Afrikaanse Lettergreepverdelingspatrone

Afrikaans Syllabification Patterns

M Fick en CJ Swanepoel
Departement Besluitkunde,Universiteit van Suid-Afrika
Posbus 392, Unisa, 0003
E-pos: fickm@unisa.ac.za en swanecj@unisa.ac.za

Tilla Fick

Chris Swanepoel

Tilla Fick behaal in 1973 die graad BSc (Wiskunde en Wiskundige Statistiek) aan die Universiteit van Pretoria. Sy is sedert 1984 verbonde aan die Departement Besluitkunde (voorheen Kwantitatiewe Bestuur) aan die Universiteit van Suid-Afrika. In 2002 behaal sy die graad MSc (Operasionele Navorsing) aan Unisa en is tans besig met 'n PhD in Operasionele Navorsing.	**Tilla Fick** obtained the degree BSc (Mathematics and Mathematical Statistics) in 1973 from the University of Pretoria. Since 1984 she has been working in the Department of Decision Sciences (previously Quantitative Management) at the University of South Africa. In 2002 she obtained the degree MSc (Operations Research) from Unisa and is currently busy with a PhD in Operations Research.
Chris Swanepoel verwerf in 1995 die PhD in Wiskunde aan die Universiteit van Pretoria. Hy het skoolgehou en vyf jaar as dosent in Wiskunde aan die Universiteit van Pretoria deurgebring. Sedert 1995 is hy verbonde aan die Department Besluitkunde (voorheen die Departement Kwantitatiewe Bestuur) aan die Universiteit van Suid-Afrika.	**Chris Swanepoel** obtained the PhD in Mathematics in 1995 at the University of Pretoria. He taught at a high school and spent five years as lecturer in Mathematics at the University of Pretoria. Since 1995 he is employed by the University of South Africa in the Department of Decision Sciences (formerly the Department of Quantitative Management).

ABSTRACT

Afrikaans Syllabification Patterns

In contrast to English, automatic hyphenation by computer of Afrikaans words is a problem that still needs to be addressed, since errors are still often encountered in printed text. An initial step in this task is the ability to automatically syllabify words. Since new words are created continuously by joining words, it is necessary to develop an "intelligent" technique for syllabification.

As a first phase of the research, we consider only the orthographic information of words, and disregard both syntactic and morphological information. This approach allows us to use machine-learning techniques such as artificial neural networks and decision trees that are known for their pattern recognition abilities. Both these techniques are trained with isolated patterns consisting of input patterns and corresponding outputs (or targets) that indicate whether the input pattern should be split at a certain position, or not.

In the process of compiling a list of syllabified words from which to generate training data for the syllabification problem, irregular patterns were identified. The same letter patterns are split differently in different words and complete words that are spelled identically are split differently due to meaning. We also identified irregularities in and between the different dictionaries that we used.

We examined the influence range of letters that are involved in irregularities. For example, for the r in agter-ente and vaste-rente we have to consider three letters to the left of r to be certain where the hyphen should be inserted. The influence range of the k in verstek-waarde and kleinste-kwadrate is four to the left and three to the right.

In an analysis of letter patterns in Afrikaans words we found that the letter e has the highest frequency overall (16,2% of all letters in the word list). The frequency of words starting with s is the highest, while the frequency of words ending with e is the highest. It is important to note that the frequency of words ending with s is even higher than for words starting with s. The two and three letter patterns that occur most are er (10% of all two letter patterns) and ing (4% of all three letter patterns).

In an analysis of syllables in Afrikaans words, we found that (as for complete words) syllables most often start with the letter s and end with e, while the frequency of syllables ending with s is almost as high as the frequency of syllables starting with s. This indicates that problems with hyphenation can be expected around the letter s. The two and three letter syllables that occur most often are -ge- and -ver-, respectively.

In an attempt to decide on the window length to use to generate training data for machine-learning techniques we also analysed the length of syllables. The results show that two and three letter syllables occur most often, but that four letter syllables have the most unique instances. We also analysed a spectrum of window configurations and found that the ideal configuration will have to be determined empirically.

One major problem we identified in this study is that irregular syllabification often occurs where letter patterns include the letter s. The reasons being (i) the use of the combining s when joining words, (ii) almost equal frequencies of syllables starting and ending with s and (iii) vague hyphenation rules for letter combinations containing s. To effectively address automatic syllabification in Afrikaans, it is necessary to develop more sophisticated methods to handle vagueness around the letter s.

Key concepts: Syllabification; irregularities; machine-learning techniques; artificial neural networks; decision trees; pattern recognition; training.

OPSOMMING

Vir Afrikaans is outomatiese rekenaarmatige woordafbreking ’n probleem wat aandag vereis, aangesien foute steeds dikwels in gedrukte teks voorkom. As eerste stap in hierdie taak is dit noodsaaklik om woorde outomaties in lettergrepe te verdeel. Aangesien nuwe woorde voortdurend geskep word deur woorde aanmekaar te skryf, werk bestaande tegnieke wat vir Engels ontwikkel is, nie goed in Afrikaans nie. Dit is nodig om ’n “intelligente” tegniek vir lettergreepverdeling te ontwikkel.

As eerste benadering beskou ons slegs die ortografiese inligting van woorde sonder om sintaksis of morfologie in ag te neem. Dit laat ons toe om masjienleertegnieke soos kunsmatige neurale netwerke en beslissingsbome wat bekend is vir hul patroonherkenningsvermoë vir die taak te oorweeg. Hierdie tegnieke word met geïsoleerde afrigtingspare bestaande uit invoerpatrone en ooreenstemmende uitvoere (of teikens) afgerig.

In hierdie artikel verskaf ons die motivering vir die studie en bespreek die proses wat gevolg is om data vir die afrigting van masjienleertegnieke te genereer. Ons bespreek ook probleemareas soos onreëlmatige woordafbreking en verskaf ’n ontleding van letterkombinasies (of letterpatrone) in woorde met en sonder lettergreepverdeling.

Trefwoorde: Lettergreepverdeling; onreëlmatigheid; masjienleertegnieke; kunsmatige neurale netwerke; beslissingsbome; patroonherkenning; afrigting.

1. INLEIDING

1.1. Lettergreepverdeling

Ten einde ’n onbekende woord uit te spreek, breek mens gewoonlik die woord in kleiner dele op en spreek dit so uit. Kinders leer ook lees deur gedeeltes van woorde te “klank”.[1] Hierdie klein, uitspreekbare gedeeltes van woorde word *lettergrepe* genoem en dit stem dikwels met morfeme* ooreen.

Natuurlike-taalverwerkingstake soos spraakherkenning, teks-na-spraak-sintese, teks-na-foneem-omskakeling en ook woordafbreking tydens woordverwerking is grootliks op lettergreepverdeling gebaseer. Vir al hierdie take is dit nodig om ’n tegniek te ontwikkel wat lettergreepverdeling outomaties kan doen.[9]

Alhoewel heelwat navorsing reeds oor outomatiese lettergreepverdeling in Engels en ook in Duits en Nederlands[2] gedoen is, is dit vir Afrikaans steeds ’n probleem wat aandag verdien.

Navorsing wat reeds in hierdie veld gedoen is, toon dat masjienleertegnieke belowende resultate lewer. M Fick[6] het die moontlikheid ondersoek om neurale netwerke te gebruik vir lettergreepverdeling in Afrikaans. Weens rekenaarbeperking is die neurale netwerk met slegs sowat 5 000 woorde afgerig en dit was in staat om 94,12% van woorde in ’n toetsdatastel van 510 (baie min) woorde korrek te verdeel. S Pilon *et al*[12] het die TiMBL-algoritme met sowat 40 000 woorde afgerig om woorde in lettergrepe te verdeel. Dit was in staat om 91,94% van woorde in ’n toetsdatastel korrek te verdeel.

Nie een van hierdie pogings is egter prakties bruikbaar nie. Die neurale netwerk is met baie min data afgerig en getoets, terwyl die TiMBL-algoritme foute in meer as 8% van woorde in die toetsdata gemaak het. Die moontlikheid dat foute in drukwerk sal voorkom, is dus groot. Verdere navorsing oor outomatiese lettergreepverdeling is dus nodig.

Daar bestaan twee benaderings tot lettergreepverdeling, naamlik reëlgebaseerde en datagedrewe metodes.[9] Alhoewel daar reëls vir lettergreepverdeling in Afrikaans bestaan, is dit nie moontlik om alle reëls in ’n rekenaarprogram vas te vang nie. ’n Reël soos *verdeel tussen twee identiese medeklinkers wat langs mekaar tussen klinkers staan* is programmeerbaar, maar dis nie moontlik om ’n reël soos *breek tussen betekenisvolle dele* in programmeertaal te omskryf nie. Ons konsentreer in hierdie studie op datagedrewe metodes.

In Afrikaans en ander tale soos Duits en Nederlands, wat as analities-flekterende tale beskryf kan word, kan nuwe woorde na willekeur geskep word deur twee of meer woorde aanmekaar te skryf. Nuwe letterkombinasies word dus ook voortdurend geskep, wat die bepaling van lettergreep-verdelingspunte bemoeilik. Dit is dus nodig om ’n “intelligente” metode vir lettergreepverdeling te ontwikkel.

In hierdie artikel beskryf ons ons bevindinge ten opsigte van die teoretiese grense van outomatiese lettergreepverdeling in Afrikaans. Ons neem slegs ortografiese† inligting (die letters waaruit woorde bestaan) in aanmerking sonder enige verwysing na morfologie‡ of sintaksis.§

*’n Morfeem is die kleinste linguistiese eenheid van vorm-met-betekenis.
†Ortografie dui op die spelling van woorde – die karakters waaruit woorde bestaan.
‡Morfologie dui op die struktuur van ’n woord – verbuigings- en vervoegingsvorme.
§Sintaksis dui op die struktuur van ’n sin – woordsoorte en woordvolgorde.

1.2. Woordafbreking

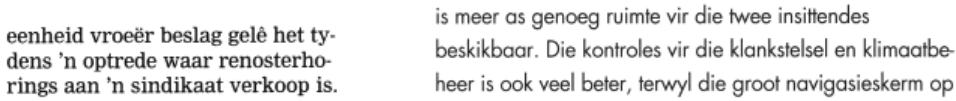
Tydens woordverwerking gebeur dit dikwels dat daar te min plek in 'n reël oor is om 'n woord volledig te laat inpas. Volgens die Afrikaanse Woordelys en Spelreëls[4] (AWS) moet so 'n woord by voorkeur volledig na die volgende reël geskuif word. Indien dit egter 'n lang woord is, sal dit by dubbelgeskouerde teks onegaligheid in drukwerk veroorsaak wanneer groot spasies tussen woorde gevorm word. By linksgeskouerde teks sal dit weer erg variërende regterkante tot gevolg hê. Woordafbreking is dus nodig – veral in koerante en tydskrifte waar teks in smal kolomme gedruk word.

Volgens die AWS moet woordafbreking in Afrikaans op grond van *betekenisvolle dele* en *lettergrepe* gedoen word, terwyl herkenbaarheid en interpreteerbaarheid in aanmerking geneem word.

Wanneer woordafbreking op grond van betekenisvolle dele gedoen word, word tussen voor- of agtervoegsels en die stamwoord afgebreek (*be-man, man-lik*) en saamgestelde woorde word tussen die samestellende woorde afgebreek (*skryf-werk, papier-dun*).

Wanneer woorde suiwer op grond van lettergreepverdeling afgebreek word, kan 'n woord soos *let-ter-greep-ver-de-ling* by enigeen van die verdelingspunte afgebreek word. Indien herkenbaarheid en interpreteerbaarheid egter in aanmerking geneem word, is vorme soos *let-tergreepverdeling* en *lettergreepverde-ling* nie aan te beveel nie.

Uit onlangse publikasies blyk dit egter dat hierdie aanbeveling nie streng nagevolg word nie, soos die voorbeelde in Figuur 1 toon.



(a) Beeld, 3 November 2009 (b) Wiel, November 2009

Figuur 1: Herkenbaarheid nie in ag geneem nie

In die AWS word gesê dat "Hoewel afbreking in drukwerk meestal outomaties deur rekenaarprogramme beheer word, word leiding hier gegee oor die korrekte afbreekstelsel." Dit is egter 'n voortydige stelling aangesien daar nog nie 'n rekenaartegniek bestaan wat woordafbreking in Afrikaans foutloos kan doen nie. Foute kom steeds dikwels in drukwerk voor, soos die voorbeelde uit onlangs gepubliseerde boeke en tydskrifte in Figuur 2 illustreer.

Woordafbreking word per hand gekontroleer voordat boeke, koerante en tydskrifte gepubliseer word. Proeflesers het bevestig dat afbrekingsfoute baie dikwels voorkom. Woordafbreking word ook dikwels vermy deur linksgeskouerde teks met groot regterkantvariasie te gebruik. Voorbeelde hiervan word in Figuur 3 getoon.

Wanneer afbreking per hand reggestel word, mag dit gebeur dat die koppelteken wat per hand ingeplaas is in die teks agterbly wanneer die uitleg verander. Figuur 4 toon 'n voorbeeld hiervan.

1.3. Masjienleertegnieke

Masjienleer is 'n onderafdeling van kunsmatige intelligensie, naamlik daardie area van rekenaarwetenskap wat daarop konsentreer om masjiene te skep wat pseudo-intelligente gedrag kan openbaar.

Een tegniek wat in masjienleer gebruik word, is om vir elke invoer 'n bekende, korrekte uitvoer (of teiken) te verskaf en die masjien dan "af te rig" om daardie uitvoer te lewer. Die hoop bestaan dan dat die masjien ook soortgelyke data wat nie in die afrigdata was nie, korrek sal hanteer.

sou bly kleef, en hy het later gesê dat dit hom gepas het dat mense nie dramatiese veranderinge van hom verwag het nie, terwyl hy reeds gewe-eet het watter rigting hy wou inslaan.

(a) *Elita en haar lewe met FW de Klerk* deur Martie Retief Meiring, 2008

In hierdie bydrae word gepoog om 'n algemene raamwerk te beskryf waarin berekenbaarheid vir 'n willekeurige topologiese ruimte X gedefinieer kan word. Die elemente van X word in dié benadering beskou as die primitiewe voorwerpe – ook vir die kodering van funksies. Ons begin deur nodige

(b) Die Suid-Afrikaanse Tydskrif vir Natuurwetenskap en Tegnologie, Desember 2008 (p. 288)

Links van hulle, oorkant die paa-
djie, sit 'n jong paartjie met 'n baba en
'n vrou langs hulle wat onbedaarlik

(c) Rooi Rose, September 2009

"Dink jy miskien jy is beter as
ons ander?" en dan het hy mi-
nagtend gesnork as hy haar Son-
dagoggende daar aflaai en vir

(d) Huisgenoot, 21 Augustus 2008

ek inkopies doen en nog steeds 'n vinnige pas kan aansit, voel ek darem nie te oud nie. Maar die een ding wat my altyd bloedjonk laat voel is die vars lug van die see. Ek kry sommer weer

(e) Finesse, Maart 2007

Figuur 2: Voorbeelde van afbrekingsfoute in tydskrifte

Maar dit maak seer om nie jou doelwitte te bereik waaraan jy so hard gewerk het nie.

Dis in hierdie tye dat jy ondersteuning nodig het. Iemand wat luister na jou gevoelens, jou gerusstel

(a) Sarie, Mei 2008

weet jy van noodhulp, sou 'n ongeluk gebeur? Brandwonde, verdrinking en val is die hoofoorsake van kindersterftes in SA. Kennis van noodhulp kan

(b) Rooi Rose, November 2009

Figuur 3: Vermyding van woordafbreking

ding om te bestudeer nie. Wetenskap-likes
is nog nie eens seker oor presies wat dit is
wat hulle moet bestudeer as dit oor die vei-

Figuur 4: Koppelteken per hand ingeplaas (Rooi Rose, September 2009)

Hierdie tegniek van "afrigting met kontrole" kan gebruik word om die rekenaar te leer om lettergreepverdeling te doen. Data vir so 'n afrigtingsproses word uit woorde wat korrek in lettergrepe verdeel is, gegenereer. (Sien Afdeling 3.)

Vir die lettergreepverdelingsprobleem ondersoek ons die masjienleertegnieke *kunsmatige neurale netwerke* en *beslissingsbome*.

- Kunsmatige neurale netwerke[7] is deur die werking van biologiese senuweestelsels geïnspireer. Dit bestaan uit lae kunsmatige *neurone* (verwerkingseenhede) wat onderling deur geweegde skakels verbind is. Invoer van buite, of vanaf ander neurone, word deur die neurone verwerk en na die volgende laag neurone aangestuur om uiteindelik 'n uitvoer te lewer. Hierdie uitvoer word met die teikenuitvoer vergelyk en gewigte op die skakels word iteratief aangepas totdat die uitvoer genoegsaam met die teiken ooreenstem. 'n Afgerigte kunsmatige neurale netwerk kan dan gebruik word om die taak waarvoor dit afgerig is, doeltreffend uit te voer.

- Beslissingsbome[3] is 'n masjienleertegniek waar voorbeelde van korrekte optrede gebruik word om 'n boomstruktuur te ontwikkel wat die uitkoms van elke item in 'n datastel deur middel van 'n reeks verdelings verteenwoordig. Daar word tussen twee soorte beslissingsbome onderskei, naamlik *klassifikasiebome*, waar die voorspelde uitkoms die klas is waaraan 'n data-item behoort, en *regressiebome*, waar die voorspelde uitkoms 'n reële getal is.

  'n Beslissingsboom word vir 'n datastel ontwikkel deur die data rekursief te verdeel sodat die onsekerheid oor die klas waaraan 'n data-item behoort, afneem soos in die boom afbeweeg word. By elke vertakking word besluit watter eienskap getoets moet word en wat die toets moet wees. Die volgorde van toetse moet sodanig gekies word dat die onsekerheid oor die klas waaraan 'n data-item behoort so gou as moontlik sal afneem. Twee verdelingskriteria wat dikwels gebruik word, is *Inligtingswins* en die *Gini*-indeks.

In 'n vergelykende studie is 'n neurale netwerk en 'n beslissingsboom afgerig met afrigtingsdata wat uit 10 000 woorde gegenereer is. Beide tegnieke is getoets met 'n datastel van 1 000 woorde wat nie in die afrigtingsdata was nie. Die resultate word in Tabel 1 getoon. Let daarop dat elke posisie tussen die letters van 'n woord 'n moontlike verdelingsgeleentheid is.

TABEL 1: Prestasie van 'n neurale netwerk en 'n beslissingsboom

	Neurale netwerk		**Beslissingsboom**	
Woorde	Aantal	Persentasie	Aantal	Persentasie
Korrek	804	80,4%	668	66,8%
Lettergrepe gemis	102	10,2%	43	4,3%
Lettergrepe foutief	94	9,4%	289	28,9%
Verdelingsgeleenthede	Aantal	Persentasie	Aantal	Persentasie
Korrek	9 250	97,4%	9 082	95,6%
Gemis	156	1,6%	89	0,9%
Foutief	96	1,0%	331	3,5%

Alhoewel die neurale netwerk 97,4% van alle moontlike verdelingspunte korrek geklassifiseer het (ons beskou dit as die akkuraatheid van die tegniek), was daar verdelingsfoute in 9,4% van die woorde. Ons beskou die gevalle waar 'n verdeling gemis is (waar 'n koppelteken bloot weggelaat

is) nie as werklike foute nie, aangesien dit nie foute in drukwerk sal veroorsaak nie, maar dalk die spasiëring mag beïnvloed. Die beslissingsboom se akkuraatheid is 95,6%, maar dit het verdelingsfoute in 28,9% van woorde gemaak, wat beslis nie aanvaarbaar is nie. Dit is dus duidelik dat gepoog moet word om akkuraatheid so hoog as moontlik te kry terwyl die persentasie foutiewe verdelingsgeleenthede na nul streef.

Alle masjienleertegnieke vereis groot hoeveelhede afrigtingsdata. Vir die lettergreepverdelingsprobleem is dit dus belangrik om soveel as moontlik woorde wat in lettergrepe verdeel is, te versamel.

2. DATA

'n Lys Afrikaanse woorde wat soveel as moontlik saamgestelde woorde insluit, is saamgestel. Die volgende bronne is gebruik om teks in elektroniese formaat te bekom:

- ▷ Elektroniese woordeboeke soos die elektroniese HAT;
- ▷ Elektroniese woordelyste vanaf die internet;
- ▷ Elektroniese teks van bekende Afrikaanse boeke en koerante;
- ▷ Internetpublikasies en nuusbriewe soos LitNet,[8] Soldeer[13] en Die Vrye Afrikaan.[5]

'n Lys woorde is uit hierdie groot korpus gegenereer. Die frekwensie van woorde in die korpus word nie in aanmerking geneem nie en slegs 'n enkele weergawe van elke woord is in die lys opgeneem.

Ten einde anderstalige woorde uit die lys te verwyder, het ons Engelse, Franse, Nederlandse en Duitse woordelyste van die internet bekom en met ons lys vergelyk. Ons het aanvaar dat woorde wat uniek aan ons woordelys is, wel Afrikaanse woorde is. Net so is woorde uniek aan die anderstalige lys as nie-Afrikaans aanvaar. Woorde wat in beide ons lys en, byvoorbeeld, die Engelse lys voorgekom het, is per hand deurgegaan om woorde wat in beide Afrikaans en Engels geldig is (soos *word*, *stand*, *program*), te ondervang.

Woorde wat met hoofletters begin, is aanvanklik so in die lys opgeneem. Dis moeilik om te onderskei tussen eiename en woorde wat aan die begin van 'n sin met 'n hoofletter verskyn. Aangesien 'n woord nie anders afgebreek word omdat dit met 'n hoofletter geskryf word nie, is besluit om alle woorde, eiename ingesluit, met kleinletters in die lys op te neem.

2.1. Verdeling in lettergrepe

'n Kunsmatige neurale netwerk wat in 'n vorige studie[6] ontwikkel is, is as eerste iterasie gebruik om die woorde in ons lys in lettergrepe te verdeel. Hierdie neurale netwerk is met relatief min data ($\pm$5 000 woorde) afgerig en ons het nie verwag dat dit goeie resultate sou lewer nie.

Foutiewe lettergreepverdeling is per hand reggestel. Verder is foute en onreëlmatighede programmaties geïdentifiseer. Ons het byvoorbeeld die frekwensie van letterpatrone* van verskillende lengtes in die woordelys bepaal. Letterpatrone wat slegs een keer in die lys voorgekom het, het op óf 'n fout, óf 'n uitsondering gedui. Foute is reggestel en uitsonderings aangeteken. Uitsonderings is dikwels die gevolg van woorde of eiename wat hul oorsprong in 'n ander taal het, soos byvoorbeeld, *Jür-gen*, *ren-dez-vous*, *bou-clé*.

'n Lys van 184 244 woorde wat grootliks korrek in lettergrepe verdeel is, is saamgestel.

*Ons gebruik die benaming "letterpatroon" vir 'n bepaalde opeenvolging van letters in 'n woord.

2.2. Onreëlmatighede

Tydens die ontfouting van die woordelys is verskeie soorte onreëlmatighede geïdentifiseer. Ons beskou dit as ’n onreëlmatigheid wanneer ’n letterpatroon in verskillende woorde verskillend in lettergrepe verdeel word. Hierdie verskynsel kan tot probleme met rekenaarmatige lettergreepverdeling lei.

2.2.1. Onreëlmatighede in woorde

Identiese letterpatrone in verskillende woorde word soms na gelang van betekenis, uitspraak of herkoms verskillend in lettergrepe verdeel. Voorbeelde van sulke onreëlmatighede word in Tabel 2 getoon.

TABEL 2: Onreëlmatighede in woorddele

Letterpatroon	Weergawe 1	Weergawe 2
dienste	ver-dien-ste	kerk-diens-te
ekspe	eks-pe-di-sie	ek-spert
ekst	eks-tro-vert	ek-sta-se
eum	li-no-le-um	mu-seum
gnos	ag-nos-ti-ci	gnos-ti-ci
komste	in-kom-ste	by-een-koms-te
sju	bro-sjure	dis-junk
tjie	kaart-jie	paar-tjie

In hierdie gevalle is dit dikwels moontlik om uit die nabye konteks (letters rondom die verdelingspunt) te bepaal waar lettergreepverdeling moet plaasvind.

Ook volledige woorde wat presies dieselfde gespel word, maar op grond van betekenis verskillend in lettergrepe verdeel word, is geïdentifiseer. Voorbeelde hiervan word in Tabel 3 getoon.

TABEL 3: Ononderskeibare onreëlmatighede

Woord	Betekenis 1	Betekenis 2
eselskop	e-sels-kop	e-sel-skop
geklik	ge-klik	gek-lik
kateter	ka-te-ter	kat-e-ter
proeslag	proe-slag	proes-lag
pronkertjie	pronk-ert-jie	pron-ker-tjie
sandaal	san-daal	sand-aal
trompop	trom-pop	tromp-op

Dit is moeilik, en dikwels onmoontlik, om sulke woorde rekenaarmatig korrek te verdeel, selfs al sou sintaktiese inligting gebruik word.

2.2.2. Onreëlmatighede rondom die letter *s*

Onreëlmatighede kom dikwels voor waar letterpatrone die letter *s* insluit. Faktore wat tot hierdie verskynsel bydra, is die verbindings-*s* wat by saamgestelde woorde gebruik word, asook die opsionaliteit in woordafbrekingsreëls vir letterpatrone wat 'n *s* insluit. Reël 1.8 van die AWS[4] stel dit dat

> "Indien drie konsonantletters langs mekaar tussen vokaalletters staan en die eerste konsonantletter 'n *s* is, kan daar voor of ná die *s* afgebreek word."

Die opsionaliteit van hierdie reël het noodwendig onreëlmatigheid rondom die letter *s* tot gevolg wat probleme by die afrigting van masjienleertegnieke kan veroorsaak. Vir die woord *distrik* is beide *dis-trik* en *di-strik* byvoorbeeld aanvaarbaar.

In Tabel 4 word voorbeelde van onreëlmatighede as gevolg van die verbindings-*s* getoon.

TABEL 4: Onreëlmatighede as gevolg van verbindings-*s*

beroep-sport	groeps-portret
edel-staal	handels-taal
garing-soort	bewarings-oord
gevoel-sintuig	gevoels-indruk
omgewing-sake	omgewings-aktiwiteit
voorsitter-stoel	voorsitters-toespraak

Vergelyk byvoorbeeld die woorde *beroep-sport* en *groeps-portret*. Die onderstreepte gedeeltes is identies in die twee woorde, maar die verdeling verskil. Eers wanneer ons ten minste vyf letters links of regs van die letter *s* beskou, is dit duidelik waar die woord verdeel moet word. Ons noem hierdie aantal letters links/regs van die onreëlmatige letter wat met sekerheid aandui waar die woord verdeel moet word, die *invloedreikwydte* van die letter. 'n Ontleding van invloedreikwydtes kan moontlik onreëlmatighede, en sodoende ook verwarring by outomatiese lettergreepverdeling beperk. Meer hieroor in Afdeling 6.

2.2.3. Onreëlmatighede in bronne

Woordeboeke is gebruik om lettergreepverdeling te kontroleer. Ons het die HAT[10] en Pharos[11] as bronne gebruik, aangesien beide lettergreepverdeling aandui. In Tabel 5 word voorbeelde van onreëlmatighede wat binne, en ook tussen die twee woordeboeke voorkom, getoon.

Let daarop dat die woordeboeke die hoofklem in woorde verskillend aandui: Pharos onderstreep die klinker wat die hoofklem kry, terwyl die HAT 'n apostroof gebruik om aan te toon dat die voorafgaande lettergreep die hoofklem kry. In beide woordeboeke word lettergreepverdeling deur 'n verhewe punt aangedui, terwyl die apostroof by die HAT ook lettergreepverdeling aandui. In die tabel dui ons lettergreepverdeling soos dit in die bronne voorkom, deurgaans met 'n koppelteken aan.

TABEL 5: Onreëlmatighede binne en tussen woordeboeke

HAT		Pharos	
	–	en-kou-stiek	en-kous-ties*
guns-teling	gun-stig	guns-tig	
(Is-ma-el)	Is-rael	Is-ra-el	Is-rae-liet
ma-erheid/ver-ma-er	maer	ma-er	
(si-ne-re-se)	sin-er-gis-me	si-ner-gis-me	
	sin-es-te-sie	si-ne-ste-sie	
	ta-xi**	tax-i	
vors-tin	vor-ste-dom	vors-te-dom	vor-ste-lik
(in teks: wor-stel)	wors-tel	wor-stel	

* Hierdie onreëlmatigheid is daaraan te wyte dat die klem op verskillende plekke in die woorde val, naamlik *en-kou-stiek* en *en-kous-ties*. Ons beskou dit as 'n uitsonderlike geval.

** Volgens Reël 1.16 in die AWS moet verdeling ná die *x* plaasvind.

3. AFRIGTINGSPATRONE

Ons ondersoek die grense van wat moontlik is as mens slegs die opeenvolging van karakters en die patrone wat daardeur gevorm word, gebruik om lettergreepverdeling te doen. Ons gebruik dus net die patrone van letters in die omgewing van elke moontlike verdelingspunt in 'n woord om te besluit of dit geldig is om 'n woord in daardie posisie te verdeel, of nie. Ons laat dus morfologiese en sintaktiese inligting buite rekening. Dit mag egter in verdere studie nodig wees om ook hierdie inligting te gebruik ten einde tekortkominge in die patroonbenadering aan te spreek.

Vir die afrigting van masjienleertegnieke benodig ons patrone van bepaalde lengte met ooreenstemmende teikenwaardes wat aandui of verdeling by 'n sekere posisie in 'n patroon geldig is of nie. Om sulke patrone te genereer, word 'n "venster" van 'n sekere lengte van links na regs oor 'n woord geskuif, terwyl ooreenstemmende teikenwaardes uit die korrekte lettergreepverdeling van die woord gegenereer word.

In Figuur 5 word die proses geïllustreer om patrone en ooreenstemmende teikenwaardes uit die woord *ver-de-ling* te onttrek.

Venster								Teiken
			v	e	r	d	e	0
		v	e	r	d	e	l	0
	v	e	r	d	e	l	i	1
v	e	r	d	e	l	i	n	0
e	r	d	e	l	i	n	g	1
r	d	e	l	i	n	g		0
d	e	l	i	n	g			0
e	l	i	n	g				0

Figuur 5: Patrone van lengte 8 met ooreenstemmende teikens

Die patroonlengte (venstergrootte) is agt karakters; die verdelingspunt is tussen karakters vier en vyf; drie spasies voor en ná die woord verseker dat elke posisie in die woord beskou word; 'n

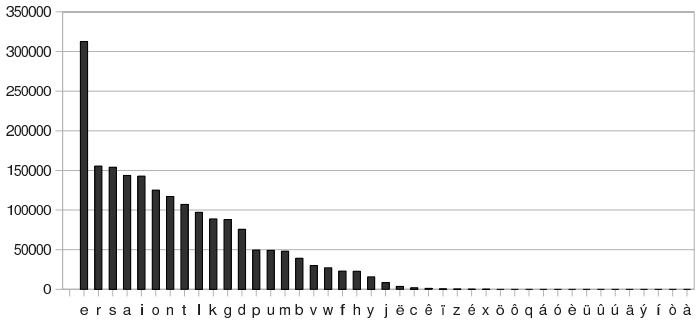
teikenwaarde van een dui aan dat verdeling in die bepaalde punt geldig is, terwyl ’n nul aandui dat dit nie geldig is nie. Uit die woord *verdeling* is agt afrigtingspatrone gegenereer – een vir elke posisie tussen die letters van die woord.

4. PATROONONTLEDING

Ten einde ’n beter begrip te kry van die patrone wat letters in Afrikaanse woorde vorm, het ons letterpatrone van verskillende lengtes ontleed.

4.1. Letterfrekwensie

Die frekwensie van die 43 letters wat in Afrikaans gebruik word, is ontleed. Figuur 6 toon hierdie frekwensieverdeling.



Figuur 6: Letters wat in Afrikaanse woorde voorkom

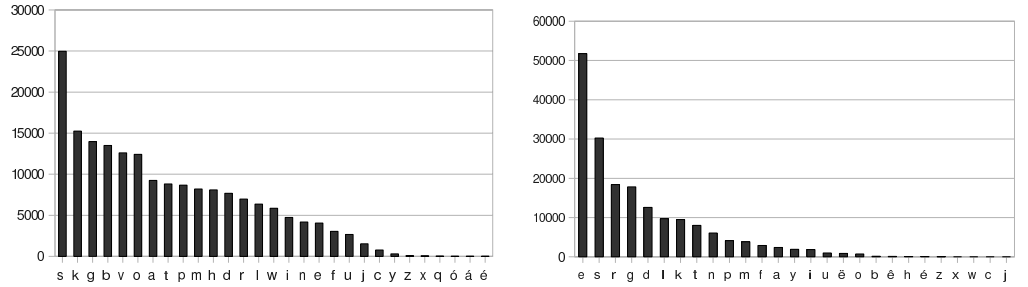
Die klinker *e* het die hoogste frekwensie (16,2% van die totale aantal letters in die woordelys) terwyl die medeklinkers *r* en *s* die meeste voorkom. Die letters *z*, *x* en *q*, asook die letters met diakritiese tekens soos *û*, *ä* en *ò* kom relatief min voor.

Die frekwensieverdeling van letters wat aan die begin en einde van woorde voorkom, word in Figuur 7 getoon.

Die frekwensie van woorde wat met die letter *e* eindig is die hoogste, terwyl woorde wat met ’n *e* begin baie minder gereeld voorkom. Die letter *b* kom weer baie dikwels aan die begin van woorde voor, maar byna nooit aan die einde van woorde nie. Die letter *s* kom egter byna net so gereeld aan die begin as aan die einde van woorde voor.

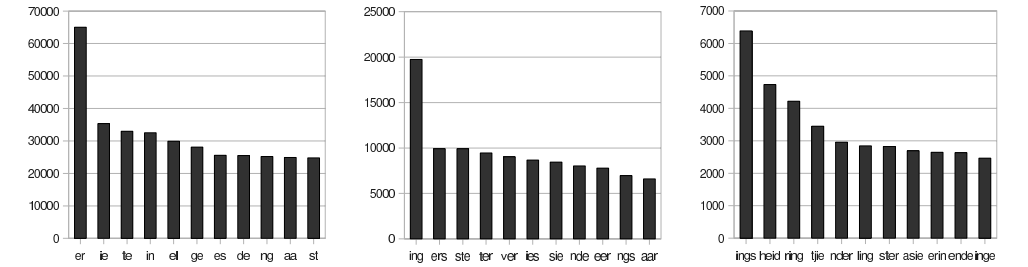
4.2. Letterpatrone

Letterpatrone van lengtes een tot agt is uit die woordelys (sonder lettergreepaanduiding) onttrek. In Figuur 8 word die frekwensieverdelings van onderskeidelik twee-, drie- en vierletterpatrone getoon. Voortaan sluit ons by patrone wat uit twee of meer letters bestaan net die patrone met die hoogste frekwensie in.



(a) Begin van woorde (b) Einde van woorde

Figuur 7: Letters waarmee woorde begin en eindig



(a) Tweeletterpatrone (b) Drieletterpatrone (c) Vierletterpatrone

Figuur 8: Frekwensieverdeling van letterpatrone

4.3. Lettergreepontleding

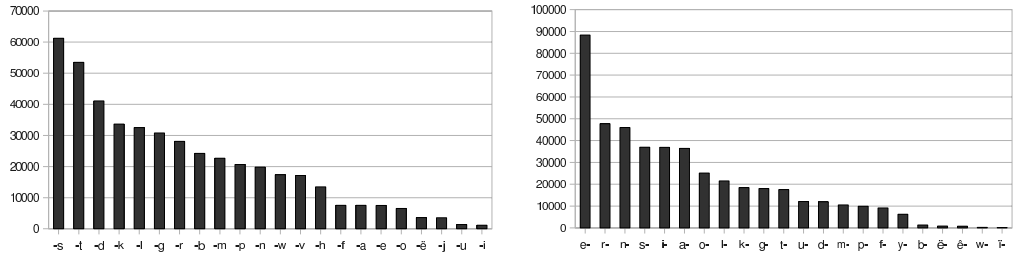
Aangesien ons in letterpatrone rondom verdelingspunte belangstel, het ons dit op verskillende maniere ondersoek. Ons het die frekwensies van letterpatrone aan die begin en einde van lettergrepe, asook van volledige lettergrepe binne en aan die begin en die einde van woorde bepaal. Verder het ons ook die frekwensie van verskillende lettergreeplengtes ondersoek.

4.3.1. Letterpatrone aan begin en einde van lettergrepe

Die frekwensieverdeling van verskillende letterpatrone waarmee lettergrepe *binne woorde* begin en eindig, is bepaal. Hier beskou ons dus nie lettergrepe aan die begin en einde van woorde nie.

Hierdie patrone word daaraan uitgeken dat letters aan die begin van sulke lettergrepe ná ’n koppelteken staan en aan die einde van lettergrepe vóór ’n koppelteken. Byvoorbeeld, die begin van die tweede lettergreep in *ver-deel* word aangedui deur *-d* en die einde van die eerste lettergreep deur *r-*. In Figuur 9 word die frekwensie van enkelletters aan die einde en aan die begin van lettergrepe getoon.

Soos by die frekwensies van letters aan die begin en einde van woorde sien ons dat lettergrepe baie dikwels met ’n *e* eindig, maar selde daarmee begin. Die letter *s* kom egter aan die begin sowel as aan die einde van lettergrepe met hoë frekwensie voor wat daarop dui dat probleme met verdeling

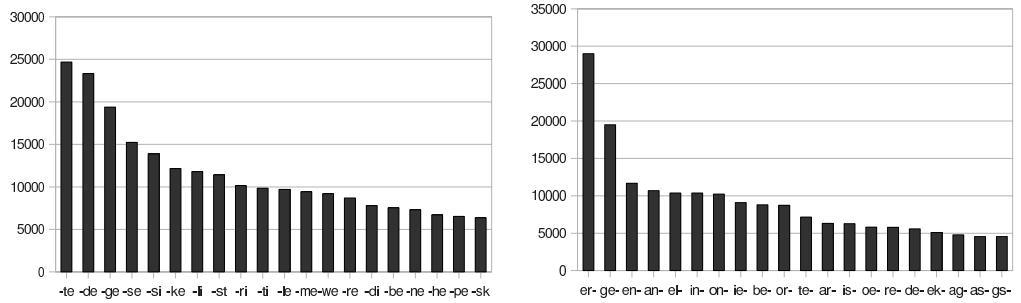


(a) Begin van lettergrepe (b) Einde van lettergrepe

Figuur 9: Letters aan begin en einde van lettergrepe

rondom die letter *s* verwag kan word.

In Figuur 10 word die frekwensieverdeling van tweeletterpatrone aan die begin en einde van lettergrepe getoon. Hier word letterpatrone, soos voorheen, daaraan uitgeken dat dit aan die einde van lettergrepe vóór die koppelteken staan en aan die begin van lettergrepe ná die koppelteken. Byvoorbeeld, die tweeletterpatroon aan die begin van die tweede lettergreep in *ver-deel* is *-de* en aan die einde van die eerste lettergreep is dit *er-*.



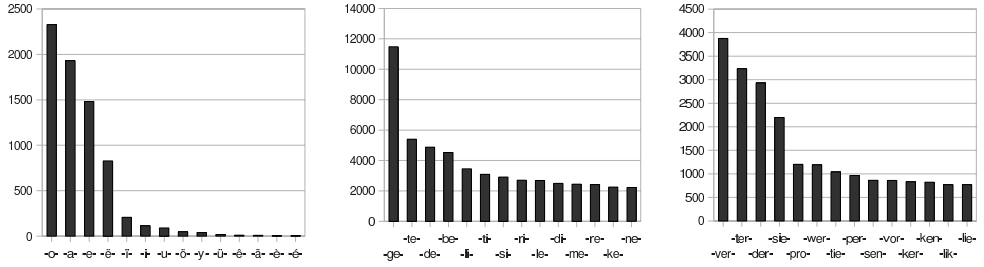
(a) Begin van lettergrepe (b) Einde van lettergrepe

Figuur 10: Tweeletterpatrone aan begin en einde van lettergrepe

4.4. Volledige lettergrepe

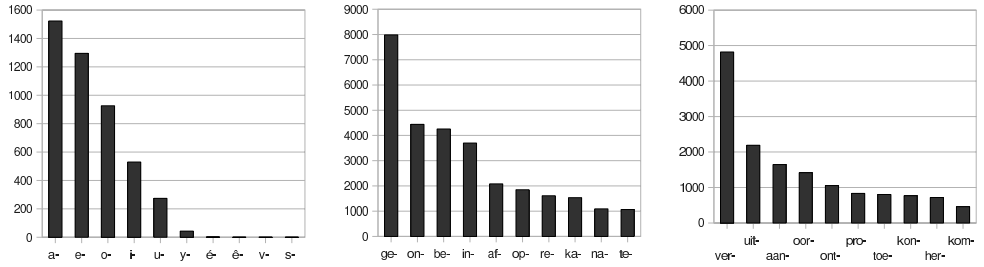
Figuur 11 toon frekwensieverdelings van volledige lettergrepe wat binne woorde (tussen koppeltekens) voorkom. Byvoorbeeld, vir die woord *let-ter-greep* is die volledige lettergreep binne die woord *-ter-*.

Ook volledige lettergrepe aan die begin en einde van woorde is beskou. In Figure 12 en 13 word onderskeidelik die frekwensieverdeling van lettergrepe aan die begin en einde van woorde getoon. Die volledige lettergreep aan die begin van ’n woord staan vóór die eerste koppelteken en aan die einde van die woord ná die laaste koppelteken. Byvoorbeeld, vir die woord *let-ter-gre-pe* is die volledige lettergreep aan die begin van die woord *let-* en aan die einde van die woord *-pe*.



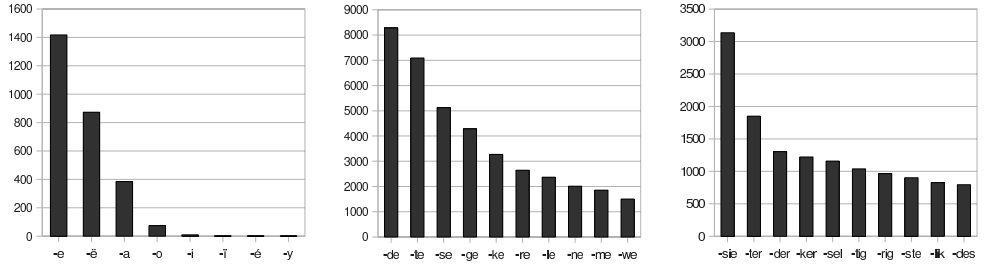
(a) Enkelletters (b) Twee letters (c) Drie letters

Figuur 11: Volledige lettergrepe binne woorde



(a) Enkelletters (b) Twee letters (c) Drie letters

Figuur 12: Lettergrepe aan die begin van woorde



(a) Enkelletters (b) Twee letters (c) Drie letters

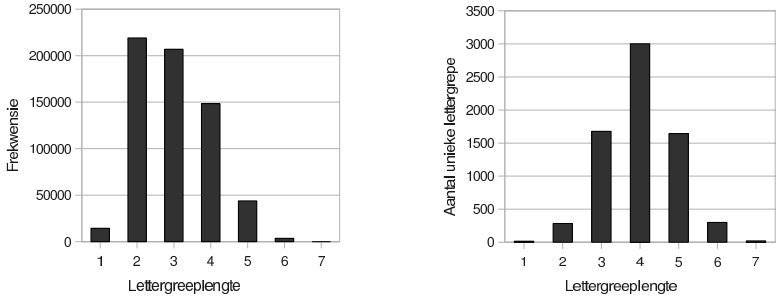
Figuur 13: Lettergrepe aan die einde van woorde

4.5. Lettergreeplengte

'n Ontleding van die lengte van lettergrepe in ons woordelys het die resultate in Figuur 14 gelewer. In Figuur 14(a) word die frekwensie waarmee lettergrepe van verskillende lengtes voorkom, getoon. Hieruit is dit duidelik dat kort lettergrepe – twee en drie letters lank – die meeste voorkom.

In Figuur 14(b) word die aantal unieke lettergrepe vir verskillende lettergreeplengtes getoon.

Hieruit is dit duidelik dat die grootste verskeidenheid by lettergrepe van lengte vier voorkom, terwyl lettergrepe van lengtes drie en vyf byna dieselfde aantal unieke lettergrepe het. Hierdie inligting is waardevol vir die besluit oor die venstergrootte wat vir patroongenerasie gebruik moet word.



(a) Lettergreepfrekwensie (b) Unieke lettergrepe

Figuur 14: Lettergreepfrekwensie

5. VENSTERKONFIGURASIE

In ’n poging om te bepaal wat die ideale patroonlengte en verdeling van letters voor en ná verdelingspunte vir Afrikaans is, is die volgende data uit die woordelys gegenereer: vensters en ooreenstemmende teikens; die unieke vensters in die lys; onreëlmatige patrone – identiese vensters met verskillende teikens; en die woorde wat by onreëlmatige patrone betrokke is. Uit die 184 244 woorde in ons lys is 1 749 045 vensters gegenereer – een vir elke moontlike verdelingspunt in elke woord.

Die resultate vir verskillende vensterkonfigurasies word in Tabel 6 getoon. Die vensterkonfigurasie 5–3 dui byvoorbeeld op ’n venster wat agt letters lank is met vyf letters voor die verdelingspunt en drie letters daarna.

Die tabel toon dat ’n 5–5-vensterkonfigurasie die minste onreëlmatighede (36) tot gevolg het, wat slegs 268 (0,15%) van die woorde in die lys betrek. Dit lewer egter die meeste unieke vensters, naamlik 56,49% van die totale aantal vensters. Daarteenoor lewer ’n 2–6-vensterkonfigurasie die minste unieke vensters (35,02% van die totale aantal vensters) maar dit het baie onreëlmatighede (920) tot gevolg wat 10 909 (5,92%) woorde betrek.

Die afrigting van kunsmatige neurale netwerke is ’n berekeningsintensiewe taak en tans is selfs die kragtigste persoonlike rekenaars se geheue onvoldoende om groot genoeg datastelle vir sinvolle afrigting te hanteer. Dit is dus belangrik om die aantal unieke vensters hiervoor te beperk.

Minder rekenaargeheue is egter nodig om beslissingsbome te ontwikkel en dis moontlik om redelik groot datastelle op ’n standaard persoonlike rekenaar te gebruik. Dit is dus vir beslissingsbome belangriker om die aantal uitsonderings te beperk.

Aangesien ons dieselfde data vir beide tegnieke wil gebruik ten einde hul prestasie te vergelyk, moet die ideale vensterkonfigurasie empiries bepaal word.

TABEL 6: Resultate vir verskeie vensterkonfigurasies

Totale venster-wydte	Venster-konfigu-rasie	Aantal unieke vensters	% van totale vensters	Aantal onreël-matighede	% van unieke vensters	Aantal woorde betrek	% van alle woorde
8	4–4	697 657	39,89	275	0,04	1 655	0,90
8	5–3	677 329	38,73	664	0,10	3 845	2,09
8	3–5	673 779	38,52	390	0,06	3 945	2,14
8	6–2	614 581	35,14	1 732	0,28	13 122	7,12
8	2–6	612 590	35,02	920	0,15	10 909	5,92
9	5–4	842 874	48,19	125	0,01	891	0,48
9	4–5	841 286	48,10	119	0,01	665	0,36
9	6–3	790 757	45,21	525	0,07	2 986	1,62
9	3–6	789 622	45,15	341	0,04	3 658	1,98
9	7–2	694 703	39,72	1 607	0,23	12 365	6,71
9	2–7	701 139	40,09	885	0,13	10 637	5,77
10	5–5	987 982	56,49	36	0,00	268	0,15
10	6–4	957 812	54,76	84	0,01	383	0,21
10	4–6	958 579	54,81	108	0,01	858	0,47
10	7–3	872 371	49,88	480	0,06	2 780	1,51
10	3–7	879 609	50,29	332	0,04	3 625	1,97

6. INVLOEDREIKWYDTE

As gevolg van die aanmekaarskryf van samestellings in Afrikaans, asook die gebruik van verbindingsletters (soos *s, e, en, er* en *ens*) word nuwe letterpatrone dikwels geskep. Nuwe letterpatrone wat so gevorm word, is dikwels identies aan bekende letterpatrone, maar met onreëlmatige lettergreepverdeling. Voorbeelde hiervan word in Tabel 7 getoon. Om die rekenaar in staat te stel om tussen sulke letterpatrone te onderskei, moet die venster wat vir patroongenerasie gebruik word, die letters insluit wat met sekerheid aandui waar die woord verdeel moet word.

Ons noem die aantal letters links en/of regs van die letter wat by die onreëlmatigheid betrokke is, soos die *r* in *vasterente* en *agterente* en die *s* in *doodskreet* en *doodskree*, wat met sekerheid aandui waar afbreking moet plaasvind, die *invloedreikwydte* van die letter.

In Tabel 7 word voorbeelde van die invloedreikwydte getoon. Die onderstreepte gedeelte is telkens identies in die onderskeie woorde, terwyl die onreëlmatige letter vet gedruk word. Die aantal letters wat links en regs van hierdie letter beskou moet word om die onreëlmatigheid te ondervang, word onderskeidelik in die kolomme aan die linker- en regterkant getoon.

Hieruit blyk dit dat ’n vensterkonfigurasie van ten minste 8–5 nodig sal wees om hierdie onreëlmatighede te ondervang. Uit Afdeling 5 is dit egter duidelik dat dit nie wenslik is om so ’n groot vensterwydte (13) te gebruik nie, aangesien die aantal unieke vensters toeneem namate die vensterwydte toeneem.

Alhoewel die aantal onreëlmatighede wel afneem soos die vensterwydte toeneem, veroorsaak die toename in die hoeveelheid afrigtingsdata dat die proses om verdelingspunte te bepaal, lomp word. Ander tegnieke sal dus ondersoek moet word om die probleem van invloedreikwydtes aan te spreek.

TABEL 7: Invloedreikwydte links en regs van onreëlmatige letter

Links	Woord 1	Woord 2	Regs
3	agte**r**-ente	vaste-**r**ente	–
4	verste**k**-waarde	kleinste-**k**wadrate	3
5	dien**s**-teraad	verdien-**s**tereëling	4
5	rooste**r**-oond	testoste-**r**oonvlak	4
6	karaktereie**n**-aardigheid	trekkereie-**n**aar	4
7	vervoerdien**s**-te	verdien-**s**te	–
8	regering**s**-taal	legering-**s**taal	–
–	dood**s**-kreet	dood-**s**kree	5

7. SLOTBESPREKING

Die oogmerk van hierdie navorsing is om 'n bydrae te lewer tot die ontwikkeling van 'n outomatiese lettergreepverdeler, en uiteindelik ook van 'n betroubare woordafbreker vir Afrikaans. Tans word woordafbreking per hand nagegaan voordat boeke, tydskrifte en koerante gedruk word.

Ons benadering tot lettergreepverdeling is om woorde bloot as patrone van letters te beskou sonder om morfologie, sintaksis of die wyer konteks waarin dit voorkom, in ag te neem. Die doel van hierdie artikel is om die grense te bepaal van wat met hierdie eenvoudige dog doeltreffende benadering bereik kan word.

Ons het reeds die moontlikheid ondersoek om die masjienleertegnieke kunsmatige neurale netwerke en beslissingsbome vir outomatiese lettergreepverdeling te gebruik. Beide tegnieke lewer aanvaarbare resultate, maar dis duidelik dat nie een van die twee op sy eie 'n perfekte oplossing bied nie. Die moontlikheid word ondersoek om 'n kombinasie van tegnieke te gebruik of selfs om 'n nuwe benadering te vind.

'n Spesifieke uitdaging vir verdere navorsing is om lettergreepverdeling rondom die letter *s* te hanteer. Van die belangrikste redes vir die gereelde voorkoms van foute in die omgewing van die letter *s* is dat lettergrepe byna net so gereeld daarmee begin as eindig en dat die verbindings-*s* en opsionaliteit in afbrekingsreëls dikwels tot onreëlmatighede lei. Om die prestasie van masjienleertegnieke te verbeter, sal meer gesofistikeerde metodes ontwikkel moet word om hierdie vaagheid rondom die letter *s* te hanteer.

Ons wil die anonieme keurders wat waardevolle bydraes tot hierdie artikel gelewer het, van harte bedank.

VERWYSINGS

[1] S. Bartlett, G. Kondrak, en C. Cherry. Automatic Syllabification with Structured SVMs for Letter-To-Phoneme Conversion. In *ACL-08: HLT*, bladsye 568–576. Association for Computational Linguistics, 2008. Beskikbaar by: `http://aclweb.org/anthology-new/P/P08/P08-1065.pdf` [4 November 2009].

[2] S. Bartlett, G. Kondrak, en C. Cherry. On the Syllabification of Phonemes. In *NAACL '09: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, bladsye 308–316. Association for Computational Linguistics, 2009. Beskikbaar by: `http://portal.acm.org/citation.cfm?id=1620754.1620799` [5 November 2009].

[3] L. Breiman, J. Friedman, R. Ohlsen, en C. Stone. *Classification and Regression Trees*. Wadsworth International Group, 1984.

[4] Die Taalkommissie van die Suid-Afrikaanse Akademie vir Wetenskap en Kuns. *Afrikaanse woordelys en spelreëls*. Pharos Woordeboeke, NB-Uitgewers Beperk, Kaapstad, 2009.

[5] FAK. Die Vrye Afrikaan. 2008. Beskikbaar by: `http://www.vryeafrikaan.co.za` [20 Mei 2008].

[6] M. Fick. Neurale Netwerke as moontlike Woordafkappingstegniek vir Afrikaans. *Die Suid-Afrikaanse Tydskrif vir Natuurwetenskap en Tegnologie*, 22(1), 2003.

[7] M. T. Hagan, H. B. Demuth, en M. Beale. *Neural Network Design*. PWS Publishing Company, 1996.

[8] LitNet. Stem vir Afrikaans! 2008. Beskikbaar by: `http://www.litnet.co.za` [26 Mei 2008].

[9] Y. Marchand, C. R. Adsett, en R. I. Damper. Automatic Syllabification in English: A Comparison of Different Algorithms. *Language and Speech*, 52(1):1–27, 2009. Beskikbaar by: `http://las.sagepub.com/cgi/content/refs/52/1/1` [4 November 2009].

[10] F. F. Odendaal en R. H. Gouws. *HAT – Handwoordeboek van die Afrikaanse Taal*. Pearson Education South Africa, 2009.

[11] Pharos. *Afrikaans-Engels Woordeboek/English-Afrikaans Dictionary*. Pharos Woordeboeke, NB-Uitgewers Beperk, 2005.

[12] S. Pilon, M. J. Puttkamer en G. B. van Huyssteen. *Die ontwikkeling van 'n woordafbreker en kompositumanaliseerder vir Afrikaans*. Literator: Journal of Literary Criticism, comparative linguistics and literary studies, 2008. Beskikbaar by: `http://findarticles.com/p/articles/mi_7020/is_1_29/ai_n31116556/`[6 April 2010].

[13] Solidariteit. Soldeer. 2008. Beskikbaar by: `http://www.solidaritysa.co.za` [26 Mei 2008].