Stem-, Spraak- en Taalpathologie
1997, Vol. 6, No. 3, pp. 148-163

0924-7025/97/0603-0148

Een efficiënte validiteitsbepaling van fonetische transcripties

P. Hettinga en W. Vieregge

Vakgroep Taal en Spraak, Katholieke Universiteit Nijmegen

Dit artikel beschrijft de ontwikkeling van een efficiënte manier om fonetische transcripties te evalueren voortkomend uit een onderzoek naar de uitspraak van het Nederlands door Servo-Kroatische moedertaalsprekers (zie Hettinga, 1996). Voor dit onderzoek zijn door een niet-ervaren transcribent 1007 transcripten gemaakt. In zijn algemeenheid kan de *overeenkomst* tussen een transcriptie en een consensustranscriptie vervaardigd door een aantal ervaren transcribenten worden geïnterpreteerd als de *validiteit* van deze transcriptie omdat een door experts vervaardigde consensustranscriptie het meest objectief haalbare is (Vieregge, 1987, 1992). Echter, dit is een zeer tijdrovende en derhalve kostbare zaak, redenen waarom getracht werd in bovenvermelde scriptie deze situatie op een meer *efficiënte manier* te benaderen: van een steekproef (15% van de 1007 transcripten) is door een getraind-ervaren transcribent een zogenaamde intrapersonele consensustranscriptie gemaakt, dat wil zeggen een transcriptie die als volgt tot stand is gekomen: door dezelfde transcribent worden op twee verschillende tijdstippen t1 en t2 van dezelfde uitingen transcripties gemaakt die vervolgens op tijdstip t3 alleen voor die gevallen gecorrigeerd worden waar geen overeenstemming bestaat tussen de transcripties op de tijdstippen t1 en t2. Slechts de gevallen die op tijdstip t1 en t2 identiek zijn (de 100% score) worden gebruikt om de validiteit te bepalen. In dit artikel wordt de validiteit met behulp van twee maten nader bepaald: *overeenkomstpercentage* ('percentage agreement') en *afstandsgemiddelde* ('average distance') die in dit onderzoek worden vergeleken met resultaten uit de literatuur. Daaruit blijkt dat de transcripten, die aan de studie van Hettinga (1996) ten grondslag liggen, consistent en valide zijn.

Inleiding

Het vervaardigen van een fonetische transcriptie is een veel gebruikte onderzoeksmethode die in verschillende taalwetenschappelijke disciplines wordt toegepast. Hoewel bekend is dat transcriptie geen absoluut objectief meetinstru-

Correspondentieadres: Prof. Dr. W. Vieregge, Vakgroep Taal en Spraak, Katholieke Universiteit Nijmegen, Postbus 3103, 6500 HD Nijmegen.

ment is wordt doorgaans verzuimd dit instrument te evalueren, dat wil zeggen de *intra- en interpersonele overeenkomst* en de *validiteit* ervan te bepalen. Segmentele transcripties stellen geen exacte reproducties van spraakuitingen voor maar representaties van spraakuitingen, waarbij een continu veranderend spraaksignaal gereduceerd wordt tot een lineaire sequentie van discrete symbolen. Transcripten van dezelfde spraakuiting die gemaakt zijn door verschillende transcribenten (interpersoneel) of door één transcribent op verschillende tijdstippen (intrapersoneel), laten vaak aanzienlijke verschillen zien. Zelfs consensustranscripties (een transcriptie die door meerdere personen in overleg unaniem tot stand is gekomen) van ervaren transcribenten die op verschillende tijdstippen zijn vervaardigd kunnen verschillen (Vieregge, 1992; Vieregge & Broeders, 1995). Enkele factoren die deze variatie in segmentele transcripties kunnen veroorzaken zijn de volgende (Cucchiarini, 1993): Het transcriptieproces, opgevat als meetinstrument, brengt variatie met zich mee omdat transcripties – zoals boven vermeld – per definitie het resultaat van data-reductie zijn. Verder wordt elke transcriptie door het menselijke auditieve waarnemingsmechanisme bepaald waardoor geen perfecte resultaten bereikt kunnen worden; aan de auditieve perceptie en aandacht van mensen zijn grenzen verbonden; bovendien zijn mensen door motivatie, concentratie en dergelijke sterk beïnvloedbaar. Verder kunnen factoren zoals de moedertaal van een transcribent, de bekendheid met de te transcriberen taal en de mate van ervaring in het *analytisch luisteren* het transcriptieresultaat beïnvloeden. Aangezien het transcriptieproces dus een subjectief meetinstrument is moet de noodzaak van een objectieve evaluatie ervan worden onderstreept.

Overeenkomstmaten in plaats van betrouwbaarheid en validiteit

Een meetinstrument wordt gewoonlijk geëvalueerd door de betrouwbaarheid en validiteit ervan te bepalen. Betrouwbaarheid houdt in de mate van consistentie, geobserveerd tussen herhaalde metingen van hetzelfde object. Bij validiteit vraagt men zich af in hoeverre het meetinstrument meet wat het bedoelt te meten. Om het transcriptieproces te evalueren moeten dus de betrouwbaarbaarheid en de validiteit worden bepaald. Hieronder wordt uiteengezet waarom de betrouwbaarheid en de validiteit van transcripties principieel niet te bepalen zijn. Tevens wordt een alternatief geboden om toch uitspraken te kunnen doen over de overeenkomst (in plaats van betrouwbaarheid) en de validiteit van transcripties.

Om de betrouwbaarheid van transcripties te bepalen zou de mate van consistentie moeten worden geobserveerd tussen herhaalde transcripties (inter- en intrapersoneel) van spraakuitingen. Cucchiarini (1993, 1996) geeft aan dat dit niet mogelijk is omdat betrouwbaarheid berekeningen eist op intervalniveau, zoals het bepalen van gemiddelden en standaarddeviaties van het gemiddelde, terwijl fonetische symbolen *nominale variabelen* zijn waarvoor aan deze eisen niet voldaan kan worden. De fonetische symbolen kunnen immers gelijk of verschillend zijn maar het berekenen van bijvoorbeeld een gemiddelde van verschillende

klanken is niet mogelijk. Dat betekent dan ook dat de betrouwbaarheid bijvoorbeeld niet volgens Cohen's Kappa kan worden berekend. Een alternatief voor de betrouwbaarheid is het bepalen van de overeenkomst van twee transcripties van een zelfde spraakuiting omdat dit op nominaal niveau wel mogelijk is. Cucchiarini (1993, 1996) heeft voor haar onderzoek twee maten voor de bepaling van de overeenkomst gehanteerd, namelijk het *overeenkomstpercentage* ('percentage agreement') en het *afstandsgemiddelde* ('average distance'). Beide maten zijn bedoeld om fonetische transcripties van hetzelfde spraakfragment op overeenkomsten en verschillen te kunnen vergelijken. Het overeenkomstpercentage bepaalt of twee transcripties van een zelfde spraakuiting gelijk of verschillend zijn. Dit betreft als het ware een alles-of-niets bepaling. Het afstandsgemiddelde daarentegen bepaalt de grootte van het fonetisch-artikulatorische verschil tussen twee transcripties van een zelfde spraakklank. Zo is de afstand tussen bijvoorbeeld een [i] en een [a] intuïtief gezien groter dan die tussen een [i] en een [e], een verschil dus dat gerelateerd is aan de grootte van de artikulatorische afstand tussen deze klanken. De laatstgenoemde maat maakt daarbij gebruik van het feit dat fonetische symbolen classificeerbaar zijn met behulp van distinctieve kenmerken. In het vervolg wordt achtereenvolgens het overeenkomstpercentage en het afstandsgemiddelde nader toegelicht.

Het overeenkomstpercentage

Het overeenkomstpercentage wordt met behulp van de volgende formule berekend:

$$OP = \frac{\text{aantal overeenkomsten}}{\text{aantal verschillen + aantal overeenkomsten}} \cdot 100 \quad [\%]. \qquad (1)$$

Een hoog percentage betekent een grote overeenkomst tussen de transcripties. Cucchiarini (1993) beschrijft een aantal aspecten waarmee bij het berekenen van het overeenkomstpercentage rekening moet worden gehouden. Ten eerste moet erop worden gelet dat de juiste symbolen met elkaar worden vergeleken, dat betekent o.a. dat de twee te vergelijken transcriptiereeksen juist moeten worden opgelijnd voor het geval dat deze niet even lang zijn. Daarbij doen zich verschillende mogelijkheden voor: twee symbolen kunnen overeenstemmen, een symbool kan worden gesubstitueerd door een ander symbool, of deletie dan wel insertie van een symbool kan plaatsvinden. Ten tweede is de kans dat fonetisch enge transcripties overeenkomen lager dan fonetisch brede transcripties. Bij fonetisch enge transcripties wordt immers een keuze gemaakt uit meer fonetische symbolen die bovendien met elkaar gecombineerd kunnen worden. Ten derde is bij het berekenen van het overeenkomstpercentage geen rekening gehouden met het feit dat fonetische symbolen meerwaardige nominale variabelen zijn. Daardoor is bijvoorbeeld het overeenkomstpercentage tussen de symbolen [b] en [p] even groot als tussen [p] en [z] wat niet strookt met onze kennis over de articulatie van spraakklanken: het verschil tussen [b] en [p] kan worden beschreven

langs één dimensie, te weten stemhebbendheid. Echter, voor het beschrijven van het verschil tussen [p] en [z] hebben wij twee dimensies nodig: stemhebbendheid en manier van articulatie (stemloze explosief versus stemhebbende fricatief).

Het afstandsgemiddelde

De tweede overeenkomstmaat, het afstandsgemiddelde, houdt rekening met de fonetisch-artikulatorische afstand tussen transcriptiesymbolen door deze te analyseren in termen van distinctieve kenmerken. Vieregge et al. (1984) heeft één matrix voor consonanten en één voor vocalen ontwikkeld waarin de afstanden tussen fonetische symbolen in getallen zijn uitgedrukt. Deze matrices zijn gebaseerd op experimenten waarin proefpersonen de mate waarin twee spraakklanken verschillen met behulp van een tienpuntschaal beoordeelden. Ook voor de diacritische tekens is een afstandssysteem ontworpen (Vieregge, 1987, pp. 24-27. Zie Appendix c). Als gevolg van aanwezige diakritische tekens kan de afstand tussen twee symbolen kleiner of groter worden. Omdat de hierarchische structurering van distinctieve kenmerken van taal tot taal blijkt te variëren moet voor het bepalen van het afstandsgemiddelde per taal een afstandenmatrix worden vervaardigd. Cucchiarini (1993, 1996) heeft het door Vieregge et al. (1984) ontworpen afstandssysteem verbeterd en meer genuanceerd. De matrices die door Cucchiarini (1993) zijn gebruikt hebben betrekking op nederlandse spraakklanken. Het afstandsgemiddelde wordt met behulp van de volgende formule berekend (Cucchiarini 1996, p. 150):

$$AG = 1/N \sum_{i=1}^{N} d_i, \qquad (2)$$

met AG = afstandsgemiddelde, N = aantal klinker-(of medeklinker)paren, d = artikulatorische afstand tussen de symboolparen in kwestie (klinker of medeklinkers), en i = index. Een hoge AG-score betekent een grote afstand tussen twee transcriptiesymbolen wat wederom betekent dat de overeenkomst tussen deze twee symbolen klein is. Een lage AG-score daarentegen betekent een relatief goede overeenkomst. Op deze manier kan de mate van overeenkomst cijfermatig worden gerelateerd aan de fonetische afstand tussen twee symbolen.

Om een antwoord te krijgen op de vraag of deze twee objectieve overeenkomstmaten daadwerkelijk de realiteit weerspiegelen bij het evalueren van transcriptiemateriaal is door Biemans (1993) onderzocht of deze maten overeenkomen met beoordelingen van 19 transcriptie-experts uit verschillende Europese landen (zeven kwamen uit Duitsland, vier uit Nederland, drie uit Groot-Brittanië, twee uit Denemarken, twee uit Noorwegen, en één uit Finland). De fonetici beoordeelden de overeenkomsten en verschillen van twee uitgeschreven transcripties van een zelfde spraaksegment op een 10-puntsschaal. De 50 beoordeelde spraaksegmenten vormden een subset van spraaksegmenten uit het onderzoeksmateriaal van Cucchiarini (1993, 1996). Biemans laat zien dat de oordelen die de fonetici geven over het verschil tussen twee transcripties van hetzelfde spraakfragment onderling vrij goed overeenkomen. Bovendien vond Biemans dat de oordelen

van de fonetici *redelijk hoog correleren* met de waarden van het overeenkomstpercentage en *vrij hoog* met de waarden van het afstandsgemiddelde. Dus zowel het overeenkomstpercentage als het afstandsgemiddelde zijn goede objectieve maten voor het evalueren van transcripties.

Cucchiarini (1993) constateert dat beide overeenkomstmaten voor- en nadelen hebben. Zo is het voordeel van het overeenkomstpercentage dat dit eenvoudiger te berekenen is dan het afstandsgemiddelde. Een ander voordeel is de taalonafhankelijkheid van het overeenkomstpercentage in tegenstelling tot het afstandsgemiddelde. Een voordeel van het afstandsgemiddelde daarentegen is dat rekening wordt gehouden met optredende 'gewogen' verschillen in tegenstelling tot het overeenkomstpercentage. Beide overeenkomstmaten zijn overigens niet compleet onafhankelijk van elkaar omdat elke toename in het overeenkomstpercentage bij zal dragen aan het vermeerderen van het afstandsgemiddelde. Maar ondanks het feit dat ze aan elkaar gerelateerd zijn, geven ze toch verschillende informatie over de afstanden tussen twee transcripties. Door beide metingen te gebruiken wordt de relatie zichtbaar tussen het aantal en de grootte van de transcriptiediscrepanties. Omdat deze twee maten elkaar aanvullen is aan te bevelen ze naast elkaar te gebruiken.

Om tenslotte de validiteit van transcripties te kunnen bepalen moet – zoals al gezegd – worden nagegaan in hoeverre de transcriptie meet wat zij bedoelt te meten. Echter, omdat het transcriptieproces geen absoluut objectief meetinstrument is, kan er ook geen definitie worden gegeven van een ideale transcriptie die als referentiepunt kan gelden. Door het ontbreken van een dergelijk objectief referentiepunt is het meten van validiteit onmogelijk. Als alternatief stelt Vieregge (1987, p. 31) voor om gebruik te maken van een consensustranscriptie die gemaakt is door een groep getraind-ervaren fonetici. Een consensustranscriptie is een transcriptie die gemaakt is door een groep personen nadat zij een consensus hebben bereikt over elk symbool dat de transcriptie bevat. Hierdoor worden de 'meetfouten' geminimaliseerd zodat de ideale transcriptie wordt benaderd. Bovendien zou daardoor de overeenkomst tussen transcribenten verhoogd worden (Shriberg, Kwiatkowski, & Hoffmann, 1984). Dit vermoeden wordt bevestigd in een experimenteel onderzoek van Vieregge en Broeders (1995). In hun onderzoek is naar voren gekomen dat overeenkomst tussen transcripties inderdaad noodzakelijk is om als referentiepunt voor validiteit te kunnen voldoen. Zij voegen daar echter aan toe dat het essentieel is dat de transcribenten competent zijn. Vieregge en Broeders stellen vast dat als referentiepunt voor validiteit de 100%-overeenkomsten tussen twee of meer – in de tijd ver uit elkaar liggende – consensustranscripties, gemaakt door ervaren transcribenten, moet worden genomen.

Een efficiënte evaluatie van transcripties

In het scriptieonderzoek naar de uitspraak van het Nederlands door servo-kroatische moedertaalsprekers (Hettinga, 1996) zijn uitingen van twee vrouwelijke

Servo-Kroatische moedertaalsprekers nader onderzocht die op het moment van het onderzoek 2.9 jaar in Nederland vertoefden, dus 2.9 jaar met het Nederlands als vreemde taal geconfronteerd werden. Het spraakmateriaal werd samengesteld als volgt: een leestekst bevatte 175 spraakvariabelen bestaande uit consonanten, consonantclusters, vocalen en alle belangrijke regels voor fonetisch-fonologische processen in lopende spraak en wel in verschillende posities in een woord (initiaal, mediaal en finaal). Quasi spontane spraak werd hieraan toegevoegd door een elicitatieprocedure (plaatjes beschrijven) en door een interview. Uit deze twee spraakbronnen werden vervolgens 1007 tokens geselecteerd die op narrow niveau werden getransribeerd. Als voorbeeld volgen hieronder een aantal verschillende variabelen, aangeduid met /.../, en de bijbehorende transcripties [...]. Deze transcripties laten zien hoe de variabelen in kwestie door de servo-kroatische moedertaalsprekers zijn gerealiseerd.

/-ŋ-/:	[ŋg], [ŋʏ]
/h/:	[ʏ], [ɦ̤]
/spl-/:	[sɬ], [s̲ɬ]
/Y/:	[u], [ʉ]
/ə/:	[ĕ], [Ĭ], [Y]
/œy/:	[a.u], [ɑu]
/ɛ i/:	[ai], [ɑi]

normale schwa-insertie tussen de /l/ en een niet-homorgane consonant: [0] = nulsymbool (schwa wordt niet gerealiseerd)

Zoals eerder is opgemerkt betekent een grondige evaluatie van het transcriptieproces een grote tijdsinvestering. Derhalve is in het scriptie-onderzoek naar de uitspraak van het Nederlands door servo-kroatische moedertaalsprekers (Hettinga, 1996) een methode ontwikkeld die ook bij aanwezigheid van slechts één getraind-ervaren transcribent kan leiden tot een bevredigend resultaat. Deze methode kan in vergelijkbaar onderzoek, waarin transcripties centraal staan, worden toegepast. De fonetisch enge transcripties zijn op de volgende wijze geëvalueerd. Allereerst werden alle te analyseren klanken en klankcombinaties door één niet-ervaren transcribent (20 uur ervaring in het transcriberen van normale en 20 uur in het describeren van pathologische spraak) getranscribeerd. Vervolgens is uit deze transcriptie een steekproef van 15% getrokken. Om een gelijke verdeling van consonanten, consonantclusters, vocalen en spraakfragmenten beïnvloed door fonetisch-fonologische regels in lopende spraak van de twee servo-kroatische sprekers in twee spreekstijlen te krijgen, zijn 16 aparte steekproefen (2 sprekers × 2 spreekstijlen × 4 spraakklankklassen) van 15% ad random getrokken. In Tabel 1 wordt deze steekproeftrekking schematisch weergegeven.

Een getraind-ervaren transcribent heeft van de 151 transcripten (=15%) uit de streekproef een intra-consensustranscriptie gemaakt. Dat wil zeggen een transcriptie die door hem op twee verschillende tijdstippen is gemaakt aan de hand van herhaald beluisteren en die op tijdstip 3 alleen voor die gevallen werd gecorri-

Tabel 1. De steekproeftrekking.

Spreker A								
opgelezen spraak				quasi spontane spraak				
a	b	c	d	a	b	c	d	
96	56	132	64	47	18	63	31	Totaal = 507
14	8	20	10	7	3	9	5	15% = 76
Spreker B								
opgelezen spraak				quasi spontane spraak				
a	b	c	d	a	b	c	d	
96	56	131	65	43	18	61	30	Totaal = 500
14	8	20	10	6	3	9	5	15% = 75
								Totaal =1007
								15% = 151

Noot. a = aantal consonanten, b = aantal consonantclusters, c = aantal vocalen, d = aantal spraakfragmenten beïnvloed door fonetisch-fonologische regels in lopende spraak. Totaal = totaal aantal getranscribeerde klanken, 15% = de steekproeftrekking.

geerd waar geen overeenstemming bestond tussen de transcripties op de tijdstippen t1 en t2. Gekozen is voor deze intra-consensustranscriptie omdat het om praktische redenen niet mogelijk bleek te zijn om een inter-consensustranscriptie tussen twee of meer getraind-ervaren transcribenten te verkrijgen zoals voorgesteld is door Vieregge en Broeders (1995), een situatie die zich bij dergelijk onderzoek vaker kan voordoen. Immers, getraind-ervaren transcribenten zijn schaars. Tussen de verschillende tijdstippen (t1–t2, t2–t3) moet in principe een periode liggen die herinneringseffecten voorkomt. In dit onderzoek bleek een periode van twee weken te voldoen.

Vervolgens is het alternatief voor betrouwbaarheid berekend aan de hand van het overeenkomstpercentage en het afstandsgemiddelde tussen de gehele intra-consensustranscriptie (ervaren transcribent) en de bijbehorende transcriptie (niet-ervaren transcribent). Als referentiepunt voor de validiteit is uitgegaan van die gevallen in de intra-consensustranscriptie van de ervaren transcribent waar *op tijdstippen t1 en t2 volledige overeenstemming* bestond, te weten de zogenaamde 100% scores. De overeenkomst tussen die 100% scores en de bijbehorende transcripten van de niet-ervaren transcribent, *geïnterpreteerd als validiteit,* werd ook bepaald aan de hand van het overeenkomstpercentage en het afstandsgemiddelde.

Omdat uitsluitend nederlandse klanken zijn getranscribeerd kon gebruik worden gemaakt van de afstandenmatrices van Cucchiarini (1993, 1996), zie appendix a en b. Uit deze matrices kunnen de afstanden tussen twee klanken worden herleid. Zo is bijvoorbeeld de afstand tussen de spraakklanken [p] en [b] 1, tussen [n] en [ŋ] 3, tussen [a] en [i] 3.5 en tussen [v] en [v̥] 1.

Resultaten

Alternatief voor betrouwbaarheid

De resultaten van de overeenkomst tussen de *gehele* intra-consensustranscriptie van de getraind-ervaren transcribent en de bijbehorende transcripties van de niet-ervaren transcribent staan in Tabel 2 vermeld.

De gemiddelden van het overeenkomstpercentage (37.8%) en het afstandsgemiddelde (0.8) van dit onderzoek zijn vergeleken met andere onderzoeksresultaten om een indruk te krijgen van wat acceptabel is. Hoewel sprake is van een vergelijking tussen transcripties die gemaakt zijn door verschillende personen, een interpersonele transcriptie, blijkt het toch mogelijk te zijn om deze te vergelijken met het resultaat van een overeenkomst tussen transcripties die door één persoon zijn gemaakt, een intrapersonele transcriptie. Uit onderzoek blijkt namelijk dat inter- en intrapersonele transcripties in overeenkomstpercentage niet significant van elkaar verschillen (Shriberg & Lof, 1991). Allereerst is het *overeenkomstpercentage* van 37.8% vergeleken met het *overeenkomstpercentage* van de twee intrapersonele transcripties die gemaakt zijn door de ervaren transcribent op de tijdstippen t1 en t2. Van de transcripties die gemaakt zijn op tijdstip t1 blijkt 42.6% precies hetzelfde te zijn als op tijdstip t2. Vervolgens is het overeen-

Tabel 2. De overeenkomst tussen de *gehele* intra-consensustranscriptie van de steekproef van de getraind-ervaren transcribent en de bijbehorende transcripties van de niet-ervaren transcribent.

	%	x	N
Spreker A: opgelezen spraak	44.2	0.7	52
Spreker A: quasi spontane spraak	41.4	0.8	24
Spreker B: opgelezen spraak	28.6	0.9	52
Spreker B: quasi spontane spraak	37.0	0.6	23
Gemiddeld:	37.8	0.8	37.8
Totaal:			151

Noot. % = overeenkomstpercentage, x = afstandsgemiddelde, N = aantal transcripties. De waarden zijn afgerond op 1 cijfer achter de komma.

komstpercentage vergeleken met resultaten uit het onderzoek van Cucchiarini (1993, 1996). Zij heeft onder andere verschilpercentages (dit is het complement van het overeenkomstpercentage) van interpersonele transcripties berekend en wel apart voor vocalen en consonanten in verschillende experimentele condities. Deze condities betroffen de invloed van de bekendheid met de taal (Tjechisch, Limburgs en Nederlands), de spreekstijl (opgelezen en spontane spraak) en de invloed van de al of niet aanwezige context waarin de te transcriberen klanken werden aangeboden (wel/geen context). Ons overeenkomstpercentage van 37.8% is vergeleken met de experimentele conditie Limburgs met aanwezige context omdat deze het beste de condities van ons onderzoek benadert. Met andere woorden: qua bekendheid met de taal in kwestie is het Limburgs dialect eerder vergelijkbaar met het Nederlands door servo-kroatische moedertaalsprekers dan met het Nederlands door Nederlanders. Cucchiarini (1993, p. 135 fig. 7.8b, p. 138 fig. 7.10b) laat zien dat consonanten en vocalen in die conditie een verschilpercentage van respectievelijk 25% en 42% hebben, wat gemiddeld 34% is. Het verschilpercentage 34% is gelijk aan 66% overeenkomstpercentage. Hieruit blijkt dat het overeenkomstpercentage van 37.8% iets lager is dan de intrapersonele transcriptie van de steekproef tussen tijdstippen t1 en t2 door de ervaren transcribent, namelijk 42.6%, en dat dit percentage beduidend lager is dan dat uit het onderzoek van Cuccchiarini, namelijk 66%. Het moet worden benadrukt dat de resultaten van het onderhavige onderzoek en dat van Cucchiarini (1993) slechts in zeer groffe lijnen, d.w.z. wat de orde van grootte betreft, vergeleken mag worden omdat de experimentele condities sterk van elkaar verschillen. Een belangrijk verschil komt bijvoorbeeld tot stand door het feit dat een getraind-ervaren transcribent gewoonlijk meer diacritica toepast dan een niet-ervaren transcribent. Alleen al hierdoor kan het overeenkomstpercentage drastisch dalen. De relatief hoge overeenkomstpercentages in Cucchiarini's onderzoek zijn derhalve mogelijk het gevolg van het feit dat Cucchiarini's proefpersonen niet-ervaren transcribenten waren.

Wat het *afstandsgemiddelde* betreft kan uit tabel 2 worden opgemaakt dat dit tussen de transcripties van de getraind-ervaren en niet-ervaren transcribent gemiddeld 0.8 is. In het onderzoek van Cucchiarini daarentegen is het afstandsgemiddelde 0.9 (Cucchiarini 1993: 135 fig. 7.8a, 138 fig. 7.10a). Dus hoewel in het huidige onderzoek meer transcripties van elkaar verschillen vergeleken met het onderzoek van Cucchiarini, blijkt de mate waarin deze transcripties verschillen juist minder groot te zijn. Een afstandsgemiddelde van 0.8 is voor fonetisch enge transcripties klein als men nagaat dat voor de basissymbolen met mogelijke diacritica de maximale afstand bij consonanten 9 en bij vocalen 8 is (zie Appendix a, b en c). Dus een verschil van 0.8 is kleiner dan het verschil tussen [p] en [b]. Cucchiarini (1993, p. 159, 1996) concludeert dat het afstandsgemiddelde een realistischer beeld geeft dan het overeenkomstpercentage omdat de laatst genoemde maat slechts alles-of-niets meet, terwijl het afstandsgemiddelde de fonetisch-artikulatorische afstanden mede in beschouwing neemt. Dit komt overigens overeen met het al genoemde onderzoek van Biemans (1993) waar een hogere

correlatie tussen de 19 Europese, ervaren transcribenten voor het afstandsgemiddelde dan voor het overeenkomstpercentage werd gevonden.

Alternatief voor validiteit

De alternatieve maat voor de *validiteit*, dat wil zeggen de overeenkomst tussen de 100% scores van transcripties op tijdstippen t1 en t2 van de getraind-ervaren transcribent en de bijbehorende transcripties van de niet-ervaren transcribent, staan in tabel 3 vermeld.

Het gemiddelde van het overeenkomstpercentage voor validiteit (51.5%) is vergeleken met de validiteit in het onderzoek van Vieregge en Broeders (1995). In dat onderzoek is de validiteit berekend door het overeenkomstpercentage te bepalen tussen consensustranscripties die gemaakt zijn door negen onervaren transcribentparen en de 100% overeenkomstscores van de interpersonele consensustranscriptie door twee getraind-ervaren transcribenten. Het overeenkomstpercentage was gemiddeld 58.7%, met een variatie tussen de paren van 49.0% tot en met 68.7%. Men moet er echter wel rekening mee houden dat beide onderzoeken in onderzoeksopzet verschilden. Om praktische redenen is immers in het huidige onderzoek door de getraind-ervaren transcribent een intrapersonele consensustranscriptie en door de niet-ervaren transcribent individueel een transcriptie gemaakt. Daarentegen is in het onderzoek van Vieregge en Broeders door twee getraind-ervaren transcribenten een interpersonele consensustranscriptie gemaakt en hebben de niet-ervaren transcribenten paarsgewijs een consensustranscriptie kunnen maken. Verder is in beide onderzoeken de spraak van Nederlandssprekenden getranscribeerd, maar in het onderzoek van Vieregge en Broeders is het Nederlands de moedertaal en in ons onderzoek de tweede taal. Een laatste verschil is dat in het onderzoek van Vieregge en Broeders maar zes variabelen zijn onderzocht (/x/, /v/ /z/, schwa-deletie na /l,r/, assimilatie van stem voor /b,d/ en /n/-deletie na schwa), waarvan twee variabelen binair zijn, namelijk

Tabel 3. Het alternatief voor validiteit gemeten als overeenkomstpercentage (%) en afstandsgemiddelde (x).

	%	x	N
Spreker A: opgelezen spraak	50	0.6	22
Spreker A: quasi spontane spraak	61.1	0.4	15
Spreker B: opgelezen spraak	35	0.6	16
Spreker B: quasi spontane spraak	60	0.6	8
Gemiddeld:	51.5	0.6	15.3
Totaal:			61

Noot. N = aantal transcripties. De waarden zijn afgerond op 1 cijfer achter de komma.

schwa-deletie en /n/-deletie. Het onderhavige onderzoek betreft veel meer variabelen die diverse varianten (tokens) toonden. Wanneer rekening wordt gehouden met bovenvermelde feiten is het verschil in overeenkomstpercentage, namelijk 51.5% in het onderhavige onderzoek tegenover 58.7% in het onderzoek van Vieregge en Broeders, relatief gezien klein.

In Tabel 3 is te zien dat het afstandsgemiddelde 0.6 is, dat wil zeggen dat de transcripties van de niet-ervaren transcribent met een gemiddelde afstand van 0.6 afwijken van de 100% score van de ervaren transcribent. Dus terwijl ongeveer de helft (49.5%) van de transcripties niet met elkaar overeenkomt, is de afstand van de afwijking gemiddeld slechts 0.6. Met betrekking tot validiteit van transcripties geeft derhalve het afstandsgemiddelde ook een realistischer beeld dan het overeenkomstpercentage.

Conclusie

Uit de resultaten blijkt dat het transcriptieproces in dit onderzoek, vergeleken met ander soortgelijk onderzoek, voldoende betrouwbaar en valide is. Het grote *aantal* verschillen tussen transcripties en de *mate* waarin deze van elkaar verschillen zijn inherent aan de beperkingen die transcriptie als 'meetinstrument' met zich meebrengt. De "validiteitsscores" van transcripties zijn hoger dan de "betrouwbaarheidsscores", namelijk: 51.5% overeenkomstpercentage en 0.6 afstandsgemiddelde geïnterpreteerd als validiteit tegenover 37.8% overeenkomstpercentage en 0.8 afstandsgemiddelde geïnterpreteerd als betrouwbaarheid. Dus de overeenkomst tussen de transcripties van de niet-ervaren transcribent en de 100% scores van de intraconsensustranscriptie van de getraind-ervaren transcribent is hoger dan de overeenkomst tussen de transcripties van de niet-ervaren transcribent en de transcripties uit de gehele intraconsensustranscriptie van de getraind-ervaren transcribent. Dat betekent dat voor die gevallen waarin de getraind-ervaren transcribent 100% scoort ook de niet-ervaren tarnscribent hoger scoort, namelijk 51.5% in plaats van 37.8%.

Hoewel beide overeenkomstmaten aan elkaar gerelateerd zijn, blijkt dat er toch duidelijk verschil is tussen deze twee maten. Dit onderzoek laat namelijk zien dat veel transcripties van elkaar verschillen maar dat de mate waarin zij gemiddeld verschillen gering is. Daarom geeft het afstandgemiddelde een realistischer beeld bij het evalueren van transcripties dan het overeenkomstpercentage. Waarschijnlijk speelt het feit dat we te maken hebben met een fonetisch enge transcriptie daarbij een rol, want minimale verschillen zorgen al snel voor lage overeenkomstpercentages. Maar ook tussen sprekers en verschillende spreekstijlen kunnen grote variaties worden gevonden bij het vergelijken van beide overeenkomstmaten. In tabel 3 is bijvoorbeeld te zien dat bij het zelfde afstandsgemiddelde van 0.6 een overeenkomstpercentage van 50%, 35% en 60% hoort bij respectievelijk spreker A in opgelezen spraak, spreker B in opgelezen spraak en spreker B in quasi spontane spraak.

De voorgestelde twee maten, te weten overeenkomstpercentage en afstandsgemiddelde, vullen elkaar aan en kunnen worden ingezet om op een efficiënte manier, dat wil zeggen met inspanning van slechts één ervaren en één niet-ervaren transcribent, alternatieven voor de betrouwbaarheid en de validiteit te meten.

Summary

This article deals with an efficient evaluation procedure of the transcription process, which was developed in a thesis entitled 'The pronunciation of Dutch by Servo-Croatian speakers' within the Speech and Language Pathology programme of the Faculty of Arts at the University of Nijmegen (Hettinga, 1996). An untrained transcriber made 1007 transcripts. Generally speaking the *agreement* between transcripts and the consensus transcription of experts (trained transcribers) can be interpreted as the *validity* of these transcripts, because such a consensus transcription is the most objective transcription one may obtain (see Vieregge 1987, 1992). However, the proposed procedure to get insight into the validity is very time consuming. This is why an efficient manner of validity determination is proposed in this article. From the 1007 transcripts made by the untrained transcriber a randomly selected subset (15%) was transcribed twice by a trained transcriber resulting in an *intrapersonal consensus transcription* by correcting those cases in which the first transcription was not the same as the second one. Only those cases in which the first transcript was identical with the second one (the 100% score) were used to calculate the validity. In this study the validity is determined by means of two measures: percentage agreement and average distance. Comparing the resulting percentage agreement and average distance values from this investigation with those found in the literature leads to the conclusion that the transcripts used in this study are consistent and valid.

Literatuur

Biemans, M. (1993). *Het verschil tussen twee transcripties van hetzelfde spraakfragment; een onderzoek naar de overeenkomst tussen oordelen van fonetici en twee theoretische maten.* Intern verslag, Katholieke Universiteit Nijmegen.

Cucchiarini, C. (1993.) *Phonetic transcription: A methodological and empirical study.* Proefschrift Faculteit der Letteren, Katholieke Unversiteit Nijmegen.

Cuchiarini, C. (1996). Assessing transcription agreement: Methodological aspects. *Clinical Linguistics and Phonetics, 10,* 131-155.

Hettinga, P.M. (1996). *De uitspraak van het Nederlands door Servo-Kroatische moedertaalsprekers.* Doctoraalscriptie opleiding Spraak- en Taalpathologie, Katholieke Universiteit Nijmegen.

Shriberg, L.D., & Lof, L. (1991). Reliability studies in broad and narrow phonetic transcription. *Clinical Linguistics and Phonetics, 5,* 225-279.

Shriberg, L.D., Kwiatkowski, J., & Hoffmann, K. (1984). A procedure for phonetic transcription by consensus. *Journal of Speech and Hearing Research, 27,* 456-465.

Vieregge, W.H., Rietveld, A.C.M., & Jansen, C.I.E. (1984). A distinctive feature based system for the evaluation of segmental transcription in Dutch. *Proceedings of the Xth International Congress of Phonetic Sciences* (pp. 654-659). Dordrecht: Foris Publications.

Vieregge, W.H. (1987). Basic aspects of phonetic segmental transcription. In A. Almeida & A. Braun (Red.), *Probleme der phonetischen Transkription* (pp. 5-55). Stuttgart: Franz Steiner.

Vieregge, W.H. (1992). Das Konzept der auditiven Aufmerksamkeitsspanne beim analytischen Hören. In *Phonetik und Dialektologie. Joachim Göschel zum 60.Geburtstag.* Schriften der Universität Marburg, Nr. 64, 54-75, Marburg.

Vieregge, W.H. & A.P.A. Broeders (1995). Agreement in consensus transcriptions of trained and untrained transcribers. In K. Elenius & P. Branderud (Red.), *Proceedings of the XIIIth International Congress of Phonetic Sciences* (ICPhS 95), *3*, 174-177.

Appendix

Feature matrices (uit: Cucchiarini, 1993, 195-196).
The feature matrices used as input to the transcription alignment program are shown below. On the basis of these matrices, the program calculates the distances between consonants and vowels, respectively.

a. Consonant feature matrix

cons.	place	voice	nas	stop	glide	lat	fric	trill	high	distr.
p	1			1						1
b	1	1		1						1
t	2			1						
d	2	1		1						
c	3			1					1	1
k	4			1					1	
g	4	1		1					1	
f	1	1					1			
s	2						1			
z	2	1					1			
≈	2						1		1	1
ʒ	2	1					1		1	1
x	4						1		1	1
ʏ	4	1					1		1	
χ	4						1			
∂	4	1					1			
m	1	1	1							1
ɱ	1	1	1							
n	2	1	1							
ŋ	4	1	1						1	
ɲ	3	1	1						1	1
l	2	1				1				
ɬ	4	1				1				
r	2	1						1		
±	2	1								
ʀ	4	1						1		
w	1	1			1					1
ʊ	1	1			1					
j	3	1			1				1	
h	5						1			
?	5			1						
ç	3						1		1	1
≠	2	1			1					
ɟ	3	1		1					1	1
≠	2	1					1	1		
ɦ	5	1					1			
∂	1.5	1					1			
θ	1.5						1			

b. Vowel feature matrix

Vowel	front/back	tongue hight	lip rounding
a	1.5	1	
æ	1	1.5	
ɛ	1	2	
œ	1	2	2
e	1	3	
ø	1	3	1
i	1	4	
y	1.5	4	1
ɪ	1.5	3	
Y	1.5	3	1
ɵ	2	2.5	
ɜ	2	2	
ɐ	2	1.5	
U	2.5	3	1
u	3	4	1
o	3	3	1
ɔ	3	2.5	1
ʌ	3	2	
ɒ	3	1	1
ɑ	3	1	

c. Numbers assigned to diacritical marks (uit: Vieregge, 1987, 24).

diacritical mark + meaning			number
1.	x̃	nasalisation	2
2.	ẍ	central articulated	1
3.	x̚	unreleased explosion	.5
4.	x’	glottal stop simultaneously	1
5.	x̥	devoicing	1
4.	x̬	voicing	1
7.	x̝	closed variety	.5
8.	x̞	open variety	.5
9.	x̟	advanced variety	.5
10.	x̠	retracted variety	.5
11.	x̪	dental articulation	.5
12.	x̹	lips more rounded	.5
13.	x̜	lips more spread	.5

14.	x::	very long segment	4
15.	x:	long segment	2
16.	x.	half long segment	1
17.	x̆	weak element of a diphthong; weak realisation of a segment	.5
18.	x^h	aspiration	.5
19.	x^w	labialisation	1

List 1. Numbers assigned to differences between a symbol [x] and a symbol with a diacritical mark [xD]; the difference is indicated by [x/xD], for instance [e/ẽ] = 2.