Naar een accentflora voor het Nederlands [1]

Vincent J. van Heuven & Marlon de Graaf
Fonetisch Laboratorium, Universiteit Leiden Centre for Linguistics,
Postbus 9515, 2300 RA Leiden

1. Inleiding

Forensische achtergrond. Stem- en spreekgedrag vormen soms belangrijke aanwijzingen voor de oplossing van misdrijven. Bij gemaskerde overvallen is het stemgeluid van de misdadiger, zoals dat leeft in de herinnering van de slachtoffers, een aanknopingspunt bij de opsporing. Soms zijn zelfs geluidsregistraties beschikbaar van misdadigers. Wie anoniem bedreigd wordt, kan de telefoongesprekken opnemen op band. Aan de hand van akoestische metingen en/of luisteroordelen van experts, kan later worden vastgesteld of de stem van een verdachte dezelfde is als de opgenomen anonieme stem. Maar ook eerder al, bij het zoeken naar een verdachte, kunnen de geluidsopnamen een rol spelen. Aan het stemgeluid kunnen fysieke eigenschappen van de spreker worden vastgesteld. Zo lukt het bijna altijd om uit te maken of de spreker een man of een vrouw is. Minder nauwkeurig, maar altijd beter dan verwacht zou worden op basis van toeval, kunnen we de leeftijd van een spreker voorspellen, of zijn lichaamsgrootte, gewicht of zelfs ras. Daarnaast verklapt het spreekgedrag mogelijk een aantal psychosociale eigenschappen van de spreker. Meestal is gemakkelijk te horen of een spreker een Nederlander is of een buitenlander – en in dat laatste geval kan ook nog vaak worden vastgesteld welke taal de spreker van huis uit heeft, wat weer een aanwijzing vormt voor zijn (etnische) herkomst. Ook bevat het spreekgedrag aanwijzingen aan de hand waarvan experts menen te kunnen vaststellen of iemand de waarheid spreekt dan wel bluft of anderszins liegt.

Regionale herkomst. Ons onderzoek gaat over de mogelijkheid om aan de hand van een korte geluidsopname vast te stellen waar een moedertaalspreker van het Nederlands zijn jeugd heeft doorgebracht. We willen de regionale herkomst van sprekers vaststellen aan de hand van hun uitspraak van het Standaard Nederlands. Bij regionale variatie in een taal, bij voorbeeld van het Nederlands, wordt onderscheid gemaakt tussen dialecten en accenten. Het Nederlands taalgebied kent veel dialecten, d.w.z. variëteiten die langs meerdere taalkundige dimensies verschillen van de standaardtaal. Dialecten hebben andere woorden met andere klinkers en medeklinkers, met afwijkende morfologische eigenschappen, en kennen vaak ook afwijkingen in de zinsbouw en zinsmelodie. Als nu iemand die is opgegroeid in een dialectsprekende omgeving, zich bedient van het Standaardnederlands, dan spreekt hij dat met een regionaal accent. Zo'n regionaal accent wijkt alleen af van de standaardtaal in klankvormelijk opzicht, d.w.z. in de uitspraak van de klinkers en medeklinkers en eventueel in zinsmelodie en ritmiek.

Expert-luisteraars. Er zijn twee groepen Nederlanders die in staat geacht mogen worden om, aan de hand van een geluidsfragment, van een spreker vast te stellen of deze afkomstig is uit een bepaalde regio binnen Nederland. De eerste groep wordt gevormd door de inwoners van die regio zelf. Een geboren en getogen Almeloër zou in staat moeten zijn om op het gehoor vast te stellen of een willekeurige andere spreker een mede-Almeloër is of niet. Een logistiek nadeel van deze methode is natuurlijk dat voor iedere verschillende spreker een andere groep beoordelaars

moet worden ingeschakeld. Een efficiënt alternatief biedt dan de professionele dialectoloog, die op het gehoor zou moeten kunnen bepalen wat de dialectachtergrond is van een spreker zoals die doorklinkt in de standaardtaal.

De professionele, klassieke dialectoloog is een uitstervend ras. Niettemin is het, ook voor forensische toepassingen, belangrijk dat het vermogen om van een willekeurige Nederlander op het gehoor vast te stellen waar deze vandaan komt, blijft voortbestaan. Daarom is, bij wijze van verkenning, dit onderzoek uitgevoerd naar de mogelijkheid om de kennis van de professionele dialectoloog te vangen in een expertsysteem.

Accentflora. Het soort expertsysteem dat ons voor ogen staat, lijkt op de flora zoals die gebruikt wordt door veldbioloog en amateur-botanicus. Een flora is een plantengids met behulp waarvan de gebruiker een in het veld aangetroffen plant kan identificeren (‘determineren’). Onhandig zijn gidsen waarin alle planten in alfabetische volgorde zijn opgenomen, met een foto of tekening. Wie niet weet hoe de plant heet, is gedwongen de hele gids door te ploegen en per plaatje te beslissen of dit wel of niet overeenkomt met het plantje in het veld. Daarom is al sinds jaar en dag de flora beschikbaar, die de gebruiker in een systematisch vraag-antwoordspel in enkele stappen leidt naar de correcte determinatie van het object. Omdat talen en dialecten vaak worden vergeleken met levende organismen (‘dode’ talen, ‘uitstervende’ talen, ‘bedreigde’ talen, ‘levenskrachtige’ talen), is de benaming ‘accentflora’ zeker op zijn plaats. De eis die wij aan onze accentflora stellen is dat de gebruiker met een gering aantal vragen naar specifieke kenmerken van een beluisterde taalvariëteit trefzeker de identiteit van het fragment kan vaststellen.

Vocalisme versus consonantisme. Over de dialecten van het Nederlands (en van de Germaanse talen in het algemeen) wordt wel beweerd dat de onderlinge verschillen vooral te vinden zijn in de klinkersystemen terwijl de variatie in de medeklinkers veel geringer zou zijn. Wij kennen geen onderzoek waarin de relatieve bijdrage van klinker- versus medeklinkerverschillen aan de veelheid van dialecten van het Nederlands (of andere talen) systematisch gekwantificeerd is.[2] Een fundamentele vraag die we aan de hand van onze onderzoeksresultaten willen beantwoorden, is of de regionale accenten van het Nederlands sterker gedifferentieerd zijn in hun vocalisme dan in het consonantisme.

Corpus Gesproken Nederlands. Op dit ogenblik wordt in Nederland en België een grote hoeveelheid spraak (ca. 10 miljoen woorden) verzameld. Deze materiaalverzameling is het Corpus Gesproken Nederlands, of kortweg CGN. De sprekers in dit CGN zijn getrokken met gelijkmatige spreiding over het Nederlandse en Belgische grondgebied in een verhouding van 2:1. Het CGN-materiaal is verdeeld over veertien verschillende spraaktypen, variërend van voorgelezen teksten (nieuwsberichten) tot vrije conversatie tussen bekenden. Het aantal sprekers per spraaktype is variabel. Wanneer relevant, is de regionale herkomst in het CGN per spreker vastgelegd met behulp van de eerste drie cijfers van de postcode, hetgeen het herkomstgebied beperkt tot een straal van enkele kilometers. Voor details over het CGN verwijzen wij verder naar de website (lands.let.kun.nl/cgn/home.htm). Onderdeel van de opdracht die wij onszelf gesteld hebben, is dat we willen nagaan in hoeverre het CGN bruikbaar is om de verkenning uit te voeren.

2. Methode

Aanpak. De beoogde accentflora is in feite een beslissingsboom die een gebruiker in staat stelt een object te identificeren als een uniek element uit een eindige verzameling aan de hand van informatie over een beperkt aantal onderscheidende kenmerken van die objecten. Regionale accenten van het Nederlands verschillen per definitie (zie boven) alleen in hun uitspraakkenmerken. Vanuit praktisch oogpunt lijkt het daarom het meest doenlijk om per regionale variant alle hoorbare afwijkingen ten opzichte van het ABN (de niet regionaal gekleurde uitspraak van het Standaardnederlands) op het gehoor vast te stellen en te benoemen in termen van articulatorisch-fonetische dimensies. Tevens kunnen we de omstandigheid benutten dat de *r* in het ABN vrije variatie toestaat tussen tongpunt- en huigarticulatie, terwijl veel regionale variëteiten van het Nederlands een unieke keus maken uit de twee typen *r*. Als gevolg van de gekozen segmentele aanpak vallen afwijkingen in prosodie (ritmiek en intonatie) af: de fonetiek is op dit moment niet in staat de prosodische verschillen tussen de regionale variëteiten van het Nederlands, onderling noch in relatie tot het ABN, te parametriseren.[3]

Wanneer de verschillen benoemd en geteld zijn, kan vervolgens worden nagegaan welke verschillen optimaal differentiëren tussen de regionale variëteiten van het Nederlands. In het ideale geval zal blijken dat iedere regionale variëteit op een unieke wijze afwijkt van het ABN. Als onze hypothese correct is, zou bovendien moeten blijken dat de differentiatie tussen de regionale variëteiten sterker is in de klinkerafwijkingen dan in de medeklinkerafwijkingen.

Selectie van spraaktype. Allereerst is besloten het onderzoek te beperken tot alleen spraak geproduceerd door Nederlanders. Het Belgisch deel van het CGN-corpus is niet in de beschouwing betrokken. Wij hebben voorts besloten het onderzoek te richten op spontaan geproduceerde, interactieve spraak. In het CGN zijn drie spraaktypen opgenomen die in deze categorie vallen, te weten ‘face-to-face’ conversatie tussen twee of meer bekenden (3 miljoen woorden, waarvan inmiddels ruim 1 miljoen beschikbaar als uitgeschreven tekst), telefoondialogen (eveneens 3 miljoen woorden voorzien, maar nog geen uitgeschreven tekst beschikbaar) en zakelijke onderhandelingen (175.000 woorden voorzien, waarvan nu 136.000 uitgeschreven). Wij gaan ervan uit dat de regionale achtergrond het sterkst zal doorklinken in de minst formele gesprekssituatie, d.w.z. de vrije conversatie tussen vrienden en bekenden. Bovendien is van dat type spraak het meeste (uitgeschreven) materiaal beschikbaar in het CGN, waardoor de (regionale) verscheidenheid aan sprekers maximaal kan zijn.

Selectie van sprekers en spraaksamples. Om de verkenning te beperken tot duidelijke gevallen van regionaal accent hebben we alleen sprekers toegelaten die voldeden aan de eis eis dat zij geboren en minstens tot hun 16e levensjaar getogen zijn in dezelfde woonplaats, zoals aangegeven in de meta-informatie van het CGN. In het gekozen spraaktype bevindt zich materiaal van 40 mannen en 42 vrouwen die voldoen aan het gestelde criterium.[4]

Dit materiaal is door de CGN-organisatie verrijkt met een uitgeschreven tekst en automatisch voorzien van een opdeling in interpauzale eenheden (IPU ofwel interpausal unit), d.w.z. stukken verbonden spraak die lopen van pauze (d.w.z. fysieke stilte) tot pauze. Van iedere spreker hebben wij de IPU’s met hinderlijke achtergrondgeluiden verwijderd, en daarna de IPU’s per spreker geordend op lengte (uitgedrukt in aantal woorden. De langste IPU’s zijn vervolgens aaneengeregen totdat per spreker een minuut ononderbroken spraak verkregen was.

De 82 resulterende spraakstalen van een minuut elk hebben wij door twee taalweten-

schappers laten beluisteren die beroepshalve grote ervaring hebben in het determineren van de regionale herkomst van Nederlandse sprekers. Hun opdracht was om van iedere spreker op het gehoor vast te stellen of het plausibel was dat de spreker geboren en getogen was in de plaats die door ons op grond van de CGN-meta-informatie was aangegeven. Alleen als beide deskundigen van mening waren dat de spreker hoorbaar afkomstig was uit de opgegeven woonplaats, werd de spraakstaal in onze uiteindelijke steekproef opgenomen.[5] Als gevolg van deze procedure hebben wij de sprekers beperkt tot die gevallen waarvan de experts zelf redelijk zeker zijn. Het leek ons in het kader van een eerste verkenning onrealistisch een expertsysteem te implementeren dat een grotere nauwkeurigheid nastreeft dan de te modelleren experts zelf kunnen bereiken. Als gevolg van deze nadere selectie is ons onderzoek beperkt tot alleen vrij duidelijke gevallen van regionaal accent, n.l. 13 mannen en 11 vrouwen, verspreid over 17 woonplaatsen (zie tabel 1).

Kenmerkbeschrijving. Van de resterende 24 sprekers (één minuut spraak per spreker) is een gedetailleerde beschrijving gemaakt, waarbij ieder foneem dat hoorbaar afweek van de ABN-norm, werd geïdentificeerd en beschreven in fonetische termen. Hierbij werd de eerste annotatie uitgevoerd door de tweede auteur, waarna de eerste auteur de annotaties op het gehoor controleerde, en zonodig aanvulde. Bij de beschrijving van verschillen tussen spraakstaal en de ABN norm, mochten de transcribenten gebruik maken van alle fonetische dimensies die omschreven worden in het handboek van de International Phonetic Association (IPA, zie ook Rietveld & van Heuven 2001: 390-396), en daarbij ook gradaties aangeven. Daarnaast werd bij iedere *r* aangegeven wat zijn articulatieplaats was (tongpunt versus huig).[6] Bij inhoudelijke discrepanties tussen de transcribenten werd consensus bereikt in gezamenlijke sessies waarbij een derde foneticus aanwezig was om in onbesliste gevallen de knoop door te hakken via meerderheid van stemmen.

De consensustranscriptie van de afwijkingen en *r*-varianten werd vervolgens terminologisch geharmoniseerd, waarna de voorkomens van afwijkende klanken en specifieke *r*-realisaties per spreker geteld werden. Voor een aantal woonplaatsen waren meerdere sprekers beschikbaar. In zulke situaties is per woonplaats de frequentie van de verschijnselen gemiddeld over het aantal sprekers. Per verschijnsel is vervolgens vastgesteld hoe vaak zich dit in een woonplaats voordeed. Wanneer een verschijnsel slechts een of twee keer werd aangetroffen in de minuut spraak, werd het als een toevalstreffer buiten beschouwing gelaten.[7]

3. Op zoek naar differentiërende kenmerken

Op deze manier hebben we een lijst van in totaal 34 verschillende verschijnselen opgesteld, die bij minstens één regionaal accent kenmerkend waren. Van deze verschijnselen hadden er 21 betrekking op afwijkingen ten opzichte van het Standaardnederlands in de klinkers, en waren er 13 die medeklinkerafwijkingen of *r*-realisaties betroffen. Dit is een eerste aanwijzing dat de regionale accenten van het Nederlands sterker van elkaar en van de standaardtaal afwijken in het klinker- dan in het medeklinkersysteem.

De 21 kenmerkende klinkerverschijnselen staan opgesomd in tabel 1. In de kolommen staan de 17 woonplaatsen. Het aantal opgenomen sprekers is per woonplaats aangegeven (m = man, v = vrouw). In de cellen van de matrix staat de (gemiddelde, zie boven) absolute frequentie waarmee het betreffende verschijnsel is geconstateerd in de minuut spraak van iedere spreker.

Woonplaatsen staan meer naar links in de tabel naarmate het regionaal accent dat daar gesproken wordt, in meer opzichten afwijkt van het ABN. De klinkerverschijnselen zijn van

boven naar beneden geordend van meer naar minder voorkomend in de 17 woonplaatsen.[8]

Tabel 1. Overzicht van geconstateerde afwijkingen t.o.v. het ABN in het klinkersysteem van 17 regionale accenten.

	Klinkerkenmerken	Haaksbergen, O (m)	Werkendam, NB (m)	Nijmegen, G (mmm)	Tholen, Z (m)	Oldenzaal, O (m)	Maastricht, L (v)	Rijswijk, ZH (v)	Arnhem, G (mm)	Dinteloord, Z (m)	Heerlen, L (v)	Nuenen, NB (v)	Alblasserdam, ZH (m)	Hellendoorn, O (m)	Hulsel, NB (vvv)	Tilburg, NB (m)	Clinge, Z (v)	Cadzand, Z (v)
1.	*ee* monoftong	10		5		3	8					4		6				
2.	*e* vernauwd				6				4			3			9	7		
3.	*ei* monoftong; gecentreerd	4	6	6		7	4											
4.	*oo* monoftong	6		5		8			3					4				
5.	*aa* teruggetrokken			4	4				3	8								
6.	*a* teruggetrokken			3	3		3	12										
7.	*a* front	12	9			5												
8.	*ee* diftongisch(er)		6							7			7					
9.	*e* verwijd						3			6	8							
10.	*ei* inzet verwijd				6					7								
11.	*i* vernauwd								6			4						
12.	*ui* vernauwd; weinig liprond.	4						3										
13.	*aa* naar voren; vernauwd												6					
14.	*e* genasaleerd				5													
15.	*eu* vernauwd; weinig liprond.							3										
16.	*au* inzet vernauwd		4															
17.	*ooi* inzet vernauwd		3															
18.	*oo* diftongisch(er)							3										
19.	*o* vernauwd										3							
20.	*a* vernauwd										3							
21.	*au* monoftong; vernauwd	3																
Kenmerken afwijkend		**6**	**5**	**5**	**5**	**4**	**4**	**4**	**4**	**4**	**3**	**3**	**2**	**2**	**1**	**1**	**0**	**0**

Uit tabel 1 blijkt dat 13 regionale accenten gekenmerkt worden door een unieke set klinkerafwijkingen van het ABN. Het accent van Haaksbergen wijkt in meer klinkeropzichten af van het ABN (namelijk zes) dan dat van enige andere woonplaats. Slechts vier accenten zijn niet uniek onderscheiden maar delen hun afwijkingen paarsgewijs; deze vier zijn in de tabel grijs gemaakt. Dit zijn de vier accenten waarvan de klinkeruitspraak in slechts één opzicht van het ABN afwijkt (Tilburg en Hulsel, waar de ongespannen *e* vernauwder (geslotener) wordt uitgesproken) of zelfs in geen enkel opzicht (Clinge, Cadzand).

Tabel 2 presenteert op analoge wijze de gegevens voor de *r*-varianten en de medeklinkerafwijkingen. In deze tabel staan eerst de *r*-realisaties vermeld, daarna volgen de medeklinkerafwijkingen. Bij de opgaaf van bijzondere *r*-realisaties en afwijkingen van het ABN hebben we de annotaties gedecomponeerd in individuele fonetische kenmerken. Per regel in de matrix staat aangegeven hoe veel keer een bijzonder of afwijkend kenmerk is aangetroffen in de minuut spraak (eventueel gemiddeld over de sprekers uit dezelfde woonplaats).

Tabel 2. Overzicht van *r*-realisaties en van afwijkende medeklinkerkenmerken in 17 regionale accenten.

	Medeklinkerkenmerken	Haaksbergen,, O (m)	Heerlen, L (v)	Tilburg, NB (m)	Nijmegen, G	Hulsel, NB (vvv)	Nuenen, NB (v)	Hellendoorn, O (m)	Arnhem, G (mm)	Cadzand, Z (v)	Alblasserdam, ZH	Maastricht, L (v)	Rijswijk, ZH (v)	Werkendam, NB	Tholen, Z (m)	Oldenzaal, O (m)	Dinteloord, Z (m)	Clinge, Z (v)
1.	*r* tongpunt	17		26	5	12	24	27	12	14	21			24	25	16	20	7
2.	*r* huig		10		11				11			10	7					
3.	*r* +tril	8	4	16	4	12	21	21	7	8	7	10	3	17	11	10	7	7
4.	*r* –tril	9	6	10	12		3	8	16	6	14		4	7	14	6	13	
5.	*g* stemhebbend	11	14	38	17	21	20		11			11						
6.	*g* velair	16	14		17	21	20	20	20		15	11	15	16	15	14	11	11
7.	*g* post-palataal			38														
8.	*g* faryngaal									19								
9.	*g* geschraapt							12										
10.	*l* gevelariseerd		3								9		4					
11.	*n* syllabisch	5								8								
12.	*v,z* stemloos				13													
13.	*w* bilabiaal			4														
	Kenmerken afwijkend	**3**	**3**	**3**	**3**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**1**	**1**	**1**	**1**	**1**	**1**

Bij de *r*-varianten constateren we dat het voornaamste onderscheid ligt in de articulatieplaats: er lijkt een vrij scherpe tweedeling tussen regionale variëteiten met een tongpunt-*r* en die met een huig-*r*. Alleen in Arnhem en Nijmegen komen tongpunt- en huig-*r* naast elkaar voor, wat niet ongebruikelijk is voor de grote(re) steden. Alle andere regionale accenten kennen slechts één articulatieplaats voor de *r*. Uit Nijmegen en Arnhem zijn meerdere sprekers (resp. drie en twee) in het materiaal opgenomen, zodat de variatie tussen tongpunt- en huig-*r* dan het gevolg kan zijn van middeling over sprekers die alleen een tongpunt-*r* bezigen of alleen een huig-*r*.

Binnen de accenten met een tongpunt-*r* geldt dat de getrilde en de getapte variant bijna altijd naast elkaar voorkomen. In Hulsel en Clinge komt echter uitsluitend de getrilde tongpunt-*r* voor. Ook bij de huig-*r* accenten gaan de getrilde en de niet-getrilde variant meestal samen. Alleen de Maastrichtse spreker kent uitsluitend de getrilde huig- *r*.

Bij de consonantafwijkingen van het ABN neemt de uitspraak van de achter-fricatief een bijzondere plaats in. Wij gaan ervan uit dat de oppositie tussen de traditioneel stemhebbende (en dan velaire) *g* en de stemloze (en dan uvulaire) *ch* in het ABN niet meer bestaat, ook niet intervocalisch of in anlaut na [+stem].[9] Wie het onderscheid (nog) heeft, spreekt dus niet volkomen ABN maar laat een gewestelijke invloed toe. Het blijkt dan dat in op één na alle onderzochte woonplaatsen de *g* afwijkend wordt uitgesproken van de *ch*. Dikwijls is de articulatie meer naar voren (dus velair of zelfs postpalataal), soms ook meer naar achteren (faryngaal).

In het medeklinkersysteem blijkt geen enkel accent in meer dan drie eigenschappen af te wijken van het ABN. Dit is opnieuw een aanwijzing dat het vocalisme bepalender is voor de regionale variëteiten van het Nederlands dan het consonantisme. Bovendien blijken er vier accenten te zijn die niet van elkaar onderscheiden kunnen worden in het medeklinkersysteem, ook niet als we de *r*-realisaties mede in de vergelijking betrekken: dit zijn de accenten van Werken-

dam, Dinteloord, Oldenzaal en Tholen. Bij het klinkersysteem kon de keus altijd ingeperkt worden tot twee alternatieven. Ook in dit opzicht is het vocalisme dus sterker dan het consonantisme.

Wanneer we echter zowel gebruik mogen maken van klinkerafwijkingen als van medeklinkerverschijnselen, dan blijken alle 17 de regionale variëteiten uniek onderscheiden, zowel onderling als van het ABN. De vier plaatsen die niet onderscheiden zijn in de medeklinkereigenschappen, overlappen in het geheel niet met de twee maal twee plaatsen die hun klinkereigenschappen delen.

4. Het beslissingsalgoritme

Nu we hebben vastgesteld dat, en hoe, de 17 regionale variëteiten alle uniek van elkaar en van het ABN onderscheiden zijn, komen we toe aan de opgave om een maximaal efficiënt en robuust beslisalgoritme op te stellen waarmee de 17 accenten gedetermineerd kunnen worden.

Het algoritme is maximaal efficiënt als het de gebruiker toestaat om met een gemiddeld minimum aantal vragen te komen tot de juiste oplossing. Nu zijn er accenten die uniek gekenmerkt zijn doordat zij van het ABN afwijken in één eigenschap die in een enkel ander accent voorkomt. Zo komt de bilabiale *w* in ons materiaal als afwijking voor alleen in het accent van Tilburg. Toch is het niet efficiënt om de beslisboom dan te beginnen met de vraag of in het spraakfragment de *w* bilabiaal wordt uitgesproken. Zouden de dat wel doen, dan wordt het aantal te beantwoorden vragen voor alle 16 andere accenten één hoger dan anders het geval zou zijn geweest. We moeten dus op zoek naar binaire vragen (ja/nee-vragen) die het aantal kandidaten zo veel mogelijk halveren, dus naar vragen die voor de helft van de 17 variëteiten positief beantwoord kunnen worden en voor de andere helft negatief. Iets technischer, we stellen de vragen in de beslisboom in volgorde van de entropie van de onderscheidende kenmerken.[10] Omwille van de robuustheid van het beslissysteem krijgen vragen naar kenmerken die (bijna) altijd in het regionaal accent aanwezig zijn, voorrang boven vragen naar verschijnselen die minder vaak optreden. De vragen aan de hand waarvan de beslisboom wordt afgelopen zijn hiërarchisch geordend, d.w.z. een lager geordende vraag (vervolgvraag) is alleen van toepassing als hoger geordende vragen daar via een specifiek pad van met “ja” en/of “nee” beslissingen naartoe hebben geleid. Bij de vragen die gesteld worden maken we geen onderscheid meer tussen vragen naar klinkerafwijkingen, medeklinkerafwijkingen en *r*-varianten; vragen naar deze eigenschappen kunnen op iedere positie in de hiërarchie gesteld worden.

In figuur 1 is de accentflora georganiseerd als een binair vertakkende beslisboom. Om de boom compact weer te geven staat bij iedere splitsing het positieve antwoord op een impliciet gehouden vraag vermeld als “+ kenmerk X”, terwijl op hetzelfde hiërarchische niveau elders in de boom bij “– kenmerk X” de tak begint die correspondeert met het negatieve antwoord. Het toegevoegde kenmerk achter de gedetermineerde plaatsnaam dient als (redundante) controlevraag.

De eerste vraag die het algoritme stelt is: “Heeft het accent een stemhebbende *g*?” Als het antwoord op deze vraag “ja” is, dan blijven er acht mogelijke woonplaatsen over: Arnhem, Nijmegen, Tilburg, Maastricht, Heerlen, Nuenen, Haaksbergen en Hulsel. De overige negen mogelijkheden zijn nu uitgeschakeld. Deze vraag levert bij 17 te determineren objecten dus de maximale entropie op en wordt daarom als eerste gesteld. Voor de ingeperkte groep van acht wordt nu de vervolgvraag gesteld: “Is de *ee* gemonoftongeerd?” Indien “ja” dan blijven nog maar vier plaatsen over: Nijmegen, Haaksbergen, Maastricht en Nuenen, de andere vier vallen af. Ook deze vraag bereikt de maximale entropie: een vier-vier splitsing bij acht mogelijkheden. Voor de

resterende vier plaatsen stelt het algoritme nu de vraag: “Is de *oo* een monoftong?” Indien ja, dan blijven over Nijmegen en Haaksbergen, de andere twee plaatsen vallen af; opnieuw is de maximale entropie gehaald: een twee-twee splitsing bij vier alternatieven. Om de unieke keus te kunnen maken uit het overgebleven paar, stelt het algoritme de laatste vraag: “Is er sprake van verstemlozing van *v* en *z*?” Indien ja, dan komt de spreker uit Nijmegen, zo niet dan komt hij uit Haaksbergen. Zo wordt met vier binaire vragen de regionale herkomst van een Nijmegenaar of een Haaksbergenaar gedetermineerd.

+ *g* stemhebbend		
+ *ee* monoftong		
+ *oo* monoftong		
+ *v,z* stemloos	→ **Nijmegen**	[+ *g* velair]
– *v,z* stemloos	→ **Haaksbergen**	[+ tongpunt-*r*]
– *oo* monoftong		
+ huig-*r*	→ **Maastricht**	[+ *e* vernauwd]
– huig-*r*	→ **Nuenen**	[+ *e* vernauwd]
– *ee* monoftong		
+ *e* vernauwd		
+ *i* vernauwd	→ **Arnhem**	[+ *aa* teruggetrokken]
– *i* vernauwd	→ **Hulsel**	[+ tongpunt-*r*]
– *e* vernauwd		
+ huig-*r*	→ **Heerlen**	[+ *l* gevelariseerd]
– huig-*r*	→ **Tilburg**	[+ *g* postpalataal]
– *g* stemhebbend		
+ *ee* diftong		
+ *oo* monoftong		
+ *ei* monoftong, gecentreerd		
+ *ui* vernauwd, ontrond	→ **Haaksbergen**	[+ *n* syllabisch]
– *ui* vernauwd, ontrond	→ **Oldenzaal**	[+ *a* naar voren]
– *ei* monoftong, gecentreerd	→ **Hellendoorn**	[+ *g* geschraapt]
– *oo* monoftong		
+ *a* fronted	→ **Werkendam**	[+ *ooi* inzet vernauwd]
– *a* fronted		
+ *l* gevelariseerd	→ **Alblasserdam**	[+ tongpunt-*r*]
– *l* gevelariseerd	→ **Dinteloord**	[+ *ei* inzet verwijd]
– *ee* diftong		
+ *a* teruggetrokken		
+ huig-*r*	→ **Rijswijk**	[+ *l* gevelariseerd]
– huig-*r*	→ **Tholen**	[+ *ei* inzet verwijd]
– *a* teruggetrokken		
+ *g* faryngaal	→ **Cadzand**	[+ *n* syllabisch]
– *g* faryngaal	→ **Clinge**	[+ *g* velair]

Figuur 1. Binaire beslisboom voor 17 regionale uitspraken van het Standaard Nederlands.

De meeste woonplaatsen worden gedetermineerd met vier vragen, in een enkel geval is een vijfde vraag nodig, maar nooit een zesde. Veel efficiënter zou het algoritme nauwelijks kunnen zijn (zie ook noot 10).

5. Slot

In deze conclusieparagraaf komen we terug op een aantal vragen die, expliciet of impliciet, aan de orde zijn gesteld in de inleiding van dit artikel.

Voor ons als taalkundigen is het belangrijk te constateren dat de regionale accenten van het Nederlands, in ieder geval de 17 accenten die wij uit het CGN-subcorpus spontane dia- en multiloog hebben geselecteerd, uniek onderscheiden kunnen worden van elkaar en van het ABN. Unieke onderscheidbaarheid geldt dus niet alleen voor de dialecten van het Nederlands (per definitie, als twee variëteiten niet onderscheiden zijn, kunnen zij nooit aparte dialecten zijn) maar ook voor de regionale uitspraak van de standaardtaal. Bovendien hebben we kunnen vaststellen dat de eigenschappen in het klinkersysteem hierbij sterker differentiëren tussen de regionale accenten dan de eigenschappen van het medeklinkersysteem. Onze voorspelling dat het vocalisme sterker differentieert dan het consonantisme is hiermee bevestigd.

De belangrijkste praktische conclusie kan zijn dat het vooralsnog zeer wel mogelijk lijkt om te komen tot een efficiënt en robuust expertsysteem waarmee regionale variëteiten van het Standaardnederlands geïdentificeerd kunnen worden. Voorwaarde daarbij is dat de gebruiker de beschikking heeft over een spraakstaal van een minuut ononderbroken spraak van een individu, waarin de afwijkingen van het ABN zijn aangegeven. Het beschrijven van de afwijkingen moet worden gedaan door een fonetisch geschoolde luisteraar met een scherp besef van de uitspraaknorm voor het ABN. In de praktijk zal dit voorwerk dus moeten worden gedaan door een foneticus en moedertaalspreker van het Nederlands. Het is echter niet nodig dat deze transcribent enige parate kennis heeft van de Nederlandse dialecten of van regionale accenten van de standaardtaal. Deze laatste kennis kan worden geïmplementeerd in het expertsysteem.

De tweede praktische conclusie is dat de materiaalverzameling in het Corpus Gesproken Nederlands slechts in beperkte mate bruikbaar is geweest bij de beantwoording van onze vraagstelling. Zelfs in de meest omvangrijke component van het CGN, het subcorpus spontane dia- en multiloog, zijn per woonplaats slechts enkele sprekers beschikbaar, meestal slechts één. Voor de meeste toepassingen van het CGN is vergaande spreiding van sprekers over het Nederlandse taalgebied gewenst, maar voor ons doel is het daarenboven ook noodzakelijk om uit iedere woonplaats meerdere sprekers te kunnen bestuderen, liefst gelijkelijk verdeeld over mannen en vrouwen. Zolang we slechts de beschikking hebben over één spreker per woonplaats zullen we nooit met voldoende zekerheid kunnen vaststellen of de geconstateerde afwijkingen van het ABN idiosyncratisch zijn of het plaatselijk accent reflecteren. Bij vergelijking van meerdere sprekers uit dezelfde woonplaats kan de stabiliteit van de afwijkingen gemakkelijk in kaart worden gebracht. Inmiddels is uit onderzoek van Van Bezooijen & Ytsma (2000) naar de eigenschappen van plaatselijke accenten gebleken dat de mate van afwijkendheid ten opzichte van het ABN soms sterk varieert tussen de sprekers van een plaatselijk accent. Onduidelijk is vooralsnog of de sprekers in hun afwijkingen implicationeel geordend zijn. In dat geval is het voldoende als de minst afwijkende spreker toch nog uniek onderscheiden is van zijn soortgenoten in de andere plaatselijke accenten; alle andere sprekers zijn dan *a fortiori* ook onderscheiden. Om dit te doen voor het gehele Nederlandse taalgebied is een substantiële onderzoeksinspanning. Niettemin weten we op grond van de hierboven beschreven verkenning dat zo'n onderneming kansrijk is.

Noten

[1] Deze bijdrage is gebaseerd op een doctoraalscriptie n.a.v. de afstudeerstage van de tweede auteur aan het Nederlands Forensisch Instituut te Rijswijk. Wij danken Mw. dr. Tina Cambier en Drs. Ton Broeders (beiden NFI) voor hun aandeel in de stagebegeleiding. Tevens zijn wij dank verschuldigd aan Dr. Ton Goeman (Meertens Instituut te Amsterdam) voor zijn expert-beoordeling van spraakfragmenten uit het Corpus Gesproken Nederlands en aan Mw. dr. Renée van Bezooijen (KU Nijmegen) voor commentaar op het manuscript.

[2] Het werk van Hoppenbrouwers & Hoppenbrouwers (2001) bevat wel kwantitatieve gegevens waarmee vraag of de Nederlandse dialecten (anders dan regionale accenten) sterker gedifferentieerd zijn in het klinkersysteem dan in het medeklinkersysteem, minstens ten dele beantwoord zou kunnen worden. De auteurs zelf stellen de vraag echter niet.

[3] Er zijn aanwijzingen dat de Nederlandse dialecten hoorbaar en herkenbaar verschillen in hun repertoire aan zinsmelodieën (van Bezooijen & Gooskens, 1999).

[4] Ons criterium is strenger dan de voorwaarde die de CGN-organisatie heeft gehanteerd voor toelating van een spreker tot het subcorpus face-to-face conversatie. In het Nederlandse deel van dit subcorpus is spraak opgenomen van in totaal 154 mannen en 158 vrouwen die voldoen aan het CGN-criterium dat zij vanaf hun 4e (of eerder) tot hun 16e levensjaar (of later) in dezelfde plaats gewoond hebben.

[5] In de praktijk is een eerste selectie gemaakt door Ton Broeders van het NFI. Alleen de sprekers die door hem beoordeeld waren als herkenbaar afkomstig uit de opgegeven woonplaats, werden in tweede instantie voorgelegd aan Ton Goeman, dialectoloog aan het Meertens Instituut.

[6] Deze kenmerkbeschrijving van de klankafwijkingen en *r*-varianten lijkt op de aanpak van Collins & Mees (1999). Deze auteurs geven per klinker- en medeklinkerfoneem in de standaardtaal (resp. pp. 131-138 en 189-202) een (niet-uitputtend) overzicht van regionale uitspraakvarianten, voor de (middel-) grote steden en per provincie.

[7] Wij zijn er gemakshalve vanuit gegaan dat een spraakfragment van een minuut alle fonemen van de taal voldoende kans geeft om een aantal keren voor te komen. In een minuut spraak wordt ca. 900 keer een foneem gerealiseerd (op basis van gemiddeld drie fonemen per lettergreep, en vijf lettergrepen per seconde). De gebruiksfrequenties van de ca. 35 fonemen convergeren al bij kleine steekproeven naar een vaste verdeling (zgn. fonetisch gebalanceerde tekst). Wij hebben er daarom van afgezien om in detail de relatieve frequentie van de verschijnselen te bepalen.

[8] Wij kiezen er in dit artikel voor om de spraakklanken van het Nederlands niet voor te stellen met hun IPA fonetische tekens. Omwille van de toegankelijkheid van de tekst duiden we de klanken (fonemen) aan met de schrijfwijze waarmee zij gespeld worden in een gesloten lettergreep.

[9] De oppositie tussen de twee achter-fricatieven is uiteraard geneutraliseerd (tot de stemloze uvulaire variant) in de contexten waarin ook de stemhebbend/stemloos-oppositie in de andere fricatiefparen (*f*~*v* en *s*~*z*) geneutraliseerd wordt. Wij noteerden alleen afwijkingen in de realisatie van *g* in niet-neutraliserende contexten.

[10] Voor een uitvoerige(r) uitleg van de notie entropie en de berekening van deze informatiemaat in bits, zie van Heuven (1978) en verwijzingen aldaar. Als alle kenmerken maximale entropie zouden bevatten, dan zouden we tussen de vier en de vijf ja/nee-vragen moeten stellen om een unieke keus te kunnen maken uit 17 mogelijkheden: met vier vragen kan in het gunstigste geval een eenduidige keus gemaakt worden uit $2^4 = 16$ mogelijkheden, met vijf vragen uit $2^5 = 32$ mogelijkheden.

Literatuur

Bezooijen, R. van & C. Gooskens (1999), Identification of language varieties: the contribution of different linguistic levels, *Journal of language and social psychology*, 18, 31-48.

Bezooijen, R. van & Ytsma (2000), Accents of Dutch. Personality impression, divergence, and identifiability. In R. Belemans, R. Vandekerckhove (red.) *Variation in (sub)standard language*: 105-129, Amsterdam: Benjamins.

Collins, B. & I.M. Mees (1999), *The Phonetics of English and Dutch*. Leiden/Boston/Köln: Brill.

Heuven, V.J. van (1978), *Spelling en lezen. Hoe tragisch zijn de werkwoordvormen?*, Assen: Van Gorcum.

Hoppenbrouwers, C. & G. Hoppenbrouwers (2001), *De indeling van de Nederlandse streektalen: dialecten van 156 steden en dorpen geklasseerd volgens de FFM*, Assen: Van Gorcum.

Oostdijk, N., W. Goedertier, F. Van Eynde, L. Boves, J. Martens, M. Moortgat & H. Baayen (2002). *Proceedings LREC (Las Palmas).*

Rietveld, A.C.M. & V.J. van Heuven (2001), *Algemene Fonetiek*, Coutinho, Bussum.

Dit artikel is verschenen in: Tom Koole, Jacomine Nortier en Bert Tahitu (red.) *Artikelen van de Vierde Sociolinguistische Conferentie*, Delft, Eburon, 2003.