewe waarde van die voorstelle uit elke datastel. Verskillende implementasiekeuses en parameters, byvoorbeeld die soortgelykheidsdrempel, kan ook hiermee vergelyk word. Indien die stelsel 'n ander meganisme gebruik om te besluit of 'n voorstel aangebied moet word, kan dit met alternatiewe vergelyk word in eksperimente.

Tot op hede het geen enkele evaluasiemetode homself as voorkeurmetode onderskei nie. Afdeling 3.1 bied 'n oorsig van bestaande metodes vir die evaluasie van vertaalgeheuestelsels. Daar word nie in hierdie hoofstuk probeer om 'n meta-evaluasie te doen om die beste onder hulle te kies nie. Die fokus is eerder op die onderliggende soortgelykheidsmate wat deur hulle almal gebruik word. Hierdie mate is nie almal eie aan die veld van vertaalgeheuestelsels nie — sommige word in ander velde soos masjienvertaling, spraakherkenning en bio-informatika gebruik. 'n Noemenswaardige aspek van outomatiese evaluering van vertaalgeheuestelsels is dat soortgelykheidsmate nie net gebruik word tydens die normale werking van die stelsel (onttrekking) nie, maar ook dikwels as deel van outomatiese evaluasiemetodes. Dit is dus belangrik om te onderskei tussen hierdie twee gebruike van die soortgelykheidsmate: as onttrekkingsmaat en as evaluasiemaat.

Afdeling 3.2 is 'n bespreking van verskeie van hierdie soortgelykheidsmate en sommige implementasiebesonderhede word in afdeling 3.3 genoem. Vanaf afdeling 3.4 word 'n belangrike vraag ondersoek, wat ook deur [82] uitgespreek is: *Het 'n evaluasiemaat 'n voorkeur vir 'n onttrekkingsmaat wat identies of soortgelyk is?*[3] Met ander woorde, toon 'n evaluasiemaat sydigheid wanneer verskillende onttrekkingsmate vergelyk word? Hierdie ondersoek dien as noodsaaklike agtergrond vir die eerste navorsingsvraag (sien afdeling 1.2) deurdat dit die probleme met huidige metodes vir die evaluasie van vertaalgeheuestelsels uitwys. In afdeling 3.5 word resultate bespreek en met die relevante literatuur vergelyk. Afdeling 3.6 bevat 'n samevatting

3 "... raises the question whether an evaluation metric favors a fuzzy matching metric which is identical or similar to it."

van die bydraes van die hoofstuk en verduidelik die belangrikheid daarvan in die lig van die eerste navorsingsvraag.

3.1 BESTAANDE EVALUASIEMETODES

Relevante vorige werk op die gebied van die evaluasie van vertaalgeheuestelsels is in afdeling 2.5 bespreek. Die belangrikste vorige werk aan volledig outomatiese evaluasie word vervolgens in meer besonderhede aangebied en krities bespreek.

3.1.1 *Metodes gebaseer op die evaluasie van inligtingherwinning*

Sommige metodes vir die evaluasie van vertaalgeheuestelsels het inspirasie gevind in die veld van inligtingherwinningstelsels.*

(en) *information retrieval systems*

In [84] oorweeg die skrywers 'n swartboks-evaluasiemetode vir 'n vertaalgeheuestelsel — dit oorweeg slegs die toevoer en afvoer van die stelsel, nie die inhoud van die vertaalgeheue nie. Hulle metode bied 'n manier om 'n goeie waarde vir die drempel f vir wasige passing te bepaal om in die vertaalgeheuestelsel te gebruik. Dié metode definieer beide 'n presisieagtige en herroepingagtige maat. Al twee word bereken met 'n skatting van die aantal sleuteldrukke* wat dien as soortgelykheidsmaat (sien afdeling 3.2.3). Die evaluasie met dié twee mate terwyl f gevarieer word, bied 'n manier om 'n optimale waarde vir f te kies deur vir die gewenste kombinasie van presisie en herroeping te optimeer.

(en) *key-strokes*

Nog 'n evaluasiemetode word voorgestel in [7]. Verskillende soortgelykheidsmate word geëvalueer vir gebruik in onttrekking. Die metode gebruik 10-voudige kruisvalidasie oor die hele datastel — telkens word alle segmente uit een tiende van die datastel gebruik as navrae teen die oorblywende 90% wat as die vertaalgeheue dien. Dié metode gebruik twee soortgelykheidsmate tydens evaluasie en bepaal daarmee die nuttigste doeltaalvoorstel: die 3-bewerkingredigeersoortgelykheid

oor 2-gramme (sien afdeling 3.2.2) en die WSC-maat[4] [8]. Die beste resultaat wat onttrek word, word geklassifiseer as korrek of foutief, afhangend daarvan of dit die nuttigste doeltaalvoorstel het. Dit gebruik dus volledige kennis van die doelsegmente in die vertaalgeheue om die nuttigste doeltaalvoorstel te identifiseer. Hierdie metode kan dus nie deur 'n eindgebruiker gebruik word om 'n diens aanlyn, soos MyMemory, te evalueer nie omdat toegang tot die totale vertaalgeheue onmoontlik is.

3.1.2 *Korrelasie met 'n verwysingsmaat*

Onlangse werk het 'n metode voorgestel wat die prestasie van onttrekkingsmate vergelyk [82]. Hiervolgens word prestasie op twee maniere geëvalueer. Die eerste toets hoe goed die onttrekkingsmaat die kwaliteit van voorstelle skat, en die tweede toets die waarde van die voorstelle self.

(en) *test set, evaluation set*

Vir elke segment in die toetsstel* word twee hoeveelhede bereken: (1) die soortgelykheid tussen die bronteksnavraag en die *bronteks* van die beste voorstel wat aangebied word, soos bereken deur die onttrekkingsmaat, en (2) die evaluasietelling onder 'n maat $\mathrm{Sim}_{\mathrm{TER}}$ (gebaseer op TER [74]) van die *doeltaalvoorstel* as dit vergelyk word met 'n verwysingsvertaling uit die toetsstel. Die eersgenoemde hoeveelheid is dus afhanklik van die onttrekkingsmaat, maar die berekening met $\mathrm{Sim}_{\mathrm{TER}}$ op die twee doeltaalsegmente is onafhanklik van die onttrekkingsmaat. Die Pearson-korrelasie tussen dié twee hoeveelhede word gebruik om die prestasie van die onttrekkingsmaat te peil. In hierdie opsig word $\mathrm{Sim}_{\mathrm{TER}}$ gebruik as 'n verwysingsmaat—daar word geëvalueer in watter mate die onttrekkingsmaat se siening van die brontekste se soortgelykheid ooreenstem met wat $\mathrm{Sim}_{\mathrm{TER}}$ rapporteer oor die doeltekste.

4 *Weighted Sequencial Correspondence* (WSC) analiseer teks nie net sekwensieel soos in die Levenshtein-afstand nie, maar ook die mate waartoe passende segmente aaneenlopend is. Na my beste wete geniet WSC nie wye gebruik nie en word dus nie in verdere besonderhede hier bespreek nie.

Hierdie eerste deel van hulle evaluasiemetode evalueer dus die akkuraatheid waarmee ’n soortgelykheidsmaat die waarde van sy voorstelle skat (hetsy goeie of slegte voorstelle), eerder as dat die waarde van die voorstelle vir die vertaler direk geevalueer word. ’n Hoër korrelasie dui op ’n onttrekkingsmaat wat met hoër akkuraatheid die waarde van sy voorstelle skat. In hierdie deel van die metode word die effek van die soortgelykheidsdrempel nie oorweeg nie — daar word ’n voorstel vir elke toetssegment gegenereer. Dit toets dus die akkuraatheid van die onttrekkingsmaat se soortgelykheidskatting ook in gevalle van baie lae soortgelykheid. Hierdie voorstelle — moontlik selfs ’n meerderheid — word nie deur programme vir rekenaargesteunde vertaling aangebied nie. Dus word die akkuraatheid van die onttrekkingsmaat in ’n groot mate geëvalueer by voorstelle wat nie gebruik sal word nie. Hierteenoor kan die tweede deel van die metode wel gesien word as ’n manier om hierdie drempelwaarde van die stelsel te evalueer.

Vir die res van hierdie evaluasiemetode word die gemiddelde evaluasietelling (die gemiddeld van die tweede hoeveelheid wat bo genoem is) gebruik. Dit is ’n poging om die inherente waarde van die doeltaalvoorstelle vir die vertaler te evalueer eerder as bloot die akkuraatheid van die onttrekkingsmaat. Skoenlushersteekproefneming* word gebruik om die statistiese beduidendheid van verskille tussen die evaluasietellings van kompeterende onttrekkingsmate te bepaal.

(en) *bootstrap resampling*

In die tweede deel van die metode ondersoek die skrywers die effek op die gemiddelde evaluasietelling soos wat voorstelle vir ’n kleiner of groter getal van die toetssegmente onttrek word. Die berekening van die gemiddelde evaluasietelling vir N toetssegmente behels dat die N segmente met voorstelle wat die beste telling volgens die onttrekkingsmaat het, gekies word. Omdat die toetssegmente volgens die onttrekkingsmaat se soortgelykheid gesorteer word, sal hierdie N voorstelle almal ’n soortgelykheid hê wat gelyk aan of beter is as die N^{de} toetssegment s’n. Die soortgelykheid van die N^{de} segment kan dus gesien word as ’n soortgelykheidsdrempel. Hoe meer seg-

mente van voorstelle voorsien word (hoër N), hoe laer is die drempel.

Hierdie ondersoek na die gemiddelde evaluasietelling terwyl N gevarieer word, is dus vergelykbaar met 'n ondersoek na die soortgelykheidsdrempel. N dien hier as 'n plaasvervanger waarvolgens die stelsel bepaal of 'n spesifieke inskrywing uit die geheue voorgestel word of nie. Aangesien die gemiddelde evaluasietelling asimptoties monotoon toeneem soos N toeneem [82, p. 159], sal só 'n evaluasie 'n arbitrêr lae soortgelykheidsdrempel voorstel sodat so veel moontlik voorstelle aangebied word. So 'n keuse van f strook nie met die praktyk dat 'n te lae drempel produktiwiteit kelder nie. O'Brien beskryf haar ondervinding in dié opsig soos volg [65]:

> Alhoewel dit van een organisasie of vertaler na die volgende wissel, wil vertalers selde met wasige voorstelle onder 75% werk omdat hulle die werk wat nodig is, groter ag as om bloot die sin sonder voorstelle uit die vertaalgeheue te vertaal.[5]

3.1.3 *Metodes gebaseer op die evaluasie van masjienvertaling*

'n Vertaalgeheuestelsel kan as 'n masjienvertaalstelsel geëvalueer word [71] deur gebruik te maak van bekende metodes soos BLEU, NIST, METEOR en WER. Aangesien hierdie evaluasiemetodes afhang van die feit dat afvoer gegenereer word vir elke segment, moet sulke eksperimente gedoen word sonder om resultate met 'n soortgelykheidsdrempel te filtreer. Hierdie metodes kan dus nie die werking van 'n vertaalgeheuestelsel evalueer soos wat dit in die praktyk gebruik word nie — die metodes is eenvoudig ontwerp met 'n ander toepassing in gedagte. Daarom is hierdie metodes vir die evaluasie van masjienvertaalstelsels nie verder oorweeg in hierdie navorsing nie. Sekere bevindinge in dié artikel is egter wel relevant tot die ondersoek

5 "While this varies from one organisation or translator to the next, translators rarely want to deal with Fuzzy Match values below 75% because they deem the work involved to be greater than simply translating the sentence with no prompts from the Translation Memory."

in hierdie hoofstuk, en in die bespreking in afdeling 3.5 sal daar weer hierna verwys word. 'n Meer breedvoerige bespreking van die verskille tussen vertaalgeheuestelsels en masjienvertaalstelsels is reeds op bladsy 8 aangebied.

Sommige evaluasiemetodes vir masjienvertaling werk op die vlak van die toetsstel, soos bv. BLEU [66]. Verskeie van hulle kan wel as soortgelykheidsmate op segmentvlak gesien word of daarvoor aangepas word (soos al gedoen is met TER [82]), soortgelyk aan die mate wat in die volgende afdeling bespreek word.

3.2 SOORTGELYKHEIDSMATE

Daar is vroeër genoem dat soortgelykheidsmate twee rolle kan vervul in 'n ondersoek na vertaalgeheuestelsels: as onttrekkingsmaat en as evaluasiemaat. Telkens verskaf so 'n maat 'n aanduiding van hoe soortgelyk twee segmente teks is. In hierdie afdeling word 'n verskeidenheid soortgelykheidsmate met verskillende eienskappe bespreek wat later in die hoofstuk gebruik word. Elke maat het 'n ander manier om soortgelykheid te bepaal.

Mate wat soortgelykheid tussen stringe meet, word meestal gebaseer op 'n tipe redigeerafstand. 'n *Afstandfunksie** (soos die bekende Levenshtein-afstand) meet die verskil tussen twee stringe. Identiese stringe het 'n afstand van nul. Hoe minder die soortgelykheid tussen twee stringe, hoe groter die afstand tussen hulle. 'n *Soortgelykheidsmaat** meet die ooreenkoms tussen twee stringe wat gewoonlik as 'n breuk tussen 0,0 en 1,0 uitgedruk word, of as 'n persentasie tussen 0% en 100%. Identiese stringe het 'n soortgelykheid van 1,0 of 100%. Om 'n afstand d na 'n soortgelykheidsmaat om te skakel, word 'n *normaliseringskonstante*, sê l, gebruik om die afstand te beperk tot die interval $[0, 1]$. Die soortgelykheidsmaat word dan gedefinieer as $1 - d/l$. Om te verseker dat die minimumwaarde van die soortgelykheidsmaat nul is, moet die waarde van l die maksimum- moontlike waarde van d aanneem.

(en) *distance function*

(en) *similarity metric*

Verskillende stringpare se soortgelykheidsmate kan vergelyk word d.m.v. die gebruik van sulke genormaliseerde waardes, ongeag of hulle kort of lank is. Sonder normalisering kan 'n afstand van 6 dui op twee baie soortgelyke stringe, of twee baie verskillende stringe, afhangend van hul lengtes. Tydens die evaluasie van vertaalgeheuestelsels word die soortgelykheidsmate globaal geneem (bv. as 'n gemiddeld oor 'n toetsversameling). In so 'n geval kan 'n ongenormaliseerde telling veroorsaak dat lang stringe die uiteindelike resultaat oorheers en daardeur die resultate verdraai en die interpretasie daarvan kompliseer.

'n Soortgelykheidsdrempel beperk die lengtes van voorstelle op 'n voorspelbare manier weens die manier waarop die afstandfunksie en normalisering werk. Dit is 'n belangrike aspek vir die optimering van implementasies wat werk met groot datastelle, aangesien dit maklik is om kandidate volgens hulle lengtes te filtreer en daarmee die meeste van die geheue te elimineer vir oorweging wanneer voorstelle vir 'n spesifieke segment onttrek word. Dit word in groter detail in afdeling 3.3.1 bespreek.

Aangesien die soortgelykheidsmate en hul normalisering verskillend werk, kan daar verskille wees tussen die soortgelykheidtellings vir 'n stringpaar soos dit gemeet word deur twee verskillende soortgelykheidsmate. Die verskillende soortgelykheidsmate affekteer dus nie net die rangbepaling van voorstelle nie, maar ook of 'n voorstel enigsins aan die gebruiker gebied word wanneer dit met 'n gegewe soortgelykheidsdrempel ingeperk word.

Hierdie afdeling beskryf verskeie mate wat werk op 'n paar stringe. Die karakter kan beskou word as die laagste vlak van oorweging. 'n Ander moontlikheid is om op woordvlak te werk. 'n Verdere moontlikheid is n-gramme van karakters of selfs woorde. 'n Vorige studie wat ondersoek ingestel het na verskillende vlakke van oorweging het die volgendes oorweeg [7]: 1-gramme, 2-gramme en gemengde 1-gram/2-gram-kombinasies op karakter- en woordvlak. Sommige van die mate wat hieronder beskryf word, kan maklik vir enige van hierdie vlakke aangepas word, met uitsonderings wat hier onder aan-

gedui word. In die komende afdelings van hierdie hoofstuk word daar ondersoek ingestel op die vlak van karakters, karakter-2-gramme en woordvlak (waar dit van toepassing is).

Die mate wat in die ondersoek ingesluit word, is die volgende: 4-bewerkingredigeersoortgelykheid,* 3-bewerkingredigeersoortgelykheid,* 'n skatting van sleuteldrukke, en n-gram-presisie. In die res van hierdie afdeling word hierdie mate in meer detail bespreek. Verskeie ander soortgelykheidsmate sou bespreek kon word. Die seleksie hier onder is gebaseer op prestasie in vorige literatuur, en is gekies om 'n wye omvang alternatiewe te weerspieël sodat mate met verskillende eienskappe verteenwoordig word. So byvoorbeeld is die mate WSC en Sato92 in [7] as niekompeterend aangedui afgesien daarvan dat hulle ook 'n ordegrootte meer looptyd as ander deelnemende mate vereis het, en hulle word dus nie hier ingesluit nie. Die Damerau-Levenshtein-afstand [22] is spesifiek vir toepassings in speltoetsers ontwerp en word daarom ook nie oorweeg nie. Soos in afdeling 3.1.3 verduidelik is, word evaluasiemetodes vir masjienvertaling ook nie oorweeg nie. Ons beweer nie dat die versameling soortgelykheidsmate wat hier oorweeg word volledig is nie, maar ons glo dit is divers genoeg om die vraag op bladsy 32 te beantwoord. Ons versameling bevat mate wat werk op die vlak van karakters, karakter-2-gramme en woorde. Dit bevat kommutatiewe en niekommutatiewe mate. Verder verteenwoordig n-gram-presisie mate sonder volle sensitiwiteit vir volgorde (dit werk met woordversamelings* en woord-n-gramversamelings).

(en) *4 operation edit similarity*

(en) *3 operation edit similarity*

(en) *sets of words*

Die volgende notasie gaan in die beskrywings van die soortgelykheidsmate hier onder gebruik word: A en B verwys na die twee stringe wat vergelyk word (onderskeidelik die navraag en die kandidaatstring), $|A|$ dui op die lengte van die string A in terme van die granulariteit wat oorweeg word (bv. karakterlengte as daar op karaktervlak gewerk word). By n-gram-presisie (afdeling 3.2.4 hier onder) verwys $|A_n|$ na die kardinaliteit van die versameling woord-n-gramme.

3.2.1 *4-bewerkingredigeersoortgelykheid*

Hierdie maat is gebaseer op seker een van die bekendste afstandfunksies, gewoonlik bekend as die Levenshtein-afstand [54]. Die Levenshtein-afstand is die getal invoegings, skrappings en vervangings* wat nodig is om die een string in die ander een te omskep. Dit word "4-bewerkingredigeerafstand" genoem met die veronderstelling dat die identiteitsbewerking (geen verandering) die vierde bewerking is.[6]

(en) *insertions, deletions and substitutions*

Die maksimumafstand tussen twee stringe is die lengte van die langste string, en sal aangetref word in die geval van 'n vergelyking met die leë string, of wanneer die twee stringe geen gemene karakters het nie. Die soortgelykheidsmaat word daarom gedefinieer as

$$\mathrm{sim}_{4ops}(A, B) = 1 - \frac{\mathrm{edit}_{4ops}(A, B)}{\mathrm{maks}(|A|, |B|)}.$$

Alhoewel die implementasiebesonderhede van eiendomsagteware* vir rekenaargesteunde vertaling verborge is, dui 'n informele ondersoek na stukke Vry en Oopbronsagteware* daarop dat dié maat deur die volgende implementasies van vertaalgeheuestelsels gebruik word: OmegaT,[7] Okapi FrameWork,[8] Virtaal[9] en die Amagama-vertaalgeheuebediener.[10]

(en) *proprietary software*

(en) *Free and Open Source Software, FOSS*

Hierdie maat is kommutatief, dit wil sê, $\mathrm{sim}_{4ops}(A, B) = \mathrm{sim}_{4ops}(B, A)$. Hier is 'n paar voorbeelde wat op verskillende vlakke van granulariteit werk:

- karakters: $\mathrm{sim}_{4ops}(\text{“}metaphor\text{”}, \text{“}metamorphosis\text{”}) = 1 - \frac{6}{13} = 0.538$
- 2-gramme : $\mathrm{sim}_{4ops}(\text{“}metaphor\text{”}, \text{“}metamorphosis\text{”}) = 1 - \frac{7}{12} = 0.417$

6 Hierdie definisie neem eenheidkoste aan, met ander woorde 'n afstand van 1 vir elke bewerking behalwe identiteit. Verskillende gewigte kan toegeken word aan verskillende bewerkings, maar word nie verder in hierdie navorsing oorweeg nie.

7 http://omegat.org/

8 http://okapi.opentag.com/

9 http://virtaal.translatehouse.org/

10 http://amagama.translatehouse.org/

- woorde: $\text{sim}_{4\text{ops}}(\text{“}metaphor\text{”}, \text{“}metamorphosis\text{”}) = 1 - \frac{1}{1} = 0$
- karakters: $\text{sim}_{4\text{ops}}(\text{“}A\ number\text{”}, \text{“}A\ number\ of\ tests\text{”}) = 1 - \frac{9}{17} = 0.471$
- 2-gramme: $\text{sim}_{4\text{ops}}(\text{“}A\ number\text{”}, \text{“}A\ number\ of\ tests\text{”}) = 1 - \frac{9}{16} = 0.438$
- woorde: $\text{sim}_{4\text{ops}}(\text{“}A\ number\text{”}, \text{“}A\ number\ of\ tests\text{”}) = 1 - \frac{2}{4} = 0.5$

Twee mate wat in die evaluasie van masjienvertaling gebruik word, is verwant aan 4-bewerkingredigeersoortgelykheid: woordfouttempo* en vertaalfouttempo.* Hierdie twee mate werk noodwendig op die woordvlak, en verskil wat betref die normalisering en die modellering van 'n addisionele redigeerbewerking, nl. die skuif van een of meer woorde. Alhoewel ons nie hierdie mate ondersoek nie, ondersoek ons wel 'n ander variasie, waarna volgende gekyk word.

(en) *word error rate, WER*

(en) *translation error rate, TER*

3.2.2 *3-bewerkingredigeersoortgelykheid*

Hierdie maat is soortgelyk aan die **4-bewerkingredigeersoortgelykheid**, maar die onderliggende afstandfunksie beskou nie vervanging as een van die basiese bewerkings nie. 'n Vervanging word dus as 'n invoeging en 'n skrapping gemodelleer (twee bewerkings). Die identiteitsbewerking is die derde bewerking. Dit staan ook bekend as die *indel*-afstand (*in*sert en *del*ete).

Die maksimumafstand tussen twee stringe is die som van die lengtes van die twee stringe, en sal aangetref word wanneer die twee stringe geen gemene karakters het nie. Die soortgelykheidsmaat word daarom gedefinieer as

$$\text{sim}_{3\text{ops}}(A, B) = 1 - \frac{\text{edit}_{3\text{ops}}(A, B)}{\text{maks}(|A| + |B|)}.$$

In die evaluasie deur [7] het hierdie maat dikwels die beste presteer. Hierdie maat is kommutatief.

3.2.3 *Sleuteldruksoortgelykheid*

Hierdie soortgelykheidsmaat is gebaseer op 'n afstandfunksie wat die menslike inspanning modelleer wat nodig is om een string in 'n ander te omskep met 'n sleutelbord.* Dit modelleer dus nie net invoegings, skrappings en vervangings nie, maar ook die moeite om die redigeerwyser* te posisioneer om telkens so 'n verandering te maak. Dit neem ook die feit in ag dat dit maklik is om teks te skrap, terwyl dit moeiliker is om nuwe teks te tik (invoeging) wat dus 'n groter afstand impliseer vir invoeging.

(en) *keyboard*

(en) *cursor*

Alhoewel dit probeer om die moeite te modelleer wat nodig is om 'n voorstel te redigeer, kan die sleuteldrukafstand slegs 'n benadering wees vir 'n ideale tikster, en ignoreer dit die realiteite van menslike foute, knipbordfunksionaliteit* en in der waarheid ander funksionaliteit wat moontlik in die sagteware bestaan, soos outovoltooiing of die invoeging van plaasbare items* [3]. Dit is egter 'n poging om die menslike redigeeraktiwiteit meer akkuraat te modelleer. Deur verskillende gewigte vir die verskillende bewerkings in die 4-bewerkingredigeersoortgelykheid te gebruik, kan die gedrag benader word. As mens dit egter vergelyk met hoe die sleuteldrukafstand hier beskryf word, sal gewigte nie die konstante oorhoofse koste per bewerking kan modelleer nie, en ook nie die konstante koste van skrapping ongeag die lengte van die skrapping nie.

(en) *clipboard functionality*

(en) *placeables*

Die funksie vir die skatting van sleuteldrukke wat in [84] gedefinieer is, neem verder die gebruik van die muis aan, wat veronderstel dat die gebruiker tydens redigeerwerk hande wissel tussen die sleutelbord en die muis (of soortgelyke toestel). 'n Vereenvoudigde benadering word hier gevolg: daar word aangeneem dat die gebruiker slegs met die sleutelbord werk. Met dié benadering word die meeste van bogenoemde eienskappe steeds vasgevang.

Vergeleke met die oorspronklike beskrywing [84] los ons die eksplisiete modellering van die bewerkings skuif, ruil en aangrensende ruil,* uit. Dit word gedoen met die aanname dat hierdie bewerkings met die sleutelbord uitgevoer sal word met

(en) *move, swap and adjacent swap operations*

Tabel 3.1: Gewigte vir verskillende aksies onder die sleuteldrukmaat

Bewerking	Sleuteldrukke	Verduideliking
Invoeging	$1 + c$	Posisionering van die wyser + c karakters getik
Skrapping	2	Merk van geselekteerde area en skrapping
Vervanging	$1 + c$	Merk van geaffekteerde area + c karakters getik

'n kombinasie van die bewerkings in tabel 3.1 hier bo eerder as met die muis en sleep-en-los-funksionaliteit* soos wat in die oorspronklike publikasie voorsien is. Alhoewel hulle hierdie maat slegs vir gebruik as deel van hulle evaluasiemetode aangebied het, kan dit ook as 'n onttrekkingsmaat gebruik word. Die verwantskap tussen 'n evaluasiemaat en 'n onttrekkingsmaat word bespreek in afdeling 3.4.

(en) *drag and drop functionality*

Alhoewel 'n mens enige aantal sleuteldrukke kan gebruik om een string in 'n ander te omskep, hoef slegs afstande tot en met die lengte van die verlangde (gekorrigeerde) string oorweeg te word, aangesien dit die nodige moeite is om die verlangde string van nuuts af te tik. Die lengte van die verlangde string, $|A|$, word dus as normaliseringskonstante gebruik en die soortgelykheidsmaat word gedefinieer as

$$\mathrm{sim}_{\mathrm{keys}}(A, B) = 1 - \frac{\mathrm{keys}(A, B)}{|A|}.$$

Hierdie maat is inherent karaktergeoriënteerd en word daarom nie tersaaklik geag om op woorde of n-gramme te werk nie. Hierdie maat is nie kommutatief nie.

Hier is 'n paar voorbeelde:

- $\mathtt{keystrokes}(\textit{“contact”}, \textit{“contacted”}) = 1 - \frac{2}{7} = 0.71$ (skrapping)

- $\mathtt{keystrokes}(\textit{“contacted”}, \textit{“contact”}) = 1 - \frac{1+2}{9} = 0.67$ (invoeging)

3.2.4 *n-gram-presisie*

N-gram-presisie [15] is geïnspireer deur die n-gram-presisie onderliggend aan die BLEU-telling vir die evaluasie van masjienvertaling [66]. Dit is gebaseer op 'n eenvoudiger n-gram-presisie as BLEU en gebruik 'n rekenkundige gemiddeld eerder as 'n meetkundige een. In plaas van die bondigheidstraf* in BLEU, gebruik dit 'n gebruikergespesifiseerde parameter Z met omvang $[0, 1]$ wat beheer hoe die presisie vir elke n-gram-orde genormaliseer word. Veranderinge aan die waarde van hierdie parameter veroorsaak 'n voorkeur vir korter of langer segmente; die gevolg is dat n-gram-presisie vir n-gram-herroeping verruil word.

(en) *brevity penalty*

Volledigheidshalwe word hulle formulering hier herhaal:

$$sim_{NGP}(A, B) = \frac{1}{N} \sum_{n=1}^{N} p_n(A, B) \tag{3.1}$$

waar

$$p_n(A, B) = \frac{|A_n \cap B_n|}{Z|A_n| + (1 - Z)|B_n|}. \tag{3.2}$$

Hierdie soortgelykheidsmaat het 'n mate van ooreenkoms met **token intersection** in [7] waar dit op gemengde 1- en 2-gramme gewerk het (vergelykbaar met 'n instelling van $N = 2$ in n-gram-presisie wat hier aangebied word).

Soos in die oorspronklike publikasie stel ons $N = 4$ en $Z = 0,75$. Uit hierdie vergelykings is dit duidelik dat die normalisering nie so eenvoudig is nie, en gevolglik is dit ook nie so maklik om daaroor te redeneer nie. Dit affekteer byvoorbeeld die berekening van bo- en ondergrense op die lengte van die kandidate, wat in afdeling 3.3.1 hier onder nader bestudeer word. 'n Hoër waarde vir Z bevoordeel langer voorstelle B uit die vertaalgeheue.

Net soos in die oorspronklike artikel word hierdie maat slegs op woordvlak oorweeg. Hierdie maat is nie oor die algemeen kommutatief nie. Dit is slegs kommutatief wanneer $Z = 0,5$.

3.3 IMPLEMENTASIEBESONDERHEDE

Hierdie afdeling bevat besonderhede van aspekte van implementasie wat nodig sal wees om die resultate van hierdie hoofstuk te dupliseer.

Vir al die woordgebaseerde mate gebruik ons die BreakIterator uit die ICU-projek[11] vir tekseenheididentifisering.* Vir n-gram-presisie word tekseenhede* verwyder wat slegs spasies of leestekens bevat soos in die oorspronklike publikasie en word stambepaling* soos volg uitgevoer: vir vergelyking van die Engelse bronteks tydens onttrekking word stambepaling uitgevoer op woorde m.b.v. die Engelse stambepaler van die Snowball-projek.[12] Tydens evaluasie bereken ons die soortgelykheid tussen twee tekste in die doeltaal. Die toepaslike stambepalers uit die Snowball-projek is ook vir die doeltale gebruik.

(en) *tokenisation*

(en) *tokens*

(en) *stemming*

Die eksperimente is uitgevoer m.b.v. die oopbron Amagama-vertaalgeheuebediener.[13] Een spesifieke implementasiebesonderheid wat relevant is, is die manier waarmee kandidaatvoorstelle van die databasis onttrek word: 'n volteksindeks oor die brontekssegmente word gehou wat dit moontlik maak om 'n aanvanklike lys van kandidate te trek wat ten minste 'n minimale oorvleueling van tekseenhede het. Tekseenhede word omgeskakel na stamme (soos bo genoem) in kleinletters, en alle stopwoorde word verwyder. Dit is 'n baie losse, aanvanklike filter wat steeds toelaat dat kandidate met baie lae soortgelykheid onderwerp kan word aan die volle oorweging deur die soortgelykheidsmaat wat gebruik word. Die tellings vir soortgelykheid wat aan elke kandidaat toegeken word, word volledig deur die onttrekkingsmaat in die eksperiment bepaal — die volteksenjin het geen invloed in hierdie opsig nie aangesien dit nie weer gebruik word na die aanvanklike filtrering nie. Die gebruik van 'n volteksindeks is bloot 'n optimering om 'n groot deel van die geheue vir oorweging te elimineer met elke navraag, soort-

11 http://site.icu-project.org/
12 http://snowball.tartarus.org/
13 http://amagama.translatehouse.org/

gelyk aan die bedoeling van *approximate query coverage* (AQC) in [82]. Ander optimerings raak die grense op die lengtes van kandidate. Dit word in die volgende afdeling oorweeg.

3.3.1 *Grense op die lengte van kandidate*

Dit is verwerkingsintensief en onnodig om uitvoerig alle inskrywings in die geheue met die navraagstring te vergelyk. Weens die definisie van die redigeerafstande kan kandidate van sekere lengtes nie die nodige drempel vir soortgelykheid f haal nie, ongeag die soortgelykheid andersins. In [7] het die outeur verslag gedoen oor die onder- en bogrense op die lengtes van kandidate wat afgelei word van die lengte van die navraagstring. Hierdie onder- en bogrense kan gebruik word om die soekruimte te begrens ter wille van meer effektiewe werkverrigting. Twee afstande se grense uit sy werk wat relevant is tot ons eksperiment word in tabel 3.2 weergegee in die notasie van hierdie hoofstuk.

Tabel 3.2: Grense op die lengte van kandidate vir enkele mate

Maat	Ondergrens	Bogrens
4-bewerkingredigeersoortgelykheid	$f\|A\|$	$\|A\|/f$
3-bewerkingredigeersoortgelykheid	$f\|A\|/(2-f)$	$(2-f)\|A\|/f$

Aangesien die sleuteldruksoortgelykheid en n-gram-presisie nie in daardie studie ingesluit is nie, ondersoek ons hierdie aspek wat onontbeerlik is vir redelike werkverrigting wanneer daar op 'n groot datastel gewerk word.

3.3.1.1 *Sleuteldruksoortgelykheid*

Die ondergrens op die lengtes van die voorstelle onder die sleuteldruksoortgelykheid is vormlik verwant aan dié van 4-bewerkingredigeersoortgelykheid. Ons bereken die ondergrens met die aanname dat 'n enkele aaneenlopende substring karakters ingevoeg moet word. Dit verteenwoordig die "beste kans" wat 'n kort string het om steeds die nodige drempel vir soort-

gelykheid te haal, aangesien daar oorhoofse koste is vir elke substring wat ingevoeg word onder die sleuteldrukafstand. As c karakters ingevoeg moet word om dit om te skakel na die navraagstring, sal 4-bewerkingredigeerafstand 'n afstand van c rapporteer. In die geval van die sleuteldrukafstand is dit $c + 1$. Net so is die ondergrens bloot een hoër as dié van 4-bewerkingredigeersoortgelykheid, aangesien dit die verdere oorhoofse koste moet akkommodeer vir ten minste 'n enkele invoeging. Die ondergrens is dus $f|A| + 1$.

Teoreties is daar geen bogrens op die lengte van 'n voorstel nie, aangesien die skrapping van enige aantal aaneenlopende oorbodige karakters 'n konstante koste van 2 sal beloop (sien tabel 3.1). Om die soektog oor die databasis effens te beperk, word 'n ruim bogrens van $2|A|/f$ gebruik. Met 'n string van lengte 100 en $f = 0,4$, word kandidate tot en met 'n lengte van 500 dus oorweeg. Hierdie grens laat kandidate toe tot en met dubbeld die lengte van 4-bewerkingredigeersoortgelykheid.

As voorbeeld, oorweeg 'n navraagstring "metaphor" (8 karakters) met $f = 0,5$: die ondergrens is $0,5 \times 8 + 1 = 5$ en die bogrens is $2 \times 8/0,5 = 32$. Die soektog vir voorstelle vir "metaphor" word dus ingeperk tot inskrywings met brontekslengte tussen 5 en 32 karakters inklusief.

3.3.1.2 N-*gram-presisie*

N-gram-presisie is gedefinieer in vergelykings 3.1 en 3.2 as 'n maat wat op woordvlak werk. Hierdie maat word gekenmerk deur die gebruik van versamelings (sakke woorde) saam met hulle kardinaliteite. In hierdie afdeling poog ons om die getal kandidate uit die geheue vir B in te perk met grense wat van A afgelei word sodat alle kandidate met $sim_{NGP}(A, B) \geqslant f$ se lengtes noodwendig binne hierdie grense sal wees. Die grense sal dus verwys na woorde, nie karakters nie.

Vervolgens word die bo- en ondergrense van $|B_n|$ vir 'n enkele waarde van n van die ongelykheid $p_n \geqslant f$ afgelei. Daarna word die grense gekombineer oor alle n-gram-ordes.

3.3.1.3 *Ondergrens*

Aanvaar dat B_n aansienlik kleiner is as A_n, maar dat dit maksimaal oorvleuel (die snyding in die teller* van vergelyking 3.2 word gemaksimeer), m.a.w. $|B_n \cap A_n| = |B_n|$.

(en) *numerator*

$$
\begin{aligned}
B_n &\subset A_n \\
\therefore |B_n| &\leqslant |A_n| \\
\frac{|B_n|}{Z|A_n| + (1-Z)|B_n|} &\geqslant f \\
|B_n| &\geqslant fZ|A_n| + f(1-Z)|B_n| \\
|B_n| &\geqslant \frac{fZ|A_n|}{1-f(1-Z)}
\end{aligned}
$$

3.3.1.4 *Bogrens*

Om 'n bogrens te bereken vir $|B_n|$ neem ons aan dat B_n heelwat groter is as A_n en neem weereens maksimale oorvleueling aan.

$$
\begin{aligned}
B_n &\supset A_n \\
\therefore |B_n| &\geqslant |A_n| \\
\frac{|A_n|}{Z|A_n| + (1-Z)|B_n|} &\geqslant f \\
|A_n| &\geqslant fZ|A_n| + f(1-Z)|B_n| \\
|B_n| &\leqslant \frac{(1-fZ)|A_n|}{f(1-Z)}
\end{aligned}
$$

3.3.1.5 *Kombinasie van grense vir alle n-gram-ordes*

Die onder- en bogrense soos bo bereken, is slegs relevant vir p_n vir 'n enkele waarde van n. Ons benodig steeds grense wat korrek is vir die kombinasie van veelvuldige p_n-waardes om die grense op die volle n-gram-presisie te bereken. Die grense op $|B_n|$ in p_n vir 'n sekere waarde van n is nie meer streng nie, aangesien die n-gram-presisie 'n gemiddeld oor veelvuldige p_n-waardes is.

Die som van die ondergrense op $|B_n|$ vir elke n-gram-orde gee 'n ondergrens op $\sum_{n=1}^{N} |B_n|$. Hier is dit belangrik om $|A_n|$ afsonderlik te bereken vir elke n-gram-orde. Vir die ondergrens:

$$\sum_{n=1}^{N} |B_n| \geqslant \sum_{n=1}^{N} \frac{fZ|A_n|}{1 - f(1 - Z)}$$

En soortgelyk vir die bogrens:

$$\sum_{n=1}^{N} |B_n| \leqslant \sum_{n=1}^{N} \frac{(1 - fZ)|A_n|}{f(1 - Z)}$$

Die onder- en bogrens word al twee bepaal deur $\sum_{n=1}^{N} |A_n|$ (wat eenmalig per navraag bereken kan word) en 'n konstante faktor. Ons kan dus $\sum_{n=1}^{N} |B_n|$ saam met elke kandidaat B stoor, wat kan help om die soektog oor die vertaalgeheuedatabasis te beperk wanneer vir 'n optimale B gesoek word, selfs as dit nodig is om n-gram-presisie te bereken met wisselende waardes van f of Z.

Beskou die bronteksnavraag "I think therefore I am" as voorbeeld. Dit bestaan uit 5 tekseenhede en $\sum_{n=1}^{N} |A_n| = 4 + 4 + 3 + 2 = 13$. Vir $f = 0,5$ en $Z = 0,75$ is die ondergrens $\frac{0,5 \times 0,75 \times 13}{1 - 0,5(1 - 0,75)} = 5,57$ en die bogrens $\frac{(1 - 0,5 \times 0,75) \times 13}{0,5(1 - 0,75)} = 65$. 'n Voorstel B uit die vertaalgeheue hoef dus slegs oorweeg te word as $\sum_{n=1}^{N} |B_n|$ tussen 6 en 65 inklusief is.

Volledigheidshalwe word die bo- en ondergrense van al die relevante mate wat in hierdie hoofstuk gebruik word, aangegee in tabel 3.3. Vir al die mate buiten n-gram-presisie, word die grense aangedui in terme van blote aantal eenhede van die betrokke vlak van detail, bv. aantal karakters of aantal woorde. In die geval van n-gram-presisie word die grense aangedui as grense op $\sum_{n=1}^{N} |B_n|$.

Tabel 3.3: Grense op die lengte van kandidate vir elke maat

Maat	Ondergrens	Bogrens
4-bewerkingredigeersoortgelykheid	$f\|A\|$	$\|A\|/f$
3-bewerkingredigeersoortgelykheid	$f\|A\|/(2 - f)$	$(2 - f)\|A\|/f$
Sleuteldruksoortgelykheid	$f\|A\| + 1$	$2\|A\|/f$
N-gram-presisie	$\sum_{n=1}^{N} \frac{fZ\|A_n\|}{1 - f(1 - Z)}$	$\sum_{n=1}^{N} \frac{(1 - fZ)\|A_n\|}{f(1 - Z)}$

3.3.2 *Uitbreiding van n-gram-presisie vir enige segmentlengte*

Hier onder word 'n paar implementasiekeuses in meer besonderhede beskryf, en tekortkominge van die tegniek soos gepubliseer in [15] word uit die weg geruim sodat dit vir algemene doeleindes as 'n soortgelykheidsmaat gebruik kan word.

Die oorspronklike publikasie het slegs stringe met vyf of meer tekseenhede oorweeg [15]. Aangesien ons werklike datastelle wil gebruik, is dié beperking as te beperkend geag, veral aangesien kort stringe baie algemeen is in die GNOME-datastel—die gemiddelde segment in die GNOME-datastel het minder as vyf woorde. Datastelle word hier onder in afdeling 3.4 beskryf. Vervolgens word dus gespesifiseer hoe om dit te bereken op stringe met minder as vyf tekseenhede:

- Vir elke berekening, begin deur p_1 te bereken.

- Vir elke $n > 1$, as beide $|B_n|$ en $|A_n|$ nie 1 oorskry nie, word aangeneem dat p_n vir die laer waardes van n reeds die soortgelykheid voldoende geëvalueer het en word die berekening vroeg verlaat. Só word N dus (moontlik) tot N' verminder.

  Die regverdiging is soos volg: as $|A_n| \leqslant 1$ en $|B_n| \leqslant 1$, toets die berekening van $|A_n \cap B_n|$ (in die teller van vergelyking 3.2) bloot vir gelykheid tussen A en B (as verskille in leestekens en stambepaling geïgnoreer word). In die oorspronklike publikasie met $N = 4$ en segmente wat vyf of meer tekseenhede bevat, is dit ook die geval dat ten minste twee 4-gramme beskikbaar sou wees vir die berekening van p_4. Op dié manier tref p_n by $n = N'$ meer as bloot 'n binêre onderskeid tussen 0,0 en 1,0. By $n > N' + 1$ is p_n ook noodwendig 0,0 en sal $sim_{NGP}(A, A)$ nie 1,0 wees soos verwag word met die vergelyking van identiese stringe nie.

- As slegs N' ordes van p_n verwerk word, verander die normaliseringsfaktor in vergelyking 3.1 van $\frac{1}{N}$ na $\frac{1}{N'}$.

3.4 EKSPERIMENTELE OPSET

In 'n poging om die vraag oor 'n evaluasiemaat se voorkeur (sien bladsy 32) te beantwoord, word 'n evaluasie-eksperiment opgestel wat gebaseer is op die metode van [84], terwyl die skatting van sleuteldrukmaat met 'n ander uitgeruil word in elke geval. Aangesien hierdie metode beide 'n presisieagtige en herroepingagtige maat bied, laat dit mens toe om ondersoek in te stel na die effek op beide mate, asook 'n kombinasie daarvan in die F_1-telling. Die metode is verder verbeter met 10-voudige kruisvalidasie, soos gedoen in [7]. Aangesien ware stratifikasie nie moontlik is in hierdie soort datastel nie, verifieer ons eerder dat die verspreiding van die brontekslengtes oor die 10 voue konsekwent is deur te kontroleer dat die gemiddelde lengte in elke vou naby aan die gemiddeld van die hele datastel is. Dit dien as "semistratifikasie" in dieselfde gees as [7].

Vir elke datastel word elke deel van die eksperiment gespesifiseer met 'n tweetal soortgelykheidsmate: een vir onttrekking en een vir evaluasie (nadat resultate verkry is). Wanneer 'n maat M gedurende evaluasie gebruik word, word gesê dat die resultate "onder" die soortgelykheidsmaat M geëvalueer word. Die eksperiment voer dus uit oor die Cartesiese produk van al die volgende soortgelykheidsmate (met kort vorm in hakies):

- 4-bewerkingredigeersoortgelykheid oor karakters, karakter-2-gramme en woorde (edit4, edit4ngram, edit4word);
- 3-bewerkingredigeersoortgelykheid oor karakters, karakter-2-gramme en woorde (edit3, edit3ngram, edit3word);
- sleuteldrukskatting (slegs karakters) (keystrokes);
- n-gram-presisie (slegs woorde) (ngp).

Twee datastelle met twee taalkundig onverwante doeltale word gebruik: Die 2004_1-substel van die DGT-TM weergawe 2011 [76] en die gebruikerkoppelvlakvertalings van GNOME 3.8,[14] spesifiek vir die taalpare Engels–Frans (`en-fr`) en Engels–Hongaars (`en-hu`).

14 Beskikbaar vanaf https://l10n.gnome.org/releases/gnome-3-8/

Terwyl die DGT-korpus wetgewende teks bevat, bevat die GNOME-korpus die gebruikerkoppelvlakvertalings van die GNOME-werkskermomgewing. Die GNOME-tekste is ook anders in dié sin dat dit XML-markering, plekhouers en meer nietekstuele elemente kan bevat, soos wat algemeen is in tekste vir sagtewarelokalisering. Tabel 3.4 bied 'n opsomming van die korpusstatistiek. Die standaardafwykings* vir die gemiddelde karakterlengtes dui op die afwyking van die gemiddeldes oor die 10 voue, nie die standaardafwykings op die segmentvlak nie. Dit dui op die piepklein variansie* in die gemiddelde segmentlengte tussen die 10 voue, en toon dat die "semistratifikasie", soos bo genoem, redelik regverdig is.[15]

(en) *standard deviations*

(en) *variance*

Tabel 3.4: Korpusstatistiek

Korpus	Unieke segmente	Karakters (gemiddeld)
DGT (fr)	71 033	126,75 ± 1,225
DGT (hu)	45 964	125,61 ± 1,186
GNOME (fr)	36 493	27,51 ± 0,357
GNOME (hu)	36 008	27,65 ± 0,431

Elkeen van die voue in die kruisvalidasie dien een keer as toetsstel en dien die ander nege keer as deel van die vertaalgeheue. Slegs die voorstel met die hoogste rang volgens die soortgelykheidsmaat wat vir onttrekking gebruik word, word oorweeg (gelyke waardes word ewekansig beslis). 'n Resultaatstel vir elke soortgelykheidsmaat word op dié manier geskep en vir evaluasie voorgelê. Een verskil in die evaluasiemetode hier is dat die *doelteks* van die voorstelle vergelyk word met die verwysingsvertalings uit die toetsdata, in teenstelling met die oorspronklike metode waarvolgens brontekste vergelyk is. Die outeurs het in der waarheid opgemerk [84]:

> As die maat werklik akkuraat moet wees, moet die aantal sleuteldrukke natuurlik getel word op die

15 Alhoewel 'n ewekansige verdeling in 10 voue telkens verskillende partisies sal lewer, word statistieke soortgelyk aan dié wat hier gerapporteer word altyd verkry.

> doeltaalsegmente van die voorstelle, soos wat dit met die verlangde doeltaalvertaling vergelyk.[16]

Die rede hoekom hulle dit nie doen nie is weens 'n kommer oor subjektiwiteit in die keuse van 'n verwysingsvertaling. Sedert daardie navorsing gepubliseer is, is die gebruik van sulke verwysingsvertalings egter algemeen aanvaar in die masjienvertaalgemeenskap, en gaan ons dus voort om dit só te gebruik. Ons glo dat die 10-voudige kruisvalidasie en groter datastelle sulke knelpunte elimineer. Dit is ook goed om kennis te neem dat 'n keuse om die bronteks te gebruik eenvoudig die berekening sal dupliseer wat gedoen is toe voorstelle onttrek is, en sal dus geen manier bied om die beoogde funksie van die vertaalgeheuestelsel te evalueer nie, nl. om 'n doelteks te onttrek wat die vertaler help in die vorming van 'n doelteks vir die huidige bronsegment.

In die evaluasie van masjienvertaaleksperimente word veelvuldige verwysingsvertalings aanbeveel. Dit is 'n luukse wat ons nie in dié geval het nie.

'n Sentrale aspek van die evaluasie is om, volgens die soortgelykheidsmaat, 'n maat van "nuttigheid" vir elke voorstel te definieer. Dit maak dit moontlik om 'n geweegde "trefwaarde" oor alle voorstelle te bereken (dit word onder beskryf), eerder as om bloot "tref-" en "misslae" onder die voorstelle te tel.[17]

Gedurende evaluasie word die *geweegde presisie* P_f^w en die *geweegde ware effektiwiteit* F_f^w bereken soos dit gedefinieer is in [84], vir elkeen van die soortgelykheidsmate in die evaluasiemetode. Hierdie twee mate verteenwoordig onderskeidelik presisieag-

16 "For this measure to be strictly accurate, the key-stroke count should of course be carried out on the target-language text segments associated with the matches, as compared to the desired target translation."

17 Die konsepte van "tref" en "mis" word ontleen uit inligtingherwinning waar 'n resultaat van 'n soekenjin as korrek of verkeerd geklassifiseer word met hierdie terme.

tige en herroepingagtige hoeveelhede. Hulle word soos volg bereken:

$$P_f^w = \frac{h_f^w}{m_f}$$
$$F_f^w = \frac{h_f^w}{n}$$

waar n die getal segmente is wat geëvalueer word, m_f die getal hieruit is waarvoor 'n voorstel (trefslag) by soortgelykheidsdrempel f onttrek is, en h_f^w (onder gedefinieer) die geweegde "trefwaarde" by die soortgelykheidsdrempel f is. Vir die soortgelykheidsmate wat op die redigeerafstande gebaseer is, word die gewig op 'n manier bereken wat vormgewys soortgelyk is aan die veralgemeende soortgelykheidsmaat wat in afdeling 3.2 bespreek is, nl. $1 - d/l$. Die verskil is dat, terwyl die soortgelykheidsmate 'n omvang van $[0, 1]$ het, die gewigte vir evaluasie 'n omvang van $[-1, 1]$ het. Dit het 'n belangrike eienskap tot gevolg, nl. dat onbehulpsame voorstelle (met baie lae soortgelykheid aan die verwysingsvertaling) gestraf word en dus 'n negatiewe telling bydra tot h_f^w eerder as bloot 'n klein positiewe telling.

'n Veralgemening van die skema van gewigte in [84] word soos volg voorgestel:

$$\hat{W}(A, B) = 2 \times sim(A, B) - 1$$

waar $sim(A, B)$ die evaluasiemaat in gebruik is in elke geval. In die oorspronklike artikel is slegs die skatting van sleutelddrukke gebruik. Die geweegde trefwaarde is die som van die $\hat{W}$-waardes vir elke verwysingsvertaling a_i en bygaande voorstel b_i:

$$h_f^w = \sum_{i=1}^{n} \hat{W}(a_i, b_i)$$

Die $n - m_f$ segmente met geen voorstelle by f word geweeg met neutrale "nuttigheid", m.a.w. nul — hulle het geen waarde ingehou vir die vertaler nie, en geen kognitiewe las vir die vertaler meegebring nie.

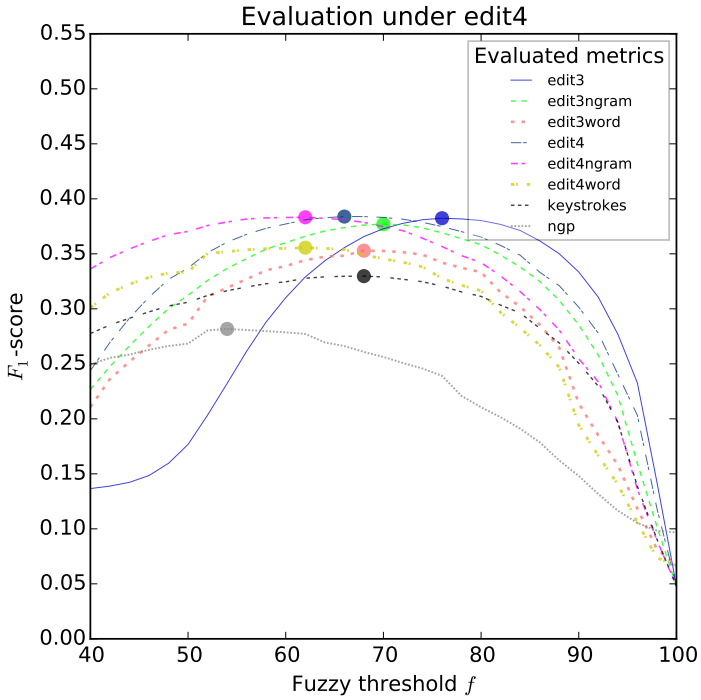
Hierdie evaluasiemetode laat verskeie vorme van ondersoek toe: dit modelleer die effek van die soortgelykheidsdrempel f en die veranderlikes P_f^w en F_f^w kan onafhanklik van mekaar bestudeer word. Aangesien dit nie die hoofdoel van hierdie navorsing is nie, beskou ons die gebalanseerde F_1-telling as voldoende instrument vir die ondersoek in hierdie hoofstuk.[18] Vir sekere toepassings is ander evaluasie-uitkomste dalk meer belangrik en kan eerder gebruik word.

3.4.1 *Die invloed van die soortgelykheidsdrempel*

Dit is aanloklik om elkeen van die eksperimente met 'n vaste soortgelykheidsdrempel uit te voer om die eksperiment te vereenvoudig. Alhoewel geen spesifieke waarde 'n voor die hand liggende keuse is nie, sou 'n keuse soos 70% nie kontroversieel wees nie, omdat dit die verstekwaarde is in sommige programme vir rekenaargesteunde vertaling. So 'n drempel sou die indruk gee dat dit die gedrag van sulke vertaalprogramme naboots. In hierdie afdeling ondersoek ons egter die effek van die soortgelykheidsdrempel, en sien dat dit as 'n veranderlike in eie reg ondersoek moet word, aangesien die prestasie van die soortgelykheidsmate hoogs afhanklik is van die waarde van die soortgelykheidsdrempel.

Ten einde die effek van veranderinge in die soortgelykheidsdrempel deeglik te ondersoek, is evaluasie-uitkomste oor 'n wye omvang van f nodig. Hiervoor filtreer ons die voorstelle wat onttrek word met 'n baie lae soortgelykheidsdrempel (sê 40%), maar doen die evaluasie iteratief vanaf daardie waarde en vermeerder dit in stappe van 2%. Elke sodanige verhoging elimineer sommige voorstelle wat nie meer die drempel haal nie, waarna die evaluasietellings weer bereken word. Dit maak dit moontlik om die verandering in evaluasie-uitkomste te sien oor 'n groot omvang van die soortgelykheidsdrempel asof vele individuele eksperimente gedoen is. Dit kan gebruik word om presisie teenoor herroeping grafies te stip, wat verdere insigte

18 $F_1 = 2 \times \frac{P_f^w \times F_f^w}{(P_f^w + F_f^w)}$



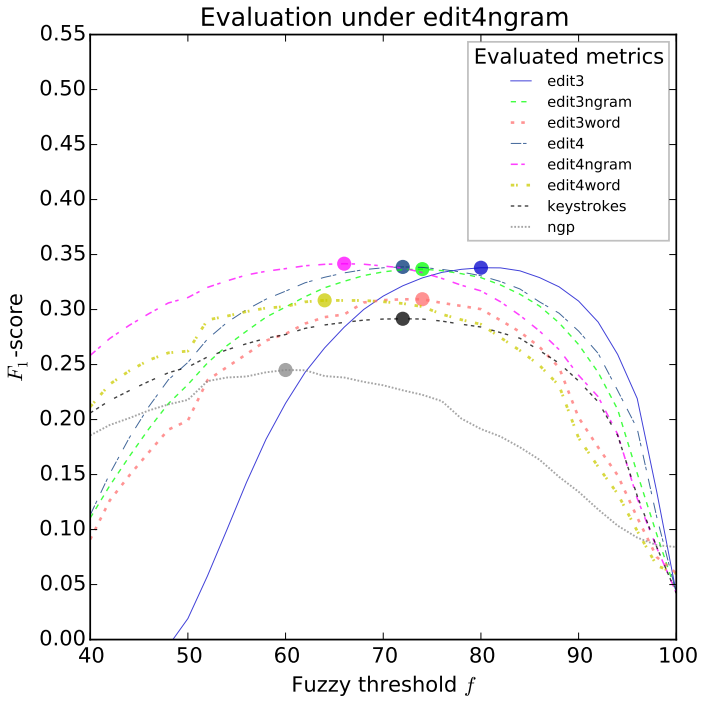
Figuur 3.1: Evaluasie van alle onttrekkingsmate op die DGT se `en-fr`-datastel onder die **edit4**-maat.

kan bied oor die verband tussen presisie, herroeping en die soortgelykheidsdrempel. Laasgenoemde val buite die bestek van hierdie studie.

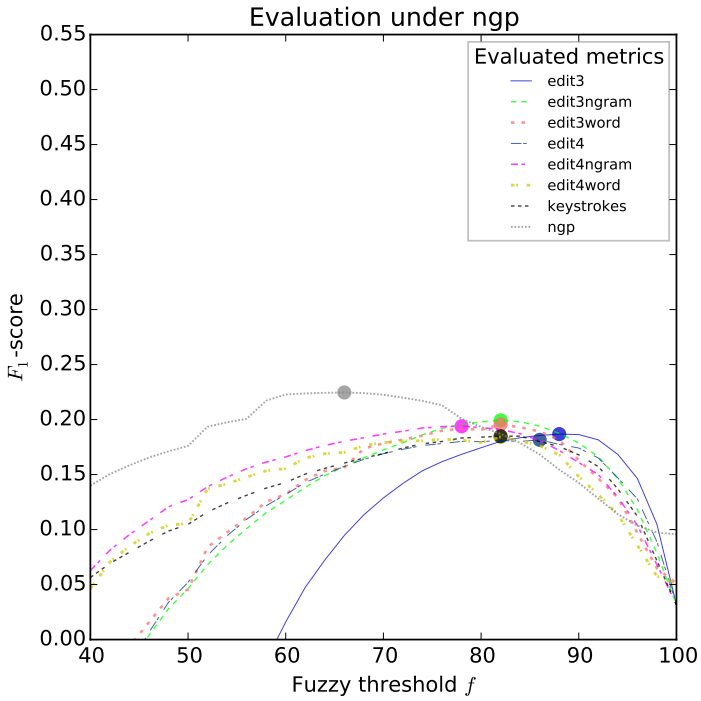
Figure 3.1–3.3 toon die F_1-tellings gestip teenoor die soortgelykheidsdrempel f, telkens onder 'n ander evaluasiemaat. Die optimale waarde vir elke kromme* word met 'n kol op die kromme aangedui. Met 'n baie hoë drempel ($f > 90\%$) is die F_1-telling laag vir alle mate. Dit weerspieël dat min voorstelle verskaf word wat aan so 'n streng vereiste van soortgelykheid voldoen. Dit verteenwoordig 'n opstelling van die vertaalgeheuestelsel met hoë presisie en lae herroeping. As die grafiek verder na links beskou word in die rigting van laer waardes van f, kan gesien word hoe die prestasie van verskillende mate verander in verhouding tot veranderinge in f.

(en) *curve*

Ons sien dat die F_1-telling van 3-bewerkingredigeersoortgelykheid oor karakters (**edit3**) sy optimum bereik by hoër waar-



Figuur 3.2: Evaluasie van alle onttrekkingsmate op die DGT se `en-fr`-datastel onder die **edit4ngram**-maat.



Figuur 3.3: Evaluasie van alle onttrekkingsmate op die DGT se `en-fr`-datastel onder die **ngp**-maat.

des van f. Aan die ander kant is n-gram-presisie (**ngp**) se optimum gewoonlik by heelwat laer waardes van f as die ander mate. Dit gebeur konsekwent en ongeag of die maat in daardie deel van die eksperiment die beste presteer.

Aangesien elke soortgelykheidsmaat sy optimale waarde by verskillende waardes van f bereik (soms ver uitmekaar), is dit belangrik om te besef dat 'n keuse van enige vaste waarde vir f om al die soortgelykheidsmate mee te vergelyk, nie elke maat by sy sterkste punt sal vergelyk nie. As 'n keuse van 'n ander (vaste) soortgelykheidsdrempel vir 'n evaluasie-eksperiment die uitkoms van evaluasie kan verander, sal dit die eksperimentele resultate ongeldig maak. Dit moet dus in ag geneem word vir die ontwerp van evaluasiemetodes. Dit bly egter belangrik om die stelsel te evalueer by 'n vaste waarde van f wat vir elke maat bepaal is, aangesien dit is hoe vertaalgeheuestelsels in die praktyk werk. Die prestasie van 'n stelsel by waardes van f ver weg van die optimale opstelling is ontoepaslik vergeleke met 'n stelsel wat verfyn* kan word tot 'n optimale opstelling (vir sover as wat die verfyning relevant is tot die taak). Die spesifieke vorm van die grafiek is dus irrelevant.

(en) *tune*

Dit is dus van kardinale belang om die verlangde telling (F_1-telling in hierdie geval) apart te optimeer vir elke soortgelykheidsmaat oor die omvang van die veranderlike f deur die waarde van die veranderlike f te verfyn op 'n aparte datastel, voordat die werklike evaluasie begin. In elkeen van die 10 voue van die finale eksperiment word 80% van die datastel gebruik as 'n vertaalgeheue, 10% om die veranderlike f te verfyn en die oorblywende 10% om die evaluasietelling te bepaal. Die gemiddeld van hierdie evaluasietellings uit elkeen van die 10 voue is die algehele telling wat in die volgende afdeling genoem word.

3.4.2 *Resultate*

In die meeste gevalle bereik die F_1-telling vir elke maat 'n duidelike maksimum binne die gegewe omvang van f — die

punt waar die soortgelykheidsdrempel f optimaal sou wees vir die betrokke onttrekkingsmaat onder die evaluasiemaat wat gebruik word. Die enigste uitsondering is **ngp** wat blykbaar in enkele gevalle steeds verbeter soos wat f in die rigting van 40% verminder, alhoewel dit meestal niekompeterende F_1-tellings het. Dit wil voorkom asof **ngp** moontlik effens hoër tellings sou kon behaal by 'n drempel onder 40% in hierdie gevalle, alhoewel dit nie lyk asof dit sy rang relatief tot ander mate sou beïnvloed nie. In hierdie gevalle word die telling by $f = 40\%$ vir die vou gebruik al is optimaliteit nie bewys nie.

Voortaan word slegs hierdie optimale waarde vir elke maat oorweeg. In die twee datastelle van die DGT-korpus kies elke soortgelykheidsmaat homself as die maat met die beste prestasie, met **edit4ngram** as enigste uitsondering wat homself nie regstreeks as die wenner kies vir die taalpaar Engels–Hongaars nie. In hierdie geval is die twee topmate (wat **edit4ngram** insluit) baie naby aan mekaar ('n verskil van minder as 0,0001).

Tabel 3.5: Resultate van evaluasie: DGT (en–hu). F_1-tellings met beste prestasie in vetdruk aangedui.

	Evaluasiemate							
Onttrekkingsmate	edit3	edit3ngram	edit3word	edit4	edit4ngram	edit4word	keystrokes	ngp
edit3	**0.400**	0.325	0.258	0.312	0.257	0.213	0.262	0.095
edit3ngram	0.397	**0.328**	0.255	0.309	0.255	0.203	0.263	0.102
edit3word	0.374	0.304	**0.282**	0.290	0.236	0.229	0.244	0.105
edit4	0.397	0.323	0.258	**0.315**	**0.259**	0.216	0.263	0.091
edit4ngram	0.398	0.325	0.260	0.314	**0.259**	0.213	0.264	0.097
edit4word	0.374	0.300	0.281	0.293	0.239	**0.237**	0.245	0.099
keystrokes	0.364	0.296	0.231	0.272	0.222	0.176	**0.268**	0.091
ngp	0.318	0.251	0.189	0.221	0.170	0.133	0.187	**0.119**

In die geval van die GNOME-korpus vind ons 'n soortgelyke situasie — in die Engels–Franse datastel kies elkeen van die mate homself as wenner. In die Engels–Hongaarse datastel kies drie van die agt mate (**edit3**, **edit3word** en **edit4**) nie hulself nie, maar die F_1-tellings vir hierdie drie mate is baie naby aan die

Tabel 3.6: Resultate van evaluasie: DGT (en-fr). F_1-tellings met beste prestasie in vetdruk aangedui.

	Evaluasiemate							
Onttrekkingsmate	edit3	edit3ngram	edit3word	edit4	edit4ngram	edit4word	keystrokes	ngp
edit3	**0.449**	0.384	0.324	0.380	0.336	0.290	0.338	0.197
edit3ngram	0.447	**0.390**	0.324	0.375	0.335	0.282	0.340	0.209
edit3word	0.419	0.357	**0.349**	0.351	0.308	0.307	0.316	0.205
edit4	0.445	0.379	0.322	**0.383**	0.338	0.291	0.337	0.191
edit4ngram	0.447	0.386	0.324	0.382	**0.340**	0.290	0.340	0.204
edit4word	0.418	0.351	0.343	0.354	0.307	**0.310**	0.308	0.193
keystrokes	0.404	0.351	0.296	0.327	0.290	0.248	**0.353**	0.195
ngp	0.355	0.301	0.258	0.277	0.242	0.211	0.265	**0.234**

wenmaat. As die konsekwentheid van die voorkeur in die korpusse in ag geneem word, asook die piepklein verskil tussen die gevalle waar die voorkeur nie volkome was nie, neem ons aan dat die vier genoemde gevalle nie ons resultate ongeldig maak nie. Die volle resultate vir die vier datastelle word in tabelle 3.5–3.8 gewys. Rye dui op die afvoer van die vertaalgeheuestelsel soos dit met elkeen van die soortgelykheidsmate onttrek word, en elke kolom dui die evaluasie met 'n enkele soortgelykheidsmaat aan. Gevalle waar 'n maat nie homself gekies het nie, word in skuinsdruk aangedui. Die wenmaat in 'n kolom word in vetdruk aangedui, en 'n vetgedrukte inskrywing op die diagonaal dui op 'n maat wat homself kies.

Afgesien van die selfvoorkeur is daar gewoonlik 'n verskil in prestasie tussen die woordgebaseerde mate en die karaktergebaseerde mate wanneer daar onder een van die karaktergebaseerde mate geëvalueer word. Dit is opmerklik dat hierdie verskil minder ooglopend is wanneer daar onder een van die woordgebaseerde mate geëvalueer word. Dit wil dus voorkom asof daar ten minste 'n effense sydigheid is tussen die woordgebaseerde mate en die karaktergebaseerde mate. In hierdie opsig lyk dit of die 2-gram-mate en die sleuteldrukmaat groter ooreenkoms het met die karaktergebaseerde mate.

Tabel 3.7: Resultate van evaluasie: GNOME (en–hu). F_1-tellings met beste prestasie in vetdruk aangedui.

	Evaluasiemate							
Onttrekkingsmate	edit3	edit3ngram	edit3word	edit4	edit4ngram	edit4word	keystrokes	ngp
edit3	**0.332**	0.244	0.061	**0.236**	0.181	0.035	0.156	0.017
edit3ngram	**0.332**	**0.252**	0.048	0.229	0.176	0.017	0.154	0.016
edit3word	0.239	0.160	*0.142*	0.145	0.101	0.102	0.104	0.026
edit4	0.322	0.231	0.057	*0.234*	0.178	0.040	0.148	0.024
edit4ngram	0.329	0.243	0.062	**0.236**	**0.183**	0.040	0.155	0.016
edit4word	0.243	0.161	**0.144**	0.157	0.105	**0.119**	0.102	0.016
keystrokes	0.289	0.206	0.067	0.177	0.126	0.037	**0.161**	0.014
ngp	0.225	0.157	-0.002	0.133	0.078	-0.043	0.076	**0.062**

edit3ngram en **edit4** het dikwels soortgelyke waardes, veral by hoër waardes van f. As 'n verandering van 'n enkele karakter tussen twee stringe oorweeg word, maak dit sin: onder **edit3ngram** affekteer so 'n verandering twee 2-gramme en onder **edit4** affekteer dit een karakter. Die verskille in normalisering beteken dat hierdie verskil meestal "uitkanselleer" wat resulteer in soortgelyke tellings.

3.5 BESPREKING

Evaluasie wat m.b.v. k-voudige kruisvalidasie toegepas word, is veronderstel om sommige bronne van sydigheid in die toetsdata te elimineer. Die ondersoek is uitgevoer oor twee uiteenlopende datastelle, en is herhaal vir taalpare met twee taalkundig onverwante doeltale. Alhoewel daar geringe variansie is in die resultate van die onderskeie datastelle, kom dieselfde resultaat na vore in al vier datastelle amper sonder uitsondering: *dat elke soortgelykheidsmaat 'n voorkeur vir homself het*. In 4 van die 32 gevalle waar die sydigheid nie volkome was nie, was die verskil met die wenmaat baie klein.

Hierdie werk verskaf 'n nuwe perspektief op die mate wat niekompeterend was in vorige studies [7, 15, 82]. 'n Maat soos

Tabel 3.8: Resultate van evaluasie: GNOME (en-fr). F_1-tellings met beste prestasie in vetdruk aangedui.

	Evaluasiemate							
Onttrekkings-mate	edit3	edit3ngram	edit3word	edit4	edit4ngram	edit4word	keystrokes	ngp
edit3	**0.372**	0.275	0.084	0.262	0.201	0.048	0.175	0.046
edit3ngram	0.364	**0.281**	0.074	0.251	0.194	0.034	0.172	0.045
edit3word	0.274	0.190	**0.172**	0.169	0.126	0.120	0.130	0.041
edit4	0.367	0.267	0.085	**0.266**	0.202	0.054	0.171	0.047
edit4ngram	0.369	0.276	0.087	0.264	**0.205**	0.052	0.176	0.052
edit4word	0.283	0.198	0.171	0.189	0.133	**0.132**	0.127	0.031
keystrokes	0.322	0.235	0.089	0.201	0.147	0.050	**0.189**	0.035
ngp	0.247	0.178	0.020	0.151	0.092	-0.038	0.087	**0.087**

woordgebaseerde 4-bewerkingredigeersoortgelykheid wat dikwels dien as 'n basislyn waarop maklik verbeter word, kon steeds die ander klop wanneer dit onder homself geëvalueer is.

Die spesifieke F_1-tellings wat verkry is, kom op sigself as redelik betekenisloos voor, aangesien die omvang van die waardes vir dieselfde stelselafvoer substansieel varieer afhangend van die soortgelykheidsmaat wat gedurende evaluasie gebruik is. Dit is daarom nie sinvol om enigiets in besonder in die spesifieke waardes te lees nie, veral wanneer resultate vergelyk word wat onder evaluasie-eksperimente met verskillende soortgelykheidsmate verkry is. Alhoewel die gebalanseerde F_1-telling gebruik is in alle resultate wat bo bespreek is, word dieselfde sydigheid waargeneem wanneer F_β-tellings gebruik is met $\beta = 0,5$ ('n voorkeur vir geweegde presisie) en $\beta = 2$ ('n voorkeur vir geweegde ware effektiwiteit).

Baldwin [7] het kommer oor so 'n sydigheid vlugtig genoem, en het gekies om die gemiddelde akkuraatheidstelling van afsonderlike evaluasies met twee soortgelykheidsmate te gebruik (variasies op edit3ngram en WSC) in 'n poging om sydigheid te verwyder. Dit is nie duidelik of daar 'n wetenskaplike basis is om te glo dat so 'n gemiddeld substansieel minder sydig

is, en of dit bloot die sydigheid van die samestellende dele in gelyke mate bevat nie. As die resultate van hierdie hoofstuk in gedagte gehou word, bring dit die vraag ter sprake of die goeie prestasie van die karaktergebaseerde mate, vergeleke met woordgebaseerde mate, in daardie studie veroorsaak is deur 'n evaluasiemetode wat twee karaktergebaseerde soortgelykheidsmate gebruik het.

Die resultate van hierdie hoofstuk bevestig die vermoedens wat uitgespreek is in 'n studie waar 'n evaluasiemaat gebaseer op TER gebruik is, en bevind is dat die mate gebaseer op TER dikwels beter as die ander presteer [82].

'n Soortgelyke sydigheid is waargeneem toe evaluasiemate vir masjienvertaling gebruik is vir beide die onttrekking en evaluasie in 'n vertaalgeheuestelsel [71], alhoewel so 'n eksperiment nie direk vergelykbaar is met ons werk hier nie (sien afdeling 3.1.3 vir 'n verduideliking). 'n Verwante saak is waargeneem toe 'n masjienvertaalstelsel geoptimeer is vir verskillende mate tydens minimumfouttempo-opleiding* [70]. Alhoewel hulle bevinding moontlik as intuïtief bestempel kan word aangesien die verfyning die geëvalueerde afvoer direk affekteer, is die verrassende resultaat in ons geval dat die sydigheid van die metode wat tydens onttrekking gebruik word oor die taalgrens oorgedra word vanaf die brontaal na die doeltaal waar die evaluasie plaasvind.

(en) *minimum error-rate training, MERT*

3.6 GEVOLGTREKKING

'n Metode is voorgestel vir die outomatiese evaluasie van vertaalgeheuestelsels gebaseer op die sterk punte van twee vorige benaderings [7, 84]. In hierdie metode gebruik ons verskeie soortgelykheidsmate om voorstelle te onttrek, en gebruik dieselfde soortgelykheidsmate weer tydens evaluasie (in alle kombinasies). Die metode bied beide 'n presisieagtige en 'n herroepingagtige mate, 'n manier om minder nuttige voorstelle te penaliseer, en k-voudige kruisvalidasie verminder probleme weens oormatige passing* en plaaslike artefakte in die data.

(en) *overfitting*

Saam met hierdie metode is 'n belangrike bydrae van hierdie hoofstuk om die probleem aan te dui waar evaluasiemetodes stringsoortgelykheidsmate gebruik om die prestasie van dieselfde mate te evalueer as tegnieke vir passing en rangbepaling van vertaalgeheuevoorstelle tydens onttrekking.

Alle moontlike kombinasies van 'n versameling van agt soortgelykheidsmate wat in vorige literatuur genoem is, is oorweeg. Die mate is divers in hoe hulle verskillende verskynsels hanteer, asook die vlak van granulariteit wat hulle oorweeg (karakter teenoor karakter-2-gram teenoor woord). Twee uiteenlopende datastelle is gebruik, in twee taalpare.

Daar is aangevoer dat dit belangrik is om die soortgelykheidsdrempel te modelleer as deel van die evaluasie-eksperiment om die manier waarop programme vir rekenaargesteunde vertaling werk, na te boots. Daar is bevind dat dit noodsaaklik is om nie 'n vaste soortgelykheidsdrempel te gebruik oor die hele eksperiment nie, aangesien dit sekere mate bo ander kan bevoordeel — elke maat moet by die punt geëvalueer word waar sy prestasie gemaksimeer word.

Die belangrikste ontdekking in hierdie hoofstuk is die sydigheid wat elke soortgelykheidsmaat tydens evaluasie vir homself wys. Hierdie ontdekking dui aan dat, sonder verdere bewyse, geen soortgelykheidsmaat inherent meer regverdig as deel van 'n evaluasiemetode is as ander nie. Dit beteken nie dat hierdie mate nie nuttig is nie. Inteendeel, die nuttigheid van verskeie van hierdie mate is reeds in bestaande produkte en literatuur bewys. In die soektog na die beste manier om vertaalgeheuestelsels te evalueer, beteken ons resultaat nie dat die beste maat nie 'n voorkeur vir homself sal hê nie. Dit is inderdaad moontlik dat een van die mate in hierdie hoofstuk die beste prestasie vir sekere toepassings lewer, en dat sy sydigheid 'n gewenste eienskap is in daardie toestande. Dit is egter nodig om ondersoek in te stel na watter mate die beste ooreenstem met objektiewe doelwitte wat nie gedefinieer word in terme van die einste soortgelykheidsmate nie. Dit is dan juis wat volgende in hoofstuk 4 ondersoek word.

4

OBJEKTIEWE WAARDE VAN SOORTGELYKHEIDSMATE

Die vorige hoofstuk het aangedui dat outomatiese evaluasie van vertaalgeheuestelsels nie sonder meer kan plaasvind sonder sorgvuldige oorweging van die soortgelykheidsmaat wat dien as evaluasiemaat nie. Dit laat ons egter steeds met die vraag van hoe 'n datastel, of enige ingrepe in die werking van 'n vertaalgeheuestelsel, geëvalueer kan word. Alhoewel een of ander vorm van menslike evaluasie 'n moontlikheid is en sekerlik as betroubaar beskou kan word, is dit noodwendig tydrowend en duur, en speel outomatiese evaluasie dus 'n belangrike rol (soos vroeër genoem). 'n Benadering is dus nodig waarvolgens 'n vertaalgeheuestelsel op een of ander manier outomaties geëvalueer kan word, met inagneming van die soortgelykheidsmate se sydigheid wat tot selfvoorkeur lei.

Dit is onrealisties om te verwag dat 'n outomatiese evaluasiemetode foutloos moet wees. 'n Kykie na die stand van sake in die outomatiese evaluasie van masjienvertaling gee die indruk dat selfs in hierdie aktiewe navorsingsveld die evaluasiemetodes as nuttige metodes beskou word, maar met gebreke.[1] 'n Vertaalgeheuestelsel bestaan ook (in 'n mindere of meerdere mate) binne die opset van subjektiewe vertaalwerk. Dus kan verwag word dat 'n goeie evaluasiemetode steeds nie in elke geval met menslike oordeel ooreen sal stem nie. Die persoonlike voorkeure van 'n vertaler, die eiesoortigheid van die taalpaar wat oorweeg word, die spesifieke sagteware vir rekenaargesteunde vertaling wat gebruik word, en aspekte van die spesifieke domein van vertaalwerk kan alles 'n invloed hê op hoe die relatiewe waarde van voorstelle uit die vertaalgeheuestelsel geskat word.

1 Vir 'n bespreking van die jaarlikse werkswinkeltaak hieroor, sien [16].

Aangesien vertaalgeheuestelsels veral vir die wins aan produktiwiteit gebruik word, stel ons ondersoek in na die moontlikheid daarvan om die redigeertyd of vertaaltyd vir 'n spesifieke segment te skat. Die verwagting is dat 'n goeie voorstel uit die vertaalgeheuestelsel slegs effens (of geensins) geredigeer hoef te word, wat dus die tyd wat per segment nodig is, sal verminder. As die tyd wat 'n vertaler per segment nodig het redelik akkuraat geskat kan word met inagneming van die voorstel, kan die relatiewe waarde van voorstelle dus só geskat word en die vertaalgeheuestelsel se werking dus geëvalueer word. In werklikheid gaan dit nie in die eerste plek oor die skatting van 'n eksakte waarde vir tyd (in sekondes) nie, maar eerder oor die ordening van mededingende voorstelle volgens hul relatiewe waarde.

In hierdie hoofstuk word soortgelykheid oorweeg tussen verskillende weergawes van die doelteks, aangesien dit hier gaan oor die *evaluasie* van 'n vertaalgeheuestelsel. Dit gaan dus nie hier oor hoe voorstelle *onttrek* word uit 'n vertaalgeheue tydens die normale werking van 'n vertaalgeheuestelsel nie.

4.1 LINEÊRE REGRESSIE

(en) *regression*

Regressie* behels die wiskundige modellering van die verband tussen veranderlikes.[2] Daar word onderskei tussen die voorspellers en die responsveranderlike. 'n Algemene regressiemodel beskou die verband soos volg:

$$\text{respons} = \text{modelfunksie} + \text{stogastiese fout} \tag{4.1}$$

Alhoewel die modelfunksie in beginsel enige vorm kan aanneem, is die eenvoudige geval van 'n lineêre kombinasie van die voorspellers 'n algemene en goed bestudeerde vorm.

(en) *dependent variable*

Lineêre regressie modelleer 'n lineêre verband tussen veranderlikes — spesifiek 'n afhanklike veranderlike* Y (bo die "re-

2 Die agtergrond en notasie hier word verskaf aan die hand van onder andere [26] en terminologie soos in [77]. Slegs enkele aspekte van regressie word hier gedek wat later in die hoofstuk gebruik word.

spons" genoem) en een of meer onafhanklike veranderlikes* $X_1 ... X_p$. Die onafhanklike veranderlikes staan ook bekend as die toevoer, die verklarende veranderlikes of die voorspellers.[3] Dit kan gebruik word vir voorspelling van Y, gegewe nuwe X-waardes, en om die sterkte van die verband tussen Y en die onderskeie X-veranderlikes te kwantifiseer. Verder kan die model aandui watter proporsie variansie van Y deur die model verklaar word.

(en) *independent variables*

'n Lineêreregressiemodel beskou die verband tussen die veranderlikes soos volg:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_p X_p + \epsilon \qquad (4.2)$$

waar ϵ die stogastiese fout (of residu) in die model voorstel, d.w.s. die verskil tussen die waarde wat die model voorspel op grond van X en die ware Y (soos gemeet). β_0 is die afsnit*, d.w.s. die waarde van Y as alle X-waardes nul is. Die oplos van die model behels dus die oplos van die parameters $\beta_0 ... \beta_p$ onderhewig aan die minimering van die som van kwadrate van die foute $\sum_{i=0}^{n} \epsilon_i^2$, gegewe n datapunte in die vorm $(Y, X_1, X_2, X_3, ... X_p)$.

(en) *intercept*

Wanneer daar verwys word na 'n lineêre model, word daar spesifiek verwys na die lineariteit in die koëffisiënte $(\beta_0 ... \beta_p)$—die parameters van die model. Die onafhanklike veranderlikes kan in enige verband tot mekaar staan, en dit is ook moontlik om onafhanklike veranderlikes in te sluit wat 'n nielineêre transformasie ondergaan het.

Indien die verskillende X-veranderlikes gestandaardiseer is sodat hulle 'n eenvormige gemiddeld en standaardafwyking het, vergemaklik dit ook die interpretasie van die β-waardes. 'n Positiewe waarde van β_j dui op 'n positiewe korrelasie tussen X_j en Y as die ander β-waardes konstant gehou word, en 'n negatiewe waarde van β_j dui op 'n negatiewe korrelasie tussen X_j en Y as die ander β-waardes konstant gehou word. 'n Groter absolute waarde vir β_j dui op 'n groter rol wat X_j speel

3 Met die term "onafhanklik" in hierdie konteks moet nie verstaan word dat dié waardes onverwant aan mekaar is nie. Dit is te verwagte dat sommige van die onafhanklike veranderlikes saam varieer.

in die skatting van Y, of 'n sterker verband tussen X_j en Y. Dit is moontlik om 'n vertrouensinterval* vir elke β-waarde te bereken, onder die aanname dat die residue uit 'n normaalverdeling* kom. Indien dié interval nul bevat, beteken dit dat die betrokke X moontlik met 'n nulfaktor in die vergelyking dien, en die nulhipotese (dat dié X geen invloed het op Y nie) kan dus nie verwerp word nie. Betekenisvolle resultate sal dus slegs afgelei kan word as die vertrouensinterval nie nul bevat nie.

(en) *confidence interval*

(en) *normal distribution*

Die *bepaaldheidskoëffisiënt** (R^2) dui die proporsie variansie van Y aan wat deur 'n regressiemodel verklaar word. 'n Hoë waarde van R^2 dui dus op 'n meer akkurate skatting van Y en 'n model wat die gedrag van Y goed modelleer.

(en) *coefficient of determination*

Aangesien 'n β_j-waarde van nul die effek van die veranderlike X_j in die model neutraliseer, kan die byvoeg van addisionele onafhanklike veranderlikes nooit die R^2 laat verlaag nie, omdat die model eerder 'n β_j-waarde van nul sal kies sodat die R^2-waarde onveranderd bly. R^2 het dus 'n nie-dalende verband met 'n toename in die aantal onafhanklike veranderlikes. Aangesien selfs die byvoeging van geheel onverwante veranderlikes nie die kwaliteit van die model kan verswak nie, word die *aangepaste bepaaldheidskoëffisiënt* eerder gebruik wat 'n model penaliseer vir 'n toename in die aantal onafhanklike veranderlikes.

In die volgende afdelings van hierdie hoofstuk word lineêre regressie gebruik om ondersoek in te stel na die verband tussen verskillende veranderlikes. Soortgelykheidsmate, asook hulle onderliggende soortgelykheidsafstande, word gebruik as onafhanklike veranderlikes waarmee die tyd per segment as afhanklike veranderlike gemodelleer word. Hiermee word gepoog om twee vrae te beantwoord:

- Watter soortgelykheidsmaat of -mate het die sterkste verband met tyd?

- Kan die soortgelykheidsmaat of -mate die variansie in tyd voldoende verklaar?

In ’n regressiemodel met al die soortgelykheidsmate as onafhanklike veranderlikes kan die β-waardes aandui watter soortgelykheidsmate die sterkste verband het met tyd. Die bepaaldheidskoëffisiënt kan aandui watter proporsie van die variansie in tyd deur die soortgelykheidsmate verklaar word, en dus ’n indruk gee van of die soortgelykheidsmate genoeg inligting gee om te voorspel hoe lank ’n vertaler met ’n segment gaan besig wees.

Die datastel wat nodig is om hierdie vrae mee te beantwoord moet die volgende inligting verskaf vir elke segment:

- Die doelteks van die voorstel (indien enige) wat aan die vertaler verskaf is.
- Die finale vertaling (doelteks) waarmee die vertaler volstaan het.
- Die tyd wat die vertaler vertoef het.

Die nodige onafhanklike veranderlikes kan almal vanaf die twee weergawes van die doelteks bereken word.

Weens die groot kommersiële en akademiese belangstelling in masjienvertaling is daar al heelwat werk gedoen om die redigering van masjienvertaalafvoer te ondersoek. Heelwat datastelle is beskikbaar. Sien byvoorbeeld die Translation Process Research Database [17]. As ons slegs belangstel in die redigeergedrag van vertalers, en daar word aangeneem dat die redigering van voorstelle van masjienvertaling en uit ’n vertaalgeheue soortgelyk is, kan hierdie datastelle gebruik word. Daar bestaan dus datastelle wat die bostaande inligting bevat, maar met een tekortkoming: die voorstelle is die afvoer van ’n masjienvertaalstelsel, nie uit ’n vertaalgeheue nie. Ons moet dus daaroor besin of dié aspek ’n beperking op die interpreteerbaarheid van die resultate sal plaas in die konteks van vertaalgeheuestelsels.

Een aspek van die datastelle blyk veral problematies te wees: onder die vele parameters wat masjienvertaalstelsels in ag neem tydens hulle werking, poog hul om afvoer van die korrekte lengte te lewer. In statistiese masjienvertaalstelsels kan

'n parameter ω per woord bepaal in watter mate die stelsel 'n voorkeur het vir korter of langer afvoer [50, p. 140]. Tydens parameteroptimering* kan 'n evaluasiemaat afvoer met die verkeerde lengte penaliseer sodat parameters (soos ω) konvergeer na waardes waarby die afvoer meer waarskynlik die regte lengte beslaan. BLEU [66], 'n gewilde maat vir die evaluasie van masjienvertaalstelsels, penaliseer afvoer wat te lank is met 'n eksplisiete term hiervoor in die formulering. Stelsels wat met BLEU of soortgelyke mate geoptimeer word, sal dus meerendeels afvoer van die korrekte lengte genereer.

(en) *tuning*

Sekere soortgelykheidsmate het verskillende beskouinge van bewerkings soos invoeging, skrapping en vervanging. Twee redigeerafstande, naamlik 3-bewerkingredigeerafstand en 4-bewerkingredigeerafstand, verskil juis ten opsigte van die bewerkings wat 'n verskil in lengte te weeg bring (invoeging en skrapping). 'n Datastel waar die vertalers meestal voorstelle van die regte lengte moes redigeer, mag dalk nie genoegsaam data verskaf om te onderskei tussen hierdie twee mate wat andersins dieselfde werk nie.

'n Studie na die effek wat voorstelle uit 'n vertaalgeheue op die vertaalproses het [69], het wel 'n datastel gegenereer wat die nodige inligting bevat. Dit word in meer besonderhede in die volgende afdeling beskryf.

4.2 AANVANGSDATA

Screen [69] stel ondersoek in na die effek van voorstelle met 'n soortgelykheid in die omvang 70%–95%.[4] Daar word gekyk of wasige voorstelle uit die vertaalgeheue die kognitiewe las tydens vertaling kan verlaag. Hiervoor is vertaalsessies in detail opgeneem waarin volledige tikgedrag ondersoek kan word.

Opsommende inligting is bekom vanaf die navorser,[5] waarin sewe professionele Walliese vertalers se hantering van elke ver-

4 Hierdie soortgelykheid is bereken volgens die vertaalprogram se eie (onbekende) soortgelykheidsmaat.

5 Dankie aan Benjamin Screen wat die data gedeel het.

taalsegment saamgevat is. Elke vertaler het dieselfde stel van 100 segmente hanteer, wat in drie substelle verdeel kan word:

- 50 segmente het voorstelle uit 'n vertaalgeheue. Die voorstelle is deur die vertaalprogram Déjà Vu X3 Professional[6] gegenereer.
- 25 segmente moes sonder enige voorstelle vertaal word.
- 25 segmente het masjienvertaalvoorstelle gehad wat deur Google Translate gegenereer is.

Vir elke segment van elke vertaler bevat die datastel die volgende elemente:

1. 'n Segmentnommer.
2. Die bronteks wat vertaal moes word.
3. Die voorstel (indien enige) vir die doelteks vir oorweging deur die vertaler.
4. Die finale doelteks waarmee die vertaler volstaan het.
5. Die redigeertyd gemeet in millisekondes.
6. Die oorsprong van die voorstel (vertaalgeheue of masjienvertaling).

Velde 3, 4 en 5 verskaf dus presies die nodige inligting vir die regressiemodel wat vroeër beskryf is. Met behulp van veld 6 kan die segmente met masjienvertaalvoorstelle uitgesluit word. As dié segmente uitgesluit word, bevat elke vertaler se datastel dus 75 segmente waarvan twee derdes die redigering van wasige voorstelle beskryf. Die datastel het die volgende in sy guns:

- Dit bevat presies die inligting wat nodig is.
- Dit is in redelik gekontroleerde omstandighede geskep.
- Elke deelnemer het dieselfde vertaal- en redigeertaak gehad. Die voorstelle was ook dieselfde in elke geval.

6 https://atril.com/product/deja-vu-x3-professional/

- Die deelnemers is redelik homogeen (al sewe is professionele vertalers vir die taalpaar Engels na Wallies).
- Dit bevat 'n redelike verdeling van segmentlengtes: 1–44 woorde in die segmente se bronteks, met 'n mediaan van 17.

Voordat die eksperiment gedoen word en die resultate ondersoek word, moet daar kennis geneem word dat die datastel sekere beperkinge het wat reeds duidelik is:

- 'n Datastel van 525 segmente behoort statisties betroubare resultate te gee, maar is klein in vergelyking met soortgelyke datastelle in die masjienvertaalgemeenskap. Vergelyk byvoorbeeld met datastelle in die Translation Process Research Database [17].
- Dit bevat net een taalpaar, dus los dit noodwendig vrae oor die veralgemeenbaarheid van die resultate.
- Die meeste van die kort segmente het geen redigeerwerk nodig gehad nie, dus is daar nie duidelikheid oor hoe goed die resultate sal veralgemeen na wasige voorstelle vir kort segmente waar redigering wel nodig is nie.

Soos hier bo beskryf, bevat die datastel nog nie die nodige onafhanklike veranderlikes om die regressiemodel op te stel nie, maar die datastel kan nou uitgebrei word.

4.3 VERRYKTE DATA

Vir die regressiemodel soos vroeër beskryf, is die waarde van elke soortgelykheidsmaat in elke segment nodig. 'n Addisionele veld word dus bygevoeg by die datastel vir elke soortgelykheidsmaat. Al die soortgelykheidsmate van hoofstuk 3 word ingesluit, asook twee wat in vroeëre werk gebruik is. Die maat **diceword** is in vorige werk geïdentifiseer as kompeterend met ander mate [7], en **terword** is in ander werk gebruik as verwysingsmaat [82]. Sien gerus afdeling 3.1 waar hierdie vorige werk reeds bespreek is.

Vir 525 segmente en 10 soortgelykheidsmate moet 525×10 waardes dus by die datastel gevoeg word. In die notasie wat bo gebruik is, is $n = 525$ en is die 10 nuwe velde die onafhanklike veranderlikes $X_1 ... X_p$ met $p = 10$.

Die soortgelykheidsmate word telkens per segment bereken tussen die voorstel (indien enige) se doelteks en die finale doelteks. In die geval waar geen voorstel gegee is nie (25 segmente per vertaler, 175 altesaam) word die finale doelteks met 'n leë string vergelyk (noodwendig 'n soortgelykheid van nul).

4.4 AANVANGSMODEL

As 'n eenvoudige aanvangseksperiment word al die onafhanklike veranderlikes wat bo beskryf is in 'n regressiemodel gekombineer. Dit sal reeds 'n aanduiding gee van hoeveel van die variansie in tyd verklaar kan word, en dit sal aandui watter van die soortgelykheidsmate die sterkste (lineêre) verband het met redigeertyd. Hierdie eksperiment dien as 'n verkennende ondersoek om te bepaal of die regressiebenadering en die data wat voorhande is, gebruik kan word soos bo genoem is.

Die regressie is geïmplementeer in Python met behulp van die StatsModels-pakket.[7] Die opsomming hier onder is die verbatim afvoer wat deur StatsModels gegenereer word om die berekende regressiemodel mee op te som. Hierdie tipe opsomming is algemeen in statistiese pakkette. Dit bevat drie dele:

- 'n opsomming van die regressiemodel in sy geheel, insluitend die sukses van die passing;
- volledige inligting oor die koëffisiënte van elke onafhanklike veranderlike wat telkens in die linkerkantste kolom genoem word, asook elke koëffisiënt se vertrouensinterval;
- 'n opsomming van die statistiese eienskappe van die residuverdeling.

7 http://www.statsmodels.org/

'n Volledige bespreking van elke element in die opsomming val buite die bestek van hierdie studie. Daar sal hoofsaaklik gefokus word op die statistieke wat in afdeling 4.1 genoem is. Ander belangrike aspekte van die opsomming, veral die eienskappe van die residuverdeling, is wel belangrik omdat dit die navorser help om te bepaal of die model betroubaar is.

Dep. Variable:	time	**R-squared:**	0.425
Model:	OLS	**Adj. R-squared:**	0.416
Method:	Least Squares	**F-statistic:**	50.84
Date:	Tue, 22 Nov 2016	**Prob (F-statistic):**	4.12e-76
Time:	17:17:26	**Log-Likelihood:**	-8105.9
No. Observations:	700	**AIC:**	1.623e+04
Df Residuals:	689	**BIC:**	1.628e+04
Df Model:	10		

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
Intercept	6.291e+04	1970.188	31.933	0.000	5.9e+04 6.68e+04
diceword	8.044e+04	8.39e+04	0.959	0.338	-8.42e+04 2.45e+05
edit3	6.603e+05	1.79e+05	3.693	0.000	3.09e+05 1.01e+06
edit3ngram	-3.403e+05	1.62e+05	-2.096	0.036	-6.59e+05 -2.15e+04
edit3word	-2.062e+05	1.41e+05	-1.465	0.143	-4.82e+05 7.01e+04
edit4	-7.441e+05	2.42e+05	-3.072	0.002	-1.22e+06 -2.69e+05
edit4ngram	5.571e+05	2.05e+05	2.721	0.007	1.55e+05 9.59e+05
edit4word	-1.532e+05	9.02e+04	-1.698	0.090	-3.3e+05 2.39e+04
keystrokes	-4.205e+04	3.13e+04	-1.342	0.180	-1.04e+05 1.95e+04
ngp	-6682.9205	1.76e+04	-0.379	0.705	-4.13e+04 2.79e+04
terword	1.421e+05	6.6e+04	2.152	0.032	1.24e+04 2.72e+05

Omnibus:	507.167	**Durbin-Watson:**	1.786
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	9307.196
Skew:	3.055	**Prob(JB):**	0.00
Kurtosis:	19.786	**Cond. No.**	1.00e+03

Hierdie aanvanklike poging is teleurstellend. Die lae aangepaste R^2-waarde van 0,416 (sien "Adj. R-squared") gee die indruk van 'n onvolledige model — minder as die helfte van die variansie in tydsduur per segment kan met hierdie model verklaar word. Die gebruik van al die veranderlikes soos wat

dit nou in die model gekombineer is, blyk dus onvoldoende te wees om die tydsduur per segment akkuraat te skat.

Die volgende voorlopige opmerkings kan wel reeds gemaak word:

- Die skeefheid* en kurtose* dui daarop dat die residuverdeling nie normaal verdeel is nie. Alhoewel geen enkele omvang vir hierdie waardes vir alle situasies gepas is nie, is beide ver buite die omvang wat verwag word vir 'n normaalverdeling (tipies skeefheid in $[-1, 1]$ en kurtose < 3). Volgens die Jarque-Bera-toets [43] kan die hipotese dat die residue uit 'n normaalverdeling kom, verwerp word. (Sien "Prob(JB)" = 0.) Dit dui daarop dat die interpretasie van aspekte van die model moeiliker gaan wees. 'n Visuele voorstelling van die residue word in figuur 4.1 getoon vir 'n enkelvoudige lineêreregressiemodel met slegs **edit3** as onafhanklike veranderlike (wat die grootste absolute t-waarde onder die onafhanklike veranderlikes gehad het in die aanvanklike passing bo).

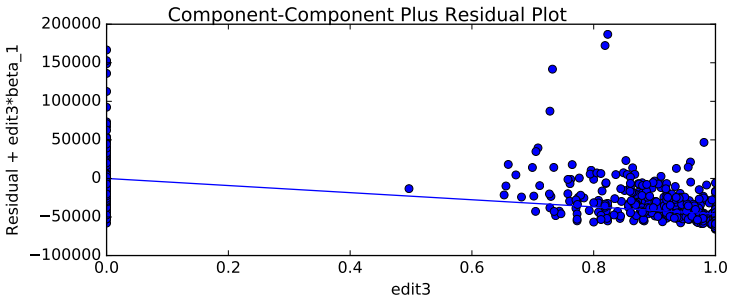
  (en) *skew*
  (en) *kurtosis*

- Die hoë geaardheidsgetal* (sien "Cond. No." ver bo 30 in die opsomming) dui op onderlinge kollineariteit,* dit wil sê, daar is hoë lineêre korrelasies tussen die veranderlikes. Kollineariteit tussen onafhanklike veranderlikes in 'n lineêreregressiemodel beteken dat die skattings vir β-waardes wisselvallig kan wees in reaksie op klein veranderinge in die model of data. Die β-waardes is van belang sodat bepaal kan word watter veranderlikes die sterkste verband het met tyd.

  (en) *condition number*
  (en) *collinearity*

- Verskeie van die onafhanklike veranderlikes se koëffisiente se 95%-vertrouensinterval bevat 0, dus bevat die beste model dalk slegs 'n subversameling van dié veranderlikes, alhoewel die model op hierdie stadium weens bogenoemde probleme nie as betroubaar geag word nie.

In die res van die hoofstuk word ondersoek ingestel na verskillende benaderings om—



Figuur 4.1: Die residue in enkelvoudige lineêre regressie met slegs **edit3** as onafhanklike veranderlike. Die groot aantal punte aan die linkerkant verteenwoordig die gevalle waar daar geen voorstel was nie en daar dus geen soortgelykheid is met die finale vertaling nie. Die edit3-waardes is andersins hoog, omdat die voorstelle wat wel gegee is van redelike hoë gehalte was (75% soortgelykheid en hoër).

- vas te stel of die data kan voldoen aan die verwagtinge van ’n lineêreregressiemodel;
- die kwaliteit van die data te verhoog;
- meer van die variansie in tyd te verklaar; en
- die betroubaarheid van die resultate te verhoog.

4.5 VERFYNDE MODEL

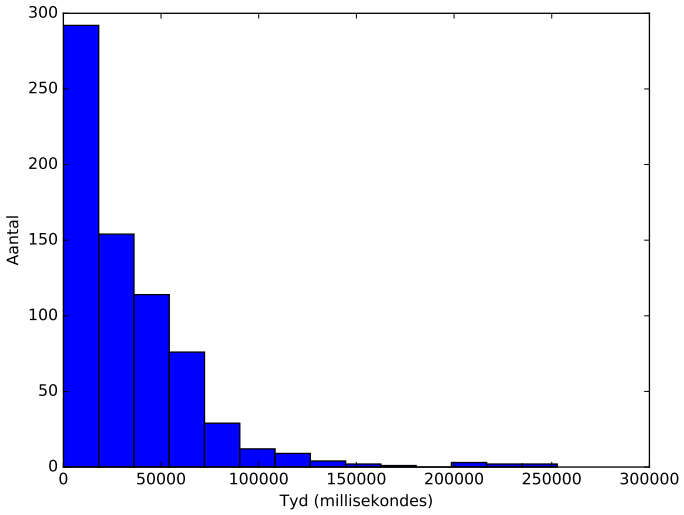
In hierdie afdeling word die regressiemodel verfyn. Ons aandag word spesifiek toegespits op veranderinge in die volgende verbeteringsareas:

- Daar is verbeterings vanuit oorwegings binne die probleemdomein, veral wat die soortgelykheidsmate betref, wat beter passing in die model moontlik maak.
- Verbeterings vanuit die statistiek. Hierdie verbeterings kan die passing van die model verbeter en die betroubaarheid van die resultate verhoog.
- Oorwegings wat die eksperimentele opset betref. Dit wil sê, deur in ag te neem hoe die vertaalproses geskied, word dit moontlik om ’n groter deel van die variansie in tyd te verklaar.

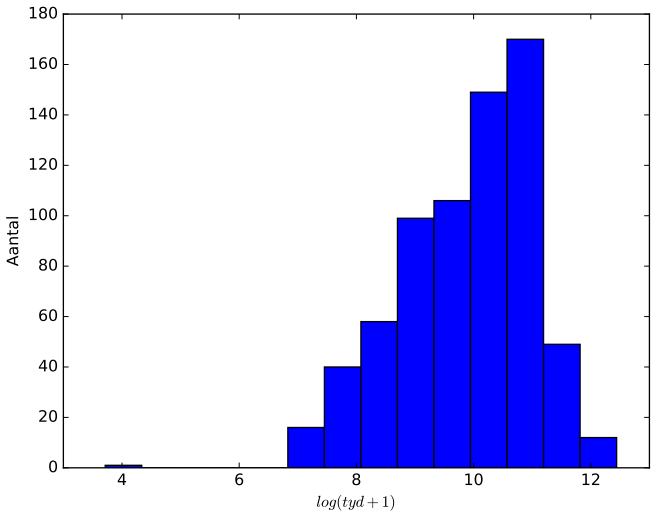
Daar is ’n verskeidenheid eienskappe van die data wat tot dusver gebruik is wat nie ideaal is vir ’n lineêreregressiemodel nie. Die skeefheid en kurtose is vroeër genoem, maar hier word nou ook ’n paar verdere aspekte bespreek.

Eerstens is die afhanklike veranderlike hoegenaamd nie normaal verdeel nie. Figuur 4.2 wys op die verdeling van dié veranderlike. Soos tipies van sinslengtes en woordlengtes in korpusse van natuurlike taal, volg die segmentlengtes en redigeertye in hierdie datastel klaarblyklik ’n Zipf-verdeling [57]. Een gevolg hiervan is heteroskedastisiteit* — die veranderlikheid in tyd is nie konsekwent in subpopulasies nie. Dit is te

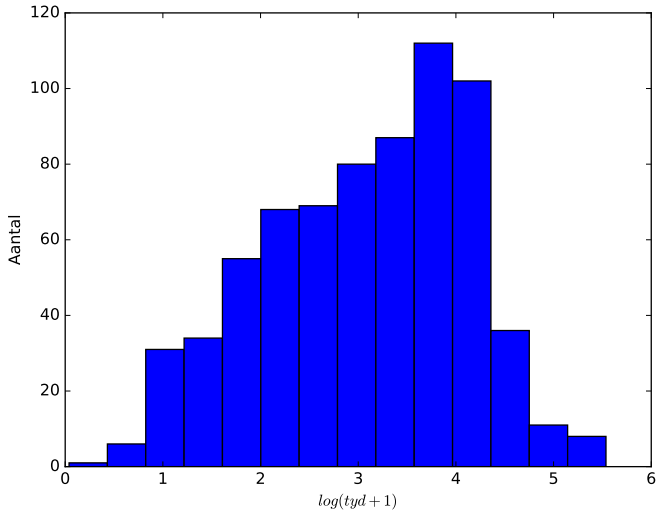
(en) *heteroscedasticity*



Figuur 4.2: Histogram van tyd



Figuur 4.3: Histogram van $\log(\text{tyd} + 1)$ met tyd in millisekondes



Figuur 4.4: Histogram van $\log(\text{tyd} + 1)$ met tyd in sekondes

verwagte dat skattings al hoe minder akkuraat sal wees vir toenemend langer segmente.

'n Transformasie van die data is 'n tipiese oplossing in regressieleer. Moontlike transformasies sluit die volgende in: magsverheffing, vierkantswortel, log en Box-Cox [26, hoofstuk 13]. In plaas van die oorspronklike veranderlikes kan die getransformeerde waardes dan gebruik word in 'n gewone lineêre model soos in vergelyking 4.2 op bladsy 69.

'n Verdeling soos geïllustreer in figuur 4.2 regverdig tipies die gebruik van 'n log-transformasie. 'n Histogram vir die transformasie van die tyd met die log-funksie word gewys in figuur 4.3. 'n Verdere histogram met die transformasie waar die tyd eers na sekondes omgeskakel is, word gewys in figuur 4.4. In al twee gevalle word die konstante 1 by die veranderlike gevoeg om te verseker dat daar nie gepoog word om 'n waarde van nul met die log-funksie te transformeer nie.

Daar is wel nadele aan so 'n transformasie: dit kompliseer die interpretasie van die model omdat die veranderlike nie meer in dieselfde eenheid gemeet word nie. Dié probleem is nie so ernstig in hierdie geval nie, omdat die model nie spesifiek vir voorspelling van tyd (in sekondes) gebruik gaan word nie.

Tabel 4.1: Die paarsgewyse lineêre korrelasies tussen die verskeie afstandmate (nie soortgelykheidsmate nie).

	edit3	edit3ngram	edit3word	edit4	edit4ngram	edit4word	keystrokes	ngp	terword
diceword	.82	.88	.97	.76	.80	.92	.66	.88	.94
edit3		.98	.88	.98	.99	.93	.78	.81	.91
edit3ngram			.93	.94	.97	.95	.82	.87	.93
edit3word				.83	.87	.97	.75	.93	.96
edit4					.99	.91	.72	.76	.88
edit4ngram						.93	.78	.80	.90
edit4word							.72	.88	.98
keystrokes								.74	.66
ngp									.88

Laastens is daar nog 'n nie-ideale eienskap van die onafhanklike veranderlikes: daar is onderlinge kollineariteit. Tabel 4.1 dui al dié paargewyse korrelasies aan. Die korrelasies in die tabel is bereken met die afstandmate eerder as die soortgelykheidsmate, en verder is al die nulwaardes verwyder asook die waardes waar daar geen voorstel was nie. Sonder hierdie aanpassings is die korrelasies nog hoër. Die hoë korrelasies is nie verbasend nie. Ons weet immers dat al die soortgelykheidsmate min of meer dieselfde fenomeen modelleer — weliswaar elk op 'n ander manier. Om die waarheid te sê, kollineariteit is eintlik te verwagte. Die kollineariteit dui daarop dat gesamentlike modelle met al die onafhanklike veranderlikes, soos hier bo, nie die gepaste instrument is nie. 'n Gepaste subversameling van die veranderlikes sal geïdentifiseer moet word. Voortaan word elke maat dus afsonderlik in sy eie model geëvalueer. Daar sal later oorweeg word om meer as een veranderlike in te sluit.

(en) *outliers*

In regressiemetodes kan uitskieters* nadelige gevolge op die model hê. Aangesien elke datapunt vir die model in ag geneem word, kan datapunte van lae kwaliteit (bv. verkeerde metings, ongemodelleerde inmenging) 'n effek op die koëffisiënte van die model hê sonder dat dit die passing verbeter. Twee moont-

like oplossings hiervoor is (1) om uitskieters te identifiseer en te verwyder uit die datastel of (2) om robuuste regressiemetodes te gebruik. In werklikheid is dit realisties om te antisipeer dat 'n vertaler met enkele segmente buitengewoon lank sal neem. Dit is dalk nodig om 'n term in 'n woordeboek na te slaan wat lank neem relatief tot die vertaling van 'n tipiese segment. As 'n vertaler se aandag om die een of ander rede afgelei word (bv. weens 'n steurnis), sal dit die tyd vir die geaffekteerde segment beïnvloed sonder dat 'n soortgelykheidsmaat dit redelikerwys sal kan modelleer. Ons volg dus benadering (1) en verwyder die datapunte waarvan die gestudentiseerde residue* se aangepaste $p < 0,05$ volgens die Bonferroni-aanpassing — StatsModels se verstekmetode vir die identifisering van uitskieters.

(en) *studentised residuals*

Kom ons oorweeg verder die betekenis van die waardes in die model. 'n Soortgelykheidsmaat lewer 'n waarde tussen nul en een (sien p. 37). 'n Soortgelykheid van 0,9 dui dus op twee segmente wat redelik dieselfde is. Dit is egter te verstane dat die tyd wat dit gaan neem om die redigeerwerk te doen van die lengtes van die twee segmente sal afhang. Twee kort segmente met 'n soortgelykheid van 0,9 benodig dalk enkele sekondes se redigeerwerk, terwyl twee lang, komplekse segmente met dieselfde soortgelykheid van 0,9 langer sal neem om te redigeer. Dit sal langer neem om twee redes: (1) die werklike redigeerwerk kan meer wees en gevolglik langer neem, en (2) dit sal langer neem om die voorstel na te gaan en die plekke vir redigering te identifiseer en die finale vertaling te kontroleer. Die tweede aspek lei tot 'n volgende waarneming: selfs in gevalle waar 'n voorstel sonder redigering aanvaar word, kan ons aanneem dat die vertaler die bronteks en die voorgestelde doelteks moet lees.

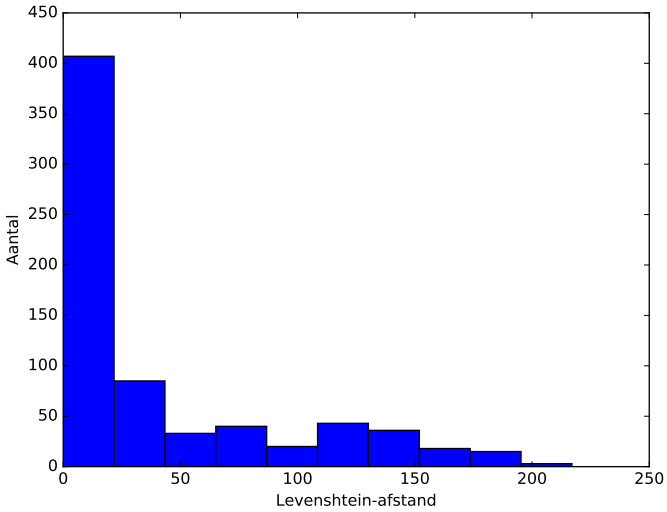
Hierdie waarnemings dwing ons om twee veranderinge aan die model te maak. Eerstens gaan 'n soortgelykheidsmaat nie meer gebruik word nie, maar eerder die onderliggende afstandmaat. Vir die soortgelykheid gebaseer op die 4-bewerkingredigeerafstand, word die redigeerafstand sonder normalisering dus eerder gebruik. Dié veranderlike sal dus 'n heelgetal wees

in die omvang $[0, \infty)$ waar 'n afstand van nul ooreenstem met 'n soortgelykheid van 1 (geen verskil). Dit behels dat in die voorbeeld bo van soortgelykheid van 0,9 duidelik onderskei sal kan word in die geval van kort segmente teenoor lang segmente. Tweedens moet die lengte van die segmente in ag geneem word in die model ongeag die soortgelykheid tussen die voorstel en die finale vertaling. Ons verryk dus die data met 'n addisionele veld vir elke afstandmaat, en verder ook nog die lengte van die finale doelsegment.

Alhoewel dit aanloklik lyk om op hierdie stadium ook die lengte van die bronteks in die model te oorweeg, is daar 'n baie hoë lineêre korrelasie tussen die lengtes van die bron- en doelteks — meer as 0,95 — dus sal dit nie 'n verskil maak aan die proporsie van die tydvariansie wat verklaar kan word nie.[8] Daar moet wel op hierdie stadium oorweeg word hoe die lengte van 'n segment bepaal moet word. In die vertaalindustrie word die grootte van 'n vertaaltaak tipies in woorde aangedui, maar vir die lengte van 'n segment lyk dit aantreklik om eerder die aantal karakters te gebruik. Só kan mens byvoorbeeld onderskei tussen een lang woord en een kort woord. Dit maak intuïtief sin dat dit langer sal neem om 'n lang woord te tik as 'n kort woord en dat lengte in karakters die beste keuse is vir die model.[8]

Hiermee word daar dus erkenning gegee daaraan dat die afhanklike veranderlike eintlik twee ineengeweefde hoeveelhede bevat, naamlik leestyd en redigeertyd. Vertalers gebruik verskillende vertaalbenaderings, en dit kan nie in die algemeen aanvaar word dat vertalers hierdie twee aspekte (lees en redigeer) los van mekaar doen nie. Dit kan dus nie as twee aparte veranderlikes gemodelleer word nie, en sal nie sonder meer eksperimenteel as twee veranderlikes gemeet kan word nie. In 'n studie wat die effek van vertaalgeheues op redigeertyd en sleuteldrukke evalueer [35], is ook gevind dat beter voorstelle nie hierdie twee aspekte ewe veel verminder nie. Die verduideliking wat aangebied is, strook ook met wat pas ge-

8 Dit is ook eksperimenteel bevestig. Die resultate word nie hier ingesluit nie.

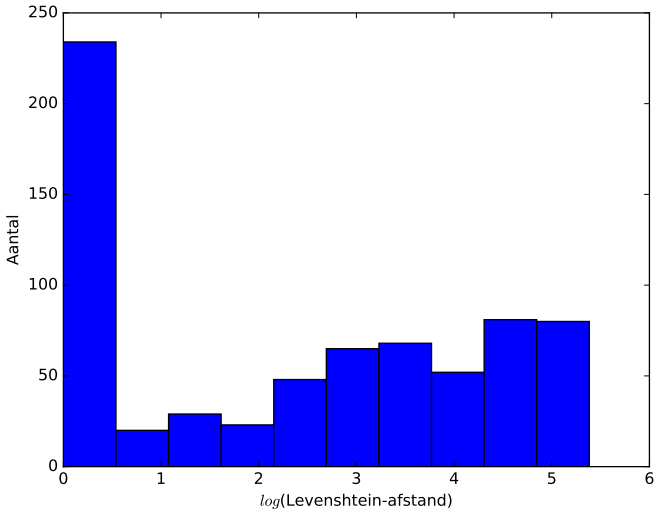


Figuur 4.5: Histogram van Levenshtein-afstand (dist_edit4)

noem is: die sleuteldrukke verteenwoordig redigering, maar die tydveranderlike bevat komponente vir lees en redigering. As aangeneem word dat die voorstelle uit die vertaalgeheue slegs die redigering werklik versnel, maak dit sin dat ons by sleuteldrukke persentasiegewys 'n groter afname sal sien as by die totale tyd, aangesien slegs een komponent van die tyd (die redigeertyd) afneem, terwyl die leestyd min of meer onveranderd bly.

Net soos vroeër met die onafhanklike veranderlike, tyd, word al die afstandmate met die log-funksie getransformeer om die verdeling se vorm te verbeter. Figuur 4.5 wys die verdeling van 4-bewerkingredigeerafstand, en die getransformeerde waarde se verdeling is in figuur 4.6.

Soos algemeen in masjienleer, wil ons verseker dat die resultate betroubaar is en redelikerwys sal veralgemeen buite die opleidingsdata. Ons modelleer dus verder die probleem met slegs ses van die vertalers se data, en evalueer dit op die oorblywende datastel van die sewende vertaler. Dit is te verwagte dat die resultate slegter sal wees. Eerstens omdat minder opleidingsdata gebruik word, en tweedens omdat die model op voorheen ongesiene data geëvalueer word. Dit gee egter 'n



Figuur 4.6: Histogram van $\log(\texttt{dist\_edit4} + 1)$

meer realistiese indruk van hoe goed die model werklik is. Met die sewe afsonderlike (maar vergelykbare) datastelle, doen ons ook sewevoudige kruisvalidasie om die betroubaarste resultate te kry.

Dit is te verwagte dat vertalers nie dieselfde werk nie, en dat party vertalers vinniger as ander sal werk (weens talent, sorgvuldigheid, ondervinding, ens.). As een vertaler konsekwent 10% vinniger werk as 'n ander een sal hierdie verskil in redigeertyd die model dwing om deels vir al twee datastelle te kompenseer. Alhoewel die twee vertalers se gedrag dieselfde is sal die model dit nie akkuraat modelleer nie, aangesien daar vir elke datapunt van die onafhanklike veranderlikes twee waardes van Y (tyd) gaan wees waarvan een 10% van die ander verskil. Dit sal die indruk gee van 'n model wat nie een van die twee vertalers se tyd akkuraat skat nie, alhoewel die twee vertalers se datapunte eintlik die model behoort te versterk om akkuraat al twee se gedrag te modelleer.

Die spoed tussen die vertalers word dus genormaliseer sodat hulle redigeertyd meer vergelykbaar is, soos ook gedoen in [35]. Die redigeertyd van elke segment word só aangepas dat dit steeds dieselfde breukdeel van die vertaler se sessie ver-

teenwoordig as vantevore. Die tydwaardes in die toetsstel word ook genormaliseer bloot om die interpreteerbaarheid van die resultate te verhoog. Ons stel immers nie daarin belang om 'n werklike tydskatting (in sekondes) te maak nie, maar bloot om te bepaal watter maat die sterkste verband het met tyd. In 'n lineêre regressie met lineêre komponente sal só 'n transformasie geen effek hê nie, aangesien dit bloot die betrokke β-waardes sal beïnvloed. In ons geval hier het dit wel 'n effek omdat daar 'n nielineêre transformasie op die data toepas word ná hierdie normalisering van redigeertyd.

Daar is nou 'n aantal wysigings op die oorspronklike regressiemodel beskryf. Hier volg nou 'n bondige opsomming, deels ook om herhaling van die eksperiment te vergemaklik:

- Die data word verryk met die lengte van die finale vertaling (in karakters), asook alle nodige soortgelykheidsmate en afstandmate.

- Die tydsduur word omgeskakel na sekondes (sonder afronding).

- Sewe modelle word opgestel telkens met ses van die vertalers se data as opleidingsdata en die laaste stel slegs vir evaluasie. In elkeen van die sewe modelle, gebeur die volgende:

  - Die tydwaardes van die ses vertalers in die opleidingsdata word genormaliseer om inherente verskille tussen die deelnemers in 'n mate te neutraliseer.

  - 'n Aanvanklike lineêre model word gebou met die finale doelteks se lengte (in karakters) en 'n afstandmaat. Die afhanklike veranderlike en al twee onafhanklike veranderlikes word met die $\log$-funksie getransformeer en gestandaardiseer.

  - Hierdie aanvanklike model word gebruik om uitskieters te identifiseer wat dan verwyder word uit die opleidingsdata.

Tabel 4.2: Die gemiddelde R^2 deur sewevoudige kruisvalidasie bepaal in elke model.

Afstandmaat	gemiddelde R^2
diceword	0,770
edit3	0,760
edit3ngram	0,764
edit3word	**0,771**
edit4	0,753
edit4ngram	0,758
edit4word	0,757
keystrokes	0,742
ngp	0,738
terword	0,714

  - Die tydwaardes in die datastelle vir elke deelnemer in die opleidingsdata word opnuut genormaliseer, en die toetsstel se tyd word ook aangepas.
  - Die model word op die toetsstel geëvalueer en R^2 word bepaal.

- Die gemiddelde R^2 van al 7 modelle word bereken. Die resultate wat in tabel 4.2 genoem word, is dus hierdie gemiddeld van die sewe modelle.

4.6 RESULTATE

Die resultate van die modelle vir elkeen van die afstandmate word aangetoon in tabel 4.2. Die beste resultate is vir die afstandmate diceword en edit3word. Dit wil ook voorkom asof die mate gebaseer op die 3-bewerkingredigeerafstand oor die algemeen beter vaar. Afgesien daarvan dat die beste twee mate woordgebaseer is, blyk daar geen duidelike voorkeur vir woordgebaseerde mate te wees teenoor dié wat op karakters of n-gramme gebaseer is nie. Die tellings vir die 4-bewerkingredigeerafstand oor woord-, n-gram- en karaktervlak is vergelykbaar.

As die model op al 7 vertalers se datastelle opgelei word en die R^2 binne-in die model bereken word (eerder as op toetsdata wat apart gehou is), vaar dié twee afstandmate ook die beste. Die afstandmaat wat met diceword ooreenstem, **dist_diceword**, vaar beste in hierdie opset. Die regressieopsomming vir die model met dist_diceword en die finale doelsegment se lengte, **tgt_ref_len**, word hier gegee:

Dep. Variable:	time	**R-squared:**	0.851
Model:	OLS	**Adj. R-squared:**	0.851
Method:	Least Squares	**F-statistic:**	1416.
Date:	Tue, 29 Nov 2016	**Prob (F-statistic):**	1.68e-205
Time:	09:21:39	**Log-Likelihood:**	-232.26
No. Observations:	498	**AIC:**	470.5
Df Residuals:	495	**BIC:**	483.2
Df Model:	2		

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
Intercept	1.6e-15	0.017	9.23e-14	1.000	-0.034 0.034
tgt_ref_len	0.3564	0.021	16.972	0.000	0.315 0.398
dist_diceword	0.6734	0.021	32.071	0.000	0.632 0.715

Omnibus:	3.467	**Durbin-Watson:**	1.632
Prob(Omnibus):	0.177	**Jarque-Bera (JB):**	3.309
Skew:	0.151	**Prob(JB):**	0.191
Kurtosis:	3.260	**Cond. No.**	1.89

Volgens die opsomming is die verdeling van die residue aansienlik beter as in die aanvangsmodel. Die skeefheid is laag (naby aan nul) en die kurtose is naby aan 3—albei dus redelike waardes vir 'n normaalverdeling. Volgens die Jarque-Bera-toets is dit moontlik dat die residue normaal verdeel is. Daar is ook 'n laer geaardheidsgetal (sien "Cond. No." in die opsomming) wat beteken dat die resultate minder sensitief is vir klein veranderinge. Ons kan dus 'n hoë mate van vertroue hê in die vertrouensintervalle.

Volgens die model sluit die 95%-vertrouensinterval van al twee onafhanklike veranderlikes nie nul in nie. Die afsnit is

onbelangrik gemaak deurdat die afhanklike veranderlikes gestandaardiseer is ná die log-transformasie.

Die waarde van R^2 kan effens verhoog word deur meer as een afstandmaat as onafhanklike veranderlike te gebruik, maar die verbetering is gering. In die geval waar 'n addisionele onafhanklike veranderlike by die model ingesluit word, moet daar ook eerder na die aangepaste R^2-waardes gekyk word. Hiervolgens is daar nie 'n verbetering met 'n addisionele afstandmaat nie. Aangesien ons weet dat verskeie afstandmate hoë lineêre korrelasie met mekaar het, (sien tabel 4.1) is dit nie verrassend nie en besluit ons dus eerder om slegs een afstandmaat in te sluit in die finale model (afgesien van die lengte in karakters).

4.7 GEVOLGTREKKING

Volgens die eksperimente hier bo kan meer as 75% van die vertalers se variansie in tyd verklaar word aan die hand van twee veranderlikes: die finale lengte van die vertaling en een van die afstandmate. Die 95%-vertrouensintervalle en regressiekoëffisiente dui daarop dat al twee veranderlikes met hoë waarskynlikheid 'n rol speel en 'n sterk verband het met tyd. Veral **diceword** en **edit3word** lewer goeie passings, alhoewel 'n hele paar van die ander mate se R^2 binne enkele persentasiepunte hiervan kom. Die soortgelyke resultate vir **diceword** en **edit3word** mag dalk aanvanklik verrassend wees — **diceword** ignoreer woordvolgorde heeltemal. Baldwin [7] het reeds aangevoer dat mate sonder ordebewustheid oorweeg moet word omdat hulle vinniger bereken kan word en in sy studie vergelykbare resultate gelewer het. Vir die datastel van hierdie hoofstuk is die verklaring eenvoudig: dié twee afstandmate het 'n baie hoë korrelasie van meer as 0,99. Dit wil sê dat hulle waardes in 'n sterk lineêre verband met mekaar staan. In hierdie datastel is sowat 'n derde van die inskrywings glad nie geredigeer nie en het dus 'n afstand van nul volgens al twee mate. Dit gee dalk 'n effens oordrewe indruk van die korrelasie tussen die twee veranderlikes, maar selfs al word dié inskrywings verwyder, is die korrelasie

steeds meer as 0,98. (Sien ook tabel 4.1 waar segmente sonder voorstelle ook verwyder is vir die berekening.) Dit blyk dat die verskille tussen dié twee afstandmate beperk word deur die sintaksis van die taal, met ander woorde, die skrapping of invoeging van 'n woord kan selde op twee plekke gebeur en sal dus meestal as dieselfde afstand weerspieël word in al twee hierdie mate.

Daar is 'n soortgelyke hoë korrelasie tussen verskeie van die afstandmate, veral onderling tussen die karaktergebaseerde mate.

Waar 'n model 'n laer R^2 behaal het, dui dit daarop dat die afstandmaat wat gebruik is, nie die tyd so goed modelleer soos die kompeterende mate nie. Hierdeur word byvoorbeeld gesien dat keystrokes, NGP en TER minder kompeterend is in die datastel. As die resultaat van hoofstuk 3 (die saak van mate se sydigheid) in gedagte gehou word, dui dit ook voorlopig daarop dat hierdie mate eweneens ook nie so geskik is vir die onttrekking van voorstelle nie.

Alhoewel die resultate van die regressiemodel dui op **diceword** of **edit3word** as moontlik die beste mate vir soortgelykheid vir die skatting van die waarde van voorstelle uit 'n vertaalgeheue, stel hierdie werk ons nie in staat om die ander as nutteloos af te maak nie. Die sterkte van die verband met tyd is naby aan mekaar. Voortaan word die resultate dus met meer as een van hierdie mate in die evaluasiemetode bereken om sodoende 'n wyer prentjie te skets van die prestasie van 'n stelsel of datastel wat geëvalueer word.

5

IDENTIFISERING VAN VUIL INSKRYWINGS IN 'N VERTAALGEHEUE DEUR MIDDEL VAN GEKONTROLEERDE LEER

In hierdie hoofstuk word 'n stelsel vir die skoonmaak van vertaalgeheues aangebied wat hoofsaaklik geredelik beskikbare komponente gebruik. 'n Aanvanklike weergawe van hierdie stelsel is ingedien by die eerste gedeelde taak vir die outomatiese skoonmaak van vertaalgeheues [13].[1] Die doel van die taak is "om outomatiese maniere te vind om vertaalgeheues skoon te maak wat om die een of ander rede nie ordentlik nagegaan is nie en verkeerde vertalings bevat."[2] Die probleem word as 'n klassifikasietaak benader wat opgelos word met masjienleer soortgelyk aan die benadering wat deur [11] gevolg is. Die gebruik van geredelik beskikbare komponente maak dit effens makliker om eksperimente op verskeie tale te doen en te fokus op die benadering en evaluasie in plaas daarvan om elke komponent individueel te optimeer.

Die opspoor van foute in teks wat deur mense geproduseer word, is lank reeds 'n aktiwiteit in natuurliketaalverwerking. Speltoetsers en grammatikatoetsers is dalk die bekendste weens hul integrasie in woordverwerkers. Die identifisering van foutiewe vertaalpare kan neerkom op 'n groot onderneming, afhangend van die tipes foute wat verwag word. Barbu [11] beskryf 'n poging om 'n vertaalgeheue skoon te maak wat saamgestel is uit verskeie bronne van variërende kwaliteit, insluitend materiaal wat deur gebruikers bygedra is en materiaal wat outomaties bekom is. Sommige van die foutiewe vertalings in die datastel sluit in deurmekaar teks of gesprekke tussen gebruikers in plaas van vertalings, verkeerde taal en gedeeltelike vertalings.

1 Hierdie hoofstuk is 'n uitgebreide weergawe van [85].
2 "... finding automatic ways of cleaning translation memories that for some reason have not been properly curated and include wrong translations."

Zariņa et al. [90] poog om nieparallelle teks te herken in korpusse wat in naam parallel is. Sulke gevalle is 'n geheel ander saak as die identifisering van meer subtiele aspekte soos swak styl, gebrekkige vloeiendheid, minder as ideale leksikale keuse of klein tipografiese foute soos spasiëring en leestekens. By kwaliteitskatting* in masjienvertaling sal weereens ander soorte foute verwag word weens die aard van masjienvertaalstelsels. Die ontwerp van 'n stelsel om verkeerde vertalings te bespeur, moet dus die verwagte foutsoorte in ag neem.

(en) *quality estimation, QE*

'n Verskeidenheid kenmerke* word onttrek uit geannoteerde data (afdeling 5.3). Die kenmerke is gebaseer op vorige werk, bestaande hulpmiddels en in afwagting van sekere foute. Ontbrekende inhoud in die bron- of doelsegment kan 'n belyningsfout of vertaalfout wees. 'n Aantal ortografiese en nietaalkundige dele van die teks, soos leestekens, getalle, spasies, URL'e, e-posadresse en XML-etikette moet meestal konsekwent wees tussen bron- en doeltekste. Waar dit nie die geval is nie dui dit dalk op 'n fout. Spelfoute en grammatikafoute is van 'n meer taalkundige aard. Vloeiendheid en leksikale keuse is dikwels van 'n meer subjektiewe aard, maar ook sulke foute word geantisipeer.

(en) *features*

Die bespreking hier bo van speltoetsers en grammatikatoetsers is nie slegs agtergrond nie. Ons stelsel kombineer verskéie kwaliteittoetsers, insluitend 'n speltoetser en 'n grammatikatoetser. Verder word reëlgebaseerde vertaaltoetsers as deel van bykomende kenmerke gebruik.

Laastens word kenmerke ingesluit wat eksterne data gebruik, beide een- en tweetalig, om probleme met vloeiendheid en leksikale keuse te bespeur. Alhoewel die insluiting van eksterne data 'n mate van veranderlikheid tot die metode toevoeg, is die gebruik van eksterne data 'n realistiese opset in die praktyk, veral vir die taalpare wat in hierdie hoofstuk gebruik word.

Ons kombineer inligting vanaf hierdie hulpmiddels as kenmerke in 'n gekontroleerde masjienleerbenadering* (afdeling 5.4) wat kompeterend was in die gedeelde taak (sien die resultate in afdeling 5.6).

(en) *supervised machine learning approach*

5.1 DIE PROBLEEM

In hierdie hoofstuk word daar gefokus op die tweede subvraag in afdeling 1.2, naamlik watter tegniek of tegnieke gebruik kan word om foutiewe inskrywings in 'n vertaalgeheue te identifiseer. Twee beskouinge van korrektheid definieer wat as 'n foutiewe vertaling beskou word: 'n nougesette interpretasie en 'n meer toegeeflike een. Dit neem in ag dat by sekere toepassings 'n vertaling met 'n enkele ortografiese fout dalk nie in dieselfde lig beskou sal word as 'n semanties verkeerde vertaling nie. In hierdie afdeling word die probleem in meer besonderhede beskryf, terwyl die datastelle van die gedeelde taak ook bekendgestel word.[3]

Die identifisering van verkeerde segmente word uitgevoer volgens 'n klassifikasiestelsel met drie etikette:

- 'n Volledig korrekte vertaling: 1.
- 'n Semanties korrekte vertaling wat slegs geringe redigering benodig: 2.
- 'n Verkeerde vertaling of 'n vertaling wat aansienlike redigering benodig: 3.

Klassifikasie-instruksies het ook genoem dat annoteerders die aantal foute relatief tot die lengte van die vertaling moet oorweeg (verskeie geringe foute in 'n kort segment sou ook in klas 3 wees).

Die taak is om ongesiene data outomaties te klassifiseer volgens drie subtake:

- Binêre klassifikasie I: om te onderskei tussen etiket 1 en die ander twee (2 en 3).
- Binêre klassifikasie II: om te onderskei tussen etiket 3 en die ander twee (1 en 2).
- Fyn klassifikasie: om segmente te klassifiseer volgens die skema wat hier bo uiteengesit is.

3 Vir volledige besonderhede, sien [13].

Die twee binêre klassifikasietake kan beskou word as onderskeiding tussen "goeie" en "slegte" vertalings onder die twee verskillende beskouinge van korrektheid (nougeset teenoor toegeeflik). Die fyn klassifikasietaak is die meer ingewikkelde geval waar tussen al drie etikette onderskei moet word.

Die datastel is 'n steekproef wat geneem is uit 'n groot vertaalgeheue oor veelvuldige domeine (wat wissel van medies en fisika tot gesprekstaal). Aanvanklik is buitengewoon kort segmente en segmente met sekere ongewenste kenmerke uitgesluit (etikette, verkeerde taal, nietriviale 1–1 belynde sinpare per segment, duplikate). Die data is verskaf vir drie taalpare (Engels–Duits, Engels–Spaans, Engels–Italiaans), en elke inskrywing in die vertaalgeheue is deur twee moedertaalsprekers met een van die drie etikette wat bo genoem is, geannoteer. Ooreenstemming tussen annoteerders* word onder bespreek.

(en) *inter-annotator agreement*

Die datastel bevat ongeveer 2000 segmente vir elke taalpaar. Ongeveer twee derdes van die data is verskaf as opleidingsdata, en die oorblywende derde is eenkant gehou vir evaluasie. Die verdeling van kategorie-etikette in die datastel is nie gelyk nie — etiket 1 (heeltemal korrek) is verreweg die algemeenste. In tabel 5.1 word die verdeling van etikette in die opleidingsdata aangetoon. Die verdeling van die datastel in opleidingstel en toetsstel is gedoen met behulp van gestratifiseerde steekproefneming.* Die kategorie-etikette in die toetsstel het dus 'n verdeling baie soortgelyk aan wat in tabel 5.1 aangedui word.

(en) *stratified sampling*

Tabel 5.1: Verdeling van die klasetikette in die opleidingsdata.

Taalpaar	Etiket 1	Etiket 2	Etiket 3
en→de	77,8%	7,2%	15,0%
en→es	68,1%	9,3%	22,6%
en→it	61,8%	18,0%	20,1%

Barbu et al. [13] dui aan dat die ooreenstemming tussen annoteerders nie hoog was nie (Cohen se kappa tussen 0,37 en 0,57), en het gevolglik slegs segmente gebruik waar die twee annoteerders vir die taal ooreengestem het. In twee gevalle het dit nie die gewenste aantal segmente gelewer nie. Vir die Engels–

Italiaanse datastel het 'n arbiter die datastel verder vergroot. In die geval van die Engels–Duitse datastel was daar nie genoeg foutiewe segmente nie, en bykomende kunsmatige foute is met die hand bygevoeg deur 'n moedertaalspreker om die tipe en verdeling van foute in die ander twee taalpare na te boots.[4]

As 'n poging om 'n vertaalgeheue skoon te maak gemik is op baie groot datastelle, kan die werkverrigting van 'n skoonmaakstelsel 'n oorweging wees. Alhoewel daar in hierdie hoofstuk op 'n klein datastel gefokus word, word die werkverrigting van klassifiseerders kortliks in afdeling 5.4 genoem. Die organiseerders van die gedeelde taak het deelnemers aangemoedig om nie van masjienvertaling gebruik te maak om segmentkwaliteit te bepaal nie. Barbu [11] het ook aangedui dat die gebruik van masjienvertaling 'n beperking kan plaas op die skaal van bewerking (die grootte van die vertaalgeheue wat hanteer kan word). Dit kan ook 'n groot finansiële uitgawe wees as daar van 'n kommersiële masjienvertaalstelsel gebruik gemaak word. In afdeling 5.6 word die benadering in hierdie hoofstuk wel vlugtig vergelyk met 'n ander stelsel wat masjienvertaling gebruik het. Die kenmerke waarvan wel gebruik gemaak is, word in afdeling 5.3 bespreek ná 'n bondige agtergrond van enkele masjienleertegnieke in die volgende afdeling.

5.2 KLASSIFIKASIE IN MASJIENLEER

Masjienleer is 'n benadering tot probleemoplossing waarin 'n rekenaarprogram self aspekte van die oplossing aanleer.[5] Alhoewel die benadering steeds deur 'n programmeerder gespesifiseer word, word sommige van die parameters vir die program outomaties geleer vanaf opleidingsdata. In die geval waar voorbeelde met die gewenste afvoer beskikbaar is om van te leer, word dit gekontroleerde leer genoem. Dit word gekontrasteer

4 Persoonlike korrespondensie met Barbu.

5 Die agtergrond hier word verskaf aan die hand van onder andere [42] en [37], en terminologie soos in [77]. Slegs enkele aspekte van masjienleer vir klassifikasieprobleme wat later in die hoofstuk gebruik word, word hier gedek.

met ongekontroleerde leer, waar die gewenste afvoer nie vir opleiding beskikbaar is nie, en die masjienleerbenadering dus self struktuur in die data moet vind.

'n Verskeidenheid probleme kan met masjienleer aangepak word. Hier word slegs klassifikasieprobleme genoem, aangesien dit die tipe probleem is wat hier ter sprake is. (Sien afdeling 5.1.)

Uit die verskeidenheid masjienleermetodes wat klassifikasie kan verrig, word die volgende oorsigtelik bespreek:

(en) *support vector classifier*
- die steunvektorklassifiseerder;*

(en) *decision tree*
- die beslissingsboom;*

(en) *random forest classifier*
- die ewekansigewoud-klassifiseerder;*

(en) *logistic regression classifier*
- die logistieseregressie-klassifiseerder.*

Al hierdie benaderings word in hierdie hoofstuk gebruik behalwe beslissingsbome. 'n Kort oorsig oor beslissingsbome is wel waardevol ter wille van 'n eenvoudiger oorsig van ewekansige woude. Hierdie benaderings kan almal as klassifiseerders dien in 'n opset met gekontroleerde leer.

In elke geval word die data gespesifiseer as 'n stel multidimensionele datapunte. Elke komponent van 'n datapunt verteenwoordig die datapunt se waarde ten opsigte van 'n spesifieke *kenmerk*. Die kenmerke kan diskreet of kontinu wees. 'n Volledige uiteensetting van die kenmerke wat in hierdie hoofstuk gebruik word, volg in afdeling 5.3.

(en) *hyperplane*

Die *steunvektorklassifiseerder* ondersoek die opleidingsdata en probeer 'n hipervlak* in die n-dimensionele ruimte te vind, sodanig dat die twee klasse punte weerskante van die hipervlak is. In werklikheid gebeur dit wel dat so 'n perfekte verdeling met 'n hipervlak nie die reël is nie. Dit gebeur wanneer die data — om watter rede ook al — nie lineêr skeibaar is nie. Die steunvektorbenadering het wel 'n manier om volgens 'n optimeringsdoel 'n hipervlak te kies wat so goed as moontlik probeer om tussen die klasse te onderskei. Die hipervlak bied dan 'n manier waarmee ongesiene datapunte geklassifiseer kan word, aangesien

dit aan die een of die ander kant van hierdie hipervlak geleë is. Met behulp van die gebruik van 'n kernfunksie* kan data op nielineêre maniere afgebeeld word op alternatiewe ruimtes waar die data dalk wel (of in 'n groter mate) lineêr skeibaar is. 'n Parameter C beheer die sydigheid-variansie-kompromis.* Vir meer inligting, sien gerus [42, p. 344].

(en) *kernel function*

(en) *bias-variance trade-off*

'n *Beslissingsboom* verdeel die kenmerkruimte van die datapunte in gebiede volgens 'n stel reëls, en ken vir elke gebied 'n enkele klassifikasie toe. Die reëls waarmee die ruimte verdeel word, word vanuit die opleidingsdata geleer. Hierdie stel reëls kan as 'n binêre boom voorgestel word, waar elke nodus van die boom 'n verdelingspunt van die kenmerkruimte voorstel wat op grond van die waarde van 'n kenmerk die ruimte verder verdeel. Beslissingsbome "is nie tipies kompeterend met die beste gekontroleerde leerbenaderings nie",[6] maar metodes wat meerdere beslissingsbome kombineer, kan wel beter presteer [42, p. 303].

'n *Ewekansige woud* is so 'n versameling beslissingsbome wat elkeen effens anders opgestel word. Skoenlussteekproefneming* word gebruik om telkens 'n subversameling van die opleidingsdata te gebruik om 'n nuwe boom te bou. By elke nodus in die boom word die keuse van kenmerke ook ewekansig beperk om die variasie tussen bome te vergroot. Die verskillende beslissingsbome wat so opgestel word, word dan in ensemble gebruik om oor die finale klassifikasie van 'n nuwe datapunt te stem.

(en) *bootstrapping*

Die *logistieseregressie-klassifiseerder* modelleer die waarskynlikheid van lidmaatskap van 'n sekere klas. Vir ongesiene datapunte kan hierdie waarskynlikheid dan bereken word, en die klassifikasie volgens die berekende waarskynlikheid gedoen word. Alhoewel hierdie klassifiseerder van die nielineêre logistiese funksie gebruik maak, is die modelleervermoëns soortgelyk aan dié van 'n steunvektorklassifiseerder [42, p. 356].

Vorige werk wat verband hou met hierdie hoofstuk het van hierdie klassifiseerders gebruik gemaak, en daarom word hulle

6 "... typically are not competitive with the best supervised learning approaches ..."

almal in hierdie hoofstuk oorweeg. Barbu [11] het 'n aantal masjienleeralgoritmes geëvalueer en het bevind dat 'n steunvektorklassifiseerder en 'n logistieseregressie-klassifiseerder die beste presteer. In 'n verwante taak rapporteer [90] die beste resultate met 'n ewekansigewoud-klassifiseerder. Al die klassifiseerders kan van dieselfde kenmerke gebruik maak. Hierdie kenmerke wat uit die opleidingsdata onttrek word, word volgende aangebied.

5.3 KENMERKE

Ons sluit sommige kenmerke in wat gebaseer is op statistiek en meting (soos kenmerke verwant aan lengtes) asook kenmerke gebaseer op eenvoudige heuristiese metodes. Voorbeelde van foute word as toeligting direk uit die opleidingsdata voorgehou in hierdie afdeling. Verskeie van die kenmerke is soortgelyk aan dié wat gebruik is vir 'n soortgelyke taak [11] asook dié van QuEst++ vir segmentvlak-kwaliteitskatting van masjienvertaling [75].

5.3.1 *Kenmerke gebaseer op lengte*

Barbu [11] het die waarde van die Gale-Church-verhouding aangedui vir 'n soortgelyke taak. Hy volg 'n formulering deur [81], en ons bereken hierdie verhouding op karakters soos oorspronklik voorgestel [30]. In die notasie hier onder is $|s|$ die lengte van die string s.

$$\frac{|s| - |t|}{\sqrt{3.4(|s| + |t|)}}$$

Ander kenmerke wat afgelei word van die lengtes van die bron- en doelsegment is ook ingesluit. Die bron- en doellengte word net so ingesluit asook 'n kenmerk wat bereken word van die verskil tussen die twee. Alhoewel masjienleertegnieke soos steunvektormasjiene kan leer om lineêre kombinasies van kenmerke te gebruik, kan metodes gebaseer op beslissingsbome

(soos ewekansige woude) wel baat vind by kombinasies van kenmerke as bykomende kenmerke. Aangesien die verskil tussen bron- en doellengtes aansienlik groter is vir langer stringe as vir korter stringe, word normalisering toegepas in 'n poging om 'n meer konsekwente kenmerk daar te stel. Ons bereken die standaardtelling (of z-telling) met die intuïsie dat kenmerkwaardes met verskillende etikette makliker lineêr geskei sal kan word. Vir 'n bronsegment s en doelsegment t word die verskil, Δ_{lengte}, soos volg bereken:

$$\Delta_{lengte} = \frac{|s| - \mu_s}{\sigma_s} - \frac{|t| - \mu_t}{\sigma_t}$$

waar μ_s en σ_s onderskeidelik die gemiddeld en standaardafwyking is van die lengtes van die bronsegmente in die opleidingsdata, en μ_t en σ_t soortgelyk is vir die lengtes van die doelsegmente. Hierdie standaardisering moet liefs op data uit 'n normaalverdeling verrig word. Aangesien die lengtes nie normaal verdeel is nie, kan ander normaliseringsmetodes of 'n transformasie op die data oorweeg word. 'n Poging om ook nog 'n sigmoïde-skaleringsfunksie te gebruik, soos voorgestel deur [25], het nie resultate verbeter nie.

Hierdie ekstreme voorbeeld toon 'n geval wat maklik geïdentifiseer word deur enige van hierdie lengtegebaseerde kenmerke:

Bronteks (en)	4.4 Special warnings
Doelteks (es)	4.4 Advertencias especiales según la especie animal a la que vaya destinado
groot lengteverskil	*(bronteks: 20 karakters, doelteks: 75 karakters)*

5.3.2 *pofilter*

Die Translate Toolkit[7] is 'n stel gereedskap en 'n toepassingsbiblioteek vir sagtewarelokalisering en lokaliseringsingenieurswerk. Dit word wyd gebruik vir die lokalisering van Vry en Oopbronsagteware (FOSS)*, en sluit funksionaliteit en aanpas-

(en) *Free and Open Source Software, FOSS*

7 http://docs.translatehouse.org/projects/translate-toolkit/

sings in vir verskeie populêre FOSS-lêerformate en -projekte. Dit word in verskeie eindgebruikertoepassings vir sagtewarelokalisering gebruik, bv. Pootle, Virtaal, Weblate, Damned Lies, Pontoon en die Wordforge-vertaalprogram. Een afdeling funksionaliteit het te make met kwaliteittoetsing van vertalings. Die funksionaliteit is gebaseer op reëls en heuristieke om vertalings te bespeur wat dalk baat sal vind by hersiening. Die toetse is bruikbaar deur 'n programmeerkoppelvlak* of 'n opdraglynprogram, `pofilter`. Dit het 'n paar ooreenkomste met TMop,[8] wat deur sommige van die deelnemers in die gedeelde taak gebruik is.

(en) *application programming interface, API*

Die meeste van die toetse vergelyk bloot die een of ander aspek van die bron- en doelteks. As 'n probleem deur 'n toets geïdentifiseer is, kan dit ernstig wees, soos foute met XML-etikette of `printf`-veranderlikes, of van 'n minder ernstige aard, bv. konsekwentheid van leestekens. Die pofilter-dokumentasie lys die kategorieë toetse as kritiek, funksioneel, kosmeties en onttrekking.[9] Vir die doeleindes van hierdie studie is slegs die eerste drie van hierdie vier kategorieë relevant. Sommige van die individuele toetse is vir projekspesifieke aspekte (soos die *gconf*-toets vir GNOME-lokalisering), of het opstelling nodig, soos *notranslatewords* ('n toets vir woorde wat onveranderd moet bly in die doeltaal). Alle projekspesifieke toetse word laat vaar asook alle toetse wat opstelling vereis. Slegs 'n subversameling van hierdie toetse word dus gebruik; elkeen as 'n kenmerk in die masjienleerbenadering.

Hierdie toetse het almal 'n binêre aard — 'n segmentpaar word beskou as foutief of nie foutief nie. Vir verskeie van die kenmerke is dit moontlik die enigste klassifikasie wat sin maak. So byvoorbeeld is spasiëring aan die einde van die segmente soortgelyk of nie soortgelyk nie, en dit is te betwyfel of 'n maat van die graad van soortgelykheid van veel waarde sal wees. Ander kenmerke kan egter wel voorsien word om die aantal moontlike probleme met die vertaling te tel, soos die aantal

8 https://github.com/hlt-mt/TMOP
9 http://docs.translatehouse.org/projects/translate-toolkit/en/latest/commands/pofilter_tests.html

problematiese XML-etikette of die aantal problematiese aanhalingstekens. In ons werk word al hierdie toetse van binêre aard gebruik as diskrete 0–1-kenmerke as die mees voor die hand liggende manier om hierdie biblioteek in te span. Aangesien sommige van hierdie kwaliteittoetse soortgelyk is aan ander kenmerke in die masjienleeropstelling, verwyder ons dié waar daar 'n kontinue kenmerk is vir 'n soortgelyke kwaliteitsprobleem. Dit sluit die volgende pofilter-toetse in: *short*, *long* en *spelling*. Die finale stel pofilter-toetse wat ingesluit is, is:[10]

- *acronyms*: of akronieme onveranderd behou is;
- *brackets*: die konsekwente gebruik van hakies;
- *doublequoting*: die konsekwente gebruik van dubbele aanhalingstekens;
- *doublespacing*: die konsekwente gebruik van dubbele spasies;
- *doublewords*: die gebruik van enige verdubbelde woord direk langs mekaar;
- *emails*: die konsekwente gebruik van e-posadresse;
- *endpunc*: die konsekwente gebruik van leestekens aan die einde van die segmente;
- *endwhitespace*: die konsekwente gebruik van wit spasie aan die einde van die segmente;
- *numbers*: die konsekwente gebruik van getalle;
- *puncspacing*: die konsekwente spasiëring by leestekens;
- *purepunc*: die konsekwentheid tussen segmente wat meestal uit leestekens bestaan;
- *sentencecount*: die konsekwentheid in die getal sinne per segment;

10 Toetse se name word deurgaans aangehaal soos hulle in die dokumentasie van pofilter gebruik word.

- *simplecaps*: aspekte van hooflettergebruik;
- *simpleplurals*: die hantering van die opsionele meervoud-aanduiding deur "(s)";
- *singlequoting*: die konsekwente gebruik van enkelaanhalingstekens;
- *startcaps*: die konsekwente gebruik van hoofletters aan die begin van die segmente;
- *startpunc*: die konsekwente gebruik van leestekens aan die begin van die segmente;
- *startwhitespace*: die konsekwente gebruik van wit spasie aan die begin van die segmente;
- *unchanged*: of die doel bloot 'n kopie van die bron is sonder verandering;
- *urls*: die konsekwente gebruik van URL'e.

Verskeie van hierdie toetse is meer geneig om foute onder die nougesette interpretasie te bespeur waar klein niesemantiese probleme geïdentifiseer moet word. Sommige van hierdie fouttoestande is nie in die opleidingsdata gevind nie, en is dus nie bruikbaar hier nie. Hulle sou definitief nuttig kon wees in 'n ander konteks.

Hier is 'n paar voorbeelde van vertaalpare waar een of meer pofilter-toetse 'n moontlike probleem identifiseer:

Bronteks (en) Doelteks (de) *acronyms*	DIN 4109 Sound insulation in buildings din 4109 schallschutz im hochbau
Bronteks (en) Doelteks (it) *acronyms, brackets, endpunc, numbers*	42/ 125 Annex II of Council Regulation (EEC) No 2377/ 90 2377/ 90 del Consiglio:

Bronteks (en)	to a clinically unacceptable level ...
Doelteks (it)	Se si osservano segni di sovraccarico circolatorio ...
startcaps	
Bronteks (en)	Page 2/ 3 ©EMEA 2006 Nobilis Influenza H5N2 has been authorised under " Exceptional Circumstances".
Doelteks (es)	2/ 3 ©EMEA 2006 la autorización de comercialización.
doublequoting	

5.3.3 *Spelling en grammatika*

Speltoetsers is 'n algemene hulptegnologie vir skryfwerk en vertaalwerk. Spelfoute kan dui op swak vertaalwerk of agterlosige skryfwerk in die algemeen. Die Vry en Oopbron-Hunspell-speltoetsers[11] word gebruik vir die betrokke tale deur die Enchant-raamwerk.[12] Hunspell verskaf verskeie eienskappe, soos ondersteuning vir samestellings en komplekse morfologie. Alhoewel ons besef dat die Hunspell-speltoetsers vir verskillende tale deur verskillende Oopbronspanne ontwikkel word, dalk met effens verskillende doelwitte en benaderings, word verwag dat die ondersteuning vir alle tale van die gedeelde taak wat in hierdie hoofstuk hanteer word, van vergelykbare (goeie) kwaliteit sal wees.

In plaas daarvan om bloot die aantal nieherkende woorde in die doelteks te tel, word verder ook tred gehou van nieherkende woorde in die bronteks. Hierdie nieherkende bronwoorde word dan verbatim toegelaat in die doelteks met die aanname dat hulle benoemde entiteite* kan wees wat die meeste speltoetsers nie sal herken nie. 'n Soortgelyke benadering word gebruik in die *spellcheck*-toets van pofilter. Die verskil is dat ons die getal moontlike foute per segment tel, terwyl pofilter slegs 'n binêre oordeel vel per segmentpaar. Die telling word genormaliseer deur dit te deel deur die aantal tekseenhede in die doelteks, en dit dien dus as 'n kontinue kenmerk.

(en) *named entities*

11 https://hunspell.github.io/
12 https://www.abisource.com/projects/enchant/

LanguageTool[13] is 'n Oopbron-proefleesprogram [62]. Weergawe 3.3 wat hier gebruik is, bied ondersteuning vir meer as 20 tale, maar in teenstelling met ons verwagting van die Hunspell-speltoetsers, is dit bekend dat die vlak van ondersteuning vir elke taal aansienlik varieer. Dit bied ondersteuning vir al vier tale wat in hierdie hoofstuk ondersoek word. Die vlak van ondersteuning vir die betrokke tale word in tabel 5.2 opgesom.[14] Afgesien van 'n blote analise van die doelteks, aktiveer ons ook "*bitext*-modus" waar kennis van valse vriende met die ander taal in ag geneem word.

Tabel 5.2: Reëls in LanguageTool

Taal	XML-reëls	Java-reëls	Valse vriende	Verwarringspare
en	1318	16	356	485
de	2076	24	126	26
es	92	1	57	0
it	135	2	37	0

Twee kenmerke word bereken gebaseer op die grammatikatoetser:

- Die getal probleme wat in die doelteks bespeur is.
- Die verskil tussen die getal probleme wat vir die bron- en doelteks gerapporteer is.

Afgesien van basiese styl- en grammatikatoetsing kan LanguageTool ook spelling toets. Omdat spelling in 'n aparte kenmerk getoets word, is hierdie funksionaliteit gedeaktiveer vir die kenmerke wat met behulp van LanguageTool bereken is. Hier volg 'n voorbeeld wat die drie kenmerke van hierdie afdeling illustreer:

Bronteks (en)	Please check our objection
Doelteks (de)	bitte prüfen sie unseren einwand
spelfoutkoers=$\frac{1}{5}$	*(hooflettergebruik by "einwand" — 1 uit 5 woorde)*
grammatikafoute=2	*(aanvanklike hoofletter, hooflettergebruik by "sie")*
grammatikaverskil=0-2	*(geen foute in bronteks, 2 in doelteks)*

13 https://languagetool.org/
14 vanaf https://www.languagetool.org/languages/

5.3.4 *Statistiese kenmerke bereken met behulp van eksterne data*

5.3.4.1 *Taalwaarskynlikheid*

Vloeiendheid word dikwels in die evaluasie van vertaling genoem. 'n Taalmodel word in statistiese masjienvertaling gebruik om afvoer meer vloeiend te maak. "'n Taalmodel is 'n waarskynlikheidsverdeling $p(s)$ oor stringe s wat beskryf hoe gereeld die string s as 'n sin in 'n domein van belang verskyn"[15] [21]. As 'n verdere kenmerk in die klassifiseerder word 'n taalmodel gebruik wat die waarskynlikheid van 'n segment teks skat. Daar kan verskeie redes wees vir 'n lae skatting van die waarskynlikheid van 'n gegewe stuk teks: dit kan byvoorbeeld gewoon oor 'n minder bespreekte onderwerp handel. Ons skat dus die waarskynlikheid vir beide die bron- en doelteks, en gebruik die verskil as die leerkenmerk. As die onderwerp van bespreking skaars is, behoort die waarskynlikheid van beide die bron- en doelteks laag te wees, en die verskil tussen hulle behoort vergelykbaar te wees met die waarskynlikheidsverskil vir teks oor 'n meer algemene onderwerp.

Die waarskynlikheid van 'n teks hang egter af van die lengte van die segment. 'n Langer segment het gewoon 'n laer waarskynlikheid. Dit veroorsaak 'n groter verskil tussen bron- en doelwaarskynlikhede vir die langer segmente — ook vir goed vertaalde segmente. Ons probeer dus om die waarskynlikheid aan te pas om die effek van segmentlengte effens te verminder. Dit is 'n poging om die verskil tussen die twee waarskynlikhede 'n meer betroubare kenmerk tussen segmente van verskillende lengtes te maak.

Uit inspeksie blyk dit dat die segmentwaarskynlikhede eksponensieel afneem met 'n toename in segmentlengte. Vir 'n bronsegment s van lengte $|s|$ verkry ons 'n waarskynlikheid $p(s)$ soos geskat deur die taalmodel, en pas dit dan aan vir die segmentlengte deur met $10^{|s|}$ te vermenigvuldig. Dit maak die aanname dat die waarskynlikheid van 'n segment normaalweg

15 "A language model is a probability distribution $p(s)$ over strings s that describes how often the string s occurs as a sentence in some domain of interest."

sou verminder met 'n vermenigvuldigingsfaktor in die orde van 0,1 vir elke addisionele tekseenheid. Alhoewel die waarde van 0,1 blykbaar die gewenste effek het om die lineêre afhanklikheid op die segmentlengte te verminder, kan 'n beter waarde moontlik met parameteroptimering bepaal word. Aangesien die waarskynlikhede as $\log_{10}$-waarskynlikhede hanteer word, beteken dit bloot dat 'n aangepaste hoeveelheid vir 'n segment s met $|s|$ tekseenhede verkry word as

$$\log_{10} p(s) + |s| \tag{5.1}$$

en soortgelyk vir die doelsegment. Hierdie hoeveelheid is nie meer 'n ware waarskynlikheid nie, maar vir bondigheid word daar in die res van die hoofstuk steeds na die taalmodelwaarskynlikheid verwys. Hierdie waardes word gestandaardiseer deur die z-tellings vir die bron en die doel te bereken, en die verskil tussen hierdie gestandaardiseerde tellings word gebruik as die kenmerk, soortgelyk aan die lengteverskil wat bo genoem is op p. 101. Die mate van sukses wat behaal word om die korrelasie met die segmentlengte te verminder, word opgesom in tabel 5.3. Ons aanpassing van die data het blykbaar nie dieselfde effek in alle taalpare nie.

Tabel 5.3: Korrelasie (ρ) tussen waarskynlikheidsverskil volgens die taalmodel en die brontekslengte voor en ná die lengteaanpassing en standaardisering. Die berekening is slegs uitgevoer op inskrywings uit die opleidingsdata in klas 1.

Taalpaar	ρ voor	ρ ná
en→de	0,3282	0,1682
en→es	0,1833	0,1952
en→it	0,2596	0,0860

Die intuïsie is dat die kenmerk sal help om die volgende tipes foute te identifiseer: gebrek aan grammatikale ooreenkoms, verkeerde woordorde, en tikfoute in algemene woorde. Sulke foute verlaag die waarskynlikheid aan een kant van die bron-

doelpaar op 'n manier wat behoort te reflekteer in die verskil ongeag hoe algemeen die onderwerp is.

Daar is vir al vier tale taalmodelle opgelei op Europarl [49]. Alhoewel dit nie 'n gebalanseerde korpus is nie, argumenteer ons dat, aangesien dit as parallelle korpus ongebalanseerd is op dieselfde manier vir elke taal, dit nie veel saak maak nie. Ons veronderstel dat die sydigheid van die taalmodelle soortgelyk sal wees in beide die bron- en doeltaal, en dat die sydigheid dus sal uitkanselleer in die waarskynlikheidsverskil tussen die twee tale se taalmodelle.

'n 5-gram-taalmodel is geïmplementeer met KenLM[16] [39] wat aangepaste Kneser-Ney-gladstryking* gebruik. Hier is twee voorbeelde met 'n groot waarskynlikheidsverskil:

(en) *modified Kneser-Ney smoothing*

Bronteks (en)	It is not only the future of Zimbabwe that is at stake.
Doelteks (de)	Es ist nicht allein die Zukunfft Simbabwes, die auf dem Spiel steht
groot verskil in waarskynlikheid	*(bron meer waarskynlik, moontlik weens spelfout "Zukunfft")*
Bronteks (en)	All food producing species All food producing species
Doelteks (es)	Todas las especies productoras de alimentos
groot verskil in waarskynlikheid	*(bronteks het lae waarskynlikheid weens duplisering)*

5.3.4.2 *Leksikale vertaalkeuse*

Volgende word die gebruik van eksterne tweetalige data oorweeg. Vanuit tweetalige data word woordbelyningsmodelle opgelei met behulp van `fast_align`[17] wat 'n geherparameteriseerde weergawe van IBM model 2 implementeer [27]. Die tweetalige belynde Europarl-korpus [49] dien hiervoor as opleidingsdata vir elke taalpaar.

Zariņa et al. [90] gebruik woordbelynings en hulle waarskynlikhede om segmente as korrek al dan nie te klassifiseer. Ons benadering is eenvoudiger en behels leksikale vertalings sonder dat woordbelyning van die geëvalueer segmente nodig is. Barbu [11] het die gebruik van 'n tweetalige woordeboek

16 http://kheafield.com/code/kenlm/
17 https://github.com/clab/fast_align

voorgestel in plaas van volskaalse masjienvertaling. Ons gebruik die geskatte leksikale vertaalwaarskynlikhede uit die eksterne korpus en onttrek vertalings met hoë waarskynlikheid in al twee rigtings ('n poging tot 'n onttrekking met hoë presisie). Uit die vertalings wat vir elke woord x gesien word, word die twee met die hoogste waarskynlikheid ondersoek. As die mees waarskynlike vertaling 'n hoë waarskynlikheid het ($\log$-waarskynlikheid hoër as -1,0) en die tweede vertaling aansienlik minder waarskynlik is, word die mees waarskynlike vertaling by die leksikon gevoeg as 'n verwagte vertaling van x. Elke inskrywing in elke rigting word gesien as 'n reël waaraan vertalings behoort te voldoen. Die omvang van die reëls wat so onttrek is, word opgesom in tabel 5.4.

Die voorkoms van die verwagte vertalings word getoets deur gebruik te maak van leksikons in al twee rigtings, en die getal "ontbrekende" woordeboekinskrywings word in al twee rigtings getel. Op hierdie manier word geen inligting oor woordorde of woordbelyning gebruik nie. Dit is aansienlik vinniger as om die vertaalwaarskynlikheid oor die hele segment te bereken volgens IBM model 2 of soortgelyke leksikale model, aangesien geen stap nodig is vir woordbelyning nie. Daar word verwag dat die nuttigheid van hierdie kenmerk beduidend beïnvloed sal word deur die opleidingsdata waaruit die tweetalige woordeboek onttrek word. Die gebruik van handgemaakte terminologie van 'n vertaalprojek kan moontlik selfs beter resultate lewer, mits dubbelsinnigheid voldoende vermy of hanteer kan word. Hier word die teks as sakke woorde gebruik sonder enige betekenisvereenduidiging.*

(en) *semantic disambiguation*

5.4 KLASSIFISEERDERS

Die onttrekte kenmerke word gebruik vir masjienleer met die klassifiseerders wat in afdeling 5.2 aangebied is. In hierdie afdeling word 'n opsomming van die ondersoek na masjienleermetodes en hul parameters gegee. Die stelsel is in Python geïmplementeer met scikit-learn [67]. Die stelselopstelling was

Tabel 5.4: Leksikale reëls wat in elke rigting onttrek is

Taalpaar	Inskrywings
en→de	86 577
de→en	239 249
en→es	84 143
es→en	123 378
en→it	82 889
it→en	118 459

identies vir alle taalpare en alle subtake. Die modelle is uit die aard van die saak afsonderlik opgelei volgens die opleidingsdata in elke subtaak.

In aanvanklike toetse is vergelykbare resultate verkry met 'n steunvektorklassifiseerder, 'n logistieseregressie-klassifiseerder, asook met 'n ewekansigewoud-benadering. Ons het verder ondersoek ingestel na die keuse van steunvektor-kerne* en ander parameters.

(en) *support vector kernels*

Vir die steunvektorklassifiseerder is verskeie kerne ondersoek.[18] Slegs die lineêre kern het 'n bruikbare klassifiseerder gelewer. Optimering van die C-parameter het nie 'n merkbare verbetering te weeg gebring nie (die verstekwaarde van C is 1). Die lineêre kern het verskeie voordele. Eerstens is dit moontlik om die geoptimeerde `LinearSVC`-implementasie in scikit-learn te gebruik, wat aansienlik vinniger is aangesien dit gebaseer is op `liblinear` wat spog met beter tydkompleksiteit as die verstekimplementasie. Tweedens maak lineêre kerne kenmerkseleksie* makliker (sien onder). Alhoewel `LinearSVC` ook nog verdere parameters bied (soos die keuse van verliesfunksie*), het aanpassings aan hierdie parameters nie veel gehelp nie. Die een parameter wat die groot verskil maak in werkverrigting is om die klassifiseerder aan te sê om die primale optimeringsprobleem op te los (wat nie moontlik is met nielineêre kerne of die standaard-`SVC`-implementasie nie). Die oplos van die primaal in plaas van die duale optimeringsprobleem is meer effektief wanneer die getal opleidingsdatapunte meer is as die aantal

(en) *feature selection*

(en) *loss function*

18 onder andere lineêr, RBF (radial basis function) en polinomies

kenmerke — wat in hierdie studie wel die geval is. Hoewel die verstek steunvektorklassifiseerder stadig was op die datastel wat in hierdie hoofstuk bespreek word, het hierdie parameter die werkverrigting vergelykbaar gemaak met die ander benaderings wat volgende bespreek word.

Die ewekansigewoud-benadering was redelik kompeterend met verstekparameters en het aantreklike werkverrigting gehad (vergeleke met die stadiger verstek-SVC-implementasie). Deur die getal bome in die ewekansige woud te vermeerder vanaf die verstek van 10, het resultate effens verbeter ten koste van looptyd. Hoewel slegs 'n klein verbetering bespeur kan word wanneer die getal bome na meer as 20 vermeerder word, sal die vermeerdering in looptyd met 'n groter getal bome uiteindelik die aantreklikheid van hierdie benadering affekteer. Tussen 10 en 20 bome is die looptyd steeds aanvaarbaar.

Die logistieseregressie-klassifiseerder was kompeterend in looptyd en kompeterend in stelselprestasie. Parameteroptimering het nie gelei tot 'n beduidende verbetering nie.

Die standaardisering tydens die berekening van lengteverskille en die taalmodelwaarskynlikheid het konsekwent die tellings verbeter, met die ewekansigewoud-benadering wat effens meer baat daarby gevind het as die ander ('n verbetering van amper 2% in F_1-telling in al twee gevalle). Die normalisering van die taalmodelwaarskynlikheid (vergelyking 5.1) bring 'n geringe verbetering. Die steunvektorklassifiseerder en die logistieseregressie-klassifiseerder het meer baat daarby gevind (verbeterings van amper 1,7% in F_1-telling). Standaardskalering is gebruik oor alle kenmerke wat die gemiddeld verwyder het en die variansie geskaleer het na 1. Alternatiewe benaderings tot kenmerkskalering het nie beduidende en konsekwente verbetering te weeg gebring nie.

5.5 BELANGRIKHEID VAN KENMERKE

Deur die klassifiseerder se prestasie te meet terwyl die getal kenmerke verminder word, kan 'n subversameling kenmerke

geïdentifiseer word wat die optimale telling lewer. Die belangrikheid van kenmerke word in hierdie afdeling op twee maniere geëvalueer.

Eerstens word ondersoek watter kenmerke gereeld gekies word as 'n klassifiseerder vrylik kan kies watter om in te sluit. Hiervoor gebruik ons rekursiewe eliminasie van kenmerke* [36], en evalueer die modelle deur 3-voudige kruisvalidasie op die geweegde F_1-maat. Tweedens ondersoek ons die rang van kenmerke wanneer die klassifiseerder geforseer word om sy keuses te orden. Vir die lineêre steunvektorklassifiseerder en die logistieseregressie-klassifiseerder gebruik die ondersoek na kenmerkbelangrikheid die koëffisiënte van die beslissingsfunksie, en vir die ewekansigewoud-klassifiseerder word Gini-belangrikheid gebruik. Sommige patrone kom na vore tussen tale en selfs by sommige van die take. Tussen die drie masjienleerbenaderings is die gedrag van die steunvektorklassifiseerder en die logistieseregressie-klassifiseerder soortgelyk. Daarteenoor verskil die gedrag van die ewekansigewoud-klassifiseerder van dié twee se gedrag. Verskillende kenmerke het verskillende invloed by die ewekansigewoud-klassifiseerder as by die ander twee. Alternatiewe beanderings tot kenmerkseleksie soos bv. genetiese algoritmes is nie oorweeg nie. Die bogenoemde benaderings is algemeen in die veld van masjienleer en die oplossingsruimte is klein genoeg daarvoor.

(en) *recursive feature eliminiation*

'n Evaluasie van die belangrikheid van verskillende kenmerke oor die verskillende subtake en taalpare is nie triviaal nie. Ewekansige getalle* word gebruik in al die klassifiseerders tydens opleiding, en op die koop toe ook in die gestratifiseerde verdeling van die opleidingsdata vir kruisvalidasie. Hoewel die evaluasietellings baie min varieer met herhaalde uitvoering met verskillende saadjies vir ewekansige getalle*, was die resultate van kenmerkseleksie nie stabiel met herhaalde uitvoering nie. Voortaan word dus slegs algemene tendense genoem. Verder kombineer ons waarnemings van die verskillende klassifiseerders en soek ook vir patrone oor die tale en take heen. Om die klein afwykings wat kan voorkom die hoof te bied, is die saadjies vir ewekansige getalle vasgemaak (of konsekwent ge-

(en) *random numbers*

(en) *random number seeds*

varieer oor herhaalde uitvoering) en verder is die toleransie vir die stopkriterium van die logistieseregressie-klassifiseerder verminder tot 10^{-5} wat, te oordeel aan die uitkoms van 'n paar toetse, al die variansie verwyder het.[19]

As die resultate van kenmerkseleksie oor take, tale of leeralgoritmes gekombineer word, kan die frekwensie waarmee kenmerke as deel van die optimale kenmerkversameling gekies word, ondersoek word. Dit maak 'n ondersoek moontlik na die konsekwentheid waarmee kenmerke met klassifikasie help ongeag die grootte van hul bydrae. Al die lengtegebaseerde kenmerke word gereeld gekies en, interessant genoeg, word die lengteverskilkenmerk (afdeling 5.3.1) gewoonlik gekies saam met die Gale-Church-verhouding, al beskryf hulle die data uit soortgelyke hoeke. Die lengteverskilkenmerk word weliswaar in minder gevalle gekies. Alhoewel die grammatikatoetser minder kragtig is vir sommige van die tale, is die twee kenmerke wat daarop gebaseer is, steeds in die meeste gevalle gekies. Die taalmodelwaarskynlikheid word ook redelik konsekwent gekies. Onder die pofilter-toetse is die twee wat konsekwentheid van hoofletters toets (*startcaps* en *simplecaps*) asook die toets vir leestekens aan die einde van 'n segment (*endpunc*) mees konsekwent gekies. Die spelkenmerk is meer gereeld gekies vir Duits. Die kenmerk vir leksikale keuse is amper altyd vir Spaans en Italiaans gekies. In die geval van Duits is leksikale keuse konsekwent gekies deur die ewekansigewoud-klassifiseerder, maar minder deur die ander klassifiseerders. Duitse samestellings veroorsaak waarskynlik meer probleme vir die belyning waardeur die effektiwiteit van hierdie kenmerk verminder word.

Deur die rang van kenmerke te ondersoek, kan 'n indruk gevorm word van die belangrikheid van kenmerke relatief tot mekaar. Al die lengtegebaseerde kenmerke het gewoonlik 'n hoë rang in al die klassifiseerders en tale. Een of al twee die grammatikakenmerke het gewoonlik 'n hoë rang vir Duits. Die

19 Dit is die enigste klassifiseerder waarvoor die saadjie vir ewekansige getalle nie volledig herproduseerbare uitvoering waarborg by herhaling nie. Sien http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

verskil in taalmodelwaarskynlikheid het altyd een van die top twee plekke by die ewekansigewoud-klassifiseerder, maar het laer rang by die ander twee klassifiseerders ('n rang van 5 of laer). Die ewekansigewoud-klassifiseerder is geneig om 'n hoër rang vir die lengtegebaseerde kenmerke te gee vergeleke met die ander twee benaderings. Onder die kenmerke wat op pofilter gebaseer is, is die een met die hoogste rang die toets vir leestekenkonsekwentheid aan die einde van segmente (*endpunc*). Die *endpunc*-kenmerk het hoër rang gehad in die eerste binêre taak en die fyn klassifikasietaak waar die meer nougesette interpretasie van foute ter sprake is. By die tweede binêre taak waar foutiewe vertalings geklassifiseer word volgens 'n meer toegeeflike interpretasie, kry hierdie kenmerk nie so 'n hoë rang nie — die kenmerke van pofilter kry meestal 'n laer rang as ander kategorieë in hierdie taak. Die eienskap vir leksikale keuse was in die top ses kenmerke vir Spaans, maar het laer rang gehad in die ander twee tale.

5.6 RESULTATE

Die resultate vir die stelsel word opgesom in tabel 5.5. Die kolom vir die F_1-telling bevat die gemiddelde F_1-telling vir die binêre take, en die geweegde F_1-telling vir die fyn klassifikasietaak. Gebaseer op tellings in kruisvalidasie op die opleidingsdata, kies ons die ewekansigewoud-klassifiseerder wat met alle kenmerke opgelei is. Soortgelyke prestasie is egter bereik met die ander benaderings. Kenmerkseleksie soos uitgevoer op die opleidingsdata soos bo beskryf in afdeling 5.5 het nie merkbare verbetering gebring in die finale toetsstel vergeleke met 'n eenvoudige passing op die totale datastel met alle kenmerke nie.

Die basislyne wat vir die gedeelde taak verskaf is, is deur alle spanne (meestal gemaklik) oorskry, en word dus nie hier ingesluit nie. Vir meer besonderhede oor die basislyne, sien [13]. Om te help om die prestasie van ons stelsel te evalueer vergeleke met alternatiewe, word die resultate ingesluit van enige deelnemer aan die gedeelde taak waar hul weergawe beter pres-

Tabel 5.5: Evaluasieresultate vir ons stelsel. Die laaste kolom word as 'n persentasie van die totale segmente aangedui. Waar enige deelnemer aan die gedeelde taak (dikwels 'n vroeër weergawe van ons stelsel*) beter presteer het in 'n spesifieke maat, word dit in hakies aangedui vir vergelyking.

Subtaak	F_1-telling	Korrek geklassifiseer
Binêr I (en-de)	0,695 *(0,71)**	82,9% *(83,9%)**
Binêr I (en-es)	0,82	85,1% *(86,0%)*
Binêr I (en-it)	0,785	80,8%
Binêr II (en-de)	0,675 *(0,68)**	87,0% *(88,3%)**
Binêr II (en-es)	0,77 *(0,805)*	85,9% *(88,2%)*
Binêr II (en-it)	0,84 *(0,845)*	91,2%
Fyn klassifikasie (en-de)	0,80	83,1% *(83,4%)**
Fyn klassifikasie (en-es)	0,78 *(0,79)*	80,1%
Fyn klassifikasie (en-it)	0,76	77,3%

teer het in enige van die twee mate. In omtrent die helfte van hierdie gevalle was die stelsel met die beste prestasie 'n vroeër weergawe van ons stelsel.

Dit is opmerklik dat die stelsel van ander deelnemers wat die beste presteer het in die tweede binêre klassifikasietaak vir Engels–Spaans en Engels–Italiaans, kenmerke gebruik het wat op masjienvertaling gebaseer is. So 'n relatief duur kenmerk bring waarde, alhoewel ons stelsel (wat nié van masjienvertaling gebruik maak nie) steeds redelik kompeterend is in hierdie twee gevalle. Dit blyk dat dit dalk die moeite werd is om masjienvertaling te oorweeg waar die berekeningskoste aanvaarbaar is.

Aangesien die presiese resultate steeds afhang van saadjies vir ewekansige getalle in die algoritmes, is daar steeds 'n mate van onstabiliteit in die stelselprestasie. Dit lyk of die datastelle klein genoeg is dat kansgebeurtenisse die gewigte van die klassifiseerders op nietriviale maniere beïnvloed. Dit is ook hoekom ons nie konsekwent 'n vorige weergawe van ons stelsel kan klop nie — hier bo word resultate gerapporteer van 'n enkele stelsel met 'n konsekwente saadjie vir ewekansige getalle. Vergeleke met die ander benaderings wat by die gedeelde taak

ingedien is, wil dit voorkom asof ons stelsel veral goed doen vir die taalpaar Engels–Duits, asook vir die fyn klassifikasietaak.

In die geval van Duits is die beter prestasie waarskynlik toe te skryf aan die beter ondersteuning vir Duits in LanguageTool (sien tabel 5.2). Die grammatikakenmerke het konsekwent 'n hoë rang gekry vir Duits. Verder is daar in die vorige afdeling bespreek hoe die speltoetskenmerk belangriker was vir Duits as vir die ander tale. Die sinne in die Engels–Duitse data was heelwat langer as die ander (die Engelse bronteks het meer as 50% meer woorde gehad as die bronteks van die ander twee taalpare). Dit lyk asof die langer segmente meer geleenthede gebied het om kwaliteitsprobleme te bespeur met die spel- en grammatikatoetsers.

Die goeie prestasie in die fyn klassifikasietaak kan toegeskryf word aan die wye verskeidenheid kategorieë vir kenmerke in afwagting van die verskeie tipes foute. 'n Foutanalise dui egter daarop dat die herroeping van klas 2 (klein probleme) min is (onder 0,40). Die verwarring tussen die drie etikette word ook moontlik vererger deur hul ongebalanseerde verteenwoordiging, met klas 2 wat die kleinste teenwoordigheid het (sien tabel 5.1).

5.7 GEVOLGTREKKING

In hierdie hoofstuk is 'n stelsel voorgehou om foutiewe inskrywings in 'n vertaalgeheue te identifiseer. Dit is gebaseer op 'n masjienleerstelsel met 29 kenmerke wat deur middel van gekontroleerde leer opgelei word. Die kenmerke gebruik inligting oor lengte, eenvoudige toetse van konsekwentheid, spelling en grammatika, asook vanaf eksterne data. Al hierdie kenmerkkategorieë blyk effektief te wees. Ons stelsel was kompeterend in 'n gedeelde taak, en daaropvolgende verfynings het in groot mate die gaping tussen ons stelsel en die beste stelsels geelimineer.

Die ongebalanseerde klasverdeling veroorsaak dat daar baie min leergeleenhthede is in die onderverteenwoordigde klasse

(sien tabel 5.1). Daar sal weer in afdeling 6.2 aandag geskenk word aan hierdie saak.

Vir verskeie van die subtake beteken die akkuraatheid wat bereik is dat 'n stelsel maklik 1 uit elke 7 inskrywings verkeerd sal klassifiseer (afhangend van die presiese taak). Of sulke akkuraatheid aanvaarbaar is, sal afhang van die gebruiksgeval. In die volgende hoofstuk sal die gebruik van skoongemaakte datastelle in twee toepassings oorweeg en geëvalueer word.

6

EVALUASIE VAN SKONER VERTAALGEHEUES

In hoofstuk 5 is 'n benadering beskryf vir die identifisering van vuil inskrywings in 'n vertaalgeheue. Hierdie werk is geevalueer as 'n klassifikasieprobleem — die evaluasie bepaal hoe akkuraat die stelsel die vuil inskrywings geïdentifiseer het. Dit staan bekend as intrinsieke evaluasie. "'n Intrinsieke evaluasiemaat is een wat die kwaliteit van 'n model meet onafhanklik van enige toepassing"[1] [44]. In hierdie hoofstuk word die werk verder gevoer deur middel van evaluasie in toepassings — ekstrinsieke evaluasie.

In die literatuuroorsig (hoofstuk 2) is verskeie toepassings van parallelle korpusse genoem. Die ekstrinsieke evaluasie in hierdie hoofstuk word aan die hand van twee van hierdie toepassings gedoen, naamlik 'n vertaalgeheuestelsel en 'n masjienvertaalstelsel. Hiermee kan die waarde van 'n skoner korpus relatief tot die aanvanklike, vuil korpus aangedui word.

Hoeveel opleidingsdata is nodig om 'n masjienvertaalstelsel op te lei? In die masjienvertaalgemeenskap word daar algemeen aanvaar dat meer data beter resultate sal lewer. Dit is te verstane in dié sin dat meer opleidingsdata 'n beter dekking van die taal het en meer betroubare statistiese modellering van minder algemene taalverskynsels moontlik maak. Gevolglik is dit nodig dat 'n korpus 'n goeie dekking het van die woordeskat wat uiteindelik tydens vertaling gebruik gaan word.

Verskillende benaderings tot masjienvertaling reageer egter verskillend op verandering in die hoeveelheid opleidingsdata. In 'n vergelyking van statistiese masjienvertaling met neurale masjienvertaling vir die taalpaar Engels–Spaans [51] is aangedui dat die neurale stelsel eers kompeterend geword het met tweetalige opleidingsdata in die omgewing van 15 miljoen En-

1 "An intrinsitic evaluation metric is one which measures the quality of a model independent of any application."

gelse woorde. Die neurale stelsel kan beter gebruik maak van groter datastelle maar vaar aansienlik swakker met minder as 10 miljoen Engelse woorde in die tweetalige opleidingsdata. As die statistiese stelsel aangevul word met 'n eentalige domeinspesifieke korpus van twee miljard woorde, is die stelsels se prestasie eers vergelykbaar by ongeveer 100 miljoen Engelse woorde.

Wanneer vuil segmente in 'n tweetalige korpus geïdentifiseer en verwyder word, is daar dus die gevaar dat die skoner, kleiner korpus 'n minder kompeterende masjienvertaalstelsel tot gevolg sal hê. Die presiese aard van die kompromis tussen die kwaliteit en die grootte van die korpus is nie onmiddellik duidelik nie. Aangesien voorlopige resultate dui daarop dat neurale masjienvertaalstelsels meer sensitief is vir opleidingsdata van swak kwaliteit [18, 20, 51], is dit 'n aantreklike toepassingsgebied vir die evaluasie van die werk in hierdie proefskrif: Enersyds kan die verwydering van foutiewe opleidingsdata moontlik die kwaliteit verbeter, en andersyds kan minder opleidingsdata die kwaliteit kelder. In hierdie lig is dit nie triviaal om die rol wat die kwaliteit en die grootte speel onafhanklik van mekaar te ondersoek nie.

In hoofstuk 5 is daar aangedui dat ewekansige getalle 'n merkbare effek het op die klassifiseerders. Dit dui op onvoldoende geannoteerde data om betroubare statistiek te verkry tydens die opleiding van die klassifiseerder. 'n Groter geannoteerde datastel is dus nodig om meer vertroue te hê in die werking van die klassifiseerder in 'n grootskaalse toepassing.

Die gepaste opleidingsdata is egter oor die algemeen nie in enige grootte beskikbaar nie. Alhoewel 'n indruk van die kwaliteit van 'n korpus as geheel gevorm kan word, byvoorbeeld deur steekproefneming, of op grond van hoe dit tot stand gekom het, gee dit steeds nie die nodige inligting op segmentvlak nie. Daar kan nie aangeneem word dat alle inskrywings in 'n laekwaliteitvertaalgeheue ook as vuil beskou moet word nie. Eweneens is alle inskrywings in 'n hoëkwaliteitvertaalgeheue nie noodwendig perfek nie.

Die gekontroleerde benadering is dus beperkend nou dat daar na toepassings gekyk word.

6.1 GEKONTROLEERDE, SEMIGEKONTROLEERDE EN ONGEKONTROLEERDE LEER

Masjienleermetodes waar opleidingsdata met die gewenste uitkomste beskikbaar is (bv. vir regressie of klassifikasie), staan bekend as gekontroleerde leer. As die gewenste afhanklike veranderlike of klassifikasie nie bekend is nie, maar die data geanaliseer moet word om datapunte te groepeer of om die verband tussen veranderlikes op te spoor, word van ongekontroleerde leerbenaderings gebruik gemaak. Die term *semigekontroleerde leer** verwys na die gevalle wat nie presies in een van bogenoemde kategorieë pas nie, byvoorbeeld as daar 'n kleiner hoeveelheid geannoteerde data is wat gekombineer word met 'n ongeannoteerde datastel [42, p. 28].

(en) *semi-supervised learning*

Vervolgens word die gekontroleerde benadering aangepas om as semigekontroleerde of ongekontroleerde leermetode te funksioneer. Alhoewel dieselfde gekontroleerdeleeralgoritmes van hoofstuk 5 soos steunvektor-, ewekansigewoud- en logistieseregressie-klassifiseerders gebruik word, sal hier verwys word na 'n ongekontroleerde benadering as geen geannoteerde data nodig is nie. Hierdie aanpassing moet self ook geëvalueer word om die prestasie relatief tot die gekontroleerde benadering te evalueer. Dit bied ook 'n kans om die verband tussen die intrinsieke en ekstrinsieke evaluasie te ondersoek deur die aangepaste benadering wat in hierdie hoofstuk gebruik gaan word eers intrinsiek te evalueer. Hierdie intrinsieke evaluasie van die aangepaste metode volg in afdeling 6.3. Die afdelings daarna bied die ekstrinsieke evaluasie aan.

6.2 AANPASSING VIR ONGEKONTROLEERDE LEER

As eerste stap in die aanpassing om sonder geannoteerde data te werk, word die moontlikheid ondersoek om met min op-

leidingsdata te werk. Hierdie opset is in werklikheid wat in hoofstuk 5 die geval was. Die ongebalanseerde klasverdeling is veral problematies — daar is min opleidingsdata in die onderverteenwoordigde klasse. (Sien tabel 5.1.) Die probleem met balans is ook realisties in enige opset waar 'n vertaalgeheue van redelike kwaliteit is — die meeste inskrywings behoort korrek te wees. As die tegniek met min geannoteerde data kan werk en die probleem met ongebalanseerde klasverdeling opgelos kan word, is 'n semigekontroleerde benadering 'n moontlikheid.

Die hantering van ongebalanseerde klasverdeling kan op 'n verskeidenheid maniere hanteer word. Leertegnieke, optimeringsdoelwitte en opleidingsdata kan aangepas word om meer gepas te wees in so 'n opset. Hersteekproefneming* kan die klasverdeling manipuleer deur datapunte van die groter klas te ignoreer of datapunte van die kleiner klas te dupliseer. SMOTE [19] en ADASYN [38] is tegnieke wat kunsmatig datapunte in die kleiner klas skep deur versteuring* van die kenmerke van bestaande datapunte in dié klas. Dit is dus 'n meer gevorderde voorbeeld van hersteekproefneming waar datapunte van die kleiner klas by die opleidingsdata gevoeg word. Hierdie twee tegnieke hou belofte in omdat 'n klein hoeveelheid geannoteerde opleidingsdata dalk voldoende sal wees. So kan die benadering as 'n semigekontroleerde leermetode optree.

(en) *resampling*

(en) *perturbation*

Eksperimente met SMOTE en ADASYN het egter nie bruikbare resultate gelewer nie. Dit het wel gedeeltelik die inspirasie verskaf vir die volgende metode wat ontwikkel is.

SMOTE en ADASYN werk deur slegs die kenmerkwaardes te manipuleer. Dit het dus geen kennis van die probleemdomein self nie. Alhoewel daar in natuurliketaalverwerking sukses behaal is met taalgenerasie d.m.v. taalmodelle, is dit steeds makliker om ongeldige teks te skep as om met sekerheid 100% geldige teks te skep. Met hierdie intuïsie in gedagte kan opleidingsdata in die foutiewe klasse maklik geskep word. Alhoewel dit triviaal is om blatante onsin as opleidingsdata vir die klassifiseerder te genereer, word daar van die klassifiseerder verwag om ook 'n redelike oordeel te vel in grensgevalle.

Daarom is dit nodig om tydens opleiding nie net blatante onsin as voorbeelde van foutiewe vertalings te hê nie.

Die benadering is dus om die korrekte vertalings (klas 1) te gebruik en die doelteks te versteur.* 'n Nuwe, foutiewe segment word gegenereer deur 'n korrekte segmentpaar te manipuleer met 'n funksie $g(s,t)$. 'n Kunsmatige fout word deur $g(s,t)$ aangebring om te verseker dat die segment se doelteks nou 'n spesifieke fout bevat. Die aangevulde data word dan gebruik soos vantevore, en die kenmerke word bereken sonder enige versteuring.

(en) *perturb*

Een van 18 funksies word ewekansig gekies om by elke segment as $g(s,t)$ te dien. Funksies is beskikbaar wat elkeen een van die volgende foute maak en gee ook telkens 'n skatting van die ernstigheid daarvan. Slegs die doelteks t word gewysig.

- Verwyder 'n letter — gering.
- Voeg 'n letter by — gering.
- Ruil een paar letters — gering.
- Ruil vier pare letters — ernstig.
- Verwyder 'n woord — ernstig.
- Verdubbel 'n woord — gering.
- Dupliseer die hele doelsegment — ernstig.
- Verwyder die eerste helfte van die doelsegment — ernstig.
- Verwyder die laaste helfte van die doelsegment — ernstig.
- Plaas 'n ewekansige getal vooraan die segment — gering.
- Vervang 'n woord met 'n soortgelyke woord wat nie 'n spelfout is nie — ernstig.
- Verwyder 'n hakie of voeg een by — gering.
- Verwyder 'n aanhalingsteken of voeg een by — gering.
- Verwyder die leesteken aan die einde van die segment as daar een was, of voeg andersins een by — gering.

- Verwyder die spasie aan die einde van die segment as daar een was, of voeg andersins een by — gering.
- Wissel die hooflettergebruik van die eerste letter in die segment — gering.
- Hergebruik die bronsegment onveranderd as doelsegment — ernstig.
- Indien daar nie 'n URL is nie, voeg 'n URL by, andersins verwyder die karakters ": //" — ernstig.

As die getal gewenste foutiewe segmente minder as die gegewe opleidingsdata is, word die getal unieke datapunte wat nodig is ewekansig* uit die gegewe opleidingsdata getrek. Indien meer foutiewe segmente verlang word, word 'n steekproef met terugplasing* geneem vir die gewenste getal segmente.

(en) *randomly*

(en) *sample with replacement*

Masjienleertegnieke soos dié wat in hoofstuk 5 bespreek is, word beïnvloed deur die balans van die klasse in die opleidingsdata. Opleidingsdata met aansienlik meer gevalle in klas x kan lei tot oormatige passing en 'n klassifiseerder wat met groter waarskynlikheid ongesiene toevoer in klas x plaas. Aangesien die hoeveelheid foutiewe segmente in die aangepaste metode willekeurig as 'n parameter beheer kan word, is dit moontlik om die stelsel se gedrag hiermee te manipuleer en oormatige passing te voorkom. Saam met die klassifikasie van foute as gering of ernstig, is daar dus maniere om die stelsel aan te pas volgens veranderende behoeftes.

Vergeleke met 'n vorige studie waar 'n gekontroleerde benadering tot hierdie probleem aangepas is [63], verskaf hierdie benadering aanpassingsmoontlikhede wat minder afhang van die opleidingsdata, aangesien die foutiewe datapunte nie uit die data self onttrek word nie. Dit bied dus die geleentheid dat 'n korpus van hoë gehalte gebruik word om die klassifiseerder mee op te lei, sodat die aanname dat alle oorspronklike inskrywings in die korpus korrek is, redelik juis is. Dit is ook 'n moontlike beperking: die sukses van hierdie benadering berus in 'n mate op die beskikbaarheid van 'n hoëkwaliteitvertaalgeheue.

Die benadering hier is ook meer aanpasbaar in dié sin dat die fyn klassifikasie steeds moontlik is, teenoor die vorige aanpassing vir ongekontroleerde leer wat slegs tot 'n binêre klassifiseerder gelei het [63].

Die aangepaste metode in hierdie hoofstuk kan ook 'n kleiner hoeveelheid geannoteerde data aanvullend tot die opleidingsdata gebruik. Dit is dus aanpasbaar om as semigekontroleerde leertegniek gebruik te word. In die res van hierdie hoofstuk word dit vir die meer algemene, moeiliker geval gebruik waar geen geannoteerde data beskikbaar is nie.

6.3 INTRINSIEKE EVALUASIE

Die klassifiseerder wat nou aangepas is om sonder geannoteerde data te werk, kan intrinsiek geëvalueer word, soos in hoofstuk 5. Alhoewel die uiteindelike doel in hierdie hoofstuk is om ekstrinsiek te evalueer, gee 'n intrinsieke evaluasie die geleentheid om die klassifikasiekwaliteit van die aangepaste benadering te vergelyk met die oorspronklike weergawe wat geannoteerde data gebruik het. Die evaluasie-eksperiment van hoofstuk 5 word nou hier herhaal met die aangepaste metode sodat die resultate vergelykbaar is.

Vir hierdie evaluasie word soveel foutiewe inskrywings kunsmatig geskep as wat in die opleidingsdata is. Die getal inskrywings wat kunsmatig geskep word, sou as 'n parameter in sigself ondersoek kon word. Die keuse om soveel te genereer as wat in die opleidingsdata is, is om die twee foutiewe klasse (klas 2 en 3) dieselfde gewig as die korrekte klas (klas 1) te gee. So is die klasverdeling meer gebalanseerd terwyl die klassifiseerder steeds 'n sterker voorkeur vir klas 1 sal hê. Hierdie verdeling van klasse behoort steeds oormatige passing te voorkom.

6.3.1 *Data en eksperiment*

In hierdie afdeling word die aangepaste stelsel geëvalueer soos die stelsel in hoofstuk 5 geëvalueer is. Dieselfde opleidingsdata en toetsstel word weer hier gebruik. Twee stelsels word voorgehou: (1) een wat vanaf 'n 100% skoon opleidingstel die kunsmatige opleidingsdata skep en (2) een wat 'n vuil datastel hiervoor gebruik.

Tabel 6.1 bevat die resultate van die eksperiment waar slegs inskrywings uit klas 1 uit die geannoteerde opleidingsdata gebruik word. Inskrywings in klasse 2 en 3 word weggelaat. Dit verteenwoordig 'n situasie waar 'n korpus van hoë kwaliteit beskikbaar is, maar nie 'n geannoteerde stel met die drie klasse nie.

Tabel 6.1: Evaluasieresultate wanneer slegs opleidingsegmente uit klas 1 gebruik word. Die prestasie van die oorspronklike gekontroleerde stelsel is in hakies.

Subtaak	F_1-telling	Korrek geklassifiseer
Binêr I (en-de)	0,75 (0,695)	83,6% (82,9%)
Binêr I (en-es)	0,77 (0,82)	81,2% (85,1%)
Binêr I (en-it)	0,75 (0,785)	78,0% (80,8%)
Binêr II (en-de)	0,53 (0,675)	85,7% (87,0%)
Binêr II (en-es)	0,705 (0,77)	84,3% (85,9%)
Binêr II (en-it)	0,57 (0,84)	81,8% (91,2%)
Fyn klassifikasie (en-de)	0,75 (0,80)	78,4% (83,1%)
Fyn klassifikasie (en-es)	0,73 (0,78)	74,6% (80,1%)
Fyn klassifikasie (en-it)	0,71 (0,76)	73,9% (77,3%)

Volgende is al die opleidingsdata gebruik met die foutiewe aanname dat alle inskrywings korrek is — dus in klas 1. Die resultate is in tabel 6.2. Dit verteenwoordig 'n situasie waar 'n korpus van gemengde kwaliteit gebruik word om die kunsmatige opleidingsdata mee te skep. In so 'n geval sal die klassifiseerder mislei word deurdat alle foutiewe inskrywings in die oorspronklike opleidingsdata as korrek beskou word en hulle versteurde waardes as foutief beskou word.

Tabel 6.2: Evaluasieresultate waar *verkeerdelik* aangeneem word dat alle opleidingsegmente in klas 1 (heeltemal korrek) is. Die prestasie van die oorspronklike gekontroleerde stelsel is in hakies.

Subtaak	F_1-telling	Korrek geklassifiseer
Binêr I (en-de)	0,67 (0,695)	81,6% (82,9%)
Binêr I (en-es)	0,65 (0,82)	75,2% (85,1%)
Binêr I (en-it)	0,66 (0,785)	72,0% (80,8%)
Binêr II (en-de)	0,505 (0,675)	85,6% (87,0%)
Binêr II (en-es)	0,47 (0,77)	77,6% (85,9%)
Binêr II (en-it)	0,50 (0,84)	80,4% (91,2%)
Fyn klassifikasie (en-de)	0,72 (0,80)	77,1% (83,1%)
Fyn klassifikasie (en-es)	0,60 (0,78)	68,7% (80,1%)
Fyn klassifikasie (en-it)	0,56 (0,76)	64,0% (77,3%)

6.3.2 *Bespreking*

Die goeie prestasie vir die taalpaar Engels–Duits by die eerste binêre taak, veral die F_1-telling, is opvallend. In die een geval is dit selfs beter as die aanvanklike gekontroleerde benadering. 'n Moontlike verklaring is dat die onderverteenwoordiging van die foutiewe klas in hierdie geval te min opleidingsdata verskaf het vir die oorspronklike gekontroleerde benadering. Die telling wat met die aangepaste metode behaal is, is vergelykbaar met die ander taalpare in dieselfde taak. Met die oorspronklike benadering was die tellings hier laer as die ander taalpare.

By die fyn klassifikasietaak is die akkuraatheid laag, veral in tabel 6.2. Die prestasie is hier naby aan 'n basislynklassifiseerder wat bloot alles aan klas 1 toeken. Die foutiewe klas 1-aanname aangaande die opleidingsdata wat in werklikheid in die foutiewe klasse is (klasse 2 en 3) affekteer veral hier die resultate negatief.

Soos wat die geval was met die geannoteerde data, presteer die aangepaste stelsel beter in die taalpaar Engels–Duits. Moontlike redes vir die beter prestasie in hierdie taalpaar is reeds in afdeling 5.6 bespreek, onder andere die meer volwasse steun wat die grammatikatoetser vir Duits bied vergeleke met

Spaans of Italiaans. In die volgende afdelings word die stelsel voortaan slegs in die taalpaar Engels–Duits geëvalueer.

6.4 EKSTRINSIEKE EVALUASIE: VERTAALGEHEUESTELSEL

In hoofstuk 3 is 'n evaluasiemetode vir vertaalgeheuestelsels aangebied. In hierdie afdeling word hierdie evaluasiemetode ingespan om die effek van die skoner datastel relatief tot die oorspronklike, vuil datastel te evalueer binne die opset van 'n vertaalgeheuestelsel. In hierdie evaluasie beskou die aangepaste klassifiseerder alle inskrywings met foute (gering en ernstig) as vuil. Dit stem dus ooreen met 'n strenger of meer nougesette siening van korrektheid.

6.4.1 *Data en eksperiment*

Die DGT-korpus [76] is vroeër in hoofstuk 3 bespreek. Die webwerf van die direkteur-generaal van vertaling van die EU noem dit die "DGT-Translation Memory".[2] Daar is dus voorsien dat hierdie hulpbron as 'n vertaalgeheue gebruik sou word.

Die korpus is aangebied met "die aanname dat die gemiddelde vertaalkwaliteit in die DGT-korpus van 'n baie hoë standaard is"[3] [76, p. 456], maar daar is ook al bevind dat daar wel belyningsfoute is [86]. Dit dien dus as 'n goeie vertaalgeheue vir hierdie eksperiment.

In hoofstuk 3 is k-voudige kruisvalidasie gebruik om die betroubaarheid van die resultate te verhoog. In hierdie afdeling word die vertaalgeheue verklein deur foutiewe inskrywings te verwyder, dus sal die toetsstel in elke vou van die k-voudige kruisvalidasie nie identies wees in die twee dele van die eksperiment nie. In hierdie afdeling kan die vertaalgeheue dus nie op hierdie manier gebruik word nie.

2 https://ec.europa.eu/jrc/en/language-technologies/dgt-translation-memory

3 "... it can be assumed that the translation quality in DGT-TM on average is of a very high standard."

Die tekste van die EU-grondwet is saamgestel tot 'n parallelle korpus van 21 tale.[4] Vir die taalpaar Engels–Duits beslaan dit 8771 segmente en dien in hierdie eksperiment as die toetsstel. Die DGT-korpus dien as gepaste vertaalgeheue by vertaling van die EU-grondwet, aangesien die twee korpusse dieselfde organisasie as oorsprong het, en die DGT-korpus heelwat regstekste bevat.

Die uitslag in hoofstuk 4 dui op **edit3word** as die soortgelykheidsmaat met die sterkste verband met redigeertyd. In hierdie afdeling gebruik ons **edit3word** en enkele ander soortgelykheidsmate wat ook 'n sterk verband met redigeertyd gehad het.

Die eksperiment gebruik die volgende mate vir onttrekking en evaluasie in die evaluasiemetode van hoofstuk 3:

- Dice-koëffisiënt op woordvlak (**diceword**).
- 3-bewerkingredigeersoortgelykheid op karaktervlak (**edit3**).
- 3-bewerkingredigeersoortgelykheid op woordvlak (**edit3word**).
- 4-bewerkingredigeersoortgelykheid op karaktervlak (**edit4**).

Die eerste drie het elkeen 'n sterk verband getoon met die tyd per segment (sien tabel 4.2) Laasgenoemde is 'n algemene basislyn en die verstekmaat vir onttrekking in die vertaalgeheuestelsel. Saam verteenwoordig hierdie mate vier uiteenlopende kyke op soortgelykheid, maar met bewese verband met die redigeertyd.

Die soortgelykheidsdrempel is verfyn op een helfte van die toetsdata, soos voorgestel in hoofstuk 3, deur te optimeer vir die F_1-telling. Die evaluasietellings is dus bereken op die oorblywende helfte van die toetsstel.

Tabelle 6.3 en 6.4 bevat die resultate vir alle kombinasies van onttrekkings- en evaluasiemate. Elke ry verteenwoordig 'n toetsstel wat met een onttrekkingsmaat gegenereer is. Elke kolom verteenwoordig evaluasie met een evaluasiemaat.

4 http://opus.lingfil.uu.se/EUconst.php

Tabel 6.3: F_1-tellings tydens evaluasie op die EU-grondwet. Vertaalgeheue is die *oorspronklike* DGT-subversameling.

	Evaluasiemate			
Onttrekkingsmate	diceword	edit3	edit3word	edit4
diceword	.070	.087	.067	.051
edit3	.081	.170	.075	.093
edit3word	.068	.086	**.061**	.050
edit4	.088	.159	.086	.082

Tabel 6.4: F_1-tellings tydens evaluasie op die EU-grondwet. Vertaalgeheue is die *skoner* DGT-subversameling.

	Evaluasiemate			
Onttrekkingsmate	diceword	edit3	edit3word	edit4
diceword	.126	.108	.123	.074
edit3	.104	.182	.104	.109
edit3word	.122	.106	**.123**	.073
edit4	.110	.176	.109	.096

6.4.2 *Bespreking*

Die konsekwente resultaat is dat elke evaluasie-uitkoms in elkeen van die toetsstelle verbeter het. Al die geëvalueerde stelle (rye van die tabelle) toon 'n verbetering onder al die evaluasiemate (kolomme van die tabelle). Waar **edit3word** vir onttrekking en evaluasie gebruik is (met vetdruk aangedui), is 'n groter verbetering sigbaar as by die ander kombinasies. Die onderlinge sydigheid wat die karakter- en woordgebaseerde metodes vir mekaar het, is weer hier duidelik, soos in hoofstuk 3.

Aangesien 'n verkleinde vertaalgeheue nie méér voorstelle kan gee as die oorspronklike nie, bied die skoner vertaalgeheue minder voorstelle en voorstelle met laer soortgelykheid aan die bronteksnavraag. Ten spyte hiervan bied dit meer waarde vir die vertaler in die opstel van die finale doelteks. Die verbetering is sigbaar ongeag watter onttrekkings- en evaluasiemaat gebruik word.

6.5 EKSTRINSIEKE EVALUASIE: MASJIENVERTAALSTELSEL

Die evaluasie van masjienvertaalstelsels is volwasse. In hierdie afdeling word die aangepaste metode ingespan om die effek van die skoner datastel relatief tot die oorspronklike, vuil datastel te evalueer binne die opset van 'n masjienvertaalstelsel. Hierdie afdeling bevat evaluasies van die toegeeflike en die streng klassifiseerders. Daar word ook ondersoek ingestel na opleidingsdata met verskillende aanvanklike vlakke van kwaliteit.

6.5.1 *Data en eksperiment*

In hierdie afdeling word 'n neurale masjienvertaalstelsel opgelei en geëvalueer op 'n standaardtoetsstel. Ons gebruik die ontwikkelingstel* en toetsstel van die nuusvertaaltaak by die werkswinkel vir masjienvertaling van 2016.[5]

(en) *development set*

5 http://www.statmt.org/wmt16/translation-task.html

Die EU-boekwinkelkorpus[6] [72] is 'n veeltalige parallelle korpus wat gebou is van publikasies uit die EU-boekwinkel[7] — 'n webwerf met publikasies van die Europese Unie. Die boekwinkelkorpus is onttrek uit PDF-dokumente, en daar was heelwat kwaliteitsprobleme om die hoof te bied tydens die samestelling. 'n Handmatige evaluasie van 'n steekproef van 200 segmente vir die taalpaar Engels–Letties dui op slegs ongeveer 60% van die segmente wat wedersydse vertalings is. Daar is dus aansienlike hoeveelheid vuil segmente in die datastel.

Die EU-boekwinkelkorpus is groot volgens die meeste hedendaagse standaarde vir parallelle korpusse: vir die meeste taalpare is daar meer as 100 miljoen woorde in meer as 'n miljoen belynde sinne. Dit is dus 'n aantreklike korpus vir sy grootte, maar die kwaliteit is nie ideaal nie. Hierdie korpus dien dus as 'n gepaste toetsgeval vir die evaluasie van 'n skoonmaakbenadering in 'n neurale masjienvertaalstelsel. Die getal segmente is in die ordegrootte wat nodig is om 'n kompeterende neurale masjienvertaalstelsel mee op te lei, maar ook op 'n stadium in die groeikurwe waar meer data nog 'n merkbare verskil behoort te maak.

(en) *News Commentary Parallel Corpus*

Die parallelle nuuskommentaarkorpus* is 'n veeltalige parallelle korpus wat geskep is as opleidingsdata vir die werkswinkel vir masjienvertaling. Dit bevat politieke en ekonomiese kommentaar vanaf die webwerf Project Syndicate.[8] Die weergawe wat hier gebruik word, is afkomstig van die Opus-projek.[9] Die Engelse kant van die Engels–Duitse taalpaar bevat 223 153 segmente wat 4 923 246 woorde bevat.[10] Dit is dus kleiner as die EU-boekwinkelkorpus, maar is 'n in-domein-korpus vir die toetsstel wat ook nuusartikels is. 'n Evaluasie van masjienvertaling met hierdie twee datastelle bied dus twee uiteenlopende toetsgevalle vir die ekstrinsieke evaluasie in hierdie hoofstuk.

6 http://opus.lingfil.uu.se/EUbookshop.php
7 http://bookshop.europa.eu/
8 https://www.project-syndicate.org/
9 http://opus.lingfil.uu.se/News-Commentary11.php
10 Woordtellings word slegs as aanduiding van grootte gegee. In hierdie hoofstuk is dit deurgaans met GNU "wc" bereken.

OpenNMT is 'n masjienvertaalstelsel wat 'n neurale benadering volg [48].[11] Dit implementeer die enkodeerder–dekodeerder-model met aandag [4]. OpenNMT word, sover moontlik, met die verstekwaardes vir instellings gebruik. Opsommenderwys behels dit die volgende:

- Die enkodeerder en dekodeerder is tweevlak-RNN-netwerke met LSTM-selle. Die woordvektorgrootte* is 500 vir bron- en doeltale.
  (en) *word embedding size*
- Opleiding gebruik SGD* as optimeringsmetode, kruisentropie as optimeringsdoelwit en 'n uitsluitwaarskynlikheid* van 0,3.
  (en) *stochastic gradient descent*
  (en) *dropout probability*
- Die dekodeerder gebruik globale aandag.*
  (en) *global attention*
- Vertaling geskied deur middel van 'n straalsoektog* met 'n straalgrootte van 5.
  (en) *beam search*

Data is voorberei met afsonderlike greeppaarenkodering* in elke taal met 30 000 stappe. Opleiding is volgehou totdat die kruisentropie nie meer verbeter het nie. In die geval van die nuuskommentaarkorpus was dit vir tussen 14–18 epogge vir die drie eksperimente.
(en) *byte-pair encoding, BPE*

Evaluasie is gedoen volgens die BLEU-mate deur middel van `multi-bleu.perl`.[12] Die ARK-programme[13] is gebruik om te toets vir statistiese beduidendheid by p=0,05.

As basislyn is die stelsel opgelei met twee miljoen segmente uit die EU-boekwinkelkorpus. Hierdie korpus is daarna twee keer afsonderlik skoongemaak met die toegeeflike en streng klassifiseerders. Aparte stelsels is opgelei met die twee skoongemaakte korpusse.

Dieselfde stappe is gevolg met die nuuskommentaarkorpus. Dieselfde instellings is ook gebruik vir die klassifiseerder.

11 http://opennmt.net/
12 https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl
13 https://www.cs.cmu.edu/~ark/MT/

Tabel 6.5: Masjienvertaalresultate van die EU-boekwinkelkorpus. *Statisties beduidend by p=0,05

Datastel	Segmente	Woorde	BLEU
Vuil (basislyn)	2 000 000	50 104 123	8,12
Skoner (toegeeflik)	1 780 674 (89,0%)	43 795 526	**9,18***
Skoner (streng)	1 476 796 (73,8%)	38 208 030	9,10*

Tabel 6.6: Masjienvertaalresultate van die nuuskommentaarkorpus. *Statisties beduidend by p=0,05

Datastel	Segmente	Woorde	BLEU
Vuil (basislyn)	223 153	4 923 246	17,15
Skoner (toegeeflik)	221 503 (99,3%)	4 888 944	**17,77***
Skoner (streng)	213 939 (95,9%)	4 719 179	16,92

6.5.2 *Bespreking*

Die evaluasie-uitslae van bogenoemde eksperimente word in tabelle 6.5 en 6.6 aangedui. Die grootte van elke opleidingstel word aangedui met die getal segmente en met die getal woorde aan die Engelse kant van die opleidingsdata.

Soos verwag uit wat bekend is oor die kwaliteit van die boekwinkelkorpus is daar baie meer vuil inskrywings geïdentifiseer (11% en 26,2%) as in die nuuskommentaarkorpus. Die nuuskommentaarkorpus is duidelik 'n korpus van hoër kwaliteit, met slegs 4,1% van die segmente wat verwyder is deur die streng klassifiseerder.

Met al twee stelle opleidingsdata word die basislyn (die vuil datastel) oorskry met skoner opleidingsdata. Die lae tellings by die boekwinkelkorpus kan daaraan toegeskryf word dat die boekwinkelkorpus minder gepas is vir die vertaling van die nuusmateriaal in die toetsstel. Met hierdie opleidingstel het beide die toegeeflike en streng klassifiseerders 'n verbetering gebring. In die geval van die nuuskommentaarkorpus het slegs die toegeeflike klassifiseerder tot beter resultate gelei. Die kleiner verbetering in hierdie geval (17,15 na 17,77 BLEU-punte) is egter ook statisties beduidend. Hierdie meer beskeie verbete-

ring is steeds noemenswaardig as in ag geneem word dat slegs 0,7% van die opleidingsdata verwyder is.

6.6 GEVOLGTREKKING

In 'n poging om 'n skoonmaakbenadering daar te stel vir praktiese gebruik, is die metode van hoofstuk 5 aangepas in hierdie hoofstuk. Die vereiste van geannoteerde opleidingsdata is hierdeur opgehef. Deur self foutiewe inskrywings kunsmatig te genereer, bied die aangepaste metode boonop beheer oor die klasverdeling in die opleidingsdata. Hierdie aanpassing vir ongekontroleerde leer is suksesvol en, in intrinsieke evaluasie-uitslae, benader dikwels die resultate wat met gekontroleerde leer behaal is.

Die skoongemaakte DGT-korpus het beter resultate gelewer in 'n vertaalgeheuestelsel, ten spyte daarvan dat die vertaalgeheue waaruit voorstelle aangebied word, verklein is. Die soortgelykheidsmate met die sterkste verband met tyd is vir die evaluasie gebruik en het konsekwent op 'n verbetering gedui.

Die hoeveelheid foutiewe inskrywings wat verwyder is uit die boekwinkelkorpus en die nuuskommentaarkorpus strook met wat bekend is oor die kwaliteit van hierdie twee hulpbronne. Die groot dele van die boekwinkelkorpus wat verwyder is (11% en 26,2%) verteenwoordig miljoene woorde se opleidingsdata. Met so 'n groot vermindering in opleidingsdata word 'n verswakking van resultate normaalweg verwag.

Alhoewel die skoner datastelle minder opleidingsdata bied, lei hulle tot beter resultate. Dit is egter nie 'n geval dat minder data noodwendig beter is nie. Inteendeel, slegs die toegeeflike klassifiseerder se afvoer het konsekwent beter resultate as die basislyn gelewer.

Die resultate hier bevestig vorige werk wat aandui dat neurale masjienvertaalstelsels sensitief is vir die kwaliteit van opleidingsdata [18, 20, 51]. Die skoner korpusse het 'n statisties beduidende verbetering in BLEU-telling te weeg gebring, en

opleiding van die stelsels is vinniger in ooreenstemming met die kleiner opleidingstelle.

Die ekstrinsieke evaluasie bevestig die waarde van die skoonmaakbenadering, soos aangedui in hierdie praktiese toepassings.

7

SLOT

Met hierdie proefskrif is aangedui hoe ’n vertaalgeheue skoner en kleiner gemaak kan word. Hiermee is ’n waardevolle bydrae gelewer tot die bestuur van vertaalgeheues deurdat foutiewe inskrywings outomaties geïdentifiseer word.

Hierdie proefskrif bevat bydraes van metodologiese sowel as praktiese waarde. Volledig outomatiese evaluasie vereenvoudig eksperimentele werk, en ’n verbetering in die betroubaarheid van hierdie resultate verhoog die vertroue wat mens in die resultate kan hê. Skoner en kleiner vertaalgeheues lei tot beter resultate in ’n vertaalgeheuestelsel en ’n masjienvertaalstelsel. Vier gepubliseerde artikels het reeds uit die navorsing vir hierdie proefskrif voortgevloei [85–88].

Hierdie laaste hoofstuk bied ’n samevatting van hierdie bydraes, asook voorstelle vir toekomstige werk. Die bydraes van hierdie studie word onder andere aangebied aan die hand van die navorsingsvrae van afdeling 1.2.

7.1 BYDRAES

Hoe kan vertaalgeheues en vertaalgeheuestelsels geëvalueer word?

Die evaluasie van ’n vertaalgeheuestelsel kán met behulp van ’n soortgelykheidsmaat geskied, maar sorg moet gedra word om die probleem van soortgelykheidsmate se sydigheid die hoof te bied. Volgens die werk in hoofstuk 4 is woordgebaseerde 3-bewerkingredigeersoortgelykheid die beste keuse hiervoor, omdat dit die sterkste verband het met die tyd per segment. Die modellering van die soortgelykheidsdrempel f is van kardinale belang sodat evaluasie-eksperimente groter ooreenstemming sal hê met programme vir rekenaargesteunde vertaling wat so ’n drempel implementeer.

Watter tegniek(e) kan gebruik word om foutiewe inskrywings in 'n vertaalgeheue te identifiseer?

Gevestigde masjienleertegnieke soos steunvektor-, ewekansigewoud- en logistieseregressie-klassifiseerders kan suksesvol gebruik word vir die identifisering van foutiewe inskrywings. 'n Wye verskeidenheid leerkenmerke is belangrik sodat 'n wye verskeidenheid foute geïdentifiseer kan word. Deur opleidingsdata vir die foutiewe klasse outomaties te genereer, is meer beheer moontlik oor die verdeling van klasse (korrek of foutief), en is geannoteerde data minder noodsaaklik, al word daar van gekontroleerde leermetodes gebruik gemaak. So kan die metode as 'n ongekontroleerde benadering funksioneer.

Is 'n skoongemaakte vertaalgeheue meer geskik vir 'n spesifieke taak as vantevore?

'n Skoongemaakte vertaalgeheue is meer geskik in 'n vertaalgeheuestelsel en ook as die opleidingsdata vir 'n neurale masjienvertaalstelsel. Die verbetering is duidelik ten spyte van 'n vermindering in die data. Dit bevestig die goeie resultate wat met intrinsieke evaluasie in die gedeelde taak behaal is. (Sien hoofstuk 5 asook [12].)

7.2 BEPERKINGE EN TOEKOMSTIGE WERK

Die evaluasie van vertaalgeheuestelsels is 'n goeie hupstoot gegee met die werk in hoofstukke 3 en 4. Enkele beperkings van die werk in hoofstuk 4 is reeds genoem, en dit sal belangrik wees om dié beperkings in die toekoms te ondersoek:

- Slegs een taalpaar is ondersoek. Na my beste wete is daar nie gepaste datastelle beskikbaar vir 'n soortgelyke studie in ander taalpare nie. Dus sou selfs die skep van 'n soortgelyke datastel vir 'n ander taalpaar van waarde wees.

- Die datastel is kleiner as vergelykbares in die masjienvertaalgemeenskap.

- Alhoewel daar inskrywings met kort brontekste was (een of twee woorde), het min van hierdie inskrywings voorstelle gehad met soortgelykheid onder 90%.

In hoofstuk 4 is die verband tussen soortgelykheidsmate en tyd ondersoek. Die verband tussen soortgelykheidsmate en ander uitkomste soos finale kwaliteit of gebruikertevredenheid kan 'n alternatiewe aansig gee op die waarde van verskillende soortgelykheidsmate vir evaluasie. 'n Geweegde gemiddeld tussen tyd en hierdie ander aspekte sou moontlik 'n bruikbare afhanklike veranderlike wees in die regressie. Weens die sydigheid wat aangedui is, behoort enige gepaste soortgelykheidsmaat vir evaluasie ook oorweeg te word vir onttrekking.

Verskeie parameters vir die ongekontroleerde benadering van hoofstuk 6 kan verder in detail ondersoek word:

- Die aantal gegenereerde foutiewe datapunte.
- Die presiese balans tussen klasse.
- Die presiese balans tussen verskillende foute wat gegenereer word.
- Die moontlikheid om tegnieke soos SMOTE aanvullend te gebruik.

Die waarde van die gebruikte kenmerke is aangedui as deel van die intrinsieke evaluasie in hoofstuk 5 en dieselfde kenmerke is gebruik by die ekstrinsieke evaluasie in toepassings in hoofstuk 6. 'n Ondersoek na kenmerkseleksie vir spesifieke toepassings of datastelle kan moontlik nog beter resultate lewer.

7.3 BREËRE BETEKENIS VAN DIE WERK

'n Besondere bydrae van hierdie proefskrif is metodologies van aard, naamlik die verbeterde evaluasiemetode vir vertaalgeheuestelsels. Verskeie metodologiese probleme wat in vorige benaderings bestaan het, is die hoof gebied. Dié metode gee akademiese en kommersiële navorsers 'n manier om meer krities te kyk na soortgelykheidsmate en hoe die voorstelle uit

'n vertaalgeheue se waarde geskat word. Deur sonder meer soortgelykheid tussen 'n voorstel se doelteks en 'n verwysingsvertaling te toets en tot gevolgtrekkings te kom, laat nie reg geskied aan die probleem van sydigheid nie. Stelselprestasie tydens evaluasie moet in verband gebring word met 'n optimeringsdoelwit met relevansie tot vertalers, soos tyd. Hiermee sal die evaluasie van meet af aan te make hê met die behoeftes en wense van vertalers, eerder as om 'n soortgelykheidsmaat in isolasie te optimeer.

Alhoewel die kwaliteit van korpusse as baie belangrik gesien word in korpustaalkunde en vertaalstudies, het dit nie veel aandag geniet in die veld van vertaalgeheues en masjienvertaling nie. Die definitiewe verbetering in die resultate by hierdie toepassings sonder dat enige van die implementasiebesonderhede van die stelsels self verander het, dui op die belangrikheid van datakwaliteit in hierdie toepassings. Hierdie verbetering is behaal ten spyte van kleiner datastelle wat bowendien eksperimentele werk kan versnel.

Die bydraes van hierdie proefskrif is van toepassing op alle vertalers wat vertaalgeheuestelsels gebruik, en dit is ook besonder relevant vir die ontwikkelaars van hierdie stelsels.

BIBLIOGRAFIE

[1] Alves, Fabio: *Tradução, cognição e tecnologia: investigando a interface entre o desempenho do tradutor e a tradução assistida por computador.* Cadernos de tradução, 2(14):185–209, 2004. https://periodicos.ufsc.br/index.php/traducao/article/download/6481/5975.

[2] Arčan, Mihael, Marco Turchi, Sara Tonelli en Paul Buitelaar: *Enhancing statistical machine translation with bilingual terminology in a CAT environment*. In *Proceedings of AMTA 2014*, volume 1, bladsye 54–68, 2014. http://www.mt-archive.info/10/AMTA-2014-Arcan.pdf.

[3] Azzano, Dino: *Placeable and localizable elements in translation memory systems*. PhD-proefskrif, Ludwig-Maximilians-Universität München, 2011. https://edoc.ub.uni-muenchen.de/13841/2/Azzano_Dino.pdf.

[4] Bahdanau, Dzmitry, Kyunghyun Cho en Yoshua Bengio: *Neural Machine Translation by Jointly Learning to Align and Translate*. CoRR, abs/1409.0473, 2014. http://arxiv.org/abs/1409.0473.

[5] Baisa, Vít, Miloš Jakubíček, Adam Kilgarriff, Vojtěch Kovář en Pavel Rychlý: *Bilingual word sketches: the translate button*. In Abel, Andrea, Chiara Vettori en Natascia Ralli (redakteurs): *Proceedings of the 16th EURALEX International Congress*, bladsye 505–513, Bolzano, Italy, Julie 2014. EURAC research, ISBN 978-88-88906-97-3. http://www.euralex.org/elx_proceedings/Euralex2014/euralex_2014_037_p_505.pdf.

[6] Baker, Mona: *Corpus-based translation studies: The challenges that lie ahead*. In Somers, Harold (redakteur): *Terminology,*

LSP and Translation: Studies in language engineering in honour of Juan C. Sager, bladsye 175–186. John Benjamins Publishing, 1996.

[7] Baldwin, Timothy: *The hare and the tortoise: speed and accuracy in translation retrieval*. Machine Translation, 23:195–240, 2009, ISSN 0922-6567. http://dx.doi.org/10.1007/s10590-009-9064-7.

[8] Baldwin, Timothy en Hozumi Tanaka: *The Effects of Word Order and Segmentation on Translation Retrieval Performance*. In *Proceedings of the 18th Conference on Computational Linguistics - Volume 1*, COLING '00, bladsye 35–41, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics, ISBN 1-55860-717-X. http://dx.doi.org/10.3115/990820.990826.

[9] Banerjee, Satanjeev en Alon Lavie: *METEOR: An automatic metric for MT evaluation with improved correlation with human judgments*. Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, bladsye 65–72, 2005. http://www.aclweb.org/anthology/W05-09.

[10] Bannard, Colin en Chris Callison-Burch: *Paraphrasing with bilingual parallel corpora*. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, bladsye 597–604. Association for Computational Linguistics, Junie 2005. http://mt-archive.info/ACL-2005-Bannard.pdf.

[11] Barbu, Eduard: *Spotting false translation segments in translation memories*. In *Proceedings of the Workshop Natural Language Processing for Translation Memories*, bladsye 9–16, Hissar, Bulgaria, September 2015. Association for Computational Linguistics. http://www.aclweb.org/anthology/W15-5202.

[12] Barbu, Eduard, Carla Parra Escartín, Luisa Bentivogli, Matteo Negri, Marco Turchi, Constantin Orasan en Marcello Federico: *The first Automatic Translation Memory Cleaning*

Shared Task. Machine Translation, 30(3):145–166, Desember 2016, ISSN 1573-0573. https://doi.org/10.1007/s10590-016-9183-x.

[13] Barbu, Eduard, Carla Parra Escartín, Luisa Bentivogli, Matteo Negri, Marco Turchi, Marcello Federico, Luca Mastrostefano en Constantin Orasan: *1st Shared Task on Automatic Translation Memory Cleaning Preparation and Lessons Learned*. In *2nd Workshop on Natural Language Processing for Translation Memories (NLP4TM 2016)*, LREC 2016, Portorož, Slovenia, 2016. http://www.lrec-conf.org/proceedings/lrec2016/workshops/LREC2016Workshop-NLP4TM_Proceedings.pdf.

[14] Barlow, Michael: *ParaConc: Concordance software for multilingual parallel corpora*. In *Proceedings of the Third International Conference on Language Resources and Evaluation. Workshop on Language Resources in Translation Work and Research*, bladsye 20–24, 2002. http://mt-archive.info/LREC-2002-Barlow.pdf.

[15] Bloodgood, Michael en Benjamin Strauss: *Translation memory retrieval methods*. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, bladsye 202–210, Gothenburg, Sweden, April 2014. Association for Computational Linguistics. http://www.aclweb.org/anthology/E14-1022.

[16] Bojar, Ondřej, Yvette Graham, Amir Kamran en Miloš Stanojević: *Results of the WMT16 Metrics Shared Task*. In *Proceedings of the First Conference on Machine Translation*, bladsye 199–231, Berlin, Germany, Augustus 2016. Association for Computational Linguistics. http://www.aclweb.org/anthology/W/W16/W16-2302.

[17] Carl, Michael, Srinivas Bangalore en Moritz Schaeffer (redakteurs): *New Directions in Empirical Translation Process Research*. Springer, 2016, ISBN 978-3-319-20357-7. http://dx.doi.org/10.1007/978-3-319-20358-4.

[18] Carpuat, Marine, Yogarshi Vyas en Xing Niu: *Detecting Cross-Lingual Semantic Divergence for Neural Machine Translation*. In *Proceedings of the First Workshop on Neural Machine Translation*, bladsye 69–79, Vancouver, Augustus 2017. Association for Computational Linguistics. http://www.aclweb.org/anthology/W17-3209.

[19] Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall en W. Philip Kegelmeyer: *SMOTE: Synthetic minority over-sampling technique*. Journal of Artificial Intelligence Research, 16(1):321–357, Junie 2002, ISSN 1076-9757. http://www.jair.org/media/953/live-953-2037-jair.pdf.

[20] Chen, Boxing, Roland Kuhn, George Foster, Colin Cherry en Fei Huang: *Bilingual methods for adaptive training data selection for machine translation*. In *Proceedings of AMTA*, bladsye 93–103, 2016. https://amtaweb.org/wp-content/uploads/2016/10/AMTA2016_Research_Proceedings_v7.pdf.

[21] Chen, Stanley F. en Joshua Goodman: *An empirical study of smoothing techniques for language modeling*. Computer Speech & Language, 13(4):359–393, 1999. http://dx.doi.org/10.1006/csla.1999.0128.

[22] Damerau, Fred J.: *A Technique for Computer Detection and Correction of Spelling Errors*. Commun. ACM, 7(3):171–176, Maart 1964, ISSN 0001-0782. http://doi.acm.org/10.1145/363958.363994.

[23] Denkowski, Michael en Alon Lavie: *Meteor Universal: Language Specific Translation Evaluation for Any Target Language*. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, bladsye 376–380, 2014. http://www.aclweb.org/anthology/W14-3348.

[24] Doddington, George: *Automatic evaluation of machine translation quality using n-gram co-occurrence statistics*. In *Proceedings of the second international conference on Human Lan-*

guage Technology Research, bladsye 138–145. Morgan Kaufmann Publishers Inc., 2002.

[25] Dougherty, Geoff: *Pattern Recognition and Classification: An Introduction*. Springer, 2013, ISBN 9781461453222.

[26] Draper, Norman R. en Harry Smith: *Applied Regression Analysis*. Wiley Series in Probability and Statistics. Wiley-Interscience, 3de uitgawe, 1998, ISBN 0471170828, 9780471170822.

[27] Dyer, Chris, Victor Chahuneau en Noah A. Smith: *A Simple, Fast, and Effective Reparameterization of IBM Model 2*. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, bladsye 644–648, Atlanta, Georgia, Junie 2013. Association for Computational Linguistics. http://www.aclweb.org/anthology/N13-1073.

[28] Esqueda, Marileide, Igor A. Lourenço da Silva en Érika Nogueira de Andrade Stupiello: *Examinando o uso dos sistemas de memória de tradução na sala de aula de tradução*. Cadernos de Tradução, 37(3):160–184, 2017. https://periodicos.ufsc.br/index.php/traducao/article/download/2175-7968.2017v37n3p160/34849.

[29] Federmann, Christian: *Appraise: an open-source toolkit for manual evaluation of MT output*. The Prague Bulletin of Mathematical Linguistics, 98:25–35, 2012. http://www.mt-archive.info/PBML-2012-Federmann-1.pdf.

[30] Gale, William A. en Kenneth W. Church: *A Program for Aligning Sentences in Bilingual Corpora*. Computational Linguistics, 19(1):75–102, Maart 1993, ISSN 0891-2017. http://dl.acm.org/citation.cfm?id=972450.972455.

[31] Gordon, Ian: *The TM Revolution-What does it really mean?* In *Translating and the Computer 19: Papers from the Aslib conference held on 13 & 14 November 1997*, London, 1997. Aslib. http://www.mt-archive.info/Aslib-1997-Gordon.pdf.

[32] Goutte, Cyril, Marine Carpuat en George Foster: *The Impact of Sentence Alignment Errors on Phrase-Based Machine Translation Performance*. In *Proceedings of AMTA-2012: The Tenth Biennial Conference of the Association for Machine Translation in the Americas.*, 2012. http://www.mt-archive.info/AMTA-2012-Goutte.pdf.

[33] Gow, Francie: *Metrics for Evaluating Translation Memory Software*. PhD-proefskrif, University of Ottawa, 2003. http://dx.doi.org/10.20381/ruor-9589.

[34] Gow, Francie: *You Must Remember This: The Copyright Conundrum of "Translation Memory" Databases*. Canadian Journal of Law and Technology, 6(3):175–192, 2007. https://ojs.library.dal.ca/CJLT/article/download/6038/5367.

[35] Gupta, Rohit, Constantin Orăsan, Marcos Zampieri, Mihaela Vela, Josef van Genabith en Ruslan Mitkov: *Improving translation memory matching and retrieval using paraphrases*. Machine Translation, 30:1–22, 2016, ISSN 1573-0573. http://dx.doi.org/10.1007/s10590-016-9180-0.

[36] Guyon, Isabelle en André Elisseeff: *An introduction to variable and feature selection*. Journal of Machine Learning Research, 3:1157–1182, 2003. http://www.jmlr.org/papers/v3/guyon03a.html.

[37] Hastie, Trevor, Robert Tibshirani en Jerome H. Friedman: *The elements of statistical learning: data mining, inference, and prediction*. Springer series in statistics. Springer, 2de uitgawe, 2009, ISBN 9780387848570.

[38] He, Haibo, Yang Bai, Edwardo A. Garcia en Shutao Li: *ADASYN: Adaptive synthetic sampling approach for imbalanced learning*. In *IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), IJCNN 2008*, bladsye 1322–1328. IEEE, 2008. http://dx.doi.org/10.1109/IJCNN.2008.4633969.

[39] Heafield, Kenneth, Ivan Pouzyrevsky, Jonathan H. Clark en Philipp Koehn: *Scalable modified Kneser-Ney language model estimation*. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, bladsye 690–696, Sofia, Bulgaria, Augustus 2013. http://kheafield.com/professional/edinburgh/estimate_paper.pdf.

[40] Hutchins, W. John: *Machine translation: past, present and future*. Ellis Horwood Limited, Chichester, 1986, ISBN 0-85312-788-3.

[41] Isabelle, Pierre, Marc Dymetman, George Foster, Jean Marc Jutras, Elliott Macklovitch, Francois Perrault, Xiaobo Ren en Michel Simard: *Translation Analysis and Translation Automation*. In *Proceedings of the 1993 Conference of the Centre for Advanced Studies on Collaborative Research: Distributed Computing - Volume 2*, CASCON '93, bladsye 1133–1147. IBM Press, 1993. http://dl.acm.org/citation.cfm?id=962367.962416.

[42] James, Gareth, Daniela Witten, Trevor Hastie en Robert Tibshirani: *An Introduction to Statistical Learning: With Applications in R*. Springer-Verlag New York, 2013, ISBN 9781461471370, 9781461471387.

[43] Jarque, Carlos M. en Anil K. Bera: *Efficient tests for normality, homoscedasticity and serial independence of regression residuals*. Economics letters, 6(3):255–259, 1980. http://www.sciencedirect.com/science/article/pii/0165176580900245.

[44] Jurafsky, Daniel en James H. Martin: *Speech and Language Processing*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2de uitgawe, 2009, ISBN 0131873210.

[45] Kay, Martin: *The proper place of men and machines in language translation*. Tegniese verslag CSL-80-11, Xerox Corporation, 1980. http://www.mt-archive.info/70/Kay-1980.pdf.

[46] Kilgarriff, Adam: *Terminology finding, parallel corpora and bilingual word sketches in the Sketch Engine*. In *Procee-*

dings of ASLIB 35th Translating and the Computer Conference, 2013. http://www.mt-archive.info/10/Aslib-2013-Kilgarriff.pdf.

[47] Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlỳ en Vít Suchomel: *The Sketch Engine: ten years on*. Lexicography, 1(1):7–36, 2014.

[48] Klein, Guillaume, Yoon Kim, Yuntian Deng, Jean Senellart en Alexander M. Rush: *OpenNMT: Open-Source Toolkit for Neural Machine Translation*. In *Proc. ACL*, 2017. https://doi.org/10.18653/v1/P17-4012.

[49] Koehn, Philipp: *Europarl: A Parallel Corpus for Statistical Machine Translation*. In *Conference Proceedings: the tenth Machine Translation Summit*, bladsye 79–86, Phuket, Thailand, 2005. AAMT, AAMT. http://mt-archive.info/MTS-2005-Koehn.pdf.

[50] Koehn, Philipp: *Statistical Machine Translation*. Cambridge University Press, New York, NY, USA, 1ste uitgawe, 2010, ISBN 0521874157, 9780521874151.

[51] Koehn, Philipp: *Neural Machine Translation*. arXiv preprint arXiv:1709.07809, 2017. https://arxiv.org/abs/1709.07809.

[52] Koehn, Philipp en Jean Senellart: *Convergence of translation memory and statistical machine translation*. In Zhechev, Ventsislav (redakteur): *Proceedings of the Second Joint EM+/CNGL Workshop "Bringing MT to the User: Research on Integrating MT in the Translation Industry" (JEC '10)*, bladsye 21–31, 2010. http://mt-archive.info/JEC-2010-Koehn.pdf.

[53] Lefever, Els, Lieve Macken en Veronique Hoste: *Language-independent Bilingual Terminology Extraction from a Multilingual Parallel Corpus*. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, bladsye 496–504, Stroudsburg,

PA, USA, 2009. Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=1609067.1609122.

[54] Levenshtein, Wladimir Iosifowitsj (Влади́мир Ио́сифович Левенште́йн): *Binary codes capable of correcting deletions, insertions, and reversals*. In *Soviet Physics Doklady*, volume 10, bladsye 707–710, 1966.

[55] Macken, Lieve, Els Lefever en Véronique Hoste: *TExSIS: Bilingual terminology extraction from parallel corpora using chunk-based alignment*. Terminology, 19(1):1–30, 2013. https://biblio.ugent.be/publication/2128573/file/6771624.

[56] Macklovitch, Elliott: *Translation technology in Canada*. In Sin-wai, Chan (redakteur): *The Routledge Encyclopedia of Translation Technology*, bladsye 267–278. Routledge, 2015.

[57] Manning, Christopher D. en Hinrich Schütze: *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA, 1999. https://nlp.stanford.edu/fsnlp/.

[58] Mapelli, Valérie, Victoria Arranz, Hélène Mazo en Khalid Choukri: *Latest developments in ELRA's services*. In Calzolari, Nicoletta (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis en Daniel Tapias (redakteurs): *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, Mei 2008. European Language Resources Association (ELRA), ISBN 2-9517408-4-0. http://www.lrec-conf.org/proceedings/lrec2008/.

[59] Matthews, Austin, Waleed Ammar, Archna Bhatia, Weston Feely, Greg Hanneman, Eva Schlinger, Swabha Swayamdipta, Yulia Tsvetkov, Alon Lavie en Chris Dyer: *The CMU Machine Translation Systems at WMT 2014*. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, bladsye 142–149. Association for Computational Linguistics, 2014. https://www.aclweb.org/anthology/W/W14/W14-33.pdf.

[60] McEnery, Tony, Richard Xiao en Yukio Tono: *Corpus-based Language Studies: An Advanced Resource Book*. Routledge applied linguistics. Routledge, 2006, ISBN 9780415286237.

[61] Melby, Alan K. en Sue Ellen Wright: *Translation Memory*. In Sin-wai, Chan (redakteur): *The Routledge Encyclopedia of Translation Technology*, bladsye 662–677. Routledge, 2015.

[62] Miłkowski, Marcin: *Developing an open-source, rule-based proofreading tool*. Software: Practice and Experience, 40(7):543–566, 2010, ISSN 1097-024X. http://dx.doi.org/10.1002/spe.971.

[63] Negri, Matteo, Duygu Ataman, Masoud Jalili Sabet, Marco Turchi en Marcello Federico: *Automatic translation memory cleaning*. Machine Translation, 31(3):93–115, Februarie 2017, ISSN 1573-0573. https://doi.org/10.1007/s10590-017-9191-5.

[64] Nie, Jian Yun: *Cross-Language Information Retrieval*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2010. http://dx.doi.org/10.2200/S00266ED1V01Y201005HLT008.

[65] O'Brien, Sharon: *Eye-tracking and translation memory matches*. Perspectives: Studies in translatology, 14(3):185–205, 2007. http://dx.doi.org/10.1080/09076760708669037.

[66] Papineni, Kishore, Salim Roukos, Todd Ward en Wei Jing Zhu: *BLEU: A Method for Automatic Evaluation of Machine Translation*. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, bladsye 311–318, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics. http://dx.doi.org/10.3115/1073083.1073135.

[67] Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg,

Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot en Édouard Duchesnay: *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 12:2825–2830, 2011. http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf.

[68] Pierce, John R., John B. Carroll, Eric P. Hamp, David G. Hays, Charles F. Hockett, Anthony G. Oettinger en Alan Perlis: *Language and Machines: Computers in Translation and Linguistics*. National Academy of Sciences/National Research Council, Washington, DC, USA, 1966. http://www.mt-archive.info/ALPAC-1966.pdf.

[69] Screen, Benjamin Alun: *What does Translation Memory do to translation? The effect of Translation Memory output on specific aspects of the translation process*. The International Journal for Translation & Interpreting Research, 8(1), 2016. http://dx.doi.org/10.12807/ti.108201.2016.a01.

[70] Servan, Christophe en Holger Schwenk: *Optimising multiple metrics with MERT*. The Prague Bulletin of Mathematical Linguistics (PBML), 2011. https://doi.org/10.2478/v10108-011-0016-z.

[71] Simard, Michel en Atsushi Fujita: *A Poor Man's Translation Memory Using Machine Translation Evaluation Metrics*. In *Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas*, 2012. http://www.mt-archive.info/AMTA-2012-Simard.pdf.

[72] Skadiņš, Raivis, Jörg Tiedemann, Roberts Rozis en Daiga Deksne: *Billions of parallel words for free: Building and using the EU bookshop corpus*. In Calzolari, Nicoletta (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk en Stelios Piperidis (redakteurs): *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, Mei 2014. European Language Resources Associ-

ation (ELRA), ISBN 978-2-9517408-8-4. http://www.lrec-conf.org/proceedings/lrec2014/pdf/846_Paper.pdf.

[73] Smith, Ross: *Copyright issues in translation memory ownership*. ASLIB Translating and the Computer, 31, 2009. http://www.mt-archive.info/Aslib-2009-Smith.pdf.

[74] Snover, Matthew, Bonnie. Dorr, Richard Schwartz, Linnea Micciulla en Linnea Makhoul: *A study of translation edit rate with targeted human annotation*. In *Proceedings of Association for Machine Translation in the Americas*, volume 200, bladsye 223–231, Augustus 2006. http://mt-archive.info/AMTA-2006-Snover.pdf.

[75] Specia, Lucia, Gustavo Paetzold en Carolina Scarton: *Multi-level translation quality prediction with QuEst++*. In *ACL-IJCNLP 2015 System Demonstrations*, bladsye 115–120, Beijing, China, 2015. Association for Computational Linguistics and The Asian Federation of Natural Language Processing. http://www.aclweb.org/anthology/P15-4020.

[76] Steinberger, Ralf, Andreas Eisele, Szymon Klocek, Spyridon Pilos en Patrick Schlüter: *DGT-TM: A freely available translation memory in 22 languages*. In Calzolari, Nicoletta (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk en Stelios Piperidis (redakteurs): *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, Mei 2012. European Language Resources Association (ELRA), ISBN 978-2-9517408-7-7. http://www.lrec-conf.org/proceedings/lrec2012/index.html.

[77] Steyn, Faans, Chris Smit en Corna Vorster (redakteurs): *Afrikaans-Engelse woordelys van statistiese terme*. Statistiese vereniging van Suid-Afrika, 2009. http://www.sastat.org.za/statistical-dictionary.

[78] Talvensaari, Tuomas: *Effects of Aligned Corpus Quality and Size in Corpus-Based CLIR*. In Macdonald, Craig,

Iadh Ounis, Vassilis Plachouras, Ian Ruthven en Ryen W. White (redakteurs): *Advances in Information Retrieval. ECIR 2008*, volume 4956 van *Lecture Notes in Computer Science*, bladsye 114–125. Springer, Berlin, Heidelberg, 2008, ISBN 978-3-540-78646-7. https://doi.org/10.1007/978-3-540-78646-7_13.

[79] Thurmair, Gregor en Vera Aleksić: *Creating term and lexicon entries from phrase tables*. In *Proceedings of the 16th EAMT Conference*, bladsye 253–260, Mei 2012. http://hltshare.fbk.eu/EAMT2012/html/Papers/58.pdf.

[80] Tiedemann, Jörg: *News from OPUS—A collection of multilingual parallel corpora with tools and interfaces*. In Nicolov, N., K. Bontcheva, G. Angelova en R. Mitkov (redakteurs): *Recent advances in natural language processing*, volume 5, bladsye 237–248, Amsterdam/Philadelphia, 2009. John Benjamins. http://stp.lingfil.uu.se/~joerg/published/ranlp-V.pdf.

[81] Tiedemann, Jörg: *Bitext Alignment*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2011. http://dx.doi.org/10.2200/S00367ED1V01Y201106HLT014.

[82] Vanallemeersch, Tom en Vincent Vandeghinste: *Assessing linguistically aware fuzzy matching in translation memories*. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*, EAMT, Antalya, Turkey, 2015. http://www.mt-archive.info/15/EAMT-2015-Vanallemeersch.pdf.

[83] Wallis, Julian: *Interactive translation vs pre-translation in the context of translation memory systems: Investigating the effects of translation method on productivity, quality and translator satisfaction*. Meesterverhandeling, University of Ottawa (Canada), 2006. http://www.ruor.uottawa.ca/handle/10393/27425.

[84] Whyman, Edward K. en Harold L. Somers: *Evaluation Metrics for a Translation Memory System*. Software—Practice and Experience, 29(14):1265–1284, Desember 1999, ISSN 0038-0644. http://dx.doi.org/10.1002/(SICI)1097-024X(19991210)29:14<1265::AID-SPE280>3.3.CO;2-S.

[85] Wolff, Friedel: *Combining off-the-shelf components to clean a translation memory*. Machine Translation, 30(3):167–181, 2016, ISSN 1573-0573. http://dx.doi.org/10.1007/s10590-016-9186-7.

[86] Wolff, Friedel, Laurette Pretorius en Paul Buitelaar: *Missed opportunities in translation memory matching*. In Calzolari, Nicoletta (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk en Stelios Piperidis (redakteurs): *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, Mei 2014. European Language Resources Association (ELRA), ISBN 978-2-9517408-8-4. http://www.lrec-conf.org/proceedings/lrec2014/pdf/1061_Paper.pdf.

[87] Wolff, Friedel, Laurette Pretorius, Loïc Dugast en Paul Buitelaar: *Self-selection bias of similarity metrics in translation memory evaluation*. Machine Translation, 30(3):129–144, 2016, ISSN 1573-0573. http://dx.doi.org/10.1007/s10590-016-9185-8.

[88] Wolff, Friedel, Laurette Pretorius, Loïc Dugast en Paul Buitelaar: *Methodological pitfalls in automated translation memory evaluation*. In *2nd Workshop on Natural Language Processing for Translation Memories (NLP4TM 2016)*, LREC 2016, Portorož, Slovenia, 2016. http://www.lrec-conf.org/proceedings/lrec2016/workshops/LREC2016Workshop-NLP4TM_Proceedings.pdf.

[89] Yarowsky, David, Grace Ngai en Richard Wicentowski: *Inducing Multilingual Text Analysis Tools via Robust Projection*

Across Aligned Corpora. In *Proceedings of the First International Conference on Human Language Technology Research*, HLT '01, bladsye 1–8, Stroudsburg, PA, USA, 2001. Association for Computational Linguistics. https://doi.org/10.3115/1072133.1072187.

[90] Zariņa, Ieva, Pēteris Ņikiforovs en Raivis Skadiņš: *Word Alignment Based Parallel Corpora Evaluation and Cleaning Using Machine Learning Techniques*. In El-Kahlout, İlknur Durgar, Mehmed Özkan, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, Fred Hollowood en Andy Way (redakteurs): *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*, bladsye 185–192, Antalya, Turkey, Mei 2015. http://aclweb.org/anthology/W15-4924.

[91] Ziemski, Michał, Marcin Junczys-Dowmunt en Bruno Pouliquen: *The United Nations parallel corpus v1.0*. In Calzolari, Nicoletta (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk en Stelios Piperidis (redakteurs): *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, Mei 2016. European Language Resources Association (ELRA), ISBN 978-2-9517408-9-1. http://www.lrec-conf.org/proceedings/lrec2016/pdf/1195_Paper.pdf.