Outomatiese lemma-identifisering vir Afrikaans

H.J. Groenewald & G.B. van Huyssteen
Sentrum vir Tekstegnologie (CTexT)
Potchefstroomkampus
Noordwes-Universiteit
POTCHEFSTROOM
E-pos: handre.groenewald@nwu.ac.za
gerhard.vanhuyssteen@nwu.ac.za

Abstract

Automatic lemmatisation for Afrikaans

Automatic lemmatisation is a general normalisation procedure in text processing, where all inflected forms of a lexical word are normalised to a single lemma (i.e. a meaningful, uninflected base form from which more complex word forms could be formed). Traditionally, lemmatisers are developed by writing language-specific rules to identify lemmas. In this article an alternative approach is investigated, namely a machine learning approach, to develop a lemmatiser for Afrikaans (LIA: "Lemma-identifiseerder vir Afrikaans"). An overview regarding the process of inflection in Afrikaans is provided with the aim of identifying the categories of inflection that are relevant for lemmatisation in Afrikaans. The format of the input and output is described with special reference to the nine inflectional categories for Afrikaans that the system should be able to handle. Then the task of lemmatisation as a classification task for machine learning is described, and a concise introduction to memory-based learning is provided. The development and evaluation of LIA is discussed in detail, and it is illustrated how the performance of the initial classifier is improved through feature selection and parameter optimisation. The best classifier reaches an accuracy of 92,8%. The article concludes with a view on some future work.

Opsomming

Outomatiese lemma-identifisering vir Afrikaans

Outomatiese lemma-identifisering ("lemmatisation") is 'n algemene normaliseringsprosedure in teksprosessering, waardeur alle geïnflekteerde vorme van 'n leksikale woord herlei word na die lemma (d.i. daardie betekenisvolle, ongeïnflekteerde basisvorm waaruit meer komplekse woordvorme gevorm kan word). Tradisioneel word lemma-identifiseerders gegrond op taalspesifieke reëls waarvolgens lemmas geïdentifiseer word. In hierdie artikel word 'n alternatiewe benadering, te wete 'n masjienleerbenadering, ondersoek om 'n lemma-identifiseerder vir Afrikaans (LIA) te ontwikkel. 'n Oorsig oor die aangeleenthede rondom fleksievorming in Afrikaans word verskaf met die doel om die fleksiekategorieë wat relevant is vir lemma-identifisering in Afrikaans te identifiseer. Hoe die toevoer- en afvoerdata van LIA daar moet uitsien, word omskryf met spesifieke verwysing na die fleksiekategorieë wat deur die sisteem hanteer moet word. Daarna word die taak van lemma-identifisering omskryf as 'n klassifiseringstaak in masjienleer en 'n bondige inleiding tot geheuegebaseerde leer word gegee. Die ontwikkeling en evaluering van LIA word vervolgens in detail bespreek en toon aan hoe die prestasie van die aanvanklike lemma-identifiseerder verbeter word deur middel van eienskapseleksie en parameteroptimalisering. Die beste klassifiseerder behaal 'n akkuraatheidsyfer van 92,8%. Die artikel sluit af met 'n vooruitskouing op toekomswerk.

1. Inleiding

Outomatiese lemma-identifisering (*lemmatisation*) is 'n algemene normaliseringsprosedure in teksprosessering waardeur alle geïnflekteerde vorme van 'n leksikale woord herlei word na die basis/lemma-/lekseem-/kanonieke vorm (Erjavec & Džeroski, 2004; Hausser, 1999; Mitkov, 2003). So sal *stoele*, *stoeltjie*, *stoeltjies* en *gestoel* byvoorbeeld deur lemma-identifisering genormaliseer word tot *stoel*, terwyl *stoelagtig* en *gestoelte* as basisvorme behoue sal bly. As sodanig word lemma-identifisering binne die konteks van hierdie studie gesien as 'n vereenvoudigde vorm van morfologiese analise (Daelemans & Strik, 2002) wat spesifiek betrekking het op die fleksieprosesse in 'n bepaalde taal.

Hierteenoor sal bostaande voorbeelde met behulp van outomatiese stamidentifisering (*stemming*) egter almal genormaliseer word na die stam *stoel*; as sodanig word stamidentifisering dus beskou as die normaliseringsprosedure waardeur die stam van 'n woord geïden-

tifiseer en onttrek word deur sowel die fleksie- as afleidingsaffikse te verwyder. Albei hierdie prosedures word algemeen in taaltegnologietoepassings gebruik, onder andere in speltoetsers, soekenjins en masjienvertaalsisteme, asook in programmatuur vir 'n algemene korpusondersoek.

'n Beperkte lemma-identifiseerder cum stamidentifiseerder vir Afrikaans, genaamd RAGEL (*Reëlgebaseerde Afrikaanse grondwoord- en lemma-identifiseerder*), is reeds aan die Noordwes-Universiteit ontwikkel en word gebruik in die *Afrikaanse Speltoetser 3.0* (CTexT, 2005). RAGEL is ontwikkel deur van tradisionele, reëlgebaseerde metodes (Gaustad & Bouma, 2002; Jongejan & Haltrup, 2005; Kraaij & Pohlmann, 1994; Plisson *et al.*, 2004; Porter, 1980) van lemma-identifisering gebruik te maak, wat behels dat taalspesifieke reëls (in die vorm van reëlmatige uitdrukkings) opgestel word waarvolgens lemmas geïdentifiseer word. Ten spyte van maandelange ontwikkelingswerk om die reëls te verfyn en te orden, behaal RAGEL 'n akkuraatheidsyfer van slegs 68% op fleksievorme.

In hierdie artikel word die geskiktheid van masjienleermetodes as alternatief tot reëlgebaseerde metodes vir die ontwikkeling van 'n suiwer lemma-identifiseerder vir Afrikaans onder die loep geneem. Die doel van hierdie ondersoek is dus om 'n effektiewe lemma-identifiseerder vir Afrikaans te ontwikkel wat nie net verbeter op die prestasie van RAGEL nie, maar wat ook kan meeding met internasionale voorpuntlemma-identifiseerders (vgl. byvoorbeeld Erjavec & Džeroski, 2004; Chrupala, 2006). Hierdie masjienleergebaseerde lemma-identifiseerder vir Afrikaans staan bekend as LIA (*Lemma-identifiseerder vir Afrikaans*).

In afdeling 2 van hierdie artikel word op die aard van die in- en afvoerdata van LIA gefokus; die konsep *lemma* word gedefinieer en daardie kategorieë van fleksie wat relevant is vir lemma-identifisering in Afrikaans word geïdentifiseer. Die fokus verskuif in afdeling 3 na 'n bondige inleiding tot masjienleer, met die oog daarop om lemma-identifisering as 'n klassifiseringstaak te omskryf. Afdeling 4 beskryf die ontwikkeling en evaluering van LIA en in afdeling 5 word aandag aan toekomswerk gegee.

2. Fleksie in morfologiese analise

Ten einde 'n outomatiese lemma-identifiseerder vir 'n bepaalde taal te ontwikkel, moet deeglik besin word oor wat die invoer en afvoer van so 'n lemma-identifiseerder moet wees. Die invoer van 'n lemma-identifiseerder is normaalweg enige woordvorm in die taal, maar

spesifiek ook geïnflekteerde woordvorme. Die lemma-identifiseerder moet dan alle geïnflekteerde woordvorme normaliseer om die lemmas, lekseme, basisvorme of kanoniese vorme (voortaan slegs *lemma*) van daardie woordvorme as afvoer te gee, terwyl woorde wat reeds lemmas is, onveranderd gelaat moet word. Aangesien definisies en interpretasies van die konsep *lemma* van taal tot taal wissel (Knowles & Don, 2004), is dit belangrik om hierdie konsep soos dit vir Afrikaans, en spesifiek in hierdie artikel prakties verstaan moet word, te omskryf.

2.1 Die konsep *lemma* in Afrikaans

In hierdie artikel word ’n lemma gedefinieer as die basiese leksikale morfeem (die kleinste betekenisvolle eenheid in die grammatika van ’n taal) wat die verskeie leksikale vorme van die morfeem verteenwoordig en wat die leksikale betekenis van die morfeem bevat (Mitkov, 2003). Die term *lemma* word gebruik as ’n generiese term vir die betekenisvolle, ongeïnflekteerde basisvorm waaruit meer komplekse woordvorme (d.i. variante) gevorm kan word (Brits *et al.*, 2006). Vir praktiese doeleindes kan die term *lemma* naastenby as sinoniem gesien word van die leksikografiese term *trefwoord* – d.i. daardie woord wat in ’n woordeboek se toegangstruktuur gebruik word om toegang tot verdere inligting oor die woord te verleen. Laasgenoemde word ook as breë riglyn in hierdie artikel toegepas: indien ’n bepaalde woordvorm as trefwoord in Afrikaanse handwoordeboeke se makro- of mikrostruktuur opgeneem word, word dit as ’n lemma beskou. Uitsonderings hierop word hieronder bespreek.

Hoewel hierdie definisie op die oog af eenvoudig voorkom, is die praktiese toepassing daarvan in Afrikaans ietwat meer kompleks. Die belangrikste rede hiervoor is dat daar in die Afrikaanse taalkundeliteratuur nog geen duidelike uitsluitsel is oor die aard van fleksie (in vergelyking met afleiding) en watter kategorieë van fleksie relevant vir Afrikaans is nie. Trouens, selfs in die internasionale literatuur is daar ook nog geen duidelikheid oor waar presies die onderskeid tussen fleksie en afleiding (as morfologiese prosesse) lê nie. Crystal (1997) identifiseer fleksie byvoorbeeld as “*...* one of the two main categories of processes of word-formation ...”, terwyl Stump (2005) daarteenoor stel: “... it is customary to distinguish inflectional morphology from word-formation ...”. Stump (2005) kom tot die gevolgtrekking dat ons waarskynlik “... a thorough rethinking of the relation of inflection to word-formation within the coming decade” sal sien. Dit is egter vir hierdie artikel voldoende om van die algemene standpunt uit te gaan dat die onderskeid tussen fleksie en

afleiding skalêr eerder as absoluut is (vgl. byvoorbeeld Booij, 2007; Bybee, 1985; Dressler, 1989; Sproat, 1992; Tuggy, 1985).

Dit is dus nodig om vervolgens eers te beskryf hoe fleksie vir die doeleindes van hierdie artikel gedefinieer word, alvorens die kategorieë van fleksie geïdentifiseer word. Die bespreking wat volg word telkens op funksionele eerder as teoretiese maatstawwe en kriteria gerig aangesien die doel van hierdie artikel prakties is: ons is dus op soek na praktiese kriteria en oplossings, eerder as om betrokke te raak by teoretiese herbesinning oor die verhouding tussen fleksie en woordvorming.

2.2 Fleksie in Afrikaans

Fleksie word oor die algemeen gedefinieer as die morfologiese markering van eienskappe op ’n lemma, wat ’n aantal vorme (grammatikale woorde) van daardie lemma tot gevolg het (Booij, 2007). Fleksie neem dus as invoer ’n lemma en gee as afvoer ’n vorm van dieselfde lemma wat gepas is binne ’n bepaalde grammatikale konteks (Sproat, 1992). As sodanig verwys fleksie dus na die proses waardeur verskillende vorme van dieselfde lemma gevorm word, terwyl verskillende lemmas gevorm word deur die proses van afleiding (Combrink, 1990).

Die Afrikaanse woorde *honde* en *hondjie* word byvoorbeeld beskou as twee verskillende geïnflekteerde vorme van die lemma *hond*. Geeneen van hierdie twee verskillende vorme het hulle eie inskrywings in ’n Afrikaanse woordeboek soos die *Verklarende Handwoordeboek van die Afrikaanse Taal* (HAT) nie. In plaas hiervan word albei vorme onder die inskrywing vir die lemma *hond* genoem. Daarteenoor kan die woord *hondagtig* beskou word as ’n produk van afleiding, aangesien dit ’n nuwe lemma tot gevolg het.

Oor wat die kriteria vir fleksie is, bestaan daar in die literatuur geen ooreenstemmigheid nie. Vergelyk byvoorbeeld Tabel 1, waar ’n opsomming gegee word van kriteria vir fleksie wat deur enkele leidende teoretici identifiseer word en waaruit duidelik blyk dat daar geen klarigheid is oor watter kriteria vir fleksie aangelê moet word nie. Ook uit die Afrikaanse taalkundeliteratuur blyk dit dat die kriteria vir fleksie in Afrikaans steeds ’n kontroversiële saak is (Jenkinson, 1993; Kempen, 1969). Volgens De Villiers (1973) kan die skeiding tussen fleksie- en afleidingsformanse as geldig vir baie Europese tale beskou word, maar dit verskil tog van taal tot taal. Afrikaans het volgens hom byvoorbeeld weinig fleksieformanse, terwyl klassieke Grieks en Latyn weer ontsaglik baie het (vgl. ook Van Schoor,

1983). Hierteenoor verwerp Combrink (1974) die term *fleksie* vir Afrikaans geheel en al wanneer hy dit beskryf as ’n “nikswerd Latinisme”.

Tabel 1: Kriteria vir fleksie

	Booij (2007)	Booij (2006)	Bauer (2003)	Tuggy (1985)	Stump (2005)	Sproat (1992)	Aronoff & Fudeman (2005)
Fleksie is ten volle produktief		x	x	x			x
Fleksie is periferaal in vergelyking met afleiding (d.i. fleksie kom ná afleiding voor)	x	x	x		x		x
Fleksie gebruik ’n geslote stel affikse			x				
Fleksie verander nie woordsoortkategorie nie	x	x	x	x	x	x	
Fleksie is relevant tot die sintaksis/grammatika	x	x	x		x	x	x
Fleksie is verpligtend	x	x					
Die paradigma is essensieel tot fleksie	x	x			x	x	
Geïnflekteerde vorme kan nie met monomorfemiese vorme vervang word nie			x				
Fleksie is beperk in betekenisskopus		x	x				
Fleksieaffikse het reëlmatige betekenis			x	x	x		
Fleksie veroorsaak geringe verandering in betekenis				x			x
Geïnflekteerde vorme word oor die algemeen nie in die leksikon gestoor nie		x					x

Behoudens hierdie onsekerheid is die outeurs wel van mening dat die volgende praktiese kriteria met die oog op outomatiese lemma-identifisering vir fleksie aangelê kan word:

- Fleksie verander nie die woordsoortkategorie van ’n lemma nie. In die voorbeeld hierbo is *hond*, *honde* en *hondjie* almal selfstandige naamwoorde, terwyl *hondagtig* (afgelei van *hond*) ’n byvoeglike naamwoord is. Die uitsondering op hierdie reël is die deelwoord, wat byna deurgaans in die internasionale literatuur as fleksie beskou word (vgl. byvoorbeeld Booij, 2007). Met betrekking tot Afrikaans merk Van Schoor (1983) tereg op dat deelwoorde “uit Nederlands na ons toe gekom [het], en daar was hulle natuurlike fleksievorme”. Ofskoon ’n mens affikse wat deelwoorde vorm sinchronies as afleidingsaffikse kan sien, beskou ons binne die konteks van hierdie artikel daardie affikse, wat die swak voltooide deelwoord in Afrikaans vorm (bv. ***ge****rooster* en ***ge****segmenteer****d***), as fleksieaffikse. Sterk voltooide deelwoorde (bv. *geswore* of *genome*) en onvoltooide deelwoorde (bv. *huilend* of *laggend*) word dan nie as fleksie hanteer nie; in toekomswerk moet daar wel hieraan aandag geskenk word.

- Fleksie kom ná afleiding voor. In die woord *hondagtig****e*** kom die fleksiesuffiks *-**e*** ná die afleidingsuffiks *-agtig* voor; afleiding vind dus voor fleksie plaas. Vir alle praktiese doeleindes kan ’n mens dus van die veronderstelling uitgaan dat fleksieaffikse op die periferie van woorde sal voorkom. In Afrikaans bied saamgestelde werkwoorde (bv. *voorkom* en *onderdompel*) ’n uitdaging, aangesien die fleksieprefiks *ge-* in hierdie gevalle aan die kern van die samestelling verbind en dus tussen die twee stamme voorkom (soos in *voor****ge****kom* en *onder****ge****dompel*). Dieselfde geld ook vir samekoppelings wat ’n ampshoedanigheid aandui, soos *adjudant****e****-generaal* of *kwartiermeester****s****-generaal*.

- Fleksie lewer nie nuwe leksikoninskrywings nie. Indien van die veronderstelling uitgegaan word dat woordeboeke ’n inventaris is van die leksikon, kan ’n mens stel dat fleksie nie nuwe trefwoorde in ’n woordeboek oplewer nie. Dusdanig sal *geswel* (PST-swel) nie as ’n trefwoord in woordeboeke verwag word nie, terwyl *geswel* (NR-swel) wel as leksikoninskrywing sou kon voorkom.[1]

2.3 Kategorieë van fleksie in Afrikaans

Teen bostaande agtergrond moet nou besluit word watter fleksiekategorieë (d.i. morfosintaktiese kategorieë; sien Booij, 2007) vir hierdie projek onderskei moet word. Oor fleksiekategorieë in Afri-

1 Sover moontlik word deurgaans die notasiesisteem en afkortings van die “Framework for Descriptive Grammars Project” gebruik (Croft, 2003).

kaans is die Afrikaanse taalkundiges dit eweneens nie met mekaar eens nie. Vergelyk Tabel 2 waar ’n opsomming gegee word van die fleksiekategorieë wat eksplisiet as fleksiekategorieë deur sommige Afrikaanse taalkundebronne erken word. Let op dat Combrink (1974) eksplisiet daardie fleksiekategorieë wat met ’n *o*-teken in die tabel gemerk is, verwerp en dus geen fleksie in Afrikaans erken nie.

Tabel 2: Fleksiekategorieë in Afrikaans

	De Klerk (1968)	Van der Walt *et al.* (1968)	Post-humus (1968)	De Villiers (1983)	Jenkin-son (1983)	Jenkin-son (1986)	Du Toit (1982)	Van Schoor (1983)	Baden-horst *et al.* (1992)	Com-brink (1974)
PL	×	×	×	×			×	×	×	o
DIM	×					×			×	o
CMPR	×	×	×	×			×	×	×	o
SUP	×	×	×	×			×	×	×	o
PST	×	×	×				×	×	×	o
PART								×		o
ATTR	×	×			×		×	×	×	o
GEN			×		×		×		×	o
INF									×	o
NEG										o
INCH										
F										
INTS										
TR									×	

Te midde van hierdie verwarring in die literatuur, is besluit om twee beginsels aan te lê wanneer besluit word oor watter fleksiekategorieë in hierdie artikel hanteer moet word:

- die kategorie moet ten minste deur meerdere bronne as ’n moontlike fleksiekategorie erken word; en
- die kategorie moet ten minste aan die breë kriteria voldoen soos in 2.2 hierbo geïdentifiseer.

Uit Tabel 2 blyk dit dat die vorming van meervoude, die vergelykende trap, die oortreffende trap, verlede tyd en die attributiewe vorm van adjektiewe deur ten minste sewe van die elf bronne as fleksieprosesse erken word. Verder word ook in ten minste vier bronne die vorming van verkleining en die deelsgenitief as fleksieprosesse erken. Al hierdie kategorieë voldoen ook aan die kriteria gestel in 2.2. Vir doeleindes van hierdie artikel word al hierdie kategorieë sondermeer as fleksiekategorieë erken.

Twee uitsonderings op bogenoemde beginsels word gemaak, te wete op die swak verlede deelwoord (sien argument hierbo) en die infinitiewe vorm van die werkwoord (byvoorbeeld *et**e*** in *iets te ete*). Laasgenoemde is onses insiens waarskynlik een van die suiwerste vorme van fleksie in Afrikaans, aangesien dit byna aan al die kriteria in Tabel 1 voldoen (miskien met die uitsondering van *produktiwiteit*, sinchronies gesien). Aanduiding van negatiwiteit (NEG – bv. ***on**moontlik* of ***nie**-nakoming*), die inchoatief (INCH – bv. ***ont**brand*), intensivering (INTS – bv. ***oer**oud* of ***aarts**lelik*), geslag (F – bv. *kelner**in*** of *werk**ster***) en oorganklikheid (TR – bv. ***be**vaar* of ***be**gaan*) word nie in hierdie artikel hanteer nie, aangesien daar onses insiens tans nie genoeg teoretiese bewyse is om dit as fleksiekategorieë te erken nie. Ons hoop dat teoretiese debatvoering in die Afrikaanse taalkunde in die toekoms meer duidelikheid hierop sal werp; vir die huidige moet egter volstaan word met keuses wat enigsins arbitrêr mag voorkom, maar wat ten minste verantwoord binne die raamwerk en beginsels van hierdie artikel blyk te wees.[2]

Om saam te vat: die volgende nege kategorieë van fleksieaffikse word vir doeleindes van hierdie artikel gedefinieer:

- Meervoud (PL) – byvoorbeeld *tafel**s***, *hond**e***, *medi**ci***, *stadi**a***, *kwajong**ens***, *woord**e***, ensovoorts
- Verkleining (DIM) – byvoorbeeld *tafel**tjie***, *hond**jie***, *boom**pie***, *kas**sie***, ensovoorts
- Vergelykend trap (CMPR) – byvoorbeeld *interessant**er***, *lelik**er***, ensovoorts

2 Vir doeleindes van hierdie artikel en vir voor die hand liggende redes word kategorieë van fleksie wat nie meer produktief in Afrikaans voorkom nie, buite rekening gelaat. Dit sluit sterk verlede deelwoorde (soos *gesproke* en *bedorwe*) en ander relikte (soos *soggens, desnoods, in der minne,* ensovoorts) dus in hierdie artikel uit.

- Oortreffende trap (SUP) – byvoorbeeld *interessant**ste***, *lelik**ste***, ensovoorts
- Attributief (ATTR) – byvoorbeeld interessant**e** persoon of *lelik**e** meisie*
- Deelsgenitief (GEN) – byvoorbeeld *iets interessant**s*** of *iemand lelik**s***
- Verlede tyd (PST) – byvoorbeeld ***ge**skop*, *om**ge**kantel*, ensovoorts
- Swak verlede deelwoord (PART) – byvoorbeeld ***ge**meng**d***, ***ge**kwalifiseer**d***, ensovoorts
- Infinitief (INF) – byvoorbeeld *iets te drink**e*** of *iets te et**e***

LIA word dus afgerig om die woorde te herken waaruit hierdie nege verskillende fleksiemorfeemklasse verwyder behoort te word, om sodoende die betrokke woord se lemma te identifiseer. Woorde wat vormlik ooreenkom met hierdie kategorieë word gebruik as afrigtingsdata tydens die masjienleerproses.

3. Masjienleer

3.1 Inleiding

Die veld van masjienleer het te make met die konstruksie van rekenaarprogramme wat outomaties verbeter (leer) met ondervinding (Mitchell, 1997; Witten & Frank, 2000; Schapire, 1992; Aloaydin, 2004; Nilsson, 1996). Masjienleer is gebaseer op die idee dat voorspellings oor die uitslag van toekomstige gebeure gemaak kan word deur die uitslag van soortgelyke gebeure in die verlede te bestudeer. Voorbeelde van suksesvolle masjienleergebaseerde stelsels wat reeds ontwikkel is, sluit in voertuie wat hulself kan bestuur, finansiële instansies wat vorige transaksies gebruik om die kredietrisiko's van kliënte te bepaal, asook verskeie toepassings op die gebied van natuurliketaalprosessering (Witten & Frank, 2000). Bogenoemde probleme is moeilik om op te los met "statiese, onleerbare stelsels", maar dit is makliker om op te los met masjienleer, aangesien masjienleer geskik is om probleme op te los op terreine waarvan ons geen of min deskundigheid het of waar ons sukkel om ons deskundigheid te verduidelik (Schapire, 1992). Dit is ook die geval met lemma-identifisering waar mense oor die deskundigheid beskik om lemmas te identifiseer, maar dit moeilik vind om die proses te verduidelik of reëls daarvoor neer te lê.

Die masjienleerproses kan soos volg gedefinieer word:

> 'n Rekenaar leer vanuit ondervinding ***O***, ten opsigte van 'n sekere tipe taak ***T*** en prestasie-aanwyser ***P***, as die prestasie in taak ***T***, soos gemeet deur ***P***, verbeter met ondervinding ***O*** (Mitchell, 1997).

Hierdie definisie impliseer dat 'n masjienleerprobleem uit drie basiese komponente moet bestaan, naamlik 'n taak **(*T*)**, prestasie-aanwyser **(*P*)** en 'n bron van ondervinding **(*O*)**. Die volgende ooreenstemmende komponente word gedefinieer vir die taak van lemma-identifisering:

- ***T***: Lemma-identifisering
- ***P***: Persentasie woorde waarvan die lemma korrek geïdentifiseer is
- ***O***: Databasis met korrek geïdentifiseerde lemmas

Om op hierdie manier te leer, kan verskeie masjienleertegnieke gebruik word, byvoorbeeld besluitnemingsbome, neurale netwerke en geheuegebaseerde leer. Daar is reeds aangetoon dat geheuegebaseerde leer geskik is vir natuurliketaalprosesseringstake (Gustafson *et al.*, 1999; Baldwin & Bond, 2003; Van Halteren *et al.*, 1998), en daarom word geheuegebaseerde leer in hierdie artikel gebruik.

3.2 Geheuegebaseerde leer

Geheuegebaseerde leer is 'n direkte afstammeling van die klassieke *k-Naastebuurpunt*-benadering (*k-Nearest Neighbour*; *k-NN*) tot die klassifikasieproses. *k-NN* is 'n kragtige patroonklassifikasie-algoritme en is die mees basiese geheuegebaseerde metode. Die veronderstelling in *k-NN* is dat al die gevalle van 'n sekere probleem voorgestel kan word as punte in 'n *n*-dimensionele ruimte. Die naaste buurpunte van 'n nuwe geval word bereken deur 'n sekere afstandsformule $\Delta(X,Y)$ te gebruik. Die klas (kategorie) van die nuwe geval word toegeken deur te kyk na die klas wat die meeste onder die naaste bure voorkom (Wagacha, 2004).

Geheuegebaseerde leer is uiters geskik vir natuurliketaalprosesseringstake, omdat elke geval in die afrigtingsdata as ewe belangrik tydens die klassifikasieproses beskou word en uitsonderings is net so belangrik soos die algemene reël (Daelemans *et al.*, 1999). TiMBL (*Tilburg Memory Based Learner*), 'n geheuegebaseerdemasjienleerstelsel, word gebruik in die ontwikkeling van LIA.

3.3 TiMBL

TiMBL is spesifiek ontwikkel met die oog daarop om natuurliketaal-prosesseringstake soos lemma-identifisering te verrig. TiMBL kan ’n verskeidenheid masjienleeralgoritmes toepas (Daelemans *et al.*, 2004). Elkeen van hierdie algoritmes kan ook op verskeie maniere verstel word om beter prestasie te lewer.

In hierdie artikel word lemma-identifisering gedefinieer as ’n klas-sifiseringstaak, waar aan elke Afrikaanse woord ’n klas toegeken word waarvolgens die betrokke woord se lemma geïdentifiseer kan word. LIA is dus ’n klassifiseerder wat op TiMBL gebaseer is.

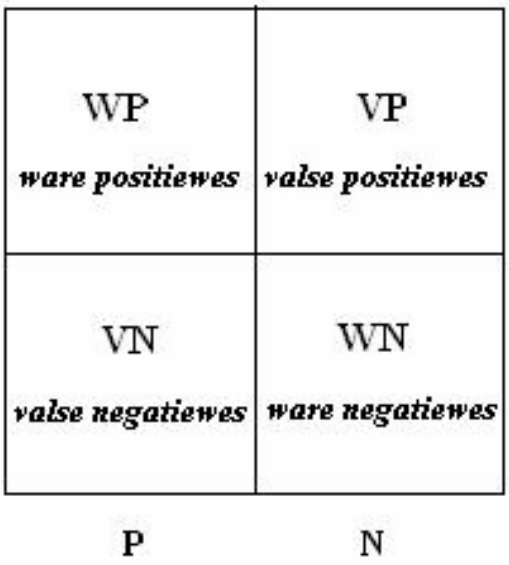
4. LIA: Lemma-identifiseerder vir Afrikaans

4.1 Ontwikkelingsproses en evalueringsmaatstawwe

Die eerste stap in die ontwikkelingsproses van LIA behels dat die stelsel afgerig word met afrigtingsdata. Tydens hierdie proses word verskeie statistiese berekenings wat as hulpmiddels tydens die klassifikasieproses dien, op die data gedoen. In die volgende stap word die datageheue gestoor as ’n aantal datapunte waarvolgens die evaluasiegevalle geklassifiseer word met behulp van die algorit-me wat gekies word. Die laaste stap in die proses behels dat die lemmas van die evaluasiegevalle geproduseer word, afhangende van die klas wat tydens klassifikasie toegeken is.

Vir elke klas in die evaluasiedata kan ’n matriks opgestel word, soos in Figuur 1 waar elke klassifikasie in een van die vier kategorieë van hierdie matriks ingedeel kan word (Daelemans *et al.*, 2004).

Figuur 1: Matriks van basiese klassifikasiemoontlikhede



Veronderstel die betrokke klas is *Lge>.* Die WP (ware positiewes) bevat die aantal gevalle wat behoort aan klas *Lge>* en wat korrek geklassifiseer is as klas *Lge>.* Die VP (valse positiewes) bevat die aantal gevalle wat aan 'n ander klas as klas *Lge>* behoort, maar wat verkeerdelik as klas *Lge>* geklassifiseer is*.* Die VN (valse negatiewes) bevat die aantal gevalle wat aan klas *Lge>* behoort, maar waaraan daar verkeerdelik 'n ander klas toegeken is. Die WN (ware negatiewe) is die aantal gevalle wat nie aan klas *Lge>* behoort nie, en wat ook nie as klas *Lge>* geklassifiseer is nie. Hierdie vier klassifikasiemoontlikhede, asook die totale aantal negatiewes (N=VP+WN) en positiewes (P=WP+VN), stel 'n mens in staat om die volgende gevorderde prestasiemetrieke te definieer:

Presisie (Akkuraatheid)

$$\Pr esisie = \frac{WP}{WP + VP}$$

Herroeping, of *True positive rate*

$$Herroeping = \frac{WP}{P}$$

f-telling

Die *f*-telling is die harmoniese gemiddelde van herroeping en presisie (Van Rijsbergen, 1979). Dit is 'n formule wat algemeen gebruik word om presisie en herroeping op te som as een prestasie-aanwyser. Die formule vir die berekening van die *f*-telling is die

volgende:

$$F-telling = \frac{2 \times presisie \times herroeping}{presisie + herroeping}$$

Aangesien uitvoersnelheid ook ter sake is in die ontwikkeling van ’n effektiewe sisteem soos ’n lemma-identifiseerder, word die snelheid in alle eksperimente ook gemeet (in sekondes).

Met betrekking tot evalueringsmetodologie word deurgaans gebruik gemaak van *N*-voudige kruisvalidering, wat behels dat die datastel telkens in *N* gelyke dele verdeel word. Daar word dan *N* aantal eksperimente uitgevoer, waartydens een van die dele vir evaluasie gebruik word, terwyl die oorblywende *N*-1 dele as afrigtingsdata gebruik word. In hierdie geval is daar besluit om ’n tienvoudige kruisvalidering te gebruik.

Soos reeds genoem, verbeter masjienleerstelsels met ondervinding. In die geval van LIA is hierdie “ondervinding” gebaseer op die hoeveelheid data waarmee die stelsel afgerig word. Die veronderstelling is dat LIA se akkuraatheid sal toeneem soos wat die hoeveelheid afrigtingsdata vermeerder. Soos blyk uit 4.2 hieronder, is die probleem met hierdie veronderstelling egter dat die annotasie van data ’n uiters tydrowende proses is; daar moet dus metodes gevind word om hierdie proses te versnel ten einde ’n sisteem te skep wat goeie prestasie lewer met so min moontlik data. Dit kan gedoen word deur onder andere relevante eienskappe (*features*) vir die data te identifiseer (sien 4.3), of deur die parameters van die masjienleeralgoritme te optimaliseer (sien 4.4).

4.2 Afrigtingsdata

Die afrigtingsdata vir LIA is onttrek uit die leksikon van *Afrikaanse Speltoetser 3.0* (CTexT, 2005) wat uit meer as 350 000 Afrikaanse woorde bestaan. Al die woorde wat vormlik ooreenstem met die gedefinieerde fleksiemorfeme (bv. alle woorde wat met die string *ge-* begin), tesame met ongeveer ’n ewe groot hoeveelheid negatiewe gevalle (woorde wat nie vormlik met die gedefinieerde fleksiemorfeme ooreenstem nie), is deur middel van eenvoudige stringpassing uit die leksikon onttrek om as afrigtingdata te dien. Hierdie proses het uiteindelik 72 226 woorde opgelewer, wat vervolgens handmatig deur assistente geannoteer is om as afrigtingsdata te dien. Hierdie proses het ongeveer drie maande geduur.

Vervolgens is ’n totaal van 271 klasse outomaties vanuit hierdie data afgelei deur middel van vergelykings tussen die oorspronklik woorde

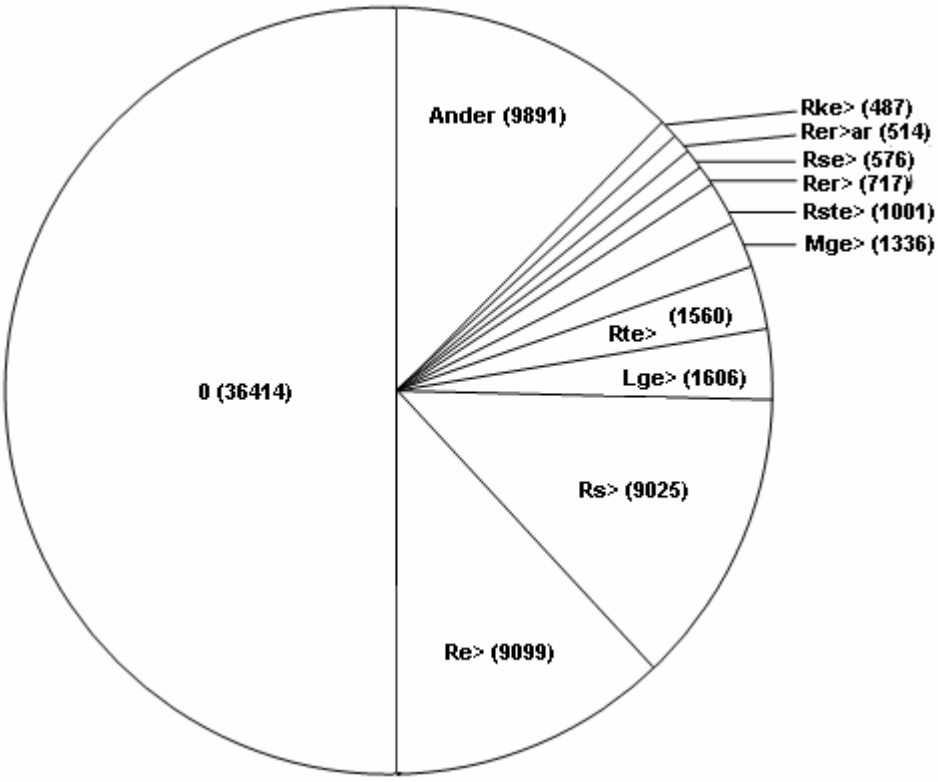
en hulle ooreenstemmende lemmas. Die klasse dui die transformasie aan wat elke woord moet ondergaan ten einde die taalkundig-korrekte lemma van die betrokke woord te onttrek. Die klasse spesifiseer die karakterstring wat verwyder moet word (bv. *-tjie* of *ge-*), die posisie van die karakterstring (bv. *L* (links), *R* (regs) of *M* (middel)), asook die moontlike string karakters wat die oorspronklike karakterstring moet vervang (bv. *-ot* wat *-te* moet vervang om *bote* te verander na *boot*). Indien 'n woord alreeds 'n lemma is, word klas 0 daaraan toegeken, wat aandui dat die betrokke woord geen transformasie tydens die lemma-identifiseringsproses moet ondergaan nie. Hierdie annotasieskema lewer klasse soos in die laaste kolom van Tabel 3.

Tabel 3: Woorde met hulle ooreenstemmende lemmas en klasse

Woord	Lemma	Klas
geel	geel	0
geslaap	slaap	Lge
hondjie	hond	Rtjie>
bote	boot	Rte>ot
omgedraaide	omdraai	MgeRde>

'n Sektordiagram van die frekwensie van die klasse in die afrigtingsdata word in Figuur 2 aangetoon. Die klasse verteenwoording nie linguistiese kategorieë nie, maar kategorieë van stringtransformasies wat nodig is om lemmas te herwin. So, byvoorbeeld, moet 'n mens nie uit Figuur 2 die afleiding maak dat die morfeme *-s* en *-e* ongeveer eweveel in die afrigtingsdata voorkom nie; as alle ander klasse waar die suffiks *-e* voorkom (bv. *Rte>* – soos in *katte* of *Rke>* – soos in *rakke*) in berekening gebring word, sal dit blyk dat die suffiks *-e* 'n veel hoër voorkomsfrekwensie het as enige van die ander fleksiemorfeme.

Figuur 2: Frekwensie van klasse



'n Uittreksel uit LIA se afrigtingsdata wat regs belyn is, word in Figuur 3 voorgestel. Elke eienskap van die betrokke woord (d.i. letters as eienskappe) word geskei met 'n komma, terwyl die klas in die laaste posisie aangedui word. Die rede waarom daar voor elke woord 'n aantal strepies is wat ook as eienskappe dien, is omdat TiMBL vereis dat elke geval 'n gelyke aantal eienskappe moet hê. Soos wat uit Figuur 3 gesien kan word, het elke afrigtingsgeval dus twintig eienskappe, ongeag die oorspronklike lengte van die woord. Die rede hiervoor is dat die meerderheid van die woorde in die data (97,79%) uit twintig of minder letters bestaan.

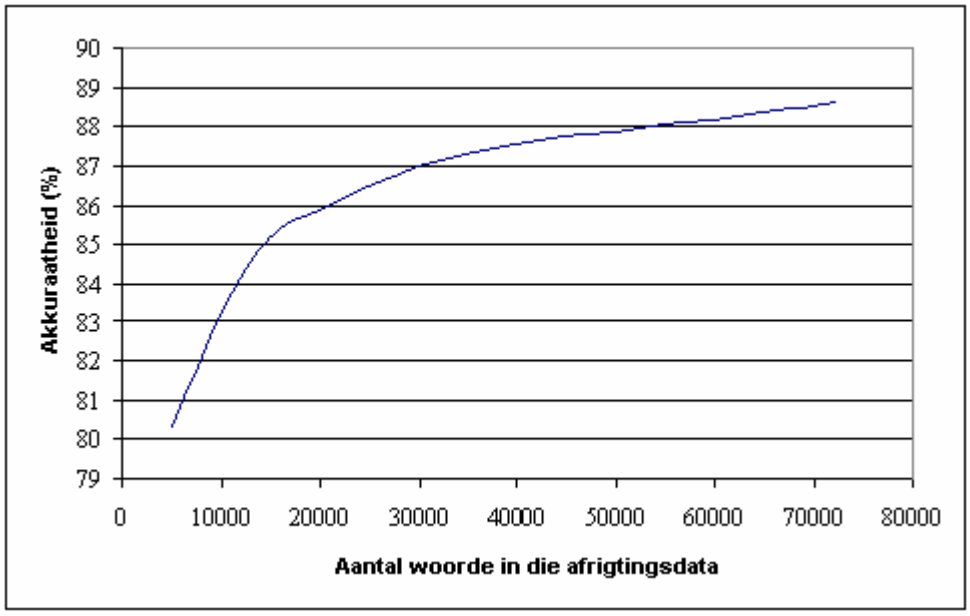
Figuur 3: Voorstelling van afrigtingsdata (regs belyn)

_,_,_,_,_,_,_,_,_,_,_,_,_,_,k,a,m,p,v,e,e,0
_,_,_,_,_,_,k,a,n,k,e,r,p,l,e,i,s,t,e,r,0
_,_,_,_,_,_,_,k,a,n,t,o,n,n,e,m,e,n,t,e,Re>
_,_,_,_,_,_,_,_,_,_,k,a,p,a,b,e,l,s,t,e,Rste>
_,_,_,_,k,a,p,i,t,a,a,l,u,i,t,g,a,w,e,s,Rs>
_,_,_,_,_,_,_,_,_,_,a,b,o,m,i,n,a,s,i,e,0
_,_,_,_,_,_,_,_,_,_,_,a,b,o,r,t,e,u,r,s,Rs>
_,_,_,_,_,_,_,_,_,_,_,a,d,d,e,r,t,j,i,e,Rtjie>
_,_,_,_,_,_,_,_,_,_,_,_,a,d,h,e,s,i,e,s,Rs>
_,_,_,_,_,_,_,_,_,_,_,a,f,d,r,u,k,r,a,m,e,Rme>am
_,_,_,_,_,_,_,_,_,_,_,a,g,n,o,s,t,i,c,i,Rci>kus

'n Eerste weergawe van LIA is afgerig met die afrigtingsdata soos in Figuur 3 en met die verstekinstellings van TiMBL. 'n Akkuraatheid-syfer van 88,61% is behaal (sien Figuur 4), wat 'n aansienlike verbetering is op die akkuraatheidsyfer van 68% van RAGEL. Figuur 4 toon aan hoe die akkuraatheid van LIA verbeter soos wat die aantal gevalle in die afrigtingsdata vermeerder (d.i. van 500 tot 72 226 afrigtingsgevalle).

Aangesien data-annotasie 'n tydrowende en duur proses is, is besluit om nie die afrigtingsdata verder uit te brei nie en eerder op alternatiewe metodes te fokus om die beskikbare data meer effektief aan te wend om sodoende die akkuraatheid van die stelsel te verhoog.

Figuur 4: Verbetering in akkuraatheid met vermeerdering van afrigtingsdata



Vir die doeleindes van hierdie artikel, is besluit om die moontlikhede te ondersoek om die akkuraatheid te verhoog deur eienskapselektering en parameteroptimalisering. Weens die feit dat daar 'n wedersydse afhanklikheid tussen eienskapselektering en parameterinstellings bestaan, is besluit om in die eksperimente slegs 'n beperkte hoeveelheid eienskappe te beskou, tesame met al die moontlike parameteropsies. Hierdie eksperimente word in die volgende subafdelings bespreek.

4.3 Eienskapselektering

Die rede waarom besluit is om met eienskapselektering te eksperimenteer, is om meer inligting in die afrigtingdata aan die klassifiseerder te verskaf sodat die akkuraatheid van die stelsel verhoog kan word. Woordsoortinligting kan byvoorbeeld tot die afrigtingsdata toegevoeg word om as addisionele inligting te dien. Erjavec en Džeroski (2004) meen selfs dat "[u]nambiguous lemmatization of words in running text is only possible if the text has been tagged with morphosyntactic information". Aangesien daar tans nog nie 'n woordsoortetiketteerder met 'n hoë akkuraatheid vir Afrikaans bestaan nie, is besluit om met eienskapsposisionering sowel as addisionele eienskappe in plaas van woordsoortinligting, te eksperimenteer. Elk van die eksperimente is aanvanklik onafhanklik van die ander uitgevoer ten einde te bepaal wat die invloed van elke verandering is.

4.3.1 Eienskapsposisionering

Aangesien die meeste fleksieaffikse in Afrikaans aan die regterkant van die woord voorkom (d.i. suffikse), is uit die staanspoor besluit om die data regs te belyn soos in Figuur 3. Op hierdie wyse kon daarin geslaag word om te verseker dat die eienskappe wat die fleksiesuffikse aandui altyd op dieselfde posisies is. *Mondjie* en *kwylmondjie* se *jie*-eienskappe kom byvoorbeeld telkens op eienskapsposisies 18, 19 en 20 voor. Dit verseker dat LIA kon leer dat daar 'n ooreenkoms tussen *mondjie* en *kwylmondjie* bestaan. Ten einde voorsiening te maak vir die prefiks *ge-*, kan die data ook links belyn word. Indien die afrigtingsdata egter links belyn word, daal die akkuraatheid van 88,60% tot 60,35% (met TiMBL se verstekinstellings).

'n Algemene fout wat LIA maak, is dat die klasse van woorde soos *geabsorbeerde* (klas: *Lge>Rde>*) verwar word met die klasse van woorde soos *verdofde* (klas: *Rde>*). Die rede hiervoor is dat woorde wat se klasse *Lge>Rde>* is, se fleksieprefiks *ge-* telkens ook op verskillende eienskapsposisies voorkom wanneer die data regs belyn is. Ten einde hierdie verwarring uit te skakel en te verseker dat woorde se unieke eienskappe deurentyd op ooreenstemmende eienskapsposisies voorgestel word, is besluit om die data voor te stel soos in Figuur 5 aangetoon word. Indien die string *ge-* aan die begin van 'n woord voorkom (bv. *gedans* of *geldwolf*), word dié deel van die string links belyn, terwyl die res van die string regs belyn word. Indien die string *-ge-* in die middel van 'n woord voorkom (bv. *ondergedompel* of *omgewing*), word die *ge* op posisies 6 en 7

geplaas, terwyl die voorafgaande string links en die daaropvolgende string regs belyn word. Alle ander gevalle word regs belyn. Hierdie voorstelling van die data verbeter die akkuraatheid van LIA vanaf 88,60% tot 91,197%.

Figuur 5: Verbeterde voorstelling van afrigtingsdata

_,_,_,n,a,a,l,d,g,a,l,v,a,n,o,m,e,t,e,r,0
_,_,_,_,_,_,_,_,_,n,a,a,t,b,o,u,l,e,r,s,Rs>
_,_,_,_,_,_,_,_,_,_,_,_,n,a,b,e,r,i,g,t,e,Rte>
_,_,_,_,_,_,_,_,_,_,n,a,b,l,o,e,i,e,r,s,Rs>
n,a,_,_,_,g,e,_,_,_,_,_,_,_,_,k,r,a,a,i,Mge>
u,i,t,_,_,g,e,_,_,_,_,_,_,_,_,k,l,e,i,Mge>
u,i,t,_,_,g,e,_,_,_,_,_,_,s,l,a,a,n,d,e,Mge>Rde>
a,f,_,_,_,g,e,_,_,_,_,_,g,r,a,d,e,e,r,Mge>
o,n,d,e,r,g,e,_,_,_,_,_,_,d,o,m,p,e,l,Mge>
o,m,_,_,_,g,e,_,_,_,_,_,_,_,_,w,i,n,g,0
g,e,_,_,_,_,_,_,_,_,_,b,o,o,m,t,e,s,Lge>Rtes>
g,e,_,_,_,_,_,_,_,_,_,_,b,r,a,a,k,t,e,Lge>Rte>
g,e,_,_,_,_,_,_,_,b,r,i,k,e,t,t,e,e,r,Lge>
g,e,_,_,_,_,_,_,_,_,_,_,_,_,d,a,n,s,Lge>
g,e,_,_,_,_,_,_,_,_,_,d,e,n,k,m,u,u,r,0
g,e,_,_,_,_,_,_,_,_,_,_,l,d,w,o,l,f,0

4.3.2 Toevoeging van addisionele eienskappe

- **Letters as eienskappe**

Aanvanklik is elke geval in die afrigtingsdata beperk tot ’n maksimum van twintig eienskappe. Alle woorde wat uit meer as twintig karakters bestaan, is verkort sodat dit in die bestek van twintig eienskappe voorgestel kon word. Die eerste drie karakters van die woord *belastingbetalersforums* (wat uit 23 karakters bestaan) is byvoorbeeld verwyder sodat die woord in twintig eienskappe kon inpas. Die oorbodige karakters is nie in alle gevalle net bloot aan die begin van ’n woord verwyder nie. In plaas hiervan is die klas van die betrokke woord as ’n aanduiding gebruik van die posisie binne die woord waar die oorbodige karakters verwyder moet word. Woorde wat met die string *ge-* begin, is byvoorbeeld verkort deur die oorbodige karakters aan die einde van elke woord te verwyder. Die eienskappe van die woorde in die afrigtingsdata is op hierdie manier verminder sodat die inherente grammatiese struktuur van die woorde in die data behou kon word, aangesien dit ’n voorvereiste vir effektiewe lemma-identifisering is.

Die langste woord in die afrigtingsdata is *geestesgesondheids-hersieningsraadslede*, wat uit 38 karakters bestaan. Ten einde enige verlies aan inligting te voorkom (d.i. lang woorde wat gereduseer word om in twintig eienskappe in te pas), is die aantal eienskappe verhoog vanaf twintig tot 38. Hierdie stap het die akkuraatheid van die stelsel marginaal verbeter, alhoewel dit die uitvoersnelheid ingrypend verleng het. Die moontlikheid om met 38 eienskappe te werk, is egter nie dadelik laat vaar nie, aangesien daar verwag is dat beter resultate later met parameteroptimalisering (sien 4.4) behaal kon word.

- **Lettergrepe as eienskappe**

Aangesien die lettergreepstruktuur van woorde ’n aanduiding kan wees van die morfologiese struktuur van die betrokke woord (bv. die prefiks *ge-* wat ’n afsonderlike sillabe vorm), is besluit dat inligting oor die lettergreepstruktuur van woorde ’n moontlike hulpmiddel kan wees in die lemma-identifiseringsproses. Daar is aanvanklik geëksperimenteer met die idee om lettergrepe by te voeg as ’n eienskap in die afrigtingsdata (sien voorbeelde in Figuur 6), maar dit het die klassifiseringsnelheid nadelig beïnvloed. Daarna is dus besluit om slegs die aantal letters in die lettergrepe as ’n addisionele eienskap by te voeg (sien voorbeelde in Figuur 7). Die gevolg van hierdie stap was dat die akkuraatheid vanaf 88,62% toegeneem het tot 90,39%.

Figuur 6: Afrigtingsdata wat lettergrepe as ’n addisionele eienskap bevat

```
gel,dig, , , , ... , , , ,g,e,l,d,i,g,0
ge,werk, , , , ... , , , ,g,e,w,e,r,k,Rge>
```

Figuur 7: Afrigtingsdata wat die aantal letters in die lettergrepe as ’n addisionele eienskap bevat

```
3,3, , , , ... , , , ,g,e,l,d,i,g,0
2,4, , , , ... , , , ,g,e,w,e,r,k,Rge>
```

- **Waarskynlikheid van die laaste letter**

Soos reeds aangedui is, bestaan die oorgrote meerderheid van Afrikaanse fleksieaffikse uit suffikse. In Afrikaans word meervoude byvoorbeeld meestal gevorm deur die letters *s* of *e* aan die einde van selfstandige naamwoorde by te voeg. Die laaste letter waarop ’n woord eindig, kan dus moontlik as aanduiding dien of ’n woord alreeds ’n lemma is of nie (’n woord wat byvoorbeeld op ’n *-s* of *-e*

eindig, is meer waarskynlik geïnflekteer as ’n woord wat op ’n *-t* eindig). Op grond hiervan is besluit om ’n addisionele eienskap by die afrigtingsdata te voeg, wat gebaseer is op die waarskynlikheid dat die woord geïnflekteer is of nie. Dit word afgelei van die laaste letter waarop die woord eindig. ’n Uittreksel van die waarskynlikhede wat met sommige letters geassosieer word, word in Tabel 4 aangetoon.

Tabel 4: Waarskynlikheid van fleksie vir woorde wat op c, d, e, p, s en t eindig

Letter	Waarskynlikheid van fleksie
c	0,0000
d	0,0946
e	0,7525
p	0,1733
s	0,6381
t	0,0710

Die waarskynlikhede in Tabel 4 is bereken deur die aantal *geïnflekteerde* woorde in die afrigtingsdata wat op die betrokke letter eindig te deel deur die totale aantal woorde in die afrigtingsdata wat op die betrokke letter eindig. Hierdie waarskynlikhede is slegs by die evaluasiedata bygevoeg en nie by die afrigtingsdata nie. In die plek hiervan is ’n eienskap by die afrigtingdata bygevoeg met ’n waarde van 1 as die betrokke geval ’n geïnflekteerde woord is en 0 as dit reeds ’n lemma is. Die rede vir die gebruik van waarskynlikhede as addisionele eienskappe is dat die afstand tussen ’n woord met ’n hoë waarskynlikheid van fleksie in die evaluasiedata en ’n woord wat definitief geïnflekteer is (fleksiewaarskynlikheid van 1), korter is as die afstand tussen ’n woord met ’n lae fleksiewaarskynlikheid en ’n geïnflekteerde woord. Die afstand tussen die geval *0.7527,k,o,p,l,a,m,p,e* in die evaluasiedata byvoorbeeld en die geval *1,s,e,i,n,l,a,m,p,e* in die afrigtingsdata, is korter as die afstand tussen *0.1733,o,l,i,e,l,a,m,p* en *1,s,e,i,n,l,a,m,p,e*. Die byvoeging van hierdie addisionele eienskap het ’n aansienlike positiewe invloed op die akkuraatheid (91,20%) gehad, maar dit het ook die klassifiseringsnelheid nadelig beïnvloed (sien resultate in 4.4).

4.4 Parameteroptimalisering

Die optimalisering van algoritmes en parameters is een van die belangrikste take wanneer toepassings ontwikkel word wat van ma-

sjienleer gebruik maak. Dit is algemeen bekend dat ’n groot variasie in die effektiwiteit (ten opsigte van akkuraatheid, klassifiseringsnelheid en geheuegebruik) van geheuegebaseerde leersisteme met die gebruik van verskillende algoritme- en parameterinstellings waargeneem kan word (Daelemans & Van den Bosch, 2005). LIA is in hierdie opsig geen uitsondering nie en die optimalisering van die masjienleeralgoritmes wat deur LIA gebruik word, word as een van die belangrikste prosesse in die ontwikkeling van LIA beskou.

Die standaardmanier om die beste instellings te verkry, is deur middel van ’n volledig omvattende soektog. ’n Volledig omvattende soektog behels dat die stelsel met elkeen van die verskillende kombinasies van klassifikasiealgoritmes en parameters afgerig word. Hierdie volledig omvattende soektogte is baie “duur” in terme van rekenaarhulpbronne, aangesien dit baie lank neem om die stelsel met elkeen van die verskillende kombinasies af te rig. Hierdie probleem vergroot in die geval van LIA waar daar ook verskillende datavoorstellings betrokke is. Daar bestaan dus ’n behoefte aan metodes vir algoritme- en parameteroptimalisering wat ten opsigte van rekenaarhulpbronne meer effektief is.

Paramsearch 1.0 (Van den Bosch, 2004) is ’n sagtewarepakket wat hierdie probleem hanteer deurdat dit die gebruik van “duur”, volledig omvattende soektogte uitskakel. *Paramsearch 1.0* implementeer begrensde steekproefneming, wat behels dat daar “kompetisies” gehou word tussen die verskillende kombinasies van instellings op groterwordende datastelle, met die doel om uiteindelik die beste instelling te bepaal. *Paramsearch 1.0* is gebruik in die plek van volledig omvattende soektogte om die beste instelling (in terme van akkuraatheid) vir LIA te bepaal.

Eksperimentering met *Paramsearch 1.0* het aangetoon dat dit nie optimaal funksioneer in die geval van LIA waar groot hoeveelhede afrigtingsdata betrokke is nie (Groenewald, 2006). Van al die verskillende kombinasies van instellings word 99% deur *Paramsearch 1.0* uitgeskakel, gebaseer op hulle prestasie op slegs 2% van die geëvalueerde afrigtingsdata. Sekere van hierdie instellings kon dalk beter presteer het indien die evaluasiedatastelle groter was. Die probleem lê in die berekening van die groottes van die datastelle wat in die begrensdesteekproefnemingsproses gebruik word: die klein verskille tussen die groottes van die datastelle wat aan die begin van die begrensdesteekproefnemingsproses gebruik word, het tot gevolg dat groot hoeveelhede instellings in die begin van die proses uitgeskakel word.

Hierdie probleem, tesame met die feit dat *Paramsearch 1.0* slegs vir twee van die klassifikasiealgoritmes in TiMBL beskikbaar is, het as aansporing gedien om ’n eie, “aangepaste” weergawe van *Paramsearch* te ontwikkel. Hierdie “aangepaste” weergawe van *Paramsearch* word *PSearch* genoem.

PSearch werk basies op dieselfde beginsels as *Paramsearch*. Die grootste verskille lê egter in die metodes waarvolgens die groottes van die opeenvolgende datastelle bereken word, asook in die feit dat *PSearch* vir al die klassifikasiealgoritmes in TiMBL beskikbaar is. *PSearch* bereken die groottes van die datastelle wat in die begrensdesteekproefnemingsproses gebruik word volgens voorafbepaalde persentasies. Dit verseker dat daar beduidende verskille is in die groottes van die opeenvolgende datastelle wat in die begrensdesteekproefnemingsproses gebruik word.

Uitvoerig eksperimentering toon dat *PSearch* beter resultate as *Paramsearch* in die geval van LIA lewer (Groenewald *et al.*, 2007). Die nadeel van *PSearch* is dat dit beduidend langer as *Paramsearch* neem om uit te voer (weens die gebruik van die groter datastelle), alhoewel dit steeds baie vinniger neem as ’n volledig omvattende soektog. In ’n poging om die klassifiseringsnelheid te verlaag, is *PSearch* uitgebrei met Aktiewe leer (*Active learning*).

Aktiewe leer behels dat slegs sekere gevalle gekies word om as afrigtingsdata te dien, eerder as om staat te maak op gevalle wat ewekansig onttrek is. Aktiewe leer word deur *PSearch* ingespan tydens die generering van die datastelle wat in die begrensdesteekproefnemingsproses gebruik word. Die weergawe van *PSearch* wat deur middel van aktiewe leer uitgebrei is, word *PSearchAL* genoem. Sowel *PSearch* as *PSearchAL* is vrylik onder ’n oopbronkodelisensie by http8//ctext.p.nwu.ac.za/Produkte/Kerntegnologie%EB.html beskikbaar.

Deur parameteroptimalisering met behulp van *PsearchAL* te doen, kan daardie parameterinstellings gebruik word wat die beste werk vir al die verskillende weergawes van LIA se afrigtingsdata (d.i. met die verskillende eienskappe soos hierbo bespreek), om finale klassifiseerders af te rig. Die resultate van die beste klassifiseerders met die onderskeie TiMBL-algoritmes word in Tabel 5 weergegee.

Tabel 5: Resultate verkry met die beste parameter-instellings en datavoorstellings vir vyf TiMBL-klassifikasiealgoritmes

Algoritme	Datavoorstelling	Gemiddelde *f*-telling	Akkuraatheid	Uitvoersnelheid
IB1	Eienskapsposisionering 20	0,923	0,928	17,560
IB2	Eienskapsposisionering 38	0,914	0,927	340,288
TRIBL	Waarskynlikheid van laaste letter	0,904	0,910	8,888
TRIBL2	Eienskapsposisionering 38	0,921	0,926	18,753
IGTREE	Aantal letters in sillabe	0,809	0,825	38,923

Tabel 5 dui aan dat die hoogste akkuraatheidsyfer van 92,8% verkry word wanneer TiMBL se IB1 klassifikasiealgoritme gebruik word. Die waarde van hierdie resultaat blyk eers wanneer dit vergelyk word met die akkuraatheidsyfer van 68% wat deur RAGEL behaal word, of die akkuraatheid van 88,61% voor die byvoeging van addisionele eienskappe en parameteroptimalisering.

Tabel 5 dui ook aan hoe groot die verskille is tussen die uitvoersnelhede van die verskillende parameters vir elkeen van die klassifikasiealgoritmes. Hierdie resultaat impliseer dat die keuse van die beste klassifikasiealgoritmes en parameters nie net deur die akkuraatheid van die klassifiseerder bepaal word nie, maar ook deur die gebruiker se voorkeure ten opsigte van klassifiseringsnelheid.

5. Gevolgtrekking

In hierdie artikel is aangetoon hoe ’n lemma-identifiseerder vir Afrikaans ontwikkel is wat ’n akkuraatheid van 92,8% behaal, sonder dat daar van woordsoortinligting gebruik gemaak is om die afrigtingsdata uit te brei. Hierdie akkuraatheidsyfer, wat goed vergelyk met internasionale voorpuntlemma-identifiseerders (Chrupala, 2006), is bereik deur middel van eienskapselektering en parameteroptimalisering.

Toekomswerk sluit in om die afrigtingsdata met meer eienskappe uit te brei, soos byvoorbeeld woordsoortetikette, asook om voorsiening vir onvoltooide deelwoorde en sterk voltooide deelwoorde te maak.

Nog ’n moontlike rigting vir toekomstige navorsing sluit in om LIA te evalueer met meer masjienleeralgoritmes om te bepaal of verdere verhogings in akkuraatheid bereik kan word. *PsearchAL* kan ook uitgebrei word na meer masjienleeralgoritmes en ander klassifiseringstake soos byvoorbeeld woordafbreking en morfologiese analise.

Geraadpleegde bronne

ALOAYDIN, E. 2004. Introduction to machine learning. Cambridge: MIT.

ARONOFF, M. & FUDEMAN, K. 2005. What is morphology? Malden: Blackwell.

BADENHORST, B., CARSTENS, A. & VAN RENSBURG, C. 1992. Basiskursus: aspekte van die Afrikaanse taalkunde. Bloemfontein: Patmos.

BALDWIN, T. & BOND, F. 2003. A plethora of methods for learning English countability. *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*: 73-80.

BAUER, L. 2003. Introducing linguistic morphology. 2nd ed. Edinburgh: Edinburgh University Press.

BOOIJ, G. 2006. Inflection and derivation. (*In* Brown, K., *ed*. Encyclopedia of language and linguistics. 5th ed. Oxford: Elsevier. p. 654-661.)

BOOIJ, G. 2007. The grammar of words. 2nd ed. Oxford: Oxford University Press.

BRITS, J.C., PRETORIUS, R.S. & VAN HUYSSTEEN, G.B. 2006. Automatic lemmatisation in Setswana: towards a prototype. *South African journal of African languages*, 25:37-47.

BYBEE, J. 1985. Morphology: a study of the relation between meaning and form. Philadelphia: Benjamins.

CHRUPALA, G. 2006. Simple data-driven context-sensitive lemmatization. http://www.computing.dcu.ie/~gchrupala/ Date of access: 3 Oct. 2007.

COMBRINK, J.G.H. 1974. Soek: Afrikaans se fleksie. (*In* Odendal, F.F., *red*. Taalkunde – ’n lewe. Kaapstad: Tafelberg. p. 21-30.)

COMBRINK, J.G.H. 1990. Afrikaanse morfologie. Pretoria: Academica.

CROFT, W. 2003. Typology and universals. Tweede uitgawe. Cambridge: Cambridge University Press.

CRYSTAL, D. 1997. A dictionary of linguistics and phonetics. 4th ed. Oxford: Blackwell.

CTEXT. 2005. Afrikaanse speltoetser 3.0, tesourus 1.0 en woordafbreker. Potchefstroom: Noordwes-Universiteit.

DAELEMANS, W. & STRIK, H. 2002. Actieplan voor het Nederlands in de taal- en spraaktechnologie. http://www.cnts.ua.ac.be/Publications/2002/DS02 Datum van gebruik: 17 Jul. 2006.

DAELEMANS, W. & VAN DEN BOSCH, A. 2005. Memory-based language processing. New York: Cambridge University Press.

DAELEMANS, W., VAN DEN BOSCH, A. & ZAVREL, J. 1999. Forgetting exceptions is harmful in language learning. *Machine learning: special issue on natural language learning*, 34(1-3):11-41.

DAELEMANS, W., VAN DEN BOSCH, A., ZAVREL, J. & VAN DER SLOOT, K. 2004. Reference guide to TiMBL. ilk.uvt.nl/downloads/pub/papers/ilk0402.pdf Date of access: 25 Apr. 2007.

DE KLERK, G.J. 1968. Die morfologie van Afrikaans. (*In* Van der Merwe, H.J.J.M., *red*. Afrikaans – sy aard en ontwikkeling. Pretoria: Van Schaik. p. 169-208.)

DE VILLIERS, M. 1973. Afrikaanse klankleer. Kaapstad: Balkema.

DRESSLER, W.U. 1989. Prototypical differences between inflection and derivation. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung*, 42:3-10.

DU TOIT, P.J. 1982. Taalleer vir onderwyser en student. Pretoria: Academica.

ERJAVEC, T. & DŽEROSKI, S. 2004. Machine learning of morphosyntactic structure: lemmatising unknown Slovene words. *Applied artificial intelligence*, 18(1):17-40.

GAUSTAD, T. & BOUMA, G. 2002. Accurate stemming of Dutch for text classification. *Language and computers*, 45(1):104-117.

GROENEWALD, H.J. 2006. Automatic lemmatisation for Afrikaans. Potchefstroom: North-West University. (Unpublished M.Ing. dissertation.)

GROENEWALD, H.J., VAN HUYSSTEEN, G.B. & PUTTKAMMER, M.J. 2007. Evaluating wrapped progressive sampling for automatic algorithmic parameter optimisation. (*In* Angelova, G., Bontcheva, K., Mitkov, R., Nikolov, N. & Nikolov, N., *eds*. Proceedings of recent advances in natural language processing. Bulgaria: Borovets. p. 251-255.)

GUSTAFSON, J., LINDBERG, N. & LUNDEBERG, M. 1999. The August spoken dialogue system. *Proceedings of Eurospeec*: 1151-1154.

HAUSSER, R. 1999. Foundation of computational linguistics: man-machine communication in natural language. Berlin: Springer.

JENKINSON, A.G. 1983. Aspekte van die morfologie van Afrikaans. *Kongresreferate Linguiste Vereniging van Suider-Afrika 1983*: 127-149.

JENKINSON, A.G. 1986. Die diminutief: fleksiemorfeem of afleidingsmorfeem? *Suid-Afrikaanse tydskrif vir taalkunde*, 4(3):13-46.

JENKINSON, A.G. 1993. Die probleem van fleksie en afleiding in Afrikaans. *South African journal of linguistics supplement*, 18:100-122.

JONGEJAN, B. & HALTRUP, D. 2005. The CST Lemmatiser. http://cst.dk/online/lemmatiser/cstlemma.pdf Date of access: 15 Nov. 2006.

KEMPEN, W. 1969. Samestelling, afleiding en woordsoortelike meerfunksionaliteit in Afrikaans. Kaapstad: Nasou.

KNOWLES, G. & DON, Z.M. 2004. The notion of a “lemma”: headwords, roots and lexical sets. *International journal of corpus linguistics,* 9(1):69-81.

KRAAIJ, W. & POHLMANN, R. 1994. Porter’s stemming algorithm for Dutch. (*In* Noordman, L.G.M. & De Vroomen, W.A.M., *eds*. Informatiewetenschap 1994: Wetenschaplike Bijdraen aan de Derde STINFON Conferentie. Tilburg: Stichting Informatiewetenschap Nederland. p.167-180.)

MITCHELL, T.M. 1997. Machine learning. Boston: McGraw-Hill.

MITKOV, R. 2003. The Oxford handbook of computational linguistics. New York: Oxford University Press.

NILSSON, N.J. 1996. Introduction to machine learning. Stanford: Stanford University.

PLISSON, J., LAVRAC, N. & MLADENIC, D. 2004. A rule based approach to word lemmatization. Proceedings of the 7th International Multi-Conference Information Society IS-2004:83-86.

PORTER, M. 1980. An algorithm for suffix stripping. *Program,* 14(3):130-137.

POSTHUMUS, M.J. 1968. Morfologie. (*In* Van der Merwe, H.J.J.M. Inleiding tot die taalkunde. Pretoria: Van Schaik. p. 101-123.)

SCHAPIRE, R.E. 1992. The analysis and design of efficient learning algorithms. Cambridge: MIT.
SPROAT, R. 1992. Morphology and computation. Cambridge: MIT.
STUMP, G.T. 2005. Word-formation and inflectional morphology. (*In* Štekauer, P. & Lieber, R., *eds*. Handbook of word-formation. Dordrecht: Springer. p. 49-71.)
TUGGY, D. 1985. The inflectional/derivational distinction. Workpapers of the Summer Institute of Linguistics at the University of North Dakota 29. Grand Forks: Summer Institute of Linguistics. p. 209-222.
VAN DEN BOSCH, A. 2004. Paramsearch 1.0 Beta Patch 24: wrapped progressive sampling search for optimizing learning algorithm parameters. (*In* Verbrugger, R., Taatgen, N. & Schomaker, L., *eds*. Proceedings of the 16th Belgian-Dutch Conference on Artificial Intelligence. Groningen: Koninklijke Vlaamse Academie van België voor Wetenschappen en Kunsten. 1(1):31-33.)
VAN HALTEREN, H., ZAVREL, J. & DAELEMANS, W. 1998. Improving accuracy in word tagging through combination of machine learning systems. *Computational linguistics*, 27(2):199-230.
VAN RIJSBERGEN, C.J. 1979. Information retrieval. London: Butterworths.
VAN SCHOOR, J.L. 1983. Die grammatika van standaard-Afrikaans. Kaapstad: Lex Patria.
WAGACHA, P.W. 2004. Instance-based learning: *k*-nearest neighbour. http://www.uonbi.ac.ke/acad_depts/ics/course_material/machine_Learning/kNN.pdf Date of access: 16 Jul. 2006.
WITTEN, F. & FRANK, E. 2000. Data mining. San Fransisco: Kaufmann.

Kernbegrippe:

Afrikaans
eienskapselektering
fleksie
lemma-identifisering
masjienleer
morfologie
natuurliketaalprosessering
parameteroptimalisering
tekstegnologie

Key concepts:

Afrikaans
feature selection
inflection
lemmatisation
machine learning
morphology
natural language processing
parameter optimisation
text technology