Daelemans, W., De Pauw, G., Durieux, G., Gillis, S., Hoste, V., Tjong Kim Sang, E. 2000. Zelflerende systemen als instrument voor de taalkunde en de taaltechnologie. . In *Met taal om de tuin geleid*, ed. S. Gillis, J. Nuyts, J. Taeldeman, pp. 81-93. Wilrijk: Universitaire Instelling Antwerpen

Zelflerende systemen als instrument voor de taalkunde en de taaltechnologie

Walter Daelemans
Guy De Pauw
Gert Durieux
Steven Gillis
Véronique Hoste
Erik Tjong Kim Sang[1]

Centrum voor Nederlandse Taal en Spraak (CNTS)
Universiteit Antwerpen, UIA

We introduceren zelflerende systemen als een operationalisering van pre-Chomskyaanse taaltheoretische concepten als analogie en inductie, en laten zien hoe ze kunnen worden toegepast in taalbeschrijving en (computer)taalkunde. Als casus bespreken we twee toepassingen: de automatische inductie van kennis over de beregeling van allomorfie bij Nederlandse diminutieven, en de rol van segmentele fonologische kennis bij de leerbaarheid van Nederlandse klemtoon.

Inleiding

De taalkunde vóór Chomsky baseerde de beschrijving en verklaring van taalkennis, taalgedrag en taalverwerking in belangrijke mate op concepten als analogie en inductie ((Bloomfield 1961; Harris 1951; Pike 1943)). Het ontbreken van een goede formalisering en operationalisering van deze concepten stelde de benadering bloot aan kritiek van aanhangers van alternatieve formalismen die wel konden bogen op een solide wiskundige basis. De theorie van formele talen en grammatica's, in de taalkunde geïntroduceerd door (Chomsky 1957; Chomsky 1965) is zo'n formalisme, en heeft de taalkunde van de laatste decennia sterk beïnvloed. Deze dominantie van regelgebaseerde taalbeschrijvingen heeft als resultaat gehad dat opposities als ***competence*** tegenover ***performance*** en ***core*** tegenover ***periphery*** een centrale rol zijn beginnen spelen in taaltheorie: fenomenen die niet konden worden verklaard vanuit een elegant, compact regelsysteem werden verbannen naar een niet nader beschreven periferie of verklaard als een (onbelangrijk) performance effect. Op die manier verklaarde de generatieve taalkunde zichzelf onbevoegd als model voor psycholinguïstiek en computertaalkunde, waar performance fenomenen centraal

[1] We willen Georges De Schutter bedanken voor de inspirerende onderzoeksomgeving die hij met het CNTS heeft gecreëerd, niet alleen als expertisecentrum voor Nederlandse Taalkunde, maar ook voor meer "exotische" disciplines als corpustaalkunde, taaltechnologie en computationele psycholinguïstiek.
Het onderzoek beschreven in dit artikel kwam mede tot stand dank zij financiele steun van FWO (WO.029.95N 'CLIF' en G.0157.97 'Linguaduct'), EU (ERBFMRXCT98-0237 'Learning Computational Grammars') en IWT (041/3819, 'Corpus Gesproken Nederlands').

staan. Een soortgelijk effect heeft de Chomskyaanse visie op taalverwerving gesorteerd: de stipulatie van het bestaan van zeer specifieke aangeboren taalkennis in het “principes en parameters” model, gebaseerd op de snelheid waarmee zeer jonge kinderen taal leren zonder expliciete instructie, verplaatst het probleem alleen maar van ontogenese naar fylogenese, en binnen de ontogenese van het extraheren van taalkennis uit taaldata naar het extraheren van ***cues*** (voor het bepalen van parameterwaarden) uit taaldata.

De computertaalkunde, in navolging van de taalkunde aanvankelijk sterk gebaseerd op met de hand gemaakte (want “aangeboren”) regels, is onder druk van toepasbaarheid en grotere rekenkracht de laatste tien jaar geleidelijk geëvolueerd naar een meer statistische, corpusgebaseerde en inductieve benadering. De laatste jaren hebben ook technieken uit de theorie van zelflerende systemen (***Machine Learning***, zie (Mitchell 1997) voor een inleiding) aan belang gewonnen. Deze technieken zijn in zekere zin vergelijkbaar met de ***discovery procedures*** uit de pre-Chomskyaanse taalkunde. Ze hebben alleszins als voordelen dat ze (i) zich baseren op werkelijk taalgebruik (door gebruik van corpora) in plaats van introspectie, (ii) aanleiding geven tot robuustere modellen die reëel taalgebruik modelleren en niet alleen de regelmatige “core”, en (iii) toelaten tegelijk acquisitie en verwerking te modelleren.

Binnen het CNTS worden deze technieken de laatste tien jaar toegepast met uiteenlopende doelstellingen die we hier aan de hand van enkele voorbeelden binnen het domein van de Nederlandse taalkunde illustreren.

- Als methode in taalkundig onderzoek. Door de door leertechnieken uit corpora geëxtraheerde kennis te analyseren en door het effect van de aanbieding van verschillende informatiebronnen aan het leeralgoritme te bestuderen, krijgen we meer inzicht in de leerbaarheid van een taalkundig fenomeen, en in de kennisstructuren die nodig zijn om het te verklaren. Een voorbeeld van deze aanpak is het onderzoek naar het geslacht van Nederlandse substantieven. In (Daelemans, Durieux, en Van den Bosch 1998) wordt aangetoond dat puur op basis van de segmentele informatie en syllabestructuur van substantieven, de klasse vrouwelijk met 92% accuraatheid voorspelbaar is, en de klasse mannelijk en onzijdig voor 81%. Een meer uitgebreid voorbeeld, diminutiefvorming in het Nederlands, bespreken we in sectie 3.
- Als model van inductieve taalverwerving. Door taaldata aan te bieden aan een leeralgoritme kunnen we taalverwerving voor specifieke problemen simuleren. De aangeboden informatie kan zodanig worden gemanipuleerd en geannoteerd dat verschillende perspectieven (van radicaal empirisch tot radicaal nativistisch) kunnen worden verkend. In een reeks studies (Daelemans, Gillis, en Durieux 1994; Gillis, Daelemans, en Durieux 1994; Gillis e.a. 1993; Gillis en Durieux 1996; Gillis, Durieux, en Daelemans 1995; Gillis, Durieux, en Daelemans 1996) hebben we aangetoond dat de toekenning van woordklemtoon in het Nederlands uit voorbeelden kan worden geleerd door een eenvoudig geheugengebaseerd leeralgoritme op basis van syllabestructuur en segmentele informatie (zonder de complexe machinerie die het Principes en Parameters model veronderstelt in de metrische fonologie). Bovendien werd een sterke correlatie gevonden

tussen de fouten die het algoritme maakt (wanneer getraind met weinig voorbeelden) en de fouten die kinderen maken bij beklemtoningstaken. We komen in sectie 4 uitgebreider terug op deze resultaten.

- Als methode in de taaltechnologie. Taaltechnologische toepassingen gebaseerd op lerende systemen hebben een aantal voordelen tegenover met de hand gemaakte (meestal regelgebaseerde) systemen: de methode is herbruikbaar voor verschillende problemen, laat snelle systeemontwikkeling toe, is robuust en accuraat, en bevat probabilistische informatie (wat van belang is voor desambiguering bij onvolledige informatie). We hebben met name geheugengebaseerde leersystemen toegepast bij uiteenlopende problemen in de Nederlandse taaltechnologie: bijv. bij de automatische desambiguering van de syntactische woordsoort van woorden in context (***Part of Speech Tagging***, Daelemans, Berck, en Gillis 1996; Daelemans e.a. 1996) en bij de automatische omzetting van de spelling van woorden in een uitspraakrepresentatie (Daelemans en Van den Bosch 1996). De hiervoor ontwikkelde methode werd eveneens gebruikt bij de ontwikkeling van de ***fonilex*** database, een lexicale databank met de Vlaamse uitspraak van Nederlandse woorden.

In het vervolg van dit artikel geven we een korte inleiding tot zelflerende systemen in het algemeen, en geheugengebaseerd leren en regelinductie in het bijzonder (Sectie 2), en gaan in op twee recente toepassingen van deze technieken in de Nederlandse (computer)taalkunde: de beregeling van de allomorfie van diminutieven (Sectie 3), en de leerbaarheid van klemtoon uit segmentele informatie (Sectie 4).

Taallerende systemen

We beperken onze beschrijving hier tot ***gesuperviseerde*** lerende systemen, algoritmen die leren op basis van de analyse van een aantal voorbeelden van de te leren taak. Die taak kan om het even wat zijn: het bepalen van de kredietwaardigheid van een persoon aan de hand van biografische gegevens, het leveren van een diagnose bij de symptomen van een patiënt, het voorspellen van de koers van een aandeel enz. Het gaat hier telkens om classificatieproblemen: gegeven een representatie van een geval in termen van de eigenschappen ervan classificeert het systeem dit probleem als behorend tot één van een eindig aantal vooraf gedefinieerde klassen.

Een zelflerend systeem bestaat uit een ***performance*** gedeelte en een ***leeralgoritme***. Het performance subsysteem zet input om in output gebruikmakend van een of andere representatie van kennis over de uit te voeren taak (bijv. regels, of prototypes): de ***competence*** van het systeem. Het leeralgoritme is een heuristisch zoekalgoritme dat op basis van voorbeelden een optimale kennisrepresentatie zoekt voor het succesvol uitvoeren van de taak. De kwaliteit van een zelflerend systeem wordt vooral gemeten in termen van de accuraatheid ervan: het percentage van de aangeboden nieuwe problemen (inputs) die het getrainde systeem correct kan oplossen. Bijkomende evaluatiecriteria zijn de efficiëntie (in geheugengebruik en snelheid van leeralgoritme en performance systeem) en begrijpelijkheid van de geleerde kennisrepresentaties. Voor taalkundige toepassingen kan dit laatste een belangrijk criterium zijn.

Een taallerend systeem is een zelflerend systeem waarbij dit model van gesuperviseerd, classificatie-gebaseerd leren wordt toegepast op taaltaken. Bijvoorbeeld, het bepalen van de morfosyntactische klasse van een ambigu woord in een zin kan worden geleerd door een aantal voorbeelden te geven van ambigue woorden en hun context als input en de correcte interpretatie als output, en het lerende systeem kan dan kennis afleiden waarmee het soortgelijke desambigueringen kan uitvoeren op nieuwe voorkomens van de aangeboden woorden, of zelfs op nieuwe woorden (zie (Daelemans 1999) voor een overzicht van verschillende leertechnieken die werden toegepast op dit probleem). In de rest van deze sectie beschrijven we twee veelgebruikte leertechnieken die we hebben gebruikt voor de taken beschreven in secties 3 en 4.

Regelinductie

Bij regelinductie zoekt het leeralgoritme een verzameling regels die optimaal de omzetting van input naar output beschrijft. Een regel combineert dan inputeigenschappen als het conditie-gedeelte en een klasse als het actie-gedeelte. De kennisrepresentatie bestaat dus uit een regelverzameling, en de performance-component past de regels toe op de input om een nieuwe output te genereren. Bij het zoeken naar goede regels voor de voorspelling van een bepaalde klasse wordt gebruik gemaakt van de gelijkenissen tussen voorbeelden met dezelfde klasse en van verschillen met de voorbeelden met een andere klasse, en ook van de relevantie van de verschillende kenmerken van de voorbeelden voor het voorspellen van de klasse. Zie (Langley 1996) en (Mitchell 1997) voor achtergrond en beschrijving van algoritmen. Bekende regelinductiesystemen zijn Ripper (Cohen 1995) en ***c4.5*** (Quinlan 1993).

Geheugengebaseerd Leren

Geheugengebaseerd Leren (Memory-Based Learning) is een leertechniek waarbij de kennis van de ***performance*** component van het systeem bestaat uit de voorbeelden zelf; in tegenstelling tot regelinductie wordt hier geen kennisstructuur geëxtraheerd uit de data, maar worden de voorbeelden zelf gebruikt bij het oplossen van nieuwe problemen. De voorbeelden zijn de ***competence*** van het systeem. De performance component lost nieuwe problemen op met de zogenaamde ***nearest neighbour rule***:

gegeven de inputbeschrijving van een nieuw probleem
- zoek de meest gelijkende voorbeelden in het geheugen
- extrapoleer de klasse van het nieuwe probleem vanuit de klassen van deze gelijkende voorbeelden

Uiteraard hangt de kwaliteit van deze aanpak af van hoe goed de gelijkenismaat is die wordt gebruikt om te beslissen dat een probleem gelijkt op een van de voorbeelden. Er bestaan verschillende statistische methodes voor het wegen van de relevantie van voorbeelden, kenmerken, of waarden van kenmerken bij het berekenen van de gelijkenis tussen patronen. Zie (Aha 1997) voor een bundel

artikelen over deze leermethode. ***TiMBL*** (Daelemans e.a. 1999) is een voor onderzoek vrij beschikbaar geheugengebaseerd softwarepakket[2].

Bij de beschouwing van deze technieken vanuit taalkundig perspectief valt het volgende op:

- Het is niet zo dat een empirische (niet-nativistische) benadering van taalleren per definitie ook niet-regelgebaseerd is; een empirische benadering van taalleren op basis van regelinductie is perfect denkbaar. Traditioneel worden empirische modellen geassocieerd met niet-regelgebaseerde benaderingen als neurale netwerken.
- De meeste zelflerende technieken (met uitzondering van evolutionaire algoritmen die gebaseerd zijn op een vorm van ***random search***), maken gebruik van analogie (gelijkenis) en inductie uit voorbeelden om taalkennis en taalgedrag te verwerven. In die zin zijn ze een operationalisering van de technieken voorgesteld in een belangrijk deel van de taalkunde vóór Chomsky.

In de volgende secties bespreken we de resultaten van de toepassing van respectievelijk regelinductie en geheugengebaseerd leren op twee problemen uit de Nederlandse taalkunde: allomorfie van diminutieven en toekenning van primaire klemtoon bij ongelede woorden.

Diminutieven

In ons onderzoek over diminutieven (Daelemans, Berck, en Gillis 1995) werd met behulp van automatische regelinductie een interessante variant van de theorie van (Trommelen 1983) uit woordenboekdata afgeleid en werden een aantal in de fonologie gepostuleerde categorieën zoals bimoraïsche klinkers eveneens “ontdekt” door het lerende systeem.

In Standaardnederlands wordt de diminutief gevormd door concatenatie van het germaanse suffix ***-tje*** aan de basisvorm van het substantief. Dit suffix vertoont allomorfie. Tabel 1 toont de allomorfen en hun frequentie in het corpus waarop de CELEX lexicale databank (Burnage 1990) is gebaseerd.

[2] Zie http://ilk.kub.nl

Allomorf	Voorbeeld	Corpusfrequentie %
-tje	kikker-tje	51
-je	wereld-je	30
-etje	roman-etje	11
-pje	lichaam-pje	4
-kje	koning-kje	4

Tabel 1: Allomorfie bij diminutiefsuffix en de frequentie ervan

Taalkundige analyses (Te Winkel 1862; Kruisinga 1915; Cohen 1958; Haverkamp-Lubbers 1971; Trommelen 1983; Booij 1995; Booij 1984; De Haas 1993) verschillen in de keuze van regels voor allomorf-selectie en in de taalkundige informatie die een rol speelt in deze regels.

We gebruikten regelinductie met behulp van ***c4.5*** (Quinlan 1993) om één van deze theorieën (Trommelen 1983) te onderzoeken. Als voorbeelden voor de taak gebruikten we een verzameling van 3950 substantieven met hun diminutiefvorm uit de ***celex*** database. De inputkenmerken voor ieder substantief waren de fonemische transcriptie van het substantief en de syllabestructuur (onset, nucleus en coda van de drie laatste syllaben), voor elke syllabe de klemtoon (+ of -). De outputklasse was de diminutiefallomorfklasse (-***tje, -je, -etje, -pje, -kje***).

Tabel 2 toont de input en output voor enkele voorbeelden. Daarbij wordt het "="-teken gebruikt om aan te duiden dat er geen waarde is voor dat kenmerk op die plaats.

Kenmerk	biezenmand	big	bijbaan
klemtoon derdelaatste	-	=	=
onset derdelaatste	b	=	=
coda derdelaatste	=	=	=
klemtoon voorlaatste	-	=	+
onset voorlaatste	z	=	b
nucleus voorlaatste	_	=	__
coda voorlaatste	=	=	=
klemtoon laatste	+	+	-
onset laatste	m	b	b
nucleus laatste	_	_	a:
coda laatste	nt	_	n
klasse	JE	ETJE	TJE

Tabel 2: Voorbeeldrepresentaties van de woorden *biezenmandje*, *big*, en *bijbaan*

Een empirisch (uit de data afgeleid) regelsysteem blijkt in staat met 97% accuraatheid de correcte diminutiefallomorf van nieuwe (niet voor training gebruikte) substantieven te voorspellen. Daarbij worden ***-tje*** en ***-je*** bijna perfect geleerd (99%), terwijl de overige allomorfen slechter worden geleerd -***kje*** en ***-pje*** 90%, ***-etje*** 84%). Dat is op zich al interessant, maar de kracht van het gebruik van leertechnieken blijkt vooral uit de mogelijkheid die ze bieden om linguïstische hypotheses te testen. Zo

testten we de hypothese van (Trommelen 1983) dat alleen de segmentele informatie over het rijm van de laatste lettergreep relevant is voor het bepalen van het diminutiefallomorf; wat impliceert dat bijv. klemtoon, een in veel alternatieve theorieën gebruikt concept, irrelevant is.

Door het zelflerende systeem te trainen met verschillende deelverzamelingen van de eerder genoemde kenmerken en de verschillen in hun accuraatheid statistisch te testen, kunnen we dergelijke linguïstische hypotheses testen. Zo stelden we vast dat het gebruik van alle informatie (over de drie lettergrepen) significant betere leerbaarheid oplevert dan het gebruik van alleen het rijm van de laatste lettergreep (contra Trommelen) voor de voorspelling van het allomorf -***etje***. Verder werd inderdaad vastgesteld dat bij beperking tot de laatste lettergreep, de informatie die onset en klemtoon bijdragen verwaarloosbaar is zoals voorspeld door Trommelen.

In Tabel 1 zien we de regels die door ***c4.5*** werden afgeleid uit de data. In dit geval zijn we geïnteresseerd in generaliseringen en concepten die de leertechniek in de data heeft ontdekt eerder dan in de accuraatheid waarmee het systeem nieuwe problemen kan oplossen.

Default klasse is *-tje*

Regel 1:

ALS coda laatste in {rk, k, s, t, lt, p, st ,χt, mt, f, ts, nt, χ, ns, rt, lf, ŋk, nst, ls, ft, rs, lk, rχ, mp, rst, lp, ks, rp, kst, b, lχ, kt}

DAN *-je*

Regel 2:

ALS coda laatste in {rm, lm}

DAN *-pje*

Regel 3:

ALS nucleus laatste in {a, e, ə, u, œy, ɛɪ, i, ø, o, y, ɑu, ɛː, ɒː}

EN coda laatste = m

DAN *-pje*

Regel 4:

ALS nucleus voorlaatste {=, ə}

EN nucleus laatste in {ɪ, ɑ, œ, ɛ, ɔ}

EN coda laatste in {n, l, r, ŋ, rn, m}

DAN *-etje*

Regel 5:

ALS nucleus laatste in {ɪ, ɑ, œ, ɛ, ɔ}

EN coda laatste in {n, l, r, m}

DAN *-etje*

Regel 6:

ALS nucleus voorlaatste in {ɪ, ɛ, i, a, ɔ, ø, ɑ, e, o, ɛɪ, œ, u, ɑu, œy}

EN coda laatste = ŋ

DAN	*-kje*

Regelverzameling geëxtraheerd uit woordenboekdata voor de beregeling van de voorspelling van de allomorfie van de diminutieven

De defaultklasse ***-tje*** wordt gekozen als geen enkele andere regel toepasbaar is. Regel 1 verwijst naar de klasse van obstruenten in de coda van de laatste syllabe als voorspeller van ***-je***. Regels 2 en 3 bepalen het gebruik van allomorf ***-pje***. In regel 3 wordt het concept van ***bimoraïsche*** vocaal "ontdekt" door de leertechniek, en interessanter, schwa wordt hier bij deze klasse gerekend zoals ook in de theorie van Trommelen (dit was niet oncontroversieel in de Nederlandse fonologie). Regel 4 beregelt de keuze voor ***-etje***. De eerste conditie verwijst naar monosyllabische woorden (nucleus voorlaatste is leeg) en woorden met een schwa als nucleus van de voorlaatste lettergreep. De tweede en derde conditie beperken de relevante gevallen verder tot woorden met een monomoraïsche vocaal in de laatste lettergreep en een nasaal of liquida op het woordeinde. Met deze regel worden een aantal moeilijke gevallen op een elegante manier opgelost uitsluitend op basis van segmentele informatie. De regels maken gebruik van slechts drie kenmerken: nucleus en coda van de laatste syllabe, zoals verwacht, maar ook (contra Trommelen) de nucleus van de voorlaatste syllabe. Geen suprasegmentele (klemtoon) kenmerken worden gebruikt.

Klemtoon

Woordklemtoon in Nederlandse ongelede woorden is een complex fenomeen. Getuige hiervan het feit dat in ongelede woorden die uit drie open syllabes bestaan, de klemtoon op de laatste syllabe (paraPLU), de voorlaatste syllabe (piJAma) of de voor-voorlaatste syllabe (VIdeo) kan vallen. De beklemtoning van monomorfematische woorden vertoont de typische karakteristieken van een linguïstisch domein: er zijn een aantal ***sterke generalizeringen*** (bvb. woordklemtoon valt op één van de laatste drie syllabes; syllabes met een sjwa worden nooit beklemtoond en woorden met een sjwa in de laatste syllabe krijgen op een paar beredeneerbare uitzonderingen na altijd klemtoon op de voorlaatste syllabe; klemtoon op de voor-voorlaatste syllabe kan wel met een open voorlaatste syllabe, maar niet met een gesloten voorlaatste), een aantal ***subregelmatigheden*** die niet uitzonderingsloos zijn (bvb. een zware eindsyllabe trekt meestal de woordklemtoon naar zich toe), en een aantal ***uitzonderingen*** die niet in algemene termen te vatten zijn. Een becijfering van hoe woorklemtoonpatronen over lexicon verdeeld zijn, is te vinden in (De Schutter 1993), die de INL lijst samengesteld door Van der Hulst en Langeweg gebruikte) en in (Daelemans, Gillis, en Durieux 1994) die het CELEX databestand raadpleegden.

De beklemtoning van Nederlandse monomorfemen werd intens bestudeerd in het kader van de Metrische Fonologie. (Kager 1989), (Trommelen en Zonneveld 1989), en anderen analyseerden het volwassenensysteem, terwijl Fikkert (1994) en Nouveau (1994) de verwerving van woordbeklemtoning in de eerste taalverwerving bestudeerden. In essentie houdt de metrische theorie in dat alle klemtoonsystemen

gecapteerd kunnen worden in termen van een beperkt aantal parametrische variaties die voorzien zijn in de universele grammatica. Een kind dat het Nederlands verwerft moet op basis van de woorden die het hoort, de correcte instelling van de metrische parameters bepalen. De correcte instelling van de parameters is uiteraard die van de volwassenentaal. De metrische theorie gaat er voorts vanuit dat voor de klemtoonbepaling het concrete segmenteel materiaal waaruit een woord bestaat van ondergeschikt belang is: woordklemtoon is een functie van de configuratie van syllabes gegroepeerd in metrische voeten die verder samengenomen worden in een prosodisch woord. Het segmenteel materiaal is in een kwantiteitsgevoelige taal zoals het Nederlands enkel van belang om het syllabegewicht te bepalen. Voor dat laatste is trouwens enkel de rijmprojectie nodig, d.w.z. enkel de segmentele opbouw van de nucleus en de coda worden gebruikt om uit te maken of een syllabe 'licht' (eindigend op VV[3]), 'zwaar' (eindigend op VC) of 'superzwaar' (eindigend op VXC) is. De Schutter (De Schutter 1993), zie ook De Schutter 1978; De Schutter 1985)) verdedigde een stelling die hier haaks op staat. Hij beargumenteert dat "zowel de kwantiteit als de kwaliteit van de laatste lettergreep in vrij sterke mate decisief zijn voor de plaatsing van de klemtoon." (De Schutter 1993: 69-70), en hij ziet aanwijzingen dat dat ook geldt voor de voorlaatste en de voor-voorlaatste syllabe.

In ons onderzoek van de beklemtoning van Nederlandse monomorfemen hebben we deze stelling experimenteel getest ((Daelemans, Gillis, en Durieux 1994)). Uit CELEX werden 4868 meersyllabische monomorfemen gehaald. Die woorden dienden als leermateriaal voor een taallerend systeem, nl. MBL (Daelemans, Gillis, en Durieux 1994). Het systeem slaat de woorden en hun klemtoonpatroon in zijn geheugen op, en wordt vervolgens getest op woorden die niet in zijn leermateriaal voorkomen. Het systeem moet m.a.w. het klemtoonpatroon van 'nieuwe' woorden voorspellen op basis van analogie met de woorden die in het leermateriaal voorkwamen (voor een operationalisering van het begrip 'analogie' verwijzen we naar Daelemans, Durieux, en Gillis (1997).

De logica van het experiment was heel eenvoudig: als de noodzakelijke en voldoende representatie voor beklemtoning enkel syllabegewichten vereist, dan zal de performantie van het systeem niet significant toenemen als in de representatie segmenteel materiaal wordt gebruikt. Als daarentegen de kwaliteit van de syllabes noodzakelijk is - zoals De Schutter beweert, dan zal de performantie van het zelflerend systeem significant toenemen.

In dit experiment werden syllabegewichten en segmenteel materiaal in het leermateriaal gecodeerd. Dat leverde drie coderingen op van elk woord:

- Codering-1: een woord werd gerepresenteerd als een string van syllabegewichten, d.i. voor elke syllabe werd het syllabegewicht ingevuld, wat overeenkomt met de representatievorm voorgeschreven in de metrische fonologie;
- Codering-2: van een woord werden de rijmprojecties gerepresenteerd, d.i. een representatie waarin de segmenten uit de nucleus en de coda van elke syllabe voorkomen;

[3] 'V' staat voor een korte vocaal, 'VV' voor een lange vocaal, 'C' voor een consnant, en 'X' voor een consonant of een vocaal.

· Codering-3: een woord werd gerepresenteerd als een string van segmenten, d.i. van elke syllabe worden de onsetconsonanten, de nucleus en de codaconsonanten weergegeven.

Van elk woord werden telkens de laatste drie syllabes geselecteerd en gecodeerd en in de twee laatste coderingen werd de syllabestructuur (onset, nucleus en coda) aangegeven. De drie coderingen van het woord ***astronaut*** zien er als volgt uit:

	Klemtoon	Woord	Codering
Codering-1	FINAAL	Astronaut	3 2 5[4]
Codering-2	FINAAL	Astronaut	A s o: = AU t[5]
Codering-3	FINAAL	Astronaut	= A s tr o: = n AU t

Het algoritme werd achtereenvolgens met deze drie coderingen getrained en getest. Daarbij werd het aantal woorden in het leermateriaal systematisch verhoogd van 500 tot 4000 in een 10-fold cross-validation design. Figuur 1 toont de evolutie van het percentage correct beklemtoonde woorden.

Voeg Figuur 1 hier in

De resultaten in Figuur 1 tonen duidelijk dat Codering-1 (syllabegewichten) tot significant slechtere resultaten leidt dan de twee andere coderingen. Codering-1 haalt een successcore van 80%, de twee andere coderingen scoren bijna 90% bij 4000 trainingsitems. Van die twee coderingen waarin de segmenten werden gebruikt, scoort Codering-3 (alle segmentele informatie) meestal net iets beter dan Codering-2 (enkel rijmprojecties), maar het verschil is telkens statistisch niet significant. De conclusie die hieruit voortvloeit is dat een codering van het leermateriaal in termen van syllabegewichten inferieur is aan de coderingen waarin het segmenteel materiaal gebruikt wordt, wat de stelling van De Schutter (1993) onderschrijft.

In dit experiment wordt bovendien aangetoond dat de beklemtoning van monomorfematische woorden kan gebeuren op basis van segmentele informatie, en dat de constructies die in de metrische fonologie worden voorgesteld niet noodzakelijk zijn.

Conclusie

[4] De volgende syllabegewichten werden gebruikt: 1 = superlicht (de nucleus = sjwa), 2 = licht (open syllabe op /VV/), 3 = zwaar (syllabe op /VC/), 4 = superzwaar (syllabe op /VCC/), 5 = superzwaar (syllabe op /VVC/).

[5] De volgende syllabegewichten werden gebruikt: 1 = superlicht (de nucleus = sjwa), 2 = licht (open syllabe op /VV/), 3 = zwaar (syllabe op /VC/), 4 = superzwaar (syllabe op /VCC/), 5 = superzwaar (syllabe op /VVC/).

Zelflerende systemen bieden een interessante methode om hypothesen over taal en taalgebruik te testen. Door het analyseren van de verschillen in leerbaarheid van een linguïstische taak bij verschillende soorten linguïstische kennis als input voor de leerder, kunnen hypothesen over de rol van die kennis worden getest. Door gebruik te maken van regelinductie kunnen bovendien uit de data geëxtraheerde generaliseringen en concepten direct vergeleken worden met de linguïstische theorie. Geheugengebaseerde leertechnieken bieden dan weer een cognitief plausibel model van regelgebaseerd taalgedrag zonder expliciete regels. Beide leertechnieken laten toe pre-Chomskyaanse theoretische concepten als analogie en inductie te operationaliseren en te onderzoeken op corpusdata.

Bibliografie

Aha, D., ed. 1997. ***Lazy Learning***. Dordrecht: Kluwer.

Bloomfield, L. 1961. ***Language***. London: Allen and Unwin.

Burnage, G. 1990. ***CELEX: A guide for users***. Nijmegen: Centre for Lexical Information.

Chomsky, N. 1957. ***Syntactic structures***. 's-Gravenhage: Mouton.

Chomsky, N. 1965. ***Aspects of the theory of syntax***. Cambridge: MIT Press.

Cohen, W. 1995. "Fast effective rule induction". Paper read at ***The12th International Conference on Machine Learning***, Lake Tahoe, California.

Daelemans, W. 1999. Machine learning approaches. In ***Syntactic wordclass tagging***, geredigeerd door H. Van Halteren. Dordrecht: Kluwer Academic Publishers.

Daelemans, W., P. Berck, en S. Gillis. 1995. Linguistics as data mining: The case of Dutch diminutives. In ***CLIN V***, geredigeerd door T. Andernach, M. Moll en A. Nijholt. Twente: Parlevink.

Daelemans, W., P. Berck, en S. Gillis. 1996. Memory-based part of speech tagging. In ***CLIN VI: Papers from the Sixth CLIN Meeting***, geredigeerd door G. Durieux, W. Daelemans en S. Gillis. Antwerpen: Center for Dutch Language and Speech.

Daelemans, W., G. Durieux, en S. Gillis. 1997. Skousen's analogical modeling algorithm: A comparison with Lazy Learning. In ***New methods in language processing***, geredigeerd door D. Jones. London: University College Press.

Daelemans, W., G. Durieux, en A. Van den Bosch. 1998. Toward inductive lexicons: A case study. In ***Proceedings LREC Workshop on Adapting Lexical and Corpus Resources to Sublanguages and Applications***, geredigeerd door P. Velardi. Granada.

Daelemans, W., S. Gillis, en G. Durieux. 1994. The acquisition of stress: A data-oriented approach. ***Computational Linguistics*** 20:421-451.

Daelemans, W., en A. Van den Bosch. 1996. Language-independent data-oriented grapheme-to-phoneme conversion. In ***Progress in speech synthesis***, geredigeerd door J. Van Santen, R. Sproat, J. Olive en J. Hirschberg. New York: Springer Verlag.

Daelemans, W., J. Zavrel, P. Berck, en S. Gillis. 1996. MBT: A memory-based part of speech tagger-generator. In ***Fourth Workshop on Very Large Corpora***, geredigeerd door E. Ejerhed en I. Dagan. Copenhagen.

Daelemans, W., J. Zavrel, K. Van der Sloot, en A. Van den Bosch. 1999. TiMBL: Tilburg Memory Based Learner, version 2.0, Reference Guide. Tilburg: ILK.

De Schutter, G. 1978. ***Aspekten van de Nederlandse klankstruktuur***. Vol. 15, ***Antwerp Papers In Linguistics***. Antwerpen: UIA.

De Schutter, G. 1985. De (niet zo) vele aksentpatronen van Nederlandse ongelede woorden. In ***Hulde-Album Marcel Hoebeke***, geredigeerd door H. Ryckeboer, J. Taeldeman en V. Vanacker. Gent: RUGent.

De Schutter, G. 1993. Klemtoonpatronen in de Nederlandse woordenschat. ***Leuvense Bijdragen*** 82:61-82.

Fikkert, P. 1994. ***On the acquisition of prosodic structure***. Dordrecht: ICG.

Gillis, S., W. Daelemans, en G. Durieux. 1994. Are children Lazy Learners? A comparison of natural and machine learning of stress. In ***Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society***, geredigeerd door A. Ram en K. Eiselt. Hillsdale: Erlbaum.

Gillis, S., W. Daelemans, G. Durieux, en A. Van den Bosch. 1993. Learnability and markedness: Dutch stress assignment. In ***Proceedings of 15th Conference of the Cognitive Science Society***. Hillsdale: Erlbaum.

Gillis, S., en G. Durieux. 1996. Data-driven approaches to phonological acquisition: An empirical test. In ***Proceedings of the UBC International Conference on Phonological Acquisition***, geredigeerd door B. Bernhardt, J. Gilbert en D. Ingram. Somerville: Cascadilla Press.

Gillis, S., G. Durieux, en W. Daelemans. 1995. A computational model of P&P: Dresher & Kaye (1990) revisited. ***Amsterdam Series in Child Language Development*** 5:135-173.

Gillis, S., G. Durieux, en W. Daelemans. 1996. How to set parameters: Analysis of a learning theory. In ***Proceedings of the Groningen Assembly on Language Acquisition***, geredigeerd door C. Koster en F. Wijnen. Groningen: Center for Language and Cognition.

Harris, Z. 1951. ***Methods in structural linguistics***. Chicago: University of Chicago Press.

Kager, R. 1989. ***A metrical theory of stress and destressing in English and Dutch***. Dordrecht: Foris.

Langley, P. 1996. ***Elements of machine learning***. San Francisco: Morgan Kaufmann.

Mitchell, T. 1997. ***Machine Learning***. New York: McGraw-Hill.

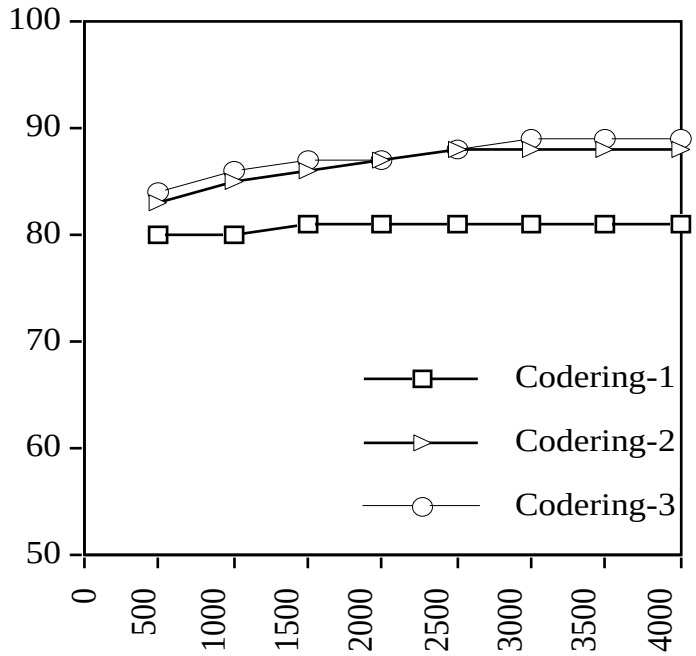
Nouveau, D. 1994. ***Language acquisition, metrical theory, and optimality: A study of Dutch word stress***. Utrecht: OTS Dissertation Series.

Pike, K. 1943. Taxemes and immediate constituents. ***Language*** 19:65-82.

Quinlan, J. R. 1993. ***C4.5: Programs for machine learning***. San Mateo: Morgan Kaufmann.

Trommelen, M. 1983. ***The syllable in Dutch, with special reference to diminutive formation***. Dordrecht: Foris.

Trommelen, M., en W. Zonneveld. 1989. ***Klemtoon en metrische fonologie***. Muiderbergh: Coutinho.



Figuur 1: Percentage correct beklemtoonde testwoorden per codering.