Menno de Jong en
Peter Jan Schellens

Methoden voor tekstevaluatie onder de loep

Ter inleiding

De begrijpelijkheid van een bijsluiter, de geschiktheid van een voorlichtingsfolder voor de doelgroep, de toegankelijkheid van een website: in diverse studies is aangetoond dat formatief evaluatieonderzoek gericht op dergelijke punten een belangrijke stap is op weg naar effectieve (papieren en digitale) teksten. De tekstontwerper staat inmiddels een groot scala aan evaluatiemethoden ter beschikking: van expertonderzoek aan de hand van het CCC-model tot doelgroeponderzoek met hardopdenkprotocollen of de plus-en-minmethode.

Tijdens het VIOT-congres in december 2002 in Antwerpen vond een symposium plaats over 'Document design en methoden voor tekstevaluatie'. Daarin stonden diverse tekstevaluatiemethoden ter discussie. Een dergelijke discussie kan natuurlijk gevoerd worden op basis van praktijkervaringen. In vier bijdragen – nu in dit themanummer opgenomen – werd echter onderzoek gepresenteerd naar de waarde van een of meer methoden. Daarmee kreeg het symposium een soort Droste-effect, dat ook dit themanummer laat zien. Hier wordt onderzoek gepresenteerd naar de kwaliteit van een aantal teksten, maar het gaat de onderzoekers niet zozeer om die teksten maar om de methoden die gebruikt worden om tekstkwaliteit in een conceptfase vast te stellen – met het oog op revisie en verbetering.

Het onderzoek naar de waarde en beperkingen van tekstevaluatiemethoden is in de afgelopen 10 jaar een interessant en geschakeerd werkterrein geworden. In *Met het oog op de lezer* (De Jong & Schellens 1995) konden we nog niet veel meer doen dan een overzicht bieden over de verschillende beschikbare methoden. Die inventarisatie bracht ook aan het licht dat we van de meeste methoden nog heel weinig wisten: hoe makkelijk of lastig zijn ze in het gebruik in de praktijk, hoe betrouwbaar zijn ze, wat zeggen de resultaten ervan precies, hoe hanteerbaar zijn de resultaten als handvat voor revisie? Bij veel van de toen besproken methoden konden we dergelijke vragen alleen maar stellen en over de antwoorden speculeren. Zo werd de inventarisatie tegelijkertijd een onderzoeksagenda voor de toekomst. Uit dit themanummer blijkt dat inmiddels hard aan die agenda wordt gewerkt.

We signaleren hier enkele ontwikkelingen. Ten eerste is het terrein van onderzoek verbreed. Het ging ons destijds nog vooral om methoden die bij het ontwerp van voorlichtingsteksten konden worden gebruikt. In dit themanummer komen naast voorlichtingsteksten (brochures maar ook bijsluiters) nu ook websites aan de orde, hier vertegenwoordigd door een online catalogus en een wereldwijd gebruikt bibliografisch systeem.

Ten tweede is inmiddels een scala aan methoden in onderzoek: in de artikelen in dit themanummer gaat het niet alleen om vertrouwde instrumenten zoals de plus-en-minmethode, het daarvan afgeleide softwareprogramma Focus en het CCC-model van Renkema, maar ook

de hardopdenkmethode wordt nu in verschillende gedaanten als object van onderzoek onder de loep genomen.

Ten derde is in dit nummer vaak niet meer alleen de vraag aan de orde of een methode geschikt is voor zijn doel, maar treden achterliggende vragen naar voren: wat verklaart het verschil in resultaten die met twee uiteenlopende methoden worden geboekt? Wat is het effect van de ingreep in het 'normale' leesproces, die een methode met zich meebrengt? Leidt de methode tot dezelfde resultaten bij proefpersonen met uiteenlopende culturele achtergrond? Zijn de cognitieve vaardigheden en terminologische categorieën die de methode vooronderstelt, bij de lezer/proefpersoon wel beschikbaar? Met dergelijke vragen verdiept het onderzoeksterrein zich en worden de theoretische vooronderstellingen die elke methode noodzakelijkerwijs met zich meebrengt, duidelijk en onderzoekbaar.

Tenslotte stellen we vast – ook al blijkt dat in een Nederlandstalig tijdschrift voornamelijk uit de aangehaalde literatuur – dat het onderzoeksterrein internationaal aan gewicht wint. We gaven eerder een overzicht van het internationale onderzoek naar probleemopsporende tekstevaluatiemethoden (De Jong & Schellens 2000, 2002). Dat onderzoek vindt plaats vanuit heel verschillende disciplinaire achtergronden (document design, interface design, human-computer interaction, ergonomie, marketing research of onderwijskunde), maar bij elkaar genomen begint zich een gezamenlijke vraagstelling en methodologie af te tekenen. Het onderzoek uit Nederland (niet alleen van de onderzoekers die in dit nummer aan het woord komen) speelt in die ontwikkeling een vooraanstaande rol.

Wat mag de lezer hier verwachten?

Henk Pander Maat en Leo Lentz rapporteren over onderzoek naar de waarde van de hardopleesmethode als methode om begripsproblemen op het spoor te komen. In vergelijking met de plus-en-minmethode en het computerprogramma Focus scoort het lezersprotocol zeer goed. Niet alleen is het aantal begripsproblemen dat met een lezersprotocol wordt opgespoord groter, ook de kans dat in een lezersprotocol reële begripsproblemen aan het licht komen blijkt groter. In een nader onderzoek met vier varianten van het lezersprotocol blijkt dat de productiviteit van de methode met name te danken is aan de onmiddellijkheid: proefpersonen geven meteen commentaar tijdens het lezen en stellen dat niet uit zoals in de plus-en-minmethode.

Sanne Elling en Leo Lentz deden onderzoek naar de waarde van het CCC-model van Renkema als alternatief voor een lezersgerichte tekstbeoordeling. De auteurs vergeleken de resultaten van experts die al dan niet het CCC-model gebruikten, met die van verschillende lezersgerichte methoden. De resultaten zijn niet bemoedigend voor experts: zij zijn het meestal niet eens over de lezersproblemen (ze voorspellen allemaal andere problemen), hun problemen komen slechts in geringe mate overeen met de problemen die doelgroepgerichte methoden aan het licht brengen en zij scoren mét CCC-model niet beter dan zonder. De 'experts' waren hier gevorderde studenten die een training van één uur met het CCC-model kregen. Mogelijk was dat te weinig, maar de resultaten bevestigen eerdere bevindingen: het blijkt niet eenvoudig om experts te brengen tot betrouwbare en goede voorspellingen van lezersproblemen.

Maaike van den Haak, Menno de Jong en Peter Jan Schellens vergelijken synchrone en retrospectieve hardopdenkprotocollen in de evaluatie van een online catalogus. In de synchrone variant verrichten de proefpersonen een aantal zoektaken hardopdenkend. In de retrospectieve variant verrichten ze die taken stil, maar geven zij achteraf commentaar bij een opname van hun verrichtingen. De methoden blijken globaal tot dezelfde resultaten te leiden, maar ze

komen anders tot stand: synchroon hardopdenken leidt tot meer fouten en omwegen in de taakuitvoering; de gevonden problemen zijn daardoor merendeels observeerbaar en niet of niet alleen uit het commentaar van de proefpersoon af te leiden. Bij retrospectief hardopdenken presteren de proefpersonen beter. De problemen komen nu merendeels pas aan het licht door het commentaar achteraf, en zijn minder uit het gedrag af te leiden. Daarmee is de reactiviteit van hardopdenkonderzoek zeker bij complexe taken aan de orde.

Miranda Hall, Menno de Jong en Michaël Steehouder snijden, in een bijdrage die niet tijdens het VIOT-congres was te horen, een heel nieuw onderwerp aan. De resultaten van verschillende pretestmethoden zouden wel eens afhankelijk kunnen zijn van de cultuur waaruit proefpersonen afkomstig zijn. Zij vonden dat de plus-en-minmethode en retrospectieve hardopdenkprotocollen, ingezet om een website te evaluaren, inderdaad anders functioneren. De plus-en-minmethode brengt bij proefpersonen uit een collectivistische cultuur minder problemen aan het licht dan de hardopdenkprotocollen. Daarnaast formuleren proefpersonen uit een collectivistische cultuur hun commentaar in de hardopdenkprotocollen indirecter dan proefpersonen uit een individualistische cultuur. Verder is interessant dat in het onderzoek de verwachte verschillen in gedrag wel werden gevonden en de verwachte verschillen in zelfrapportagematen niet. Dat wijst erop dat zelfrapportage (bijvoorbeeld in de beantwoording van vragenlijsten) veel gevoeliger is voor sociale wenselijkheid dan gedrag – een flinke complicatie in intercultureel onderzoek naar (o.a.) tekstkwaliteit.

Tenslotte buigen Leo Lentz en Menno de Jong zich over de vraag in welke termen lezers denken over tekstproblemen. In veel vragenlijsten die aan lezers worden voorgelegd, worden termen gebruikt die aan tekstonderzoek zijn ontleend; in sommige evaluatiemethoden (zoals in Focus) moeten lezers hun eigen commentaar benoemen in vooraf gegeven termen. Sluit die terminologie wel aan op categorieën en termen waarmee leken vertrouwd zijn? In hun onderzoek lieten Lentz en De Jong leken de lezersreacties uit eerder onderzoek op stapeltjes leggen en benoemen. Daaruit trekken zij de conclusie dat de termen die in het vakgebied worden gebruikt nog redelijk goed aansluiten op de lekenterminologie. De duidelijkst herkenbare klassen problemen die leken onderscheiden zijn: overbodige en ontbrekende informatie, geloofwaardigheid, formulering en structuur.

Ter afsluiting past dank en verontschuldiging. De redacteuren van de VIOT-congresbundel (Van Waes e.a., 2003) danken we hartelijk voor hun medewerking aan dit themanummer in de vorm van de bereidwillige afstand die zij van enkele zeer interessante papers voor hun bundel deden. De link met het congres en het daar gehouden symposium vormt tegelijkertijd de verontschuldiging voor het eenzijdig Utrechts-Twentse karakter van dit themanummer.

Bibliografie

Jong, M. de & P.J. Schellens (1995). *Met het oog op de lezer. Pretestmethoden voor schriftelijk voorlichtingsmateriaal*. Amsterdam: Thesis.

Jong, M. de & P.J. Schellens (2000). Toward a document evaluation methodology: What does research tell us about the validity and reliability of evaluation methods. *IEEE Transactions on professional communication, 43,* 242-260.

Jong, M. de & P.J. Schellens (2002). Tekstevaluatie. Onderzoek naar de validiteit van probleemopsporende methoden. *Tijdschrift voor Taalbeheersing, 14*, 146-166.

Waes, L. van, P. Cuvelier, G. Jacobs, I. de Ridder (red.) (2003). *Studies in Taalbeheersing* (vol.1). Van Gorcum: Assen

HENK PANDER MAAT
EN LEO LENTZ

Waarom het lezersprotocol zo'n goede methode is om begrips-problemen op te sporen

Samenvatting

In het onderzoek naar pretestmethoden is tot nu toe niet veel aandacht besteed aan een methode waarbij proefpersonen hardop denkend een tekst lezen (verder: het lezersprotocol). In dit artikel wordt deze methode vergeleken met twee andere methoden: de plus-en-minmethode en het softwareprogramma Focus. De resultaten pleiten ervoor het lezersprotocol serieus te nemen als pretestmethode. In vergelijking met de plus-en-minmethode levert het lezersprotocol met name meer begripsproblemen op; voor dat type problemen blijkt de predictieve validiteit ook groter te zijn. De verschillen met Focus zijn minder groot. In een vervolgonderzoek is nagegaan wat de verklaring is voor dit resultaat. De conclusie is dat niet het *hardop lezen* van de tekst leidt tot de hogere opbrengst van het lezersprotocol, maar dat de kracht van de methode vooral zit in het *direct verwoorden van reacties,* hetgeen bij de plus-en-minmethode niet gebeurt.

1. Inleiding

In dit artikel onderzoeken wij de waarde van een pretestmethode die tot nu toe in de literatuur nog weinig aandacht heeft gekregen: het lezersprotocol. We vergelijken die methode eerst met de plus-en-minmethode, en vervolgens met Focus, een softwareprogramma waarmee lezersreacties op een tekst kunnen worden verzameld. Tenslotte onderzoeken we verschillende varianten van het lezersprotocol, in een poging om een antwoord te krijgen op de vraag die gesteld is in de titel. De drie studies die hieronder worden besproken, zijn uitvoeriger gerapporteerd in Koppenaal (2000), Noorlander (2001) en Van Werven (2002).

Hieronder geven we eerst een korte schets van de drie onderzochte pretestmethoden, vervolgens rapporteren we de twee vergelijkende onderzoeken. Tenslotte gaan we dieper in op de specifieke kwaliteiten van het lezersprotocol.

De plus-en-minmethode. Bij de plus-en-minmethode wordt de proefpersoon uitgenodigd voor zichzelf de tekst door te nemen en plussen en minnen te zetten bij passages die positief dan wel negatief worden beoordeeld, om wat voor reden dan ook. In een nagesprek licht de proefpersoon zijn plussen en minnen toe. Ongerichte methoden van lezersonderzoek kunnen op twee manieren worden onderverdeeld (zie De Jong & Schellens 1995, 52-54). Ten eerste kan een methode meer gericht zijn op beoordeling van een tekst of op gebruik ervan. Ten tweede kan een methode synchroon zijn of retrospectief. Bij een syn-

chrone methode wordt de feedback gegeven tijdens het lezen, bij een retrospectieve methode erna. In het licht van deze onderscheidingen kan de plus-en-minmethode worden gekenschetst als een beoordelingsmethode met een gemengd synchroon-retrospectief karakter: de plussen en minnen worden synchroon gezet en later toegelicht.

Over de kwaliteit van de plus-en-minmethode is redelijk wat bekend, met name door het werk van De Jong & Schellens (De Jong 1998, De Jong & Schellens 2000). In Nederland wordt de methode erkend als een praktisch goed bruikbare methode met een redelijke predictieve validiteit. Dat laatste wil zeggen dat de problemen die aan het licht komen met behulp van de methode hoogst waarschijnlijk reële problemen zijn, en blijkens onderzoek ook kunnen worden omgezet in revisies die de tekst effectiever maken.

Anderzijds bestaat er twijfel aan de relatie tussen de minnen van een proefpersoon en de problemen die tijdens het leesproces daadwerkelijk worden ervaren. Pander Maat (1996) gebruikte een onderzoeksopzet waarin daadwerkelijk optredende begripsproblemen werden vastgesteld door na de pretest een tekstbegripstoets af te nemen. Alle foute antwoorden in die begripstoets werden beschouwd als begripsproblemen. Slechts 32 % (n=382) van deze problemen werd spontaan gerapporteerd in de voorafgaande plus-en-mintest.

Het lezersprotocol. Bij een lezersprotocol leest de proefpersoon hardop, en spreekt deze direct de gedachtes uit die de tekst oproept. De onderzoeker houdt op een observatieformulier de commentaren bij, en in een nagesprek wordt ingegaan op onduidelijke commentaren en andere opvallende gebeurtenissen tijdens het lezen, zoals haperingen en fouten. De Jong & Schellens (1995) hanteren de term hardop-leesmethode in plaats van 'lezersprotocol'. Wij nemen deze term niet over, omdat later zal blijken dat het hardop lezen van de tekst geen noodzakelijke voorwaarde is voor het verkrijgen van een lezersprotocol.

De Jong & Schellens wijzen erop dat het lezersprotocol niet mag worden verward met het gebruikersprotocol (zij spreken van de hardop-werkmethode). Dit is een gerichte methode waarbij vragen of opdrachten worden gebruikt, die hardop denkend met behulp van de tekst worden opgelost. Het lezersprotocol daarentegen is een ongerichte methode. Verder is hij synchroon en lijkt hij eerder gericht op tekstgebruik dan op tekstbeoordeling, zij het dat het om een ongerichte vorm van gebruik gaat: hardop denken geeft een indruk van het gebruik van een tekst door een lezer die alles wil begrijpen wat er staat. Deze vorm van begrijpend lezen zal in de praktijk vaak een voorbereiding zijn op het gebruik van de tekst om bepaalde vragen te beantwoorden. Bij protocollen wordt niet rechtstreeks om oordelen gevraagd, maar om een verslag van de verwerking.

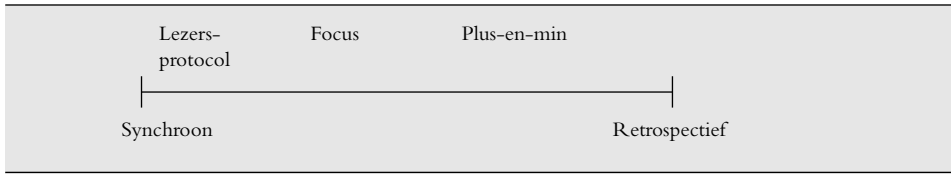
Over de kwaliteit van lezersprotocollen als pretest is nog niet al te veel bekend. De Jong & Schellens (1995, 151-157) zijn uiterst behoedzaam in hun bespreking van de methode. Wel vermoeden zij dat hardop lezen en hardop denken slecht samengaan. Inderdaad stelden Allwood & Kalén (1993) vast dat er bij hun proefpersonen interferentie optrad tussen de twee taken. Daarnaast stellen De Jong & Schellens dat de methode wellicht leidt tot een onnatuurlijke leeswijze, in die zin dat de tekst van begin tot eind wordt gelezen zonder gedeelten over te slaan of vluchtig te lezen. Maar volgens ons staat daar een belangrijk voordeel tegenover: hardop lezen en direct commentaar verschaft waarschijnlijk meer inzicht in de activiteiten van de lezer dan de plus-en-minmethode, waarin het commentaar wordt uitgesteld tot na het lezen. Dit lijkt met name voor begripsactiviteiten een voordeel op te kunnen leveren. Niet alleen wordt een begripsprobleem direct gesignaleerd, ook de interpretaties van de betreffende passage worden direct zichtbaar.

Het meeste onderzoek over hardop-denkprotocollen gaat over de hardop-werkmethode (Flower e.a. 1983, Swaney e.a. 1991). Schriver (1991) is zelfs vrij pessimistisch over het afnemen van hardop-denkprotocollen aan proefpersonen zonder specifieke opdrachten. De enige twee studies die wel specifiek ingaan op het lezersprotocol zijn Dieli (1986) en Sienot (1997). Dieli vergeleek in kleinschalig onderzoek naar de kwaliteit van een handleiding lezersprotocollen met de gebruikersprotocollen. Zij constateerde dat lezersprotocollen meer problemen aan het licht brengen met de interpretatie van specifieke tekstpassages, terwijl de hardop-werkmethode meer licht werpt op de bruikbaarheid van de tekst als geheel. Sienot onderzocht de kwaliteit van een website met behulp van lezersprotocollen en de plus-en-minmethode. Hij vond dat de plus-en-minmethode meer problemen aan het licht bracht dan het lezersprotocol; het verschil lag met name in het grotere aantal waarderingsproblemen dat de plus-en-minmethode aan het licht bracht. Voor ons onderzoek heeft de studie van Sienot een beperkte relevantie, omdat hij een speciale variant van de plus-en-minmethode onderzocht. Omdat de proefpersonen geen plussen en minnen op het beeldscherm kunnen zetten omcirkelden zij eerst met de muis een passage die zij wilden becommentariëren, en meldden zij er mondeling bij of het een plus dan wel een min betrof. Deze procedure leidde ertoe dat veel proefpersonen het niet lieten bij een plus of een min, maar direct commentaar leverden op een bepaalde passage. Daarmee ging de plus-en-minmethode enigszins lijken op het lezersprotocol, wat wellicht de productiviteit ervan heeft bevorderd.

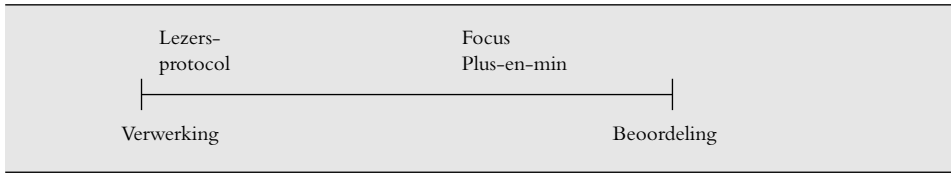
Focus. Het softwareprogramma Focus, ontwikkeld door De Jong en Lentz, stelt proefpersonen in staat teksten op een scherm te lezen, passages in de teksten aan te klikken waarop ze commentaar willen leveren en dit commentaar vervolgens zelf in te tikken in een apart commentaarblokje naast de tekst. De proefpersoon dient ieder commentaar onder te brengen in een probleemcategorie. Uit onderzoek dat De Jong & Lentz (2001) uitvoerden naar de eerste versie van het programma blijkt dat Focus praktisch goed bruikbaar is en even veel commentaar oplevert als de plus-en-minmethode.

Focus stelt de proefleider in staat om sneller commentaar te verzamelen dan de plus-en-minmethode. Misschien is er voor de proefpersoon een iets hogere drempel om commentaar te geven dan bij het lezersprotocol. Er moeten namelijk enkele computerhandelingen worden verricht, voordat het commentaar kan worden gegeven, en het commentaar wordt getikt en niet uitgesproken.

We vatten onze bespreking van de drie methoden samen met behulp van twee schalen. Op de schaal die loopt van synchrone naar retrospectieve pretestmethoden vinden we eerst het lezersprotocol, vervolgens Focus en tenslotte de plus-en-minmethode (zie Figuur 1). Daarnaast is er een schaal die loopt van tekstverwerkings- naar tekstbeoordelingsmethoden. Daarbij bedoelen we met tekstverwerking met name het opbouwen van een mentale representatie van de tekstinhoud en niet het toepassen ervan in uitvoeringshandelingen. Op de schaal die loopt van verwerking naar beoordeling ligt Focus samen met de plus-en-minmethode meer aan de kant van de beoordeling dan die van de verwerking, dit in tegenstelling tot het lezersprotocol (zie Figuur 2).



Figuur 1 *De drie methoden geplaatst op de schaal synchroon-retrospectief*



Figuur 2 *De drie methoden geplaatst op de schaal verwerking-beoordeling*

Vraagstelling. Het doel van de vergelijkende studies was om het lezersprotocol te vergelijken met de beide andere methoden. Daarbij stonden de volgende vragen centraal.

- *Hoeveelheid commentaar*: in hoeverre verschilt het lezersprotocol van de andere methoden wat betreft de opbrengst in termen van probleemdetecties?
- *Aard van het commentaar*: verschilt het lezersprotocol ten opzichte van de andere methoden in het soort commentaar dat door proefpersonen wordt geleverd? We onderscheiden commentaar ten aanzien van: begrijpelijkheid, acceptatie, overbodige informatie, ontbrekende informatie, structuur, stijl en correctheid.
- *Validiteit*: in hoeverre verschilt het lezersprotocol van de beide andere methoden wat betreft de predictieve validiteit?

Ten aanzien van de eerste vraag hadden we geen duidelijke verwachting over verschillen in opbrengst. Wat betreft de aard van het commentaar verwachtten we dat het lezersprotocol meer begripsproblemen aan het licht zou brengen dan de andere methoden. Daaruit voort vloeit de verwachting dat het lezersprotocol met name ten aanzien van dat soort problemen meer valide is dan de andere methoden. Dit is vastgesteld door de commentaren te vergelijken met de resultaten van een tekstbegripstoets.

2. Procedure

Tekstmateriaal

De proefpersonen leverden commentaar op bijsluiterteksten. In het eerste onderzoek was dat een tekst over Teveten met 900 woorden, in het tweede onderzoek ging het om een tekst over Cinnarizine van 820 woorden.

Proefpersonen
In beide studies werden voor iedere pretestmethode 15 proefpersonen gebruikt. De proefpersonen verschilden binnen iedere studie niet qua leeftijd, opleidingsniveau of geslacht. Wel waren de proefpersonen in de tweede studie jonger (gemiddeld 34 jaar) dan die in de eerste studie (gemiddeld 51 jaar). In beide studies waren hoger opgeleiden (HBO en academisch) oververtegenwoordigd (ze vormden in beide gevallen 67% van de proefpersonen).

Uitvoering pretest
De pretesten werden afgenomen zoals hierboven beschreven. Vervolgens werd een tekstbegripstoets voorgelegd, die niet van te voren was aangekondigd. In de eerste studie duurde het afnemen van het onderzoek gemiddeld 31 minuten en was er nauwelijks verschil in tijd tussen het lezersprotocol en de plus-en-minmethode. In de tweede studie duurde het lezersprotocol gemiddeld 42 minuten en Focus gemiddeld 34 minuten. Overal geldt dat ongeveer de helft van die tijd heenging met het maken van de tekstbegripstoets, zodat de eigenlijke pretest telkens slechts 15 to 20 minuten duurde.

Bij het lezersprotocol werd gewerkt met observatieformulieren, waarop de commentaren in grote lijnen genoteerd werden, evenals eigenaardigheden tijdens het leesproces zoals haperingen of herlezingen. In het nagesprek werd op die momenten nog eens ingegaan. De ervaringen met lezersprotocollen waren positief. De proefpersonen vonden het een prettige methode om mee te werken. Voor de proefleider is het invullen van de formulieren en het voeren van het nagesprek wel een arbeidsintensieve bezigheid. Er werden bandopnames gemaakt van de protocollen, maar deze hoefden nauwelijks te worden geraadpleegd. We mogen dus zeggen dat het lezersprotocol het in de pretest-praktijk zonder geluidsopnames kan stellen.

Bij de plus-en-minmethode is alleen het nagesprek arbeidsintensief voor de proefleider, en Focus is in dit opzicht de minst belastende methode, omdat de commentaren kant-en-klaar worden aangeleverd. Daar staat tegenover dat Focus een iets meer uitgebreide voorbereiding nodig heeft, omdat de tekst in het programma moet worden geplaatst en relevante commentaar-categorieën moeten worden geselecteerd. Daarnaast is natuurlijk een computer bij iedere afname nodig, en wordt van de proefpersonen enige computerervaring gevraagd.

Begripstoets als hulpmiddel om de validiteit te bepalen
In dit onderzoek werd een specifieke vorm van predictieve validiteit onderzocht: validiteit werd daarbij gedefinieerd als de samenhang tussen negatieve begripscommentaren en reële begripsproblemen *binnen dezelfde proefpersoon*. Het zou ook denkbaar zijn geweest om de predictieve validiteit binnen een relevante populatie te definiëren, zoals bijvoorbeeld gebeurt wanneer de commentaren bij een pretest onder een steekproef uit de doelgroep worden gevalideerd door een begripstoets af te nemen bij een andere steekproef van proefpersonen uit de doelgroep van de tekst. Naar onze mening echter gaat de vraag naar de validiteit van pretestresultaten *binnen* een proefpersoon vooraf aan de vraag of de pretestresultaten corresponderen met begripsproblemen in de populatie: is er geen verband tussen beide grootheden binnen een proefpersoon, dan leert dat ons iets over de manier waarop eventuele verbanden tussen beide grootheden binnen de populatie geïnterpreteerd moeten worden; met andere woorden, de predictieve validiteit binnen proefpersonen is in eerste instantie de meest interessante vraag als we willen weten wat er in een pretest werkelijk gebeurt.

Waarom het lezersprotocol zo'n goede methode is om begripsproblemen op te sporen

Om een indruk te krijgen van de reële begripsproblemen werden begripstoetsen afgenomen na de pretest. De begripstoets telde in de eerste studie 19 items en in de tweede studie 22 items. Om de proefpersonen niet te dwingen tot gokken, werd bij ieder item als antwoordmogelijkheid opgenomen "ik weet het niet". Enkele voorbeelden van vragen uit de tekstbegripstoets in het eerste onderzoek volgen hieronder.

- *Wat is een Blister?*

A Een bijsluiter.
B Een strip waarin tabletten samen zitten verpakt.
C Een recept.
D Ik weet het antwoord op deze vraag niet.
(NB. Deze term werd gebruikt in de beschrijving van de tabletten en hun verpakkingsvorm.)

- *Er staat in de tekst dat het effect van Teveten na 2-3 weken maximaal is. Wat wordt hiermee bedoeld?*

A Na 2-3 weken is de verlaging van de bloeddruk maximaal.
B Na 2-3 weken zal de bloeddruk weer langzaam toenemen.
C Na 2-3 weken is de verhoging van de bloeddruk maximaal.
D Ik weet het antwoord op deze vraag niet.

- *Mag u Teveten 600 gebruiken als u gevoelig bent voor rood ijzeroxide?*

A Nee, rood ijzeroxide is een bestanddeel van Teveten 600.
B Ja, rood ijzeroxide zit namelijk alleen in de Teveten 400.
C Ja, rood ijzeroxide zit namelijk alleen in de Teveten 300.
D Ik weet het antwoord op deze vraag niet.
(Voor het antwoord op deze vraag is enig zoekwerk nodig. Onder het kopje "wie mogen Teveten niet gebruiken" worden mensen genoemd die overgevoelig zijn voor een van de bestanddelen van het middel. Deze bestanddelen staan elders opgesomd.)

- *Bent u in staat om een auto te besturen als u Teveten gebruikt?*

A Nee, want je kunt moe of duizelig worden.
B Alleen wanneer je niet moe of duizelig wordt.
C Ja, maar als je moe of duizelig wordt, moet je wel opletten.
D Ik weet het antwoord op deze vraag niet.
(In de tekst staat: "Behandeling van hoge bloeddruk kan duizeligheid of moeheid geven. Pas dan op met autorijden en het gebruik van machines.")

In de tekstbegripstoets werd dus een vrij ruime definitie van tekstbegrip gehanteerd, waarbinnen niet alleen interpretaties van bepaalde termen vallen, maar ook van zinnen en passages als geheel en ook van informatie die verspreid door de tekst aanwezig is.

Met behulp van een toets als deze valt natuurlijk slechts een beperkt aantal interpretaties te controleren, en kan dus slechts ook een beperkt aantal commentaren gevalideerd worden. Het gaat dus eigenlijk om een steekproefsgewijze validering van de begripscommentaren. Bij de selectie van probleemdetecties hebben we ons vooral laten leiden door de mate waarin een probleem geherformuleerd kon worden in termen van een begripsvraag met een aantal realistische antwoordmogelijkheden.

De werkwijze was als volgt. Bij iedere toetsvraag wordt voor iedere proefpersoon vastgesteld of het antwoord goed of fout is, en of de proefpersoon in de pretest commentaar heeft geleverd dat een begripsprobleem signaleert met de passage waar de vraag over gaat. Deze vergelijking levert vier mogelijke uitkomsten op (zie tabel 1): een treffer, een misser, een vals alarm of een 'geen nieuws, goed nieuws'-situatie.

Tabel 1. *Mogelijke resultaten van de vergelijking tussen begripstoets en pretest*

Begripstoets	Pretest	Uitkomst
Fout antwoord (incl."ik weet het niet")	Negatief begripscommentaar	Treffer
Fout antwoord	Geen commentaar	Misser
Goed antwoord	Negatief begripscommentaar	Vals alarm
Goed antwoord	Geen commentaar	Geen nieuws, goed nieuws

Iedere begripsvraag levert een aantal treffers, een aantal missers en een aantal valse alarms. Deze aantallen zijn bij elkaar opgeteld voor de vijftien proefpersonen die een bepaalde pretestmethode gebruikten, zodat iedere methode een aantal getallen kreeg. Van belang zijn nu niet zozeer de afzonderlijke getallen binnen iedere methode als wel de verhoudingen tussen de aantallen treffers, missers en valse alarmsignalen. Deze verhouding stelt ons namelijk in staat om de trefkans en de detectiekans horend bij de methode uit te rekenen. Met de *trefkans* bedoelen wij de kans dat een probleemcommentaar tijdens de pretest correspondeert met een daadwerkelijk begripsprobleem, zoals gebleken bij de begripstoets. De trefkans geeft met andere woorden de kans aan dat in de vergaarbak van feedback echte begripsproblemen aangetroffen worden. Het is immers denkbaar dat er met een methode wel belangrijke treffers gevonden worden, maar dat die nauwelijks vindbaar zijn omdat ze verdwijnen in een geweldige hoeveelheid commentaren. De trefkans wordt berekend door het aantal treffers te vergelijken met het totale aantal verschillende begripsproblemen dat spontaan gemeld is voor zover betrekking hebbend op een passage bevraagd in de begripstoets. Dit totaal aantal commentaren bestaat uit de som van het aantal treffers en het aantal valse alarmsignalen.

Met de *detectiekans* bedoelen we de kans dat een daadwerkelijk begripsprobleem zoals gebleken in de begripstoets gesignaleerd is tijdens de pretest. Het is denkbaar dat de treffers verscholen gaan in een grote hoeveelheid commentaren (kleine trefkans), maar dat daarmee wel *alle* problemen die zich daadwerkelijk voordoen opgespoord worden (hoge detectiekans). De detectiekans wordt berekend door het aantal treffers te delen door het totaal aantal probleemsignaleringen dat in de begripstoets naar voren is gekomen. Dit totaal bestaat uit de som van de treffers en de missers.

3. Resultaten

3.1 De opbrengst van de drie methoden. De eerste onderzoeksvraag was of het lezersprotocol verschilde met de beide andere methoden ten aanzien van de opbrengst. Uit tabel 2 blijkt allereerst dat er meer negatief dan positief commentaar wordt geleverd bij alle methoden. Alleen in de eerste studie bleek dat lezersprotocollen meer negatief commentaar ople-

veren dan de plus-en-minmethode, maar minder positief commentaar (zie de statistische gegevens in de tabel; de toetsing was tweezijdig). In het tweede onderzoek bleek dat lezersprotocollen zowel meer positieve als meer negatieve reacties opleveren dan Focus.

Tabel 2. *Totale aantallen positieve en negatieve commentaren van lezersprotocol en plus-en-minmethode (studie 1), respectievelijk lezersprotocol en Focus (studie 2).*

Studie 1	**Lezersprotocol**	**Plus-en-minmethode**	**Toetsing**
positief commentaar	38	59	$Chi^2 = 4.55$ $df = 1, p < .04$
negatief commentaar	135	99	$Chi^2 = 5.54$ $df = 1, p < .02$
Studie 2	**Lezersprotocol**	**Focus**	**Toetsing**
positief commentaar	39	11	$Chi^2 = 15.68$ $df = 1, p = .00$
negatief commentaar	214	122	$Chi^2 = 25.19$ $df = 1, p = .00$

Omdat problemen bij een pretest belangrijker zijn dan positieve commentaren, beperken we ons verder tot de negatieve commentaren. Tabel 2 betreft het totaal aantal gegeven commentaren. In de praktijk van het pretesten is het ook belangrijk hoeveel *verschillende* commentaren een test oplevert. Die staan weergegeven in tabel 3.

Tabel 3. *Aantal verschillende negatieve commentaren van lezersprotocol en plus-en-minmethode (studie 1), respectievelijk lezersprotocol en Focus (studie 2).*

Studie 1	**Lezersprotocol**	**Plus-en-minmethode**	**Toetsing**
Verschillende neg. commentaren	72	56	n.s.
Studie 2	**Lezersprotocol**	**Focus**	**Toetsing**
Verschillende neg. Commentaren	88	79	n.s.

Uit tabel 3 blijkt dat de aantallen afzonderlijke problemen niet significant verschillen tussen de pretestmethoden. In de tweede studie blijkt er echter wel een verschil te bestaan tussen het lezersprotocol en Focus, en wel wat betreft de verhouding tussen het aantal verschillende problemen en het totaal aantal problemen. Bij de lezersprotocollen kwam het vaker voor dat een negatief commentaar door verschillende proefpersonen werd geuit: 214 commentaren leverden slechts 88 verschillende problemen op, dat wil zeggen dat ieder probleem ongeveer 2,4 maal genoemd werd; dezelfde ratio bedroeg slechts 1,5 bij Focus ($Chi^2 = 5.62, df = 1, p < .02$, tweezijdig getoetst). Hieruit mogen twee conclusies getrokken worden. Ten eerste is het lezersprotocol betrouwbaarder in die zin dat er tussen lezers meer overeenstemming is over problemen dan bij Focus-lezers. Ten tweede is het waarschijnlijk dat er bij lezersprotocollen minder lezers nodig zijn om een bepaald probleem aan het licht te brengen dan bij Focus het geval is.

De tweede onderzoeksvraag was hoe de probleemcommentaren verdeeld zijn over de verschillende soorten tekstproblemen (begrijpelijkheid, acceptatie, overbodige informatie, ontbrekende informatie, structuur, stijl en correctheid). In de eerste studie was er op dit punt één verschil: zoals verwacht leverden lezersprotocollen meer begripsproblemen op dan de plus-en-minmethode.

Bij vergelijking van de frequenties van begripsproblemen was het verschil zowel zichtbaar bij het totaal aantal commentaren als voor het aantal verschillende commentaren (zie tabel 4).

Tabel 4. *Aantallen begripscommentaren bij lezersprotocol en plus-en-min-methode in de eerste studie.*

	Lezersprotocol	**Plus-en-min-methode**	**Toetsing**
Totaal aantal begrips-commentaren	44	24	$Chi^2 = 5.88$, df = 1, $p < .01$
Verschillende begrips-commentaren	23	10	$Chi^2 = 5.12$, df = 1, $p < .02$

In proportionele zin (dat wil zeggen, wanneer de begripscommentaren in verhouding tot alle overige commentaren worden gezien) gold het verschil alleen voor het aantal verschillende commentaren (32% versus 18%; $Chi^2 = 3.27$, df = 1, $p < .05$, eenzijdig getoetst). We kunnen zeggen dat het verschil in opbrengst tussen beide methodes voor een belangrijk deel schuilt in het grotere aantal begripsproblemen bij de lezersprotocollen.

In de tweede studie bleken er twee verschillen te bestaan tussen de lezersprotocollen en Focus wat betreft de soorten gesignaleerde problemen.

- Ten eerste brachten de lezersprotocollen meer begripsproblemen aan het licht dan Focus (100 versus 69), maar dit verschil trad alleen op bij vergelijking van de absolute frequenties van begripscommentaren ($Chi^2 = 5.69$, df = 1, $p < .01$, eenzijdig getoetst), en dan nog alleen wanneer we keken naar het totaal aantal commentaren, niet naar het aantal verschillende commentaren. In proportionele zin waren er geen verschillen.
- Opvallend was dat lezersprotocollen relatief meer structuurproblemen aan het licht brachten dan Focus. Dat verschil trad op bij de absolute frequenties, zowel gerekend over het totaal aantal commentaren als over het aantal verschillende commentaren (zie tabel 5).

Tabel 5. *Aantallen structuurproblemen bij lezersprotocol en Focus in de tweede studie.*

	Lezersprotocol	**Focus**	**Toetsing**
Totaal aantal structuur-problemen	43	10	Chi2 = 20.55, df = 1, p = .00
Aantal verschillende structuurproblemen.	18	6	Chi2 = 6.00, df = 1, p <02

Ook proportioneel gezien was het verschil zowel significant bij de totale aantallen (20% versus 8%; $Chi^2 = 8.09$, df = 1, $p < .01$, tweezijdig getoetst) als bij het aantal verschillende commentaren (21% versus 8%; $Chi^2 = 5.59$, df = 1, $p < .02$, tweezijdig getoetst). Focus lijkt dus uit te nodigen tot een meer lokaal gerichte kritiek dan lezersprotocollen doen.

3.2 De validiteit van de begripscommentaren. Een derde onderzoeksvraag ging over de validiteit van het lezersprotocol in vergelijking met de beide andere methoden. Dit is onderzocht door de pretestcommentaren te vergelijken met de resultaten van een tekstbegripstoets. Tabel 6 geeft de trefkans weer, uitgedrukt in een percentage, in beide studies. Ter herinnering: de trefkans geeft de kans aan dat in de vergaarbak van feedback echte begripsproblemen aangetroffen worden.

Tabel 6. *Treffers en valse alarms van lezersprotocol en plus-en-minmethode (studie 1), respectievelijk lezersprotocol en Focus (studie 2).*

Studie 1	**Lezersprotocol**	**Plus-en-minmethode**
Totaal relevante commentaren	29	20
Treffer	17 (trefkans 58%)	8 (trefkans 40%)
Vals alarm	12	12
Studie 2	**Lezersprotocol**	**Focus**
Totaal relevante commentaren	41	29
Treffer	21 (trefkans 51%)	15 (trefkans 51%)
Vals alarm	20	14

Uit tabel 6 blijkt dat er slechts kleine verschillen in de trefkans tussen de verschillende methodes bestaan, die verschillen waren niet significant. In de verschillende methoden is één op de twee gevonden begripsproblemen vermoedelijk raak. Naast de trefkans onderscheidden we de detectiekans, waarmee wordt aangeduid hoe groot de dekking is van de treffers over de daadwerkelijk in de begripstoets vastgestelde problemen. Daartoe wordt de verhouding tussen treffers en missers bepaald. Tabel 7 laat de resultaten zien.

Tabel 7. *Treffers en missers van lezersprotocol en plus-en-minmethode (studie 1), respectievelijk lezersprotocol en Focus (studie 2).*

Studie 1	**Lezersprotocol**	**Plus-en-Minmethode**
Totaal aantal problemen	78	93
Treffer	17 (detectie 22%)	8 (detectie 9%)
Misser	61	85
Studie 2	**Lezersprotocol**	**Focus**
Totaal aantal problemen	117 1	16
Treffer	21 (detectie 18%)	15 (detectie 13%)
Misser	96	101

Uit tabel 7 blijkt dat alleen in de eerste studie er een verschil is in detectiekans tussen lezersprotocollen en de plus-en-minmethode (22% versus 9%; $Chi^2 = 5.91$, $df = 1$, $p < .01$, eenzijdig getoetst). Tussen lezersprotocollen en Focus is er geen verschil, noch in trefkans, noch in detectiekans. We zien ook dat de detectiekans in de beide methoden aanmerkelijk lager is dan de trefkans, waaruit geconcludeerd kan worden dat het zinvol is een ongerichte pretest altijd te combineren met een begripstoets. De kans dat daarmee nieuwe problemen zichtbaar worden lijkt immers bijzonder groot te zijn.

Er is nog een andere manier om de waarde van de begripscommentaren tijdens de pretest te bepalen aan de hand van de resultaten van de begripstoets. We kunnen namelijk voor iedere vraag van de begripstoets het aantal proefpersonen met een fout antwoord vergelijken met het aantal proefpersonen dat een probleem met de bevraagde passage signaleert. Over deze getallenparen berekenen we vervolgens de correlatie. Is die significant positief, dan gaat een groot aantal foute antwoorden samen met een groter aantal commentaren (zie tabel 8.).

Tabel 8. *Correlatie tussen aantal fouten bij iedere vraag uit de begripstoets (N=19 in studie 1; N=22 in studie 2) en het aantal keer dat het betreffende probleem in de pretest gesignaleerd is bij lezersprotocol en plus-en-minmethode (studie 1), respectievelijk lezersprotocol en Focus (studie 2).*

Studie 1	**Lezersprotocol**	**Plus-en-Minmethode**
Pearson correlatie	.57	.00
p-waarde, 2-zijdig	.01	n.s
Studie 2	**Lezersprotocol**	**Focus**
Pearson correlatie	.63	.08
p-waarde, 2-zijdig	.002	n.s

Uit tabel 8 blijkt ondubbelzinnig dat het lezersprotocol samenhang vertoont met de begripstoetsresultaten, terwijl deze samenhang ontbreekt bij de plus-en-minmethode en Focus. Men zou kunnen denken dat dit een triviaal resultaat is, omdat in lezersprotocollen meer begripscommentaren werden gegeven dan bij de twee andere methoden. Het was echter heel goed mogelijk geweest dat een methode grote aantallen begripscommentaren oplevert zonder dat er een relatie is tussen deze commentaren en de begripsprestaties op de betreffende passage.

3.3 Discussie. De eerste studie levert een duidelijk resultaat op, in die zin dat lezersprotocollen niet alleen meer commentaren opleveren dan de plus-en-minmethode, maar ook meer begripscommentaren, zowel in absolute als in relatieve zin. Daarnaast blijkt de detectiekans voor lezersprotocollen hoger te zijn en blijkt deze methode een significante correlatie op te leveren tussen aantallen fouten op begripsvragen en aantallen proefpersonen met een probleemsignalering. Het lezersprotocol is duidelijk een betere maatstaf voor begripsproblemen met de tekst dan de plus-en-minmethode.

In de tweede studie werd het lezersprotocol vergeleken met een methode die minder retrospectief van karakter is, namelijk Focus. De verschillen in deze studie zijn minder groot: lezersprotocollen leveren weliswaar meer commentaren op dan Focus, maar qua begripscommentaren is het verschil minder overtuigend. Wat betreft de validiteit van de begripscommentaren is er tussen beide methoden wel het verschil dat het lezersprotocol een correlatie toont tussen het aantal fouten op begripsvragen en het aantal proefpersonen met een probleemsignalering dienaangaande, en Focus niet.

Wij zien een aantal mogelijke verklaringen voor met name de verschillen in de eerste studie.

- Om te beginnen is er bij de plus-en-minmethode sprake van *uitstel* van commentaren, en bij het lezersprotocol vrijwel niet. Dit uitstel kan op verschillende manieren het aantal probleemcommentaren verlagen. Ten eerste kan een probleem dat ervaren werd tij-

dens het lezen, simpelweg vergeten worden wanneer het niet direct wordt uitgesproken. Ten tweede kan een probleem verdwijnen, of minder belangrijk worden, na het lezen van de vervolgtekst. Ten derde is het mogelijk dat na het lezen van een passage slechts een deel van de ervaren problemen verwoord wordt, wellicht gekozen op belang. Een argument voor het belang van de uitstelfactor is dat het verschil tussen lezersprotocollen en Focus minder groot is dan dat tussen lezersprotocollen en de plus-en-min-methode: Focus noodzaakt namelijk minder tot uitstel dan de plus-en-min-methode.

- Nog weer een ander verschil tussen beide methoden heeft te maken met het feit dat bij het lezersprotocol sprake is geweest van *verklanking* van de tekst. Dat heeft twee mogelijke gevolgen voor de opbrengst van de pretest. Ten eerste hoort de proefleider dat sommige tekstgedeeltes herhaald of haperend worden voorgelezen. Op die tekstgedeeltes kan hij terugkomen in het nagesprek. In de eerste studie werd 25% van de probleemcommentaren bij het lezersprotocol op die manier verzameld. Dit soort probleemsignalering is natuurlijk uitgesloten bij de plus-en-minmethode. In de tweede studie is dit laatste verschijnsel niet apart geanalyseerd, maar het is waarschijnlijk dat het zich ook hier heeft voorgedaan. Wanneer de proefleider conclusies kan trekken uit de wijze van voorlezen van de tekst, is het waarschijnlijk dat de proefpersoon dit zelf ook kan en dat hij zich dus bewust kan worden van problemen die anders onopgemerkt zouden zijn gebleven.
- Een derde verschil tussen de plus-en-minmethode en het lezersprotocol heeft te maken met het *tempo* waarmee de tekst verwerkt is, omdat hardop lezen zo'n twintig procent langzamer gaat dan stillezen (Rayner, 1998, 373; Ericsson, 1988, 306). Daarnaast wordt de hardop lezende proefpersoon nog verder vertraagd doordat hij commentaren geeft tijdens het lezen. Het is mogelijk dat een langzame verwerking van de tekst op zichzelf reeds meer begripsproblemen oplevert dan een snellere verwerking.

Naar deze mogelijke verklaringen hebben wij verder onderzoek gedaan door middel van het experiment waarover in paragraaf vier gerapporteerd wordt.

4. Een experimenteel onderzoek naar de kwaliteit van het lezersprotocol

4.1 Opzet. In methodevergelijkend onderzoek zijn tot dusver telkens twee heel verschillende pretestmethoden met elkaar vergeleken (zie voor een overzicht De Jong en Schellens, 2002). Omdat twee methodes op talloze punten van elkaar verschillen, valt een verschil in opbrengst niet eenduidig toe te schrijven aan bepaalde factoren. Om die reden hebben we een experimenteel vervolgonderzoek uitgevoerd waarin de opbrengst wordt vergeleken van vier verschillende varianten van het lezersprotocol. We schetsen eerst het ontwerp van het experiment en daarna gaan we in op de manier waarop we uit de resultaten conclusies kunnen trekken over de drie genoemde factoren.

In het experiment is een verkorte versie van de bijsluitertekst uit studie 2 op een beeldscherm aan proefpersonen aangeboden. Twee variabelen zijn kruislings gemanipuleerd. Ten eerste hebben we geprobeerd om het uitstel tussen lezen en commentaar geven te beïnvloeden door de tekst ofwel in alinea's ofwel in zinnen aan te bieden. In de alineaconditie kregen de proefpersonen telkens één alinea tegelijk op het scherm te zien. De meeste alinea's telden zes zinnen. Na het lezen van een tekstpassage (d.w.z na een alinea dan wel na

een zin) drukten de proefpersonen op de spatiebalk om aan te geven dat ze klaar waren met lezen. Vervolgens konden ze beginnen met commentaar leveren. De becommentarieerde passage bleef zichtbaar op het scherm. Wanneer de proefpersonen hun commentaar voltooid hadden, konden zij doorgaan naar de volgende tekstpassage door nogmaals op de spatiebalk te drukken. In de alineaconditie werd de alinea dan geheel vervangen door de volgende alinea. In de zinsconditie kregen de proefpersonen telkens wanneer zij voor de tweede maal op de spatiebalk gedrukt hadden een nieuwe zin op het scherm. De voorgaande zinnen van de alinea bleven zichtbaar. Deze verdwenen pas wanneer de laatste zin van de alinea gelezen was.

In de zinsconditie werden de proefpersonen dus aangemoedigd direct op elke zin te reageren, terwijl in de alineaconditie de proefpersonen aangemoedigd werden het commentaar uit te stellen tot de gehele alinea gelezen was. Hoewel het de proefpersonen in de alineaconditie niet verboden werd om commentaar te leveren tijdens het voorlezen van de alinea, gebeurde dit in de praktijk nauwelijks. Dat betekent dat de uitstelfactor effectief geoperationaliseerd werd met behulp van het verschil tussen de zins- en de alineaconditie.

De tweede variabele betrof het al dan niet voorlezen van de tekst, waarmee de factor *verklanking* gemanipuleerd werd. De twee onafhankelijke variabelen werden gekruist zodat het ontwerp vier cellen kreeg: alinea-stil, zin-stil, alinea-hardop, zin-hardop. Het drukken op de spatiebalk tussen lezen en hardop denken stelde ons in staat de leestijden van de proefpersonen te registreren, exclusief hun commentaartijd. Op die manier konden wij de factor *tempo* meenemen in het onderzoek, zij het niet als onafhankelijke variabele.

In totaal werkten 48 proefpersonen aan het onderzoek mee: twaalf voor elk van de vier cellen in het ontwerp. De proefpersonen waren letterenstudenten. Leeftijd en geslacht verschilden niet per cel.

4.2 Verwachtingen. Deze onderzoeksopzet kan op de volgende wijze leiden tot conclusies over het gewicht van de uitstel-, de verklankings- en de tempofactor.

- Wanneer de opbrengst van lezersprotocollen met name toe te schrijven is aan de uitstelfactor, zou de opbrengst van de zinsconditie hoger moeten zijn dan die van de alineaconditie. Immers de alineaconditie motiveert de proefpersonen om commentaar uit te stellen tot de alinea als geheel is gelezen, terwijl de zinsconditie dat niet doet.
- Wanneer de verklankingsfactor een rol speelt, zou de opbrengst van de hardoplees-conditie hoger moeten zijn dan de opbrengst van de stillees-conditie.
- Wanneer de tempofactor van belang is, zouden condities die sterk verschillen in leestijden ook verschillen in opbrengst moeten laten zien. Nu zijn er zowel leestijdverschillen te verwachten tussen hardop en stil lezen (stil lezen gaat sneller), als tussen alineagewijs en zinsgewijs lezen (alinea's lezen gaat waarschijnlijk sneller). De factor tempo is dus in dit ontwerp gecontamineerd met de verklankings- respectievelijk de uitstelfactor.

4.3 Resultaten. Wij bespreken eerst de factor *uitstel*. Uit tabel 9 blijkt dat de zinsconditie aanzienlijk meer commentaar oplevert dan de alineaconditie ($Chi^2 = 99.99$, $df = 1$, $p < .001$, tweezijdig getoetst). In tabel 10 blijkt dat dit ook geldt voor negatieve commentaren ($Chi^2 = 32.21$, $df = 1$, $p < .001$, tweezijdig getoetst). In de tabellen 11 en 12 blijkt dat dit ook geldt wanneer het aantal verschillende commentaren wordt geteld, zowel voor het totaal aantal commentaren ($Chi^2 = 53.15$, $df = 1$, $p < .001$, tweezijdig getoetst), als voor de nega-

tieve commentaren ($Chi^2 = 8.88$, df = 1, $p < .005$, tweezijdig getoetst). Dit pleit sterk voor de redenering dat de factor *uitstel* een rol speelt in de verschillende opbrengsten van enerzijds het lezersprotocol en anderzijds de plus-en-minmethode.

Tabel 9. *Totaal aantal commentaren per conditie(N=48)*

Totaal aantal commentaren	Alinea	Zin	Totaal
Hardop lezen	120	270	390
Stil lezen	139	272	411
Totaal	259	542	801

Tabel 10. *Negatieve commentaren per conditie (N=48)*

Negatieve commentaren	Alinea	Zin	Totaal
Hardop lezen	80	150	230
Stil lezen	90	142	232
Totaal	170	292	462

Tabel 11. *Aantal verschillende commentaren per conditie (N=48)*

Aantal verschillende commentaren	Alinea	Zin	Totaal
Hardop lezen	81	167	248
Stil lezen	90	168	258
Totaal	171	335	506

Tabel 12. *Verschillende negatieve commentaren per conditie* (N=48)

Verschillende negatieve commentaren	Alinea	Zin	Totaal
Hardop lezen	60	89	149
Stil lezen	61	83	144
Totaal	121	172	293

We hebben de aard van het uitsteleffect iets nader proberen te onderzoeken. Wellicht kan het uitsteleffect verklaard worden uit een gebrekkig geheugen voor problemen in zinnen die al wat langer geleden gelezen zijn. In dat geval zouden we verwachten dat er in de alineaconditie meer commentaar geleverd wordt op zinnen later in de alinea, terwijl het commentaar in de zinsconditie gelijkmatiger gespreid zou moeten zijn. Deze verwachting is

onderzocht door rangorde-correlaties te berekenen tussen de positie van zinnen in langere alinea's en het aantal commentaren op iedere zin. We vonden geen significante correlaties en kunnen daarom aannemen dat het uitstel effect niet primair een geheugeneffect is.

We hebben ook gecontroleerd of lezers in de zinsconditie problemen noemen die eenvoudig opgelost kunnen worden door de volgende zin te lezen. Inderdaad kwamen er in de zinsconditie begripsproblemen voor na het lezen van kopjes; deze problemen kwamen niet voor in de alineaconditie. Het aantal van deze kunstmatige problemen was echter laag (12). Bij de verdere analyses zijn ze buiten beschouwing gelaten.

De derde verklaring die we noemden in paragraaf 3.4 voor het uitsteleffect lag in het mogelijk selectief rapporteren door alinealezers. Het zou kunnen dat men na het lezen van een alinea zich beperkt tot de problemen die men het meest belangrijk vindt. Deze verklaring kan alleen rechtstreeks worden onderzocht door de gevonden problemen te coderen op belang, maar dat hebben we niet gedaan. Wel hebben we gekeken naar verschillen tussen alinea- en zinslezers in de aard van de commentaren.

Twee verschillen bleken significant. Allereerst blijkt dat de zinsconditie relatief meer commentaar oplevert dat vaak grappig bedoeld is, maar niet duidt op een probleem, noch op lof voor de tekst (23% versus 10%, $Chi^2 = 26.56$, $df = 1$, $p < .000$, tweezijdig getoetst). Zo merkt een proefpersoon na het lezen van de bijwerkingen op: "Ja, leuk, je hebt geen LSD meer nodig, je neemt gewoon cinnarizine." Daarentegen wordt in de alineaconditie meer commentaar gegeven op de structuur van de tekst (14% versus 7 %, $Chi^2 = 8.67$, $df = 1$, $p < .005$, tweezijdig getoetst). Deze verschillen doen zich ook voor, en in nog sterkere mate, wanneer we alleen kijken naar de negatieve commentaren (niet geclassificeerde commentaren: 13 % versus 3 %, $Chi^2 = 11.71$, $df = 1$, $p < .005$, tweezijdig getoetst; structuurcommentaren: 7% versus 20%, $Chi^2 = 14.66$, $df = 1$, $p = .000$, tweezijdig getoetst).

De zinsconditie leidt dus vaker tot enigszins geforceerde commentaren, maar dit verschil verklaart bij lange na niet het verschil in opbrengst tussen de zinsconditie en de alineaconditie: ook zonder deze categorie blijft het verschil hoog significant. Hoe de alinealezers de problemen selecteren waarover ze commentaar leveren blijft dus voorlopig een open vraag.

De tweede factor betrof de *verklanking* als mogelijke verklaring voor de hogere opbrengst van het lezersprotocol. Wanneer we de tabellen 9 tot en met 12 bekijken op de verschillen tussen hardop lezen en stil lezen, dan wordt duidelijk dat die manipulatie niet in het voordeel werkte van het hardop lezen. Het is derhalve erg onaannemelijk dat de hogere opbrengst van het lezersprotocol is toe te schrijven aan het feit dat de proefpersonen zich meer bewust worden van problemen door het verklanken van de tekst. We merken daar wel bij op dat de proefleider in onze studie geen nagesprekken voerde aan de hand van haar observaties ten aanzien van aarzelingen en herhalingen bij het voorlezen. Met andere woorden, onze resultaten geven alleen aan dat hardop lezen niet leidt tot een verhoogd probleem-bewustzijn bij de proefpersonen. Het zou nog steeds wel nuttige informatie kunnen opleveren voor de onderzoeker.

Een derde mogelijke verklaring voor het verschil in opbrengst tussen het lezersprotocol en de plus-en-minmethode lag in het verschil in tempo waarmee de tekst verwerkt wordt. Bij hardop lezen wordt de tekst langzamer verwerkt, en wellicht levert dat op zich al meer problemen op. In dit experiment is het verwerkingstempo niet gemanipuleerd, wel kunnen we

het verwerkingstempo controleren aan de hand van de leestijden. Maar eerst gaan we na of de experimentele manipulatie invloed heeft op de leestijden. We beperken ons daarbij tot de leestijd voor het gehele document, dat wil zeggen de som van de leestijden voor de afzonderlijke zinnen, respectievelijk alinea's.

Tabel 13. *Gemiddelde totale leestijden in seconden per conditie* (N=48)

Gem. totale leestijd	Alinea	Zin	Gemiddeld
Hardop lezen	202	229	215
Stil lezen	144	169	157
Gemiddeld	173	199	

In tabel 13 blijkt er verschil in leestijd te bestaan tussen hardop en stil lezen: hardop lezen gaat een stuk langzamer ($F = 26.40$, $df = 1$, $p = .000$, tweezijdig getoetst). Er is ook een significant verschil tussen alinea- versus zinsgewijs lezen ($F = 5.38$, $df = 1$, $p = .026$, tweezijdig getoetst), maar dat is duidelijk kleiner (de eta^2 voor beide effecten bedraagt respectievelijk .40 en .12).

Desondanks zou men kunnen betogen dat het verschil in opbrengst tussen de zinsconditie en de alineaconditie niet verklaard moet worden uit de uitstelfactor, maar uit verschillen in verwerkingstempo. Het zou bijvoorbeeld kunnen dat zinsgewijze lezers meer commentaren produceren omdat zij de tekst langzamer en dieper verwerken.

Om de mogelijke invloed van verwerkingstempo op het aantal commentaren te vergelijken met het effect van onze experimentele manipulaties, voerden we een regressieanalyse uit. Als afhankelijke variabelen functioneerden het totaal aantal commentaren en het totaal aantal negatieve commentaren. Als onafhankelijke variabelen functioneerden allereerst de vier cellen van het experimentele design: alinea-stil, zin-stil, alinea-hardop, zin-hardop. Als onafhankelijke variabele onderzochten we echter ook de totale leestijd van de proefpersoon, vergeleken met de gemiddelde leestijd van de andere proefpersonen van de experimentele cel.

Deze analyse leverde geen steun op voor het aannemen van een positieve relatie tussen leestijd en aantal commentaren. Om te beginnen werden er significante effecten gevonden voor het al of niet deel uitmaken van de vier experimentele cellen, maar dat is geen nieuws. Wat betreft de leestijden vonden we alleen een significant effect in de zin-stil conditie. In die conditie was er een significante *negatieve* relatie tussen leestijd enerzijds en anderzijds het totaal aantal commentaren ($b = -.142$, $t = -3.775$, $p = .001$; b is de regressie-coëfficient) en het aantal negatieve commentaren ($b = -.103$, $t = -3.803$, $p = .001$).

Dat we geen effect van leestijd vinden voor de hardop conditie, is niet zo verrassend. Het tempo waarin een tekst wordt voorgelezen hangt immers niet alleen af van het verwerkingstempo maar ook van andere zaken zoals het spreektempo van de proefpersoon en diens ambities om netjes voor te lezen. Maar we kunnen niet verklaren waarom het leestijdeffect alleen in de cel zin-stil optreedt en niet in de cel alinea-stil. Hoe dan ook, het belangrijkste is dat de richting van het effect onverwacht is: langere leestijden leiden tot minder, niet tot meer commentaren. Er is dus geen reden om aan te nemen dat de hoge productiviteit van de zinsconditie te maken heeft met een lager verwerkingstempo in die conditie.

4.4 Discussie. Het experiment heeft tot een eenduidig resultaat geleid: de uitstelfactor speelt zeker een rol bij het verklaren van verschillen in opbrengst tussen pretestmethoden, want verschillende versies van het lezersprotocol, die alleen verschillen in de mate van uitstel, verschillen in opbrengst.

Er is daarentegen geen aanwijzing gevonden voor een mogelijke rol van de factor verklanking. Dat is een verrassend resultaat. Commentaar leveren op teksten tijdens het lezen ervan blijkt goed mogelijk te zijn zonder de betreffende tekst hardop te lezen. Dat is voor ons een belangrijke reden om te spreken over *lezersprotocollen* in plaats van over de hardopleesmethode. Het is natuurlijk mogelijk dat stil lezen gevolgd door hardop becommentariëren makkelijker gaat wanneer de teksten op het scherm worden aangeboden dan wanneer ze op papier staan. Immers, op het scherm wordt de tekst stukje voor stukje aangeboden; er wordt zodoende een natuurlijk moment voor commentaar gecreëerd. Het stillezen van een papieren tekst zal dus ook op gezette tijden onderbroken moeten worden, wil de proefpersoon in een ritme komen waarin hij lezen en commentaar leveren afwisselt. Maar ook op papier lijkt het goed mogelijk dit soort rustpunten in te bouwen, bijvoorbeeld door signalen in de tekst of, als het niet anders kan, door fragmenten van de tekst op verschillende pagina's aan te bieden.

Er zijn nog twee andere verklaringen denkbaar voor de relatief hoge productiviteit van het lezersprotocol. De eerste heeft te maken met de aard van de activiteit van de proefpersonen. In de hardop-leesmethode *rapporteren* proefpersonen over hun verwerkingsactiviteiten, terwijl de proefpersonen in de plus-en-minmethode de opdracht krijgen om de tekst te *beoordelen*, evenals de proefpersonen die met Focus werken. Het is denkbaar dat beoordelende proefpersonen niet alleen, of niet zozeer afgaan op hun eigen leeservaring, maar op verwachtingen ten aanzien van de leeservaring van anderen. In dat geval speelt de proefpersoon de rol van een soort expert. Dit effect kan verschillend uitwerken op het aantal probleemcommentaren. De beoordelaar kan verwachten dat anderen minder problemen hebben met de tekst, maar ook dat zij juist meer problemen hebben. Alleen in het eerste geval zou dit effect een verklaring kunnen zijn voor de hogere opbrengst van lezersprotocollen. Deze kwestie zou verder onderzocht moeten worden, bijvoorbeeld door te experimenteren met een aangepaste variant van de methode waarin de proefpersonen de opdracht krijgen om niet te rapporteren, maar zich te beperken tot evaluaties.

Een tweede alternatieve verklaring heeft te maken met het al of niet compleet verwerken van de tekst. Een aantal protocol-lezers had de neiging om passages over te slaan die zij om een of andere reden minder interessant achtten. Zij werden echter consequent geïnstrueerd om alles te blijven lezen. Dit soort correcties is uiteraard onmogelijk bij de plus-en-minmethode, en daarom blijft het denkbaar dat de lagere productiviteit van deze methode deels te wijten is aan het feit dat de tekst niet helemaal is gelezen.

Hoewel beide alternatieve verklaringen verder onderzoek verdienen, doen zij niets af aan de conclusie dat het uitsteleffect, of liever gezegd het onmiddellijkheids-effect, zeker een zelfstandige bijdrage levert aan de goede prestaties van het lezersprotocol als pretestmethode.

5. Slot

In de studies die hierboven zijn gerapporteerd, hebben we allereerst vastgesteld dat het lezersprotocol met name voor begripsproblemen beter presteert dan de plus-en-minmethode en Focus. In de eerste twee studies bleek daarnaast dat het lezersprotocol een goed uitvoerbare methode is voor proefleiders en proefpersonen. In de laatste studie hebben we laten zien dat de kwaliteit van de methode voor een belangrijk deel ligt in de onmiddellijkheid waarmee commentaar kan worden geleverd op een tekst. Dat bij lezersprotocollen teksten verklankt kunnen worden, beïnvloedt blijkens het experiment de opbrengst waarschijnlijk niet: immers ook een "stillees-protocol" levert goede resultaten op. Ook het tempo waarin de tekst verwerkt wordt, is irrelevant voor het aantal commentaren van de proefpersoon. Verder onderzoek is nodig naar twee andere factoren die mogelijk van belang zijn voor de kwaliteit van het lezersprotocol: de compleetheid waarmee de tekst wordt verwerkt en het feit dat het lezersprotocol meer een persoonlijk verslag van eigen leeservaringen vormt, terwijl bij de andere methoden meer sprake is van een expert-beoordeling.

Wat ook de resultaten van dit vervolgonderzoek zullen zijn, we mogen nu al concluderen dat het lezersprotocol niet alleen een veelbelovende methode is voor het testen van instructieve documenten aan de hand van een specifieke lezerstaak, maar ook goed kan worden gebruikt bij het evalueren van informatieve teksten. Met andere woorden: ook het lezersprotocol verdient een plaats in de standaardbagage van professionele tekstontwerpers.

Bibliografie

Allwood, C.M. & T. Kalén (1993). User-competence and other usability aspects when introducing a patient administrative system: a case study. *Interacting with computers* 2, 167-191

Boren, M.T. & J. Ramey (2000). Thinking aloud: reconciling theory and practice. *IEEE transactions on professional communication* 3, 261-278.

Dieli, M. (1986). *Designing successful documents: an investigation of document evaluation methods.* Pittsburgh: Carnegie-Mellon University.

Ericsson, K.A. (1988). Concurrent verbal reports on text comprehension: a review. *Text* 8, 295-325.

Flower, L., J.R. Hayes & H. Swarts (1983). Revising functional documents: the scenario principle. In: P.V. Anderson, P.J. Brockmann & C.R. Miller (eds), *New essays in technical and scientific communication. Research, theory and practice.* New York: Baywood, 41-58.

Jong, de, M. (1998). Reader feedback in text design. Validity of the plus-minus method for the pretesting of public information brochures. Proefschrift Universiteit Twente, Amsterdam: Rodopi.

Jong, de, M. & P.J. Schellens (1995). *Met het oog op de lezer. Pretestmethoden voor schriftelijk voorlichtingsmateriaal.* Amsterdam: Thesis.

Jong, de, M. & P.J. Schellens (2000). Toward a document evaluation methodology: what does research tell us about the validity and reliability of evaluation methods? *IEEE transactions on professional communication* 3, 242-260.

Jong, de, M. & P.J. Schellens (2002). Tekstevaluatie. Onderzoek naar de validiteit van probleemopsporende methoden. *Tijdschrift voor Taalbeheersing* 24, nr. 2, 146-166.

Jong, de, M. & L. Lentz (2001). Focus: design and evaluation of a software tool for collecting reader feedback. *Technical Communication Quarterly* 4, 289-403.

Koppenaal, B. (2000). De hardop-leesmethode. De plus-en-minmethode & de vragenlijst. Doctoraalscriptie Universiteit Utrecht, specialisatie Communicatiekunde.

Noorlander, M. (2001). Pretesten met bijsluiterteksten. Focus & de Hardop-leesmethode. Doctoraalscriptie Universiteit Utrecht, specialisatie Communicatiekunde.

Pander Maat, H. (1996). Identifying and predicting reader problems in drug information texts. In: T. Ensink & C. Sauer (eds.), *Researching technical documents.* Groningen: Department of speech and communication Rijksuniversiteit Groningen. 17-47.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin* 124, 372-422.

Schriver, K.A. (1989). Evaluating text quality: the continuum from text-focused to reader-focused methods. *IEEE transactions on professional communication* 4, 238-255.

Schriver, K.A. (1991). Plain language through protocol-aided revision. In: E.R. Steinberg (ed.). *Plain language: principles and practice.* Detroit, Michigan: Wayne State University Press, 148-172.

Sienot, M. (1997). Pretesting websites. A comparison between the plus-minus method and the think-aloud method for the World Wide Web. *Journal of business and technical communication* 11, 469-482

Swaney, J.H., C.J. Janik, S.J. Bond & J.R. Hayes (1991). Editing for comprehension: improving the process through reading protocols. In: E.R. Steinberg (ed.), *Plain language: principles and practice.* Detroit, Michigan: Wayne State University Press, 173-203.

Vromen, N., (1998). Focus. Een evaluatie-onderzoek naar het softwareprogramma Focus, waarmee teksten in pretestsituaties beoordeeld kunnen worden. Doctoraalscriptie Universiteit Utrecht, specialisatie Communicatiekunde.

Werven, van P. (2002). Stillezen bij de hardoplees-methode. Onderzoek naar factoren die de kwaliteit van de hardoplees-methode bepalen. Doctoraalscriptie Universiteit Utrecht, specialisatie Communicatiekunde.

SANNE ELLING EN LEO LENTZ

De voorspellende kracht van het CCC-model

Samenvatting

Onderzoek naar methoden voor tekstevaluatie laat zien dat experts slecht in staat zijn om te voorspellen welke problemen lezers uit de doelgroep in een tekst hebben. Het CCC-model zou door het geven van richtlijnen de prestaties van experts kunnen verbeteren. In dit artikel wordt de waarde van het CCC-model onderzocht, met op de achtergrond de vraag in hoeverre het CCC-model een alternatief vormt voor lezergerichte evaluatiemethoden. Aan de orde komen de betrouwbaarheid en de validiteit van het model en de toegevoegde waarde van het CCC-model ten opzichte van een expertevaluatie zonder richtlijnen.

1. Inleiding

In de literatuur over tekstevaluatie speelt al jarenlang de vraag of er een alternatief is voor de tijdrovende en kostbare methoden waarbij teksten worden voorgelegd aan lezers uit de doelgroep. Daarbij wordt dan vooral gedacht aan experts die met behulp van een of ander instrument een tekst evalueren volgens vooraf omschreven criteria, in de hoop dat daarmee problemen zichtbaar worden die lezers met de tekst zouden ervaren. In Nederland is door Renkema (1996) het CCC-model gepresenteerd als een handzaam instrument voor tekstevaluatie. De precieze ambitie van het model is echter niet erg scherp gedefinieerd.

In Renkema en Wijnstekers (1997) wordt het CCC-model een tekstgerichte pretestmethode genoemd, hetgeen betekent dat het model als volwaardige evaluatiemethode kan worden ingezet. Elders formuleert Renkema echter een veel bescheidener ambitie: 'Het CCC-model is allereerst bedoeld om commentaren op een tekst te systematiseren' (Renkema, 1996, p.336). Daarmee is de ambitie teruggeschroefd naar een soort ordeningssysteem van commentaren; voor het genereren van die commentaren lijkt het model in die visie niet bruikbaar te zijn. In een op zijn oratie gebaseerde publicatie benadrukt Renkema dezelfde bescheiden ambitie: het model dient als raster om verschillende commentaren te categoriseren; 'het CCC-model is bedoeld om alle opmerkingen over kwaliteit in verhouding te kunnen zien' (Renkema, 2000, p.250). Toch vinden we in Renkema (1996) ook onderzoeksvragen die op een veel verder reikende ambitie wijzen, zoals:

- Leidt toepassing van het CCC-model tot een goede inschatting van lezersproblemen?
- Leidt tekstrevisie op basis van een CCC-analyse tot betere teksten?

Uit deze vragen blijkt wel degelijk een ambitie om met het model ook teksten te evalueren teneinde lezersproblemen te voorkomen. In dit artikel staat die ambitie centraal. Het

doel van ons onderzoek was een antwoord te formuleren op beide bovengenoemde vragen. Een belangrijk punt is in dit verband de vraag wanneer er sprake is van *een goede inschatting van lezersproblemen*. Ons eerste uitgangspunt is dat dat het geval is als experts met behulp van het model in voldoende mate overeenkomen in hun probleemdetecties. In eerder onderzoek van Lentz en De Jong (1997) bleek dat experts zonder enig model grotendeels zogenaamde unieke detecties produceerden. Ruim zeventig procent van de voorspellingen van elke individuele expert werd door geen enkele andere expert gedeeld. Als het CCC-model op dit punt tot betere scores leidt, dan is inderdaad sprake van een verbetering.
Ons tweede uitgangspunt is dat er een relatie moet zijn tussen die probleemdetecties en de problemen die lezers daadwerkelijk ervaren met de betreffende tekst. Dat betekent dat er een vergelijking gemaakt moet worden tussen enerzijds de output van experts die met het CCC-model werken en anderzijds de output van evaluatie-onderzoek waarbij lezers feedback hebben gegeven op een tekst.
Ten slotte is een derde uitgangspunt dat het model ten aanzien van betrouwbaarheid en validiteit een toegevoegde waarde moet hebben in vergelijking met experts die zonder richtlijnen een tekst beoordelen. Dat betekent dat er twee vergelijkbare groepen experts met dezelfde tekst moeten werken: de ene groep is getraind in het CCC-model, de andere groep heeft geen kennis van dat model.

Er is al menig onderzoek gedaan naar de vraag in hoeverre experts lezersproblemen kunnen voorspellen. Soms met erg teleurstellende resultaten voor de experts. De Jong en Lentz (1996), Pander Maat (1996) en Lentz en De Jong (1997) concluderen dat experts niet meer dan 15% van de lezersproblemen voorspellen. Bovendien blijkt de onderlinge overeenstemming tussen de experts zeer gering: in beide experimenten van De Jong en Lentz was ruim 70% van de voorspelde problemen een zogenaamde *unieke detectie*, een voorspelling dus die met geen enkele andere expert gedeeld werd. Maar er zijn ook andere resultaten gevonden. Dieli (1986) en Nielsen (1994) kwamen tot een score van 80% voorspelde problemen, Renkema en Wijnstekers (1997) vinden een voorspellingspercentage van 63% tot 88% en Lentz en Pander Maat (1992) kwamen zelfs tot een 100% score. Hoe vallen die verschillen te verklaren?

Er zijn drie factoren die in dit verband besproken moeten worden. Ten eerste zijn er studies waarbij de experts een instrument hanteren bij het beoordelen van de teksten naast studies waarbij de experts zonder richtlijn of heuristiek werken. Ten tweede is van belang hoe precies vastgesteld wordt of er sprake is van een correcte voorspelling. Op beide punten verschillen de studies sterk van elkaar. Ten derde verschillen de studies in de analysemethode: de een kijkt naar de prestaties van elke individuele expert, de ander kijkt naar de prestaties van de groep als geheel.

In de eerstgenoemde studies met de lage scores konden de experts geen gebruik maken van een of ander beoordelingsinstrument, terwijl in de studies met de hoge scores daartoe wel de gelegenheid was. Dat lijkt te wijzen op de kracht van zulke heuristieken, alhoewel in die tweede groep sprake is van zeer verschillende instrumenten, variërend van eenvoudige checklists met aandachtspunten voor de begrijpelijkheid van allerlei soorten documenten tot en met zeer ver uitgewerkte genrespecifieke beoordelingsinstrumenten. In feite zijn die studies daardoor onvergelijkbaar.

In de tweede plaats verschilt de methodiek sterk in de diverse studies. Zo beschouwde Dieli een probleemdetectie als een hit wanneer de expert een passage had onderstreept die door lezers daadwerkelijk als problematisch was ervaren. Genegeerd werd op die manier of het door de expert vermoede probleem enigszins overeen kwam met het lezersprobleem. Renkema en Wijnstekers (1997) vergeleken niet de eigenlijke probleemdetecties van beide methoden met elkaar, maar de oordelen over revisies die op grond van die problemen gemaakt werden. In de overige studies zijn uitspraken over de betrouwbaarheid en validiteit gebaseerd op de overeenkomst tussen probleemdetecties die de experts formuleren en de feedback uit het lezersonderzoek.

In de derde plaats maakt het een groot verschil of de resultaten van een expert-evaluatie gerapporteerd worden op groepsniveau (zoals bijvoorbeeld Dieli heeft gedaan) of op individueel niveau (zoals bijvoorbeeld Lentz en De Jong deden). Het spreekt immers vanzelf dat de scores van de groep als geheel hoger zullen zijn dan die van het gemiddelde individu.

Deze verscheidenheid aan gehanteerde methoden maakt een generaliserende conclusie over het nut van heuristieken twijfelachtig.

Het CCC-model (zie figuur 1) bestaat in feite uit een geordende reeks criteria voor tekstkwaliteit. Het model is niet gebonden aan een specifiek genre. Het belangrijkste criterium voor tekstkwaliteit is de balans tussen zender en ontvanger (correspondentie): heeft de schrijver overeenstemming weten te vinden tussen zijn eigen doelen en datgene wat de lezer verwacht of nodig heeft? De andere twee criteria zijn consistentie, het vasthouden aan eenmaal gemaakte keuzes, en correctheid, het vasthouden aan de algemene regels voor taalgebruik. Deze criteria worden toegepast op vijf niveaus: teksttype, inhoud, opbouw, formulering en presentatie. Op deze manier ontstaan er 15 ijkpunten aan de hand waarvan een tekst beoordeeld kan worden.

Tekstniveau\ Criteria	**Correspondentie**	**Consistentie**	**Correctheid**
A Teksttype	1. geschiktheid	2. genrezuiverheid	3. toepassing genreregels
B Inhoud	4. voldoende informatie	5. overeenstemming tussen feiten	6. juistheid van gegevens
C Opbouw	7. voldoende samenhang	8. consequente opbouw	9. correcte verbindingswoorden
D Formulering	10. gepaste formulering	11. eenheid van stijl	12. correcte zinsbouw en woordkeus
E Presentatie	13. gepaste toon	14. afstemming tekst en vormgeving	15. correcte spelling en interpunctie

Figuur 1. *CCC-model volgens Renkema (1996)*

Het model is in een aantal publicaties onderworpen aan onderzoek. In Renkema en Wijnstekers (1997) is de effectiviteit van het model aan de orde: er vindt een vergelijking plaats tussen de lezergerichte plus-en-minmethode enerzijds en de tekstgerichte CCC-analyse anderzijds. Hierbij is gekeken naar de revisievoorstellen die op basis van het onderzoek met de verschillende methoden werden gedaan. Uit het onderzoek blijkt dat de CCC-methode meer revisies oplevert dan de plus-en-minmethode. Steekproefsgewijs is onderzocht in hoeverre de commentaren uit de beide methoden overeen kwamen. Uit de eerste steek-

proef kwam naar voren dat de CCC-experts 88% van de lezersproblemen voorspelden, uit de tweede steekproef kwam een correct voorspeld percentage van 63% naar voren. De auteurs presenteren deze bevindingen overigens als een soort terzijde en merken op dat verder onderzoek wenselijk is.

In Renkema (2000) wordt een onderzoek gerapporteerd waarin professionele tekstschrijvers de revisies hebben beoordeeld die op basis van een CCC-evaluatie en een plus-en-minevaluatie gemaakt waren. De CCC-revisies bleken beter beoordeeld te worden. Een belangrijke kanttekening hierbij is wel dat er bij de revisies met de plus-en-minmethode gebruik is gemaakt van de feedback van slechts vier proefpersonen. Een andere kanttekening, die ook opgaat voor Renkema en Wijnstekers (1997), is dat er niet is gekeken naar de eigenlijke probleemdetecties, maar naar de daarop gebaseerde revisies.
In ons onderzoek zijn de resultaten van een evaluatie met behulp van het CCC-model vergeleken met grootschaliger lezersonderzoek, bovendien is die vergelijking niet gebaseerd op de van die resultaten afgeleide revisies, maar op de oorspronkelijke probleemdetecties. We formuleerden vier onderzoeksvragen.

1. In hoeverre is het CCC-model betrouwbaar?
Deze vraag is onderzocht door experts een tekst te laten beoordelen aan de hand van het CCC-model. Vervolgens is vastgesteld wat de onderlinge overeenstemming is in probleemdetecties: de *interbeoordelaarsbetrouwbaarheid*.

2. In hoeverre is het CCC-model predictief valide?
Om deze vraag te beantwoorden, is onderzocht in hoeverre experts in staat zijn om met het CCC-model de problemen van lezers te voorspellen. Tevens is onderzocht in welk opzicht expertproblemen en lezersproblemen van elkaar verschillen in termen van categorieën uit het CCC-model. In deze vergelijking hebben we gebruik gemaakt van de resultaten van twee methoden van lezersonderzoek die eerder door anderen (met verschillende teksten) zijn uitgevoerd (De Jong en Schellens, 1996 en Vromen, 1998).

3. Wat is de waarde van de overige detecties?
Hierbij gaat het om de problemen die wel door de experts zijn genoemd, maar die geen lezersprobleem bleken te zijn. Het is immers denkbaar dat deze probleemdetecties wel tot revisies zouden kunnen leiden waar de lezers uit de doelgroep baat bij hebben. Deze vraag is onderzocht door vier deskundigen (een deel van) de overige detecties te laten beoordelen op de aannemelijkheid van het probleem, de ernst ervan en op de vraag of het probleem wel of niet aanleiding geeft tot een revisie.

4. Wat is bij een expertonderzoek de toegevoegde waarde van het CCC-model?
Deze vraag is beantwoord door de resultaten van een expertevaluatie zonder CCC-model te vergelijken met een evaluatie waar het model wel gebruikt is. Er is een vergelijking gemaakt op de aspecten die in de eerste drie vragen aan de orde zijn, namelijk:
- betrouwbaarheid
- predictieve validiteit
- waarde overige detecties

2. Opzet van het onderzoek

De resultaten van twee expert-evaluaties (met en zonder CCC-model) zijn vergeleken met de resultaten van twee verschillende, eerder uitgevoerde, lezersonderzoeken. De daarbij gebruikte teksten, de groepen proefpersonen en de opzet van de onderzoeken komen in deze paragraaf aan de orde.

Materiaal

De eerste tekst die in het onderzoek is gebruikt, is de brochure 'Je eerste baan'. Deze brochure is in 1995 door de Belastingdienst uitgegeven. De tekst is gericht op jongeren tussen de 18 en 26 jaar die voor het eerst een baan krijgen. De tweede tekst is de bijsluiter van Cinnarizine. Dit is een geneesmiddel dat gebruikt wordt bij duizeligheid als gevolg van een stoornis in het evenwichtsorgaan, bij reisziekte en bij allergische aandoeningen. De tekst is bedoeld voor iedereen die het middel gaat gebruiken: dit zijn mensen van verschillende leeftijden, verschillend geslacht en verschillende opleidingsniveaus.

Pretestmethoden

Het lezersonderzoek naar de belastingtekst is een paar jaar terug met twee verschillende methoden uitgevoerd. Het onderzoek met de *plus-en-minmethode* is uitgevoerd door De Jong en Schellens (1996). Het onderzoek met *Focus* is uitgevoerd door Vromen (1998).

De bijsluiter is eerder beoordeeld met zowel de *hardop-leesmethode* als met *Focus* (Noorlander, 2001). Bovendien is achteraf een *begripstoets* afgenomen.

Proefpersonen

De CCC-beoordelingen zijn uitgevoerd door studenten Communicatiekunde aan de Universiteit Utrecht die in een vergevorderd stadium van hun studie waren of al waren afgestudeerd. De beoordeling van de brochure van de Belastingdienst is uitgevoerd door tien experts. De beoordeling van de bijsluiter is uitgevoerd door 18 experts met het CCC-model en door 14 experts zonder model; dit gebeurde in het kader van de cursus 'Instructieve Documenten' (onderdeel van de specialisatie). In deze cursus wordt ingegaan op diverse soorten instructieve documenten, waaronder bijsluiters, en het ontwerpen en evalueren hiervan.

De proefpersonen die mee hebben gewerkt aan de prestests kwamen uit de doelgroep van de beide teksten. Aan het onderzoek met de plus-en-minmethode naar de belastingtekst hebben 30 mensen uit de doelgroep deelgenomen met verschillende opleidingsniveaus. Het onderzoek naar dezelfde tekst met *Focus* is uitgevoerd met 21 hoger opgeleide mensen. In totaal hebben dus 51 proefpersonen uit de doelgroep feedback gegeven op de belastingtekst. De bijsluiter is beoordeeld door in totaal 30 lezers uit de doelgroep, bestaande uit lezers van diverse leeftijden en diverse opleidingsniveaus. De helft hiervan heeft de tekst beoordeeld met de *hardop-leesmethode*, de andere helft met *Focus* (Noorlander, 2001). De proefpersonen hadden geen voorkennis over het onderwerp van de bijsluiter.

De beoordelaars van de waarde van de overige detecties (geproduceerd door proefpersonen met en zonder CCC-model) waren vier medewerkers van de afdeling Taalbeheersing van de Universiteit Utrecht.

Procedure

Hieronder besteden we met name aandacht aan de procedure van het CCC-onderzoek. Voor nadere informatie over de procedures die gehanteerd zijn bij de prestest verwijzen wij naar De Jong en Schellens (1996), Vromen (1998) en Noorlander (2001). Elders in dit nummer (Pander Maat en Lentz, 2003) worden de plus-en-minmethode en Focus meer in detail toegelicht.

De experts die met het CCC-model werkten, hebben eerst een uitgebreide cursus over de werking van het CCC-model gehad waarin de vijftien punten van het model uitvoerig zijn besproken en toegelicht aan de hand van voorbeelden. Hier is ruim een uur voor uitgetrokken. Vervolgens hebben de experts het model toegepast op een oefentekst. Nadat iedereen het model goed onder de knie meende te hebben, heeft men individueel de tekst beoordeeld. De opdracht die men hierbij kreeg, was dat men zo goed mogelijk moest aangeven welke problemen lezers uit de doelgroep met de tekst zouden ervaren. Er werd expliciet gezegd dat ze zoveel mogelijk lezersproblemen uit de tekst moesten halen en zo min mogelijk problemen moesten noemen die lezers níet zouden hebben. Om de studenten extra te prikkelen werd bij de beoordeling van de bijsluiter een prijs uitgeloofd voor degene die de beste voorspeller zou blijken te zijn. Daarbij werden punten bijgeteld voor elke treffer en punten afgetrokken voor elke overige detectie.

De groep zonder model kreeg dezelfde opdracht om de lezersproblemen zo goed mogelijk te voorspellen; ook hier kon men een prijs winnen. Deze groep werd verder echter helemaal vrij gelaten in de manier van beoordelen en moest dit dus zonder richtlijnen doen.

In de CCC-groepen varieerde de tijd die men voor de beoordeling nodig had van een half uur tot ruim een uur. In de groep zonder CCC-model was men iets korter bezig met de beoordeling, namelijk tussen de 25 minuten en 45 minuten.

Voor de beoordeling van de waarde van de overige detecties die met de expert-evaluaties (met en zonder CCC-model) van de bijsluiter geproduceerd werden, is een selectie gemaakt van 34 problemen. Dat was nodig omdat de volledige lijst met overige detecties veel te lang was om voor te leggen aan beoordelaars. De selectie is gemaakt door allereerst uitsluitend problemen op te nemen die door minimaal twee experts (van de 32) zijn genoemd. Aldus ontstond een lijst met 68 problemen. Hierbij is er vanuit gegaan dat de problemen die slechts door één persoon zijn genoemd, zo onaannemelijk zijn dat ze minder kans hebben om hoog te scoren. Van deze set is vervolgens de helft voorgelegd aan vier deskundigen, allen werkzaam bij de afdeling Taalbeheersing in Utrecht. Deze deskundigen hebben bij elk probleem een uitspraak gedaan over de aannemelijkheid (op een vijfpuntsschaal van onaannemelijk naar aannemelijk), over de ernst (op een vijfpuntsschaal van niet ernstig naar zeer ernstig) en over de vraag 'wel of niet reviseren?'.

3. Resultaten

In deze paragraaf vatten we de resultaten van de verschillende deelonderzoeken samen per onderzoeksvraag en dus niet per deelonderzoek, teneinde herhaling te voorkomen. Eerst bespreken we de vraag naar de betrouwbaarheid, daarna bespreken we de validiteit. In dit deel van het onderzoek speelt de belastingtekst de hoofdrol. Voor een analyse van de toegevoegde waarde van het CCC-model ten opzichte van een expert-beoordeling zonder

model is gebruik gemaakt van de resultaten van de bijsluitertekst. Die resultaten lagen ook ten grondslag aan de analyse van de waarde van de overige detecties.

In hoeverre is het CCC-model betrouwbaar?
De betrouwbaarheid is gemeten met een analyse van de onderlinge overeenstemming in de problemen die de experts vinden in de belastingtekst. Voor 83.1% van alle probleemdetecties is er geen enkele overeenstemming. Tabel 1 geeft een overzicht van de scores.

Tabel 1. *Overeenstemming tussen experts in de beoordeling van de Belastingfolder*

Overeenstemming tussen: *n=10*	**Aantal problemen:**	**Percentage**
geen overeenstemming (unieke detectie)	103	83,1%
2 experts	17	13,7%
3 experts	2	1,6%
4 experts	0	-
5 experts	2	1,6%
6 t/m 10 experts	0	-
Totaal	124	100%

De uitkomst van de beoordeling met het CCC-model is, zo maakt tabel 1 duidelijk, in hoge mate afhankelijk van de toevallige beoordelaar. De kans dat twee onafhankelijke experts tot een zelfde reeks probleemdetecties komen lijkt minimaal te zijn. Dit resultaat is allerminst beter dan in eerder onderzoek van Lentz en De Jong (1997) is gerapporteerd, waar experts geen beschikking hadden over een beoordelingsmodel.

In hoeverre is het CCC-model predictief valide?
De predictieve validiteit is gemeten door de voorspellingen van experts (als groep) te vergelijken met de problemen die lezers daadwerkelijk met de tekst ervaren. Met behulp van de resultaten van de belastingtekst is gekeken in hoeverre er overlap is tussen de problemen die zijn gevonden door de experts met het CCC-model en:
- lezersproblemen gevonden met *plus-en-minmethode*
- lezersproblemen gevonden met *Focus*
- een set geselecteerde belangrijke lezersproblemen.

Van de 130 lezersproblemen die zijn gevonden met de plus-en-minmethode, zijn er 16 (12,3%) met het CCC-model gevonden. Van de 206 lezersproblemen die met Focus zijn gevonden, zijn er 39 (18,9%) met het CCC-model gevonden. Op individueel niveau lagen de scores uiteraard een stuk lager: ten opzichte van de plus-en-minmethode scoorde elke expert gemiddeld 2.7 hits (sd. 1.64); ten opzichte van Focus scoorde elke expert gemiddeld 5.7 hits (sd. 2.71).

Onze ervaring in eerder onderzoek is dat de hits nogal eens verloren gaan in een grote hoeveelheid false alarms. In tabel 2 staan de aantallen hits en false alarms ten opzichte van de lezersproblemen met de plus-en-minmethode en ten opzichte van de lezersproblemen gevonden met Focus.

Tabel 2. *Hits en false alarms t.o.v. plus-en-minmethode en Focus toegepast op belastingtekst*

	CCC-model t.o.v. +/- *Aant. problemen & percentage* *n=10*	**CCC-model t.o.v. Focus** *Aant. problemen & percentage* *n=10*
Aantal hits	16 (12,9%)	39 (31,5%)
Aantal false alarms	108 (87,1%)	85 (68,5%)
Totaal	124	124

Tabel 2 laat zien dat deze verhouding het meest gunstig is in de vergelijking met de Focus-methode. Daar is een op de drie met het CCC-model voorspelde problemen een hit.

Het is denkbaar dat er in de reeks lezersproblemen feedback zit die moeilijk voorspelbaar is, bijvoorbeeld omdat slechts één lezer (van de in totaal 51 lezers) dat commentaar heeft gegeven. Het is daarom interessant om naar problemen te kijken die zwaarder wegen, omdat ze vaker genoemd zijn. De set met de meest aannemelijke lezersproblemen vormt een selectie van de problemen die (met de belastingtekst) zowel bij de plus-en-minmethode als bij Focus zijn genoemd. Dit zijn 20 problemen. Van deze set is 40% wel en 60% niet gevonden met het CCC-model. De prestatie van de groep experts stijgt dus als we de lat wat hoger leggen en alleen naar de zwaarwegende problemen kijken. Deze scores liggen echter bepaald niet hoger dan de scores die in Lentz en De Jong (1997) gerapporteerd worden; daar behaalden twee groepen van tien experts (zonder beoordelingsmodel) scores tussen de 60 en 70 procent van een reeks zwaarwegende problemen die met de plus-en-minmethode gevonden waren.

Zijn er verschillen tussen experts en lezers in termen van ijkpunten uit het CCC-model?
In hoeverre hebben de experts een andere definitie van tekstkwaliteit dan de lezers uit de doelgroep? Zijn zij misschien gericht op slechts een bepaald soort problemen? In tabel 3 zijn de ijkpunten van het CCC-model opgenomen die het grootste deel van de gevonden problemen bevatten. De rangorde van de drie belangrijkste categorieën staat tussen haakjes achter de percentages.

Tabel 3. *Spreiding over ijkpunten CCC-model toegepast op de belastingtekst*

IJkpunt CCC-model	**Percentage CCC-problemen**	**Percentage +/-problemen**	**Percentage Focus-problemen**
Voldoende informatie	27,4% (1)	49,2% (1)	30,1% (2)
Juistheid van gegevens	4,8%	13,8% (3)	1,9%
Voldoende samenhang	10,5% (3)	4,6%	2,9%
Gepaste formulering	21% (2)	26,2% (2)	37,4% (1)
Eenheid van stijl	5,6%	0,8%	1,9%
Correcte zins-bouw & woordkeus	7,3%	0,8%	9,7%
Correcte spelling & interpunctie	8,1%	-	11,7% (3)

Zowel de experts met het CCC-model als de lezers uit de doelgroep noemen in de belastingtekst veel problemen die onder de ijkpunten *voldoende informatie* en *gepaste formulering* vallen. Van de expertproblemen valt bijna de helft hieronder, van de plus-minproblemen en

de focusproblemen rond de driekwart. Deze ijkpunten, en ook het ijkpunt *voldoende samenhang* (ijkpunt 7), vallen onder het criterium correspondentie. Men gaat bij het aangeven van problemen blijkbaar vooral uit van afstemming tussen tekst en lezer. Het criterium consistentie is zowel bij experts als bij lezers nauwelijks aan de orde. Problemen in de categorie correctheid worden met de plus-minmethode nauwelijks genoemd, met Focus en het CCC-model worden op dit criterium vooral op de niveaus formulering en presentatie relatief veel problemen genoemd.

Qua spreiding over de verschillende ijkpunten uit het model valt op dat deze het grootst is bij de problemen die door de experts zijn genoemd. Slechts twee ijkpunten (op het niveau teksttype) worden geen enkele keer door experts genoemd en er is één ijkpunt (afstemming tekst en vormgeving) waar slechts één expertprobleem in valt. Bij Focus is de spreiding wat minder groot, hier zijn drie ijkpunten waar geen enkel probleem in wordt genoemd en twee ijkpunten waarin één probleem wordt genoemd. Met name het niveau teksttype scoort matig (hierin valt in totaal slechts één probleem). Bij de plus-en-minmethode is de spreiding het laagst, er zijn zes ijkpunten waarin geen enkel probleem valt en drie ijkpunten waarin één probleem valt. Met name op de niveaus teksttype en presentatie worden niet of nauwelijks problemen gerapporteerd.

Dat bij de experts de spreiding in ijkpunten het grootst is, is geen verrassing. Deze beoordelaars hebben immers een uitvoerige cursus gehad over het CCC-model en hebben met alle ijkpunten geoefend. Het ligt voor de hand dat zij dan ook problemen vinden op verschillende ijkpunten. De lezers die de tekst beoordeeld hebben met Focus zijn ook gestuurd. Zij moeten bij het benoemen van een probleem, kiezen in welke categorie dit probleem valt. In dit onderzoek konden ze een keuze maken uit:

- taalfout
- punten en komma's
- begrijp ik niet
- geloof ik niet
- verkeerde volgorde
- overbodige informatie
- ik mis iets
- formulering is niet goed.

Voor een deel komen deze categorieën overeen met die van het CCC-model. Het is dan ook te verklaren dat de spreiding bij de beoordelaars met Focus wat kleiner is dan bij het CCC-model en groter is dan bij de plus-en-minmethode, waarbij men in de instructie slechts globale aanwijzingen kreeg over redenen waarom men plussen of minnen zou kunnen zetten.

Een andere verklaring voor het verschil in spreiding is dat er bij de drie methoden een ander perspectief wordt ingenomen. De CCC-experts hebben de rol van beoordelaar, zij moeten als expert voor anderen bepalen wat er wel en niet goed is aan de tekst. De plus-minlezers zijn mensen uit de doelgroep en geven dus commentaar op de tekst vanuit hun eigen perspectief als 'gebruiker'. Dit verklaart waarom zij vrijwel alleen maar problemen noemen uit de categorie correspondentie. De Focuslezers zitten tussen deze twee uitersten

in. Het zijn lezers uit de doelgroep, maar met name de hoger opgeleide lezers. Bovendien zitten Focuslezers (meer dan plus-en-minlezers) in de rol van beoordelaar omdat ze tegelijk moeten lezen en beoordelen (De Jong & Lentz, 2001).

Wat is de toegevoegde waarde van het CCC-model?
Zijn de oordelen die met een CCC-model tot stand komen meer betrouwbaar dan die welke zonder dat model geproduceerd worden? In tabel 4 staan de resultaten van een analyse van de onderlinge overeenstemming tussen de experts, die met of zonder model een beoordeling van de bijsluiter uitvoerden.

Tabel 4. *Overeenstemming tussen experts in de beoordeling van de bijsluiter*

Overeenstemming tussen:	***Met CCC-model*** *aant. problemen & percentage n=18*	***Zonder CCC-model*** *aant. problemen & percentage n=14*
geen overeenstemming (unieke detectie)	84 (64,6%)	66 (64,7%)
2 experts	23 (17,7%)	27 (26,5%)
3 experts	9 (6,9%)	3 (2,9%)
4 experts	8 (6,2%)	4 (3,9%)
5 experts	4 (3,1%)	2 (2%)
6 experts	0 -	
7 experts	1 (0,8%)	-
8 experts	1 (0,8%)	-
Totaal	130	102

Het gebrek aan overeenstemming is in beide condities vrijwel gelijk. Bijna 65% van de problemen is in beide groepen een unieke detectie. Verder wordt zowel met CCC-model als zonder model geen enkel probleem door meer dan de helft van de experts genoemd. Het gebruik van het CCC-model leidt dus niet tot meer overeenstemming tussen beoordelaars.

Leidt het dan wel tot een betere voorspelling van lezersproblemen? Voor een antwoord op die vraag hanteren we als benchmark de problemen die in het onderzoek van Noorlander (2001) naar de bijsluiter als reële problemen zijn gedefinieerd, hetgeen betekent dat er zwaarwegende argumenten zijn om te veronderstellen dat die problemen zich daadwerkelijk bij de lezers voordoen. Het eerste argument is dat deze problemen bevestigd zijn in een op de pretest aansluitende tekstbegripstoets (doordat men inderdaad het goede antwoord over de betreffende passage niet wist te geven). Dit argument leidde tot een selectie van 17 begripsproblemen. Maar niet elk probleem kan met een begripstoets gevalideerd worden; waarderingsproblemen of bijvoorbeeld acceptatieproblemen vragen om een andere manier van valideren. Daarom is er een tweede criterium opgesteld, namelijk dat ook problemen die door minimaal zes personen uit de doelgroep genoemd worden (bijna 25% van de proefpersonen) beschouwd worden als een reëel probleem. Dit leidde tot elf extra problemen. Aldus komen we tot een set van 28 reële lezersproblemen. De resultaten van beide groepen experts staan weergegeven in tabel 5.

Van deze set met 28 lezersproblemen worden er 20 gevonden door de groep CCC-beoordelaars, en 16 door de andere groep. Dit verschil is het gevolg van de grotere groep CCC-beoordelaars, hetgeen zichtbaar wordt als we kijken naar de gemiddelde score per expert.

Tabel 5. *Hoeveelheid hits van beoordelaars met en zonder CCC-model toegepast op bijsluiter*

	Met CCC-model *n = 18*	**Zonder CCC-model** *n = 14*
Aantal hits van de gehele groep t.o.v. de 28 reële lezersproblemen	20	16
Gemiddeld aantal hits per expert t.o.v. de 28 reële lezersproblemen	2.1 (sd. 1.5)	2.2 (sd .1.7)
Gemiddeld aantal hits t.o.v. totaal aantal detecties (per expert)	17.7% (sd. 0.13)	21.3 % (sd. 0.15)

Die is in beide groepen vrijwel gelijk. Het CCC-model lijkt ook op dit punt geen toegevoegde waarde te hebben. Als we naar de relatieve scores kijken lijken de experts zonder CCC-model iets zuiniger gewerkt te hebben, aangezien de verhouding tussen het aantal hits en overige detecties bij hen iets gunstiger is, maar dit verschil is niet significant. Wel merken we op dat het goed denkbaar is dat in die overige probleemdetecties wel degelijk problemen zitten die eerder door een lezer genoemd zijn. We hebben voor deze vergelijking immers een selectie gemaakt van die problemen waarvan we op goede gronden kunnen stellen dat het reële lezersproblemen zijn. We hebben om die reden de overige detecties niet als false alarm bestempeld.

In hoeverre is er een verschil in perspectief tussen experts met CCC-model en experts zonder model? Verschillen de twee groepen qua soorten problemen die ze noemen? In tabel 6 zijn de ijkpunten opgenomen waaronder de vier hoogste percentages problemen uit beide groepen vallen. Tussen haakjes is de rangorde van deze categorieën aangegeven.

Tabel 6. *Problemen uit vier meest genoemde ijkpunten CCC-model toegepast op de bijsluiter: absoluut, percentage, rangorde*

IJkpunt CCC-model	**Experts met CCC-model**			**Experts zonder CCC**		
	Abs.	*Perc.*	*Rang*	*Abs.*	*Perc.*	*Rang*
Voldoende informatie	25	19,2%	(2)	30	29,4%	(1)
Overeenstemming tussen feiten	13	10%	(4)	6	5,9%	(4)
Voldoende samenhang	18	13,8%	(3)	20	19,6%	(3)
Gepaste formulering	29	22,3%	(1)	25	24,5%	(2)
Overige ijkpunten	45			21		
Totaal	130			102		

In de tabel is te zien dat er in de rangorde van categorieën tussen experts met – en zonder model nauwelijks verschillen zijn, alleen de nummers 1 en 2 wijken van elkaar af. Net als in het onderzoek naar de folder van de Belastingdienst vallen de grootste groepen problemen onder ijkpunt 4 en ijkpunt 10. De verschillen over de gehele top vier zijn niet significant.

Ook qua spreiding zijn er nauwelijks verschillen tussen de twee groepen. In beide condities worden er op het eerste niveau (teksttype) niet of nauwelijks problemen genoemd. Alle andere ijkpunten zijn wel vertegenwoordigd in de genoemde problemen door beide groepen experts.

Wat is de waarde van de overige detecties? Wellicht noemen experts problemen die geen hit zijn, maar hebben lezers wel degelijk baat bij het oplossen van deze problemen. We onderzoeken daarom wat volgens deskundigen de kwaliteit van de overige detecties is. Van deze detecties (geproduceerd bij de beoordeling van de bijsluiter) is eerst een selectie gemaakt van problemen die door twee of meer van de 32 experts zijn genoemd. De motivering voor deze selectie hebben we in paragraaf 2 beschreven. De deskundigen, vier medewerkers van de afdeling Taalbeheersing uit Utrecht, hebben deze set van 34 problemen van een score voorzien op de variabelen aannemelijkheid, ernst, en wel of geen revisie. Een analyse van de betrouwbaarheid van deze oordelen leidde tot een matige betrouwbaarheid voor de aannemelijkheid (alpha .52), na verwijdering van één beoordelaar steeg de betrouwbaarheid tot .64. Er was een redelijke betrouwbaarheid voor het oordeel over ernst en de wenselijkheid van een revisie (beide een alpha van .63). Er was een hoge correlatie tussen deze drie variabelen, die na correctie voor de onbetrouwbaarheid tot een nieuwe variabele *eindoordeel* herleid konden worden; deze variabele heeft een minimumwaarde van 1 en een maximum van 10.

Van de 34 overige detecties zijn er 10 genoemd door alleen CCC-experts, 10 genoemd door alleen experts zonder model en 14 door experts uit beide groepen. In tabel 7 zijn de gemiddelde scores van de groepen te zien die door de beoordelaars zijn toegekend.

Tabel 7. *Gemiddelde scores van deskundigen (N=3, schaal van 1-10) op overige detecties van experts met en zonder CCC-model toegepast op bijsluiter.*

Herkomst beoordeelde overige detecties	Gemiddelde score
Uitsluitend met CCC (10)	4.7 (sd 1.7)
Uitsluitend zonder CCC (10)	4.1 (sd 2.2)
Door beide groepen experts genoemd (14)	4.7 (sd 2.0)

In elke groep is de eindscore duidelijk minder dan voldoende. Een t-toets voor onafhankelijke groepen liet geen significant verschil zien tussen de gemiddelde waardering van de CCC-problemen en de problemen die zonder CCC-model geproduceerd werden.

Mogelijk zijn deze magere scores nog enigszins geflatteerd omdat de beoordeelde detecties behoren tot een selectie van problemen die door meer dan één expert zijn genoemd. In bijvoorbeeld de CCC-conditie waren er 110 overige detecties. Hiervan zijn rond de 20 problemen overgebleven die door meer dan één persoon zijn genoemd en hiervan is de helft beoordeeld. De conclusie is dus dat de overige detecties waarover een minimale overeenstemming is en die dus al een zeker belang lijken te hebben, gemiddeld niet voldoende worden beoordeeld.

Deze toets is uitgevoerd na de vaststelling dat experts met het CCC-model niet goed in staat zijn problemen te voorspellen. We stelden vast dat mogelijkerwijs de overige detecties, die dus niet als lezersprobleem gevalideerd zijn, toch voldoende waardevol zouden kunnen zijn om tot een revisie te leiden. We trekken de conclusie dat dat vermoedelijk niet het geval is. In ieder geval worden de detecties die met het CCC-model geproduceerd zijn niet als significant waardevoller beoordeeld dan de detecties die zonder dat model geproduceerd zijn.

4. Conclusies en discussie

De overeenstemming tussen de experts die met het CCC-model een tekst beoordelen, is zeer gering. Respectievelijk 83,1% en 64% van de problemen betreft een unieke detectie. Er was geen verschil tussen experts met CCC-model en zonder dat model.
Het model zorgt er dus niet voor dat experts meer volgens een zelfde patroon gaan beoordelen en daardoor tot meer overeenstemming komen.

Ook aan de validiteit van het CCC-model kan getwijfeld worden. Met het model wordt slechts een laag percentage van de lezersproblemen voorspeld en dit gaat bovendien gepaard met een grote hoeveelheid overige detecties. De scores zijn niet hoger dan die welke in eerder onderzoek onder experts zonder CCC-model behaald werden. Het CCC-model lijkt ook op dit punt geen toegevoegde waarde te hebben.

Het percentage juist voorspelde *aannemelijke* problemen ligt met 40% wel een stuk hoger. Toch is ook dit percentage nog vrij laag te noemen. Ten eerste omdat deze set is samengesteld op basis van strenge criteria en dus de lezersproblemen lijkt te bevatten die vrij aannemelijk zijn. Ook hiervan wordt dus niet eens de helft voorspeld. In eerder onderzoek van Lentz en De Jong halen twee groepen experts op een set geselecteerde problemen hogere percentages. Verder worden de problemen die wél goed worden voorspeld maar door een zeer klein deel van de experts genoemd: gemiddeld door ongeveer twee van de tien experts. Hierbij lijkt dan ook eerder toeval en kennis van individuele experts mee te spelen dan een effectief CCC-model.

Experts en lezers hebben ten dele een verschillende definitie van tekstkwaliteit. In de beoordeling vallen de twee grootste groepen problemen onder de ijkpunten 'voldoende informatie' en 'gepaste formulering'. Het grootste verschil tussen lezers en CCC-beoordelaars is dat de problemen die de CCC-experts noemen meer verspreid liggen over het hele CCC-model.

De toegevoegde waarde van het CCC-model is te verwaarlozen wanneer we de resultaten vergelijken met een groep die zonder dat model dezelfde taak uitvoerde. Het gebrek aan overeenstemming tussen experts was in beide groepen even groot. En voor de validiteit maakte het ook niet uit of experts bij de beoordeling wel of niet gebruik maken van het CCC-model: ongeveer 20% van de expertproblemen is een hit. Tussen experts met model en zonder model zijn ook geen grote verschillen in perspectief te zien.

Over de kwaliteit van de overige detecties kan gezegd worden dat deze zeer matig is, des te meer omdat het hier gaat om de score van een selectie overige detecties die al een zeker gewicht hadden doordat ze door meerdere experts waren genoemd. Belangrijker is echter dat de beoordeling van de overige detecties van CCC-experts niet positiever is dan de oordelen over de detecties van de experts die niet over dat model beschikten.

Daarmee kunnen we de vragen uit Renkema (1996) van een antwoord voorzien. Toepassing van het CCC-model leidt niet tot een goede inschatting van lezersproblemen. Een tekstrevisie op basis van een CCC-analyse zal dan ook niet tot een betere tekst leiden (dan een revisie zonder gebruik van dat model); enerzijds omdat de oordelen over die tekst niet

betrouwbaar en valide zijn, anderzijds omdat de probleemdetecties niet waardevoller zijn dan de detecties van experts zonder CCC-model. Bij die conclusie willen we echter twee kanttekeningen maken.

Een eerste kanttekening is dat de experts misschien niet voldoende instructie hebben gekregen over het CCC-model. De verklaring voor de magere resultaten van deze experts zou dan de kwaliteit van de instructie zijn. Deze instructie betrof expliciet de werking van het CCC-model en bevatte uitleg en oefening over de 15 ijkpunten van het model. Het is uiteraard denkbaar dat een groep experts nog veel intensiever scholing krijgt op het thema van tekstkwaliteit, met bijvoorbeeld beoordelingscriteria per genre (voor bijsluiters, gebruiksaanwijzingen e.d.). Mogelijkerwijs zal dan ook de overeenstemming tussen de experts hoger worden. Deze is dan echter naar alle waarschijnlijkheid niet te danken aan meer kennis van het CCC-model, maar aan een grotere expertise in tekstkwaliteit. Het model zelf is immers niet zo ingewikkeld dat er een cursus van enkele weken voor nodig is.
De tweede kanttekening betreft de ambitie van het CCC-model. In de inleiding hebben we gesteld dat die niet altijd helder en eenduidig geformuleerd is. Eén ambitie hebben we in dit artikel van een kritische noot voorzien. Blijft over de vraag of de meer bescheiden ambitie realistischer is. Is het CCC-model een handig hulpmiddel om commentaren op een tekst te systematiseren? Renkema noemt drie criteria voor beoordeling: het model moet eenvoudig te hanteren zijn, het moet volledig zijn en de ijkpunten moeten duidelijk van elkaar te onderscheiden zijn. Op die criteria hebben wij dit model niet onderzocht. Er zou echter een vierde criterium genoemd kunnen worden: het model moet geen overbodige ijkpunten bevatten. Op dit punt hebben wij twijfels over met name de ijkpunten die onder *teksttype* vallen: er is in de evaluaties van de twee teksten geen enkel commentaar geleverd dat viel onder de ijkpunten 1) geschiktheid van het genre, en 2) genrezuiverheid. Slechts één commentaar is geplaatst onder ijkpunt 3) correcte toepassing genreregels. Er zou verder onderzocht moeten worden of het toeval is dat de teksten geen problemen bevatten op deze aspecten, of dat een beperkter CCC-model wellicht een effectiever hulpmiddel is om commentaren op een tekst te systematiseren.

Bibliografie

Dieli, M. (1986). *Designing successful documents: an investigation of document evaluation methods*. Pittsburgh, Pennsylvania: Carnegie Mellon University.

Jong, M. de (1998). *Reader feedback in text design. Validity of the plus-minus method for the pretesting of public information brochures*. Amsterdam-Atlanta: Rodopi.

Jong, M. de & L. Lentz (1996). Expert judgements versus reader feedback: a comparison of text evaluation techniques. *Journal of Technical Writing and Communication,* 26, 507-519.

Jong, M. de & L. Lentz (2001). Focus: Design and Evaluation of a Software Tool for Collecting Reader Feedback. *Technical Communication Quarterly*, 10, 387-401.

Jong, M. de & P.J. Schellens (1996). *Pretest van de brochure 'Je eerste baan'* Deelrapport 17 van het onderzoeksproject Pretesten. Vakgroep Toegepaste Taalkunde, Universiteit Twente. Enschede.

Lentz, L. & M. de Jong (1997). The evaluation of text quality: expert-focused and reader-focused methods compared. *IEEE Transactions on professional communication,* 40, 224-233.

Lentz, L. & H. Pander Maat (1992). Evaluating text quality: reader-focused or text-focused? In: H. Pander Maat & M. Steehouder (eds.) (1992). *Studies of functional text quality*. Amsterdam: Rodopi, 101-114.

Nielsen, J. (1994). Heuristic Evaluation. In: J. Nielsen & R. Mack (eds.) (1994). *Usability Inspection Methods*. New York: John Wiley & Sons, Inc., 25-62.

Noorlander, M. (2001). *Pretesten met bijsluiterteksten. Focus & de Hardop-leesmethode*. Doctoraalscriptie Taalbeheersing, Universiteit Utrecht.

Pander Maat, H. (1996). Identifying and predicting reader problems in drug information texts. In: T. Ensink & C. Sauer (eds.) (1996). *Researching technical documents*. Groningen, 17-47.

Pander Maat, H. en L. Lentz (2003). Waarom het lezersprotocol zo'n goede methode is om begripsproblemen op te sporen. *Tijdschrift voor Taalbeheersing*, 25, 202-220.

Renkema, J. (1996). Over smaak valt goed te twisten. Een evaluatiemodel voor tekstkwaliteit. *Taalbeheersing,* 18, 324-338.

Renkema, J. & M. Wijnstekers (1997). Doelgroep-onderzoek of bureau-analyse? In: H. van den Bergh e.a. (eds.) (1997). *Taalgebruik ontrafeld. Bijdragen aan het zevende VIOT-taalbeheersingscongres gehouden op 18, 19 en 20 december 1996 aan de Universiteit van Utrecht*. Dordrecht: Foris Publications, 365-373.

Renkema, J. (2000). Pretesten testen. De CCC-analyse en de plus-en-minmethode vergeleken. In: R. Neutelings e.a. (eds.) (2000), *Over de grenzen van de taalbeheersing. Bijdragen over taal, tekst en communicatie gepresenteerd op het VIOT-congres van 1999 aan de Technische Universiteit Delft*. Den Haag: Sdu Uitgevers, 273-283.

Vromen, N. (1998). *Focus. Een evaluatie-onderzoek naar het softwareprogramma Focus waarmee teksten in pretestsituaties beoordeeld kunnen worden*. Doctoraalscriptie Taalbeheersing, Universiteit Utrecht.

Maaike van den Haak
Menno de Jong en
Peter Jan Schellens

Hardopdenkprotocollen als pretestmethode

Synchroon en retrospectief hardopdenken vergeleken

Samenvatting

Hardopdenkprotocollen zijn een veelgebruikte pretestmethode. Toch is de methodologische kennis over de methode nog beperkt. In dit artikel wordt een vergelijking beschreven van synchrone en retrospectieve hardopdenkprotocollen voor de evaluatie van een online bibliotheekcatalogus. De beide methoden zijn op drie aspecten vergeleken: probleemdetecties, taakuitvoering en proefpersoonervaringen. De methoden leveren vergelijkbare verzamelingen gebruikersproblemen op, maar deze komen anders tot stand. Bij retrospectief hardopdenkonderzoek zijn de verbalisaties van proefpersonen cruciaal. Bij synchroon hardopdenkonderzoek is het leeuwendeel van de problemen observeerbaar in het proces en wordt er minder geverbaliseerd. Proefpersonen gaan vaker in de fout en zijn minder succesvol in de taakuitvoering. Dit roept vragen op over de reactiviteit van synchroon hardopdenkonderzoek, met name bij complexe taken.

*1. Introductie**

Hardopdenkprotocollen zijn een veelgebruikte methode voor de formatieve evaluatie van software, interfaces, websites en instructieve documenten. De methode houdt in dat gebruikers uit de doelgroep een aantal taken met behulp van het te evalueren communicatiemiddel uitvoeren en daarbij voortdurend hun gedachten verbaliseren. De methode heeft face validity, omdat de data die ermee verkregen worden het feitelijke gebruik weerspiegelen, en dus verder gaan dan het oordeel van proefpersonen over de gebruikersvriendelijkheid. Hardopdenkonderzoek is ingebed in een lange en gerespecteerde onderzoekstraditie, gericht op de cognitieve processen van proefpersonen tijdens de uitvoering van een breed scala aan taken – zoals schaken, probleemoplossen, schrijven, lezen en besluitvorming – met de monografie van Ericsson & Simon (1993) als belangrijke mijlpaal. Wanneer hardopdenken wordt gebruikt als pretestmethode, gaat het echter niet primair om inzicht in de cognitieve processen, maar om zicht op de kwaliteit van communicatiemiddelen of artefacten. In de afgelopen jaren zijn er diverse handboeken verschenen met uitvoerige instruc-

ties voor het uitvoeren van een hardopdenk usability test (Nielsen, 1993; Rubin, 1994; Dumas & Redish, 1999; Barnum, 2002).

De adviezen die in deze handboeken worden gegeven, worden echter nog nauwelijks ondersteund door methodologisch onderzoek (zie voor een overzicht De Jong & Schellens, 2000, 2002). Verschillende onderzoeken, waaronder dat van Jansen & Steehouder (1989), maken aannemelijk dat de hardopdenkmethode in combinatie met andere evaluatietechnieken kan leiden tot effectievere communicatiemiddelen, maar de bijdrage van het hardopdenkonderzoek is daarin niet geïsoleerd en de stap van hardopdenkresultaten naar een tekstrevisie wordt op geen enkele manier verantwoord. Recent onderzoek van Boren & Ramey (2000) maakt bovendien duidelijk dat de richtlijnen die Ericsson & Simon (1993) geven voor hardopdenkonderzoek in de praktijk niet worden opgevolgd en, zo betogen de auteurs, gezien de doelstellingen van een usability test wellicht ook niet zo strikt gehanteerd hoeven te worden. Ook woedt er in de literatuur nog steeds een discussie over de volledigheid en de mogelijke reactiviteit van hardopdenkprotocollen bij verschillende soorten taken, waarbij de resultaten van empirische vergelijkingen tussen hardopdenkende en stilwerkende proefpersonen de uitgangspunten van Ericsson & Simon (1993) lang niet altijd ondersteunen (zie bijvoorbeeld Russo, Johnson & Stephens, 1989; Short e.a., 1991; Loxterman, Beck & McKeown, 1994; Janssen, Van Waes & Van den Bergh, 1996). Hardopdenkende proefpersonen voeren hun taak soms anders, beter of slechter uit dan proefpersonen die stil mogen werken. Er zijn, met andere woorden, veel meer onzekerheden rondom hardopdenkprotocollen dan in de adviesliteratuur over usability testing wordt gesuggereerd.

Het onderzoek waarover in dit artikel wordt gerapporteerd, maakt deel uit van een groter onderzoeksproject gericht op de waarde en beperkingen van een aantal varianten van hardopdenkprotocollen als pretestmethode. De Jong & Schellens (1995) maken in hun overzicht van methoden onderscheid tussen twee varianten van deze onderzoeksmethode: hardopleesonderzoek (zonder taken) en hardopwerkonderzoek (met taken). Wij richten ons op de laatste variant. We beschrijven een eerste experiment waarin synchrone en retrospectieve hardopdenkprotocollen worden vergeleken bij de evaluatie van een online bibliotheekcatalogus. Retrospectieve hardopdenkprotocollen, ook wel 'retrospective testing' (Nielsen, 1993) of 'aided subsequent verbal protocol' (Henderson e.a., 1995) genoemd, verschillen in één opzicht van de (gebruikelijke) synchrone hardopdenkprotocollen: de proefpersonen denken niet hardop tijdens de taakuitvoering, maar voeren de taken in eerste instantie stil uit, en verbaliseren hun gedachten pas achteraf, aan de hand van een videoopname van hun taakuitvoering.

Theoretisch zijn er zowel voor- als nadelen aan het gebruik van retrospectieve hardopdenkprotocollen in plaats van synchrone hardopdenkprotocollen. Een voordeel betreft de mogelijke afname van reactiviteit: omdat de proefpersonen hun taken volledig op de eigen wijze en in het eigen tempo kunnen uitvoeren, zullen hun werkwijze en hun succes bij de taakuitvoering een betere afspiegeling vormen van hun normale taakuitvoering. Bij synchrone hardopdenkprotocollen is eerder sprake van reactiviteit: proefpersonen zouden bijvoorbeeld beter kunnen presteren dan normaal als gevolg van een meer gestructureerd werkproces, of juist slechter als gevolg van een te grote werkbelasting (Russo, Johnson & Stephens, 1989). Een tweede voordeel betreft de mogelijkheid om ook werktijden te meten. In de literatuur over usability testing worden de werktijden per taak vaak gehanteerd als indicator voor de gebruikersvriendelijkheid van een interface of website. Bij retro-

spectieve hardopdenkprotocollen kan een tijdsmeting zonder problemen als extra bron van informatie worden meegenomen; bij synchrone hardopdenkprotocollen ligt dat minder voor de hand, omdat van hardopdenken algemeen wordt aangenomen dat het de taakuitvoering in een per proefpersoon wisselende mate vertraagt. Een derde voordeel is dat proefpersonen de mogelijkheid hebben om te reflecteren op hun ervaringen tijdens de taakuitvoering, wat hen ertoe zou kunnen brengen om verbanden te leggen tussen individuele problemen en te zoeken naar meer structurele oorzaken.

Naast voordelen heeft het gebruik van retrospectieve hardopdenkprotocollen ook een aantal mogelijke nadelen. Een nadeel is dat de proefpersoonsessies aanmerkelijk langer duren dan bij synchrone hardopdenkprotocollen: de tijd verdubbelt omdat de proefpersonen niet alleen hun taken uitvoeren, maar deze ook nog achteraf bekijken. Een tweede en belangrijker nadeel is dat er gemakkelijk vertekeningen kunnen optreden in de achteraf geverbaliseerde gedachten. Het is onwaarschijnlijk dat proefpersonen zich na afloop nog al hun gedachten tijdens de taakuitvoering kunnen herinneren. Ericsson & Simon (1993) stellen dat belangrijke informatie verloren gaat in retrospectief onderzoek, wat bevestigd wordt door verschillende andere studies (e.g., Russo, Johnson & Stephens, 1989; Teague, De Jesus & Nunes-Ueno, 2001). Veel hangt echter af van de stimuli die de proefpersonen krijgen bij het verwoorden van hun gedachten. Bij retrospectief hardopdenken aan de hand van een video-opname hebben de proefpersonen meer aanknopingspunten dan bij ongeholpen retrospectieve hardopdenkprotocollen. Vertekeningen kunnen ook ontstaan wanneer proefpersonen besluiten bepaalde gedachten te verbergen, gedachten te verzinnen of te wijzigen, vanwege zelfpresentatie of sociale wenselijkheid. Hoewel dergelijke vertekeningen ook kunnen optreden bij synchrone hardopdenkprotocollen, hebben retrospectief hardopdenkende proefpersonen meer gelegenheid om hun verbalisaties te redigeren. Toch zijn ook zij voortdurend gebonden aan de gebeurtenissen zoals die op de videoband zijn vastgelegd, wat de bandbreedte om gedachten achteraf aan te passen beperkt maakt.

In de literatuur over usability testing worden synchrone en retrospectieve hardopdenkprotocollen vaak beschreven als vergelijkbare alternatieven (zie bijvoorbeeld Nielsen, 1993). Toch zijn er nog nauwelijks harde empirische gegevens beschikbaar over de twee hardopdenkvarianten. Verschillende onderzoekers hebben naar eigen zeggen een vergelijking tussen synchrone en retrospectieve hardopdenkprotocollen gemaakt, maar vergelijken synchrone hardopdenkprotocollen in feite met zuiver retrospectief onderzoek, waarbij de proefpersonen geen stimuli hebben gekregen bij het achteraf verbaliseren van hun gedachten (Branch 2000; Kuusela & Paul, 2000; Taylor & Dionne, 2000).

Tot dusver hebben slechts twee studies daadwerkelijk een vergelijking gemaakt tussen retrospectieve en synchrone hardopdenkprotocollen. Hoc & Leplat (1983) hebben de twee varianten gebruikt om een cognitief proces te onderzoeken, waarbij proefpersonen een set van letters moesten ordenen op een computerscherm met behulp van een beperkt aantal commando's. In de retrospectieve conditie werd de proefpersonen eerst gevraagd een beschrijving van hun werkproces te geven zonder stimuli, waarna ze vervolgens hardop moesten denken terwijl ze een log file met de stappen in hun taakuitvoering bekeken. Hoc & Leplat concluderen dat retrospectief onderzoek zonder stimuli beter vermeden kan worden vanwege de verstoringen en gaten in de protocollen, maar dat de retrospectieve en synchrone hardopdenk protocollen vergelijkbare resultaten opleveren. Hierbij moet wel worden opgemerkt dat zowel de taak die de proefpersonen moesten uitvoeren (deze leek op een puzzel) en de analyse van de resultaten (die meer gericht was op strategie dan op de

gevonden problemen) niet overeenkomen met die van een usability test.

Bowers & Snyder (1990) hebben de twee hardopdenkvarianten vergeleken in een usability test gericht op het gebruiken van meerdere windows op een computerscherm. Zij vonden geen significante verschillen met betrekking tot taakuitvoering en de benodigde tijd hiervoor, maar concludeerden wel dat de retrospectieve hardopdenkconditie beduidend minder verbalisaties opleverde, en dat deze verbalisaties bovendien vaak verschilden van de synchrone verbalisaties: ze waren meer gericht op verklaringen en minder op procedures. Hoewel deze resultaten zeker interessant zijn, heeft de studie een belangrijk nadeel: er wordt geen aandacht geschonken aan de hoeveelheid en soorten problemen die de proefpersonen in beide condities naar voren brachten. Omdat probleemopsporing de belangrijkste functie van usability testing is, is een cruciaal aspect niet in de vergelijking van de methodes meegenomen.

In dit artikel presenteren we een onderzoek dat gericht is op de waarde van synchrone versus retrospectieve hardopdenkprotocollen door een vergelijking te maken tussen de twee varianten met het oog op usability testing. Drie onderzoeksvragen staan centraal:

- Zijn er verschillen tussen synchrone en retrospectieve hardopdenkprotocollen met betrekking tot het aantal en type gevonden problemen?
- Zijn er verschillen tussen synchrone en retrospectieve hardopdenkprotocollen met betrekking tot taakuitvoering?
- Zijn er verschillen tussen synchrone en retrospectieve hardopdenkprotocollen met betrekking tot de ervaringen van de proefpersonen tijdens de test?

2. Opzet van het onderzoek

Testobject. Als testobject voor deze studie is gekozen voor een online bibliotheekcatalogus. Een dergelijke catalogus combineert de kenmerken van een zoekmachine met die van een website: hij heeft een taakgericht karakter, vereist de nodige online navigatie, en is vaak complex van aard, met name voor beginnende gebruikers. Online catalogi zijn met andere woorden een dankbaar object voor usability tests. Dat blijkt ook uit de literatuur op het gebied van bibliotheek- en informatiewetenschappen, waarin evaluatieonderzoek een regelmatig terugkerend thema is (bijvoorbeeld: Campbell, 2001; Battleson, Booth & Weintrop, 2001; Norlin & Winters, 2002).

De catalogus die in dit onderzoek is gebruikt, is de online catalogus van de Vrije Universiteit (UBVU). Deze is een aantal jaar geleden geïntroduceerd, en is sindsdien niet veranderd. Figuur 1 toont de homepage van de catalogus: een simpele layout, met een zoekscherm in het midden en links negen keuzeknoppen. Deze knoppen betreffen de standaard zoekmogelijkheden die catalogi doorgaans bieden, zoals eenvoudig of uitgebreid zoeken, sorteren, of bladeren. Zoals bij de meeste catalogi het geval is, heeft de UBVU ook een helpfunctie, die hulp biedt bij het gebruiken van de catalogus.

Hoewel de UBVU-catalogus in eerste instantie bestemd is voor studenten en medewerkers van de Vrije Universiteit, kunnen ook externe bezoekers toegang krijgen tot de catalogus, uiteraard met uitzondering van interne onderdelen zoals 'lenen' of 'reserveren'. Alle informatie in de catalogus is zowel in het Nederlands als in het Engels beschikbaar, behalve de helpfunctie: deze is alleen in het Engels te raadplegen.

Figuur 1. *Homepage van de UBVU-bibliotheekcatalogus*

Proefpersonen. Aan het experiment namen 40 studenten deel van de opleiding Toegepaste Communicatiewetenschap aan de Universiteit Twente. Alle studenten zaten in het tweede of derde jaar van hun studie en hadden al enige ervaring met het gebruik van online bibliotheekcatalogi. Geen van allen was echter bekend met de UBVU-catalogus. Als zodanig waren ze geschikte proefpersonen: als student behoorden ze tot de doelgroep van de catalogus maar als UT-student hadden ze geen voorkennis over en ervaring met de UBVU-catalogus.

De deelnemers zijn geworven door middel van posters en emailberichten, en kregen een financiële vergoeding voor hun deelname. Er waren geen criteria met betrekking tot geslacht of leeftijd. Uiteindelijk namen 5 mannen en 35 vrouwen deel aan het experiment, in de leeftijd van 18 tot 24. Deze 40 deelnemers werden gelijkmatig over de twee condities verdeeld: er waren geen verschillen tussen de twee groepen met betrekking tot geslacht, leeftijd en ervaring met online bibliotheekcatalogi.

Taken. Om de UBVU-catalogus te evalueren met synchrone en retrospectieve hardopdenkprotocollen werden zeven taken geformuleerd, die samen de belangrijkste functies van de catalogus omvatten. De taken konden onafhankelijk van elkaar worden uitgevoerd. De complete set taken was als volgt:

1. Zoek hoeveel publicaties de UBVU-catalogus bevat over het onderwerp 'communicatie'.
2. Zoek hoeveel publicaties de UBVU-catalogus bevat over het onderwerp 'taal' of 'interactie'.

3. Zoek hoeveel publicaties de UBVU-catalogus bevat die geschreven zijn door de auteur A. Hannay.
4. Zoek van welke auteur de UBVU-catalogus de meeste publicaties bevat over het onderwerp 'popmuziek'.
5. Zoek hoeveel Nederlandstalige publicaties de UBVU-catalogus bevat over het onderwerp 'Shakespeare'.
6. Zoek hoeveel publicaties de UBVU-catalogus bevat over het onderwerp 'telecommunicatie' die gepubliceerd zijn vanaf 1999.
7. Zoek hoeveel publicaties de UBVU-catalogus bevat over het onderwerp 'web-' (d.w.z. website, webwinkel, webcommunicatie) binnen de context van het internet.

Taak 1 tot en met 4 hadden betrekking op de zoekfunctie (eenvoudig en uitgebreid) en de sorteerfunctie van de catalogus. Taak 5 en 6 waren gericht op het inperken van zoekresultaten (op taal en op jaar van publicatie). Taak 7 had betrekking op truncatie, een bibliotheekfunctie die vergelijkbaar is met de zogenaamde 'wild card' zoekoptie.

Vragenlijsten. Naast de zeven taken kregen de proefpersonen in beide condities ook twee vragenlijsten voorgelegd. De eerste vragenlijst werd overhandigd vóór aanvang van het experiment, en bevatte vragen over de achtergrondgegevens van de proefpersonen. Naast demografische gegevens (leeftijd, geslacht en opleiding) werd gevraagd of de proefpersonen al eerder met een online bibliotheekcatalogus hadden gewerkt en of ze een cursus op dit gebied hadden gevolgd. Verder werd nagegaan of ze kennis hadden van veel voorkomende catalogusfuncties, zoals zoeken, bladeren, etc.

De tweede vragenlijst werd na afloop van het experiment aan de proefpersonen voorgelegd en betrof de ervaringen van de proefpersonen tijdens het experiment. Drie onderwerpen stonden hierbij centraal: (1) hoe hebben de proefpersonen het synchroon of retrospectief hardopdenken ervaren, (2) hebben de proefpersonen naar eigen indruk anders gewerkt dan ze normaal zouden doen, en (3) hoe hebben de proefpersonen de aanwezigheid van de onderzoeker en de opnameapparatuur ervaren? De proefpersonen moesten hun ervaringen weergeven op vijfpuntsschalen met semantische differentialen. Daarnaast bood de vragenlijst ruimte voor eventuele opmerkingen.

Onderzoeksprocedure. Het onderzoek bestond uit 40 individuele sessies, die allemaal in dezelfde testruimte plaatsvonden. Tijdens elke sessie werden video-opnamen gemaakt van het computerscherm en werd de stem van de proefpersoon opgenomen. Daarnaast was de onderzoeker aanwezig in de ruimte om te observeren en aantekingen te maken. Gemiddeld werkten de proefpersonen 20 minuten aan de zeven taken.

De synchrone hardopdenksessies verliepen als volgt. Na binnenkomst vulde de proefpersoon de eerste vragenlijst (over achtergrondkenmerken) in. Vervolgens kreeg de proefpersoon instructies over de gang van zaken tijdens het onderzoek en werden de taken overhandigd. De instructies, die met het oog op uniformiteit van papier werden voorgelezen, luidden als volgt: 'Denk hardop terwijl je de taken uitvoert en doe gewoon alsof de onderzoeker niet aanwezig is. Je kunt haar niet om hulp vragen, maar ze zal je wel eraan herinneren hardop te blijven denken als je een tijdje stilvalt. Vergeet verder niet dat we niet jou, maar de catalogus testen. Als er iets misgaat, ligt dat dus aan de catalogus, en niet aan jou.'

Nadat de proefpersoon alle zeven taken had uitgevoerd, kreeg hij/zij de tweede vragenlijst (over de ervaringen tijdens het experiment).

De retrospectieve hardopdenksessies begonnen, net als de synchrone sessies, met de eerste vragenlijst en een instructie. De proefpersoon kreeg dezelfde zeven taken voorgelegd, maar moest in dit geval deze taken in stilte uitvoeren. Wederom was het niet geoorloofd om hulp van de onderzoeker te vragen. Na voltooiing van de taken kreeg de proefpersoon een video-opname van de taakuitvoering te zien, met de vraag om hierop tijdens het lopen van de band commentaar te geven. Daarbij kon de opname niet worden stilgezet. Tot slot moest de proefpersoon weer de tweede vragenlijst invullen.

Verwerking van de onderzoeksgegevens. Nadat de veertig sessies waren voltooid, werden de verbalisaties van de proefpersonen uitgeschreven en werd ook hun navigatie binnen de site genoteerd. Uit deze navigatie en de overige handelingen van de proefpersonen werden de problemen gedestilleerd die tijdens het gebruik van de catalogus optraden. Elke handeling die afweek van het optimale handelingsverloop (dat wil zeggen: het minste aantal vereiste handelingen) bij een taak werd als probleem gemarkeerd. Daarnaast werd in de transcripten van de verbalisaties gekeken naar verbale signalen die op problemen duidden, zoals uitingen van twijfel, onwetendheid of irritatie.

De totale set aan data werd als volgt geanalyseerd. Eerst werd vastgesteld hoeveel problemen er in beide condities naar voren kwamen. Daarna werd bij iedere probleemdetectie gekeken naar de wijze waarop deze aan het licht was gekomen: door observatie van de handelingen, door analyse van de verbalisaties, of door een combinatie van observatie en verbalisaties. Tot slot werden de problemen inhoudelijk gecategoriseerd. Omdat voor de combinatie van hardopdenkgegevens en bibliotheekcatalogi nog geen standaardlijst met probleemcategorieën bestaat, zijn op basis van een globale indeling van het zoekproces en een analyse van de ontdekte gebruikersproblemen, de volgende vijf probleemtypen gehanteerd:

Lay-outproblemen:	De proefpersoon kan bepaalde informatie op het scherm van de catalogus niet of moeilijk vinden.
Terminologieproblemen:	De proefpersoon begrijpt de terminologie in de catalogus niet.
Data-invoerproblemen:	De proefpersoon weet niet hoe hij/zij een zoekopdracht moet invoeren (bijv. het invoeren van een zoekterm of het gebruik van de dropdown menu's).
Volledigheidsproblemen:	In de catalogus ontbreekt informatie die nodig is voor het effectief gebruik ervan.
Feedbackproblemen:	De catalogus geeft geen relevante feedback op (verkeerd) uitgevoerde zoekopdrachten.

Naast deze vijf probleemtypen hadden de proefpersonen ook af en toe technische problemen, zoals een internetverbinding of browser die niet werkte. Deze problemen zijn niet in de analyse meegenomen.

Met betrekking tot de taakuitvoering van de proefpersonen is gekeken naar twee indicatoren: het met goed gevolg voltooien van de opdrachten en de tijd benodigd voor het voltooien van de opdrachten. Deze indicatoren zijn zowel per taak als voor de gehele taakuitvoering (alle zeven taken) bekeken.

3. Resultaten

Paragraaf 3.1 beschrijft het aantal en soort problemen dat in beide condities naar voren is gekomen. In paragraaf 3.2 worden de resultaten met betrekking tot de taakuitvoering beschreven. In paragraaf 3.3 wordt, ten slotte, ingegaan op de ervaringen van de proefpersonen tijdens het onderzoek, zoals gemeten in de tweede vragenlijst.

3.1 Aantal en soort gevonden problemen. De analyse van de veertig sessies leverde in totaal 72 verschillende problemen op. Sommige van deze problemen werden door bijna alle (30 tot 35) proefpersonen gevonden, maar meer dan de helft van het totaal aantal verschillende problemen werd gevonden door vijf of minder van de 40 proefpersonen. Dit geeft aan dat er een behoorlijk aantal individuele problemen was: problemen die weliswaar door sommige proefpersonen werden gevonden, maar die door de meeste proefpersonen niet als problematisch werden ervaren.

Tabel 1 geeft een overzicht van het gemiddeld aantal gevonden problemen per proefpersoon. Daarbij is meteen onderscheid gemaakt naar de manier waarop de problemen aan het licht zijn gekomen: door observatie, door verbalisaties, of door een combinatie van beiden. Uit de tabel blijkt dat er geen significant verschil was in het aantal ontdekte problemen per proefpersoon in beide condities. Maar er waren wel duidelijke verschillen in de manier waarop de problemen aan het licht kwamen.

Tabel 1. *Aantal problemen per proefpersoon met synchrone en retrospectieve hardopdenkprotocollen, ontdekt door observatie, verbalisaties of een combinatie van beide*

	Synchroon		Retrospectief		
	Gemiddeld	**SD**	**Gemiddeld**	**SD**	**Significantie**
Observatie	6.7	2.2	4.0	2.0	$p<0.001$
Verbalisaties	0.5	0.7	4.5	3.4	$p<0.001$
Observatie & verbalisaties	6.7	4.0	5.1	2.2	n.s.
Totaal	13.9	3.3	13.6	4.1	n.s.

De retrospectieve hardopdenkconditie leverde aanzienlijk meer geverbaliseerde problemen op (t-test, $t=5.168$, $df=38$, $p<0.001$, Cohen's $d=1.29$). De proefpersonen in deze conditie noemden gemiddeld 4,5 problemen die niet door observatie konden worden vastgesteld (tegenover slechts 0,5 geverbaliseerde problemen in de synchrone hardopdenkconditie). Dit verschil kan worden verklaard doordat proefpersonen in de retrospectieve conditie simpelweg meer tijd hadden om problemen te verbaliseren. Zij hoefden per slot van rekening pas na afloop van de taken commentaar te geven op hun werk, waardoor ze zich volledig konden richten op het evalueren van de catalogus, en naast taakgerelateerde problemen ook andere problemen konden bespreken, vaak geformuleerd als waardeoordelen. De proefper-

sonen in de synchrone hardopdenkconditie moesten tegelijkertijd werken en hardopdenken, en konden daarom minder aandacht geven aan het verbaliseren van problemen. Als gevolg daarvan noemden zij voornamelijk problemen die direct met hun taakuitvoering te maken hadden. Dit komt ook tot uiting in het aantal problemen dat door een combinatie van verbalisaties en observatie aan het licht kwam: 93% van alle verbaliseringen in de synchrone hardopdenkconditie kwam overeen met een observeerbaar probleem in de taakuitvoering; in de retrospectieve hardopdenkconditie was dit percentage maar 54%.

Een tweede significant verschil tussen beide condities betreft het aantal problemen dat door non-verbale indicatoren, d.w.z. puur door observatie, aan het licht is gekomen. Zoals tabel 1 aangeeft, bracht de synchrone hardopdenkconditie beduidend meer geobserveerde problemen aan het licht dan de retrospectieve hardopdenkconditie (6,7 versus 4,0; t-test, $t=4.083$, $df=38$, $p<0.001$, Cohen's $d=1.63$). Blijkbaar gingen proefpersonen in de synchrone hardopdenkconditie vaker in de fout tijdens de taakuitvoering dan proefpersonen in de retrospectieve conditie. Dit verschil kan worden verklaard door de verschillende werkbelasting in beide condities: terwijl de proefpersonen in de retrospectieve hardopdenkconditie zich uitsluitend met hun taken hoefden bezig te houden, moesten de proefpersonen in de synchrone conditie hierbij ook hun gedachten verbaliseren. Het is goed denkbaar dat deze extra belasting een negatieve uitwerking had op de taakuitvoering van de proefpersonen.

Om inzicht te krijgen in de typen problemen die in beide condities naar voren kwamen, werden alle probleemdetecties gecategoriseerd. Bij wijze van voorbeeld geeft tabel 2 een indruk van concrete gebruikersproblemen zoals die in de verschillende probleemcategorieën zijn ondergebracht.

Tabel 2. *Voorbeelden van probleemtypen uit de hardopdenkprotocollen*

Lay-out	De proefpersoon heeft moeite om de knop voor gevorderd zoeken te vinden op de homepage van de catalogus De proefpersoon kan de namen van co-auteurs niet vinden in de zoekresultaten van de catalogus
Terminologie	De proefpersoon begrijpt de betekenis van het woord 'limieten' niet. De proefpersoon begrijpt de betekenis van het woord 'truncatie' niet.
Data-invoer	De proefpersoon heeft moeite met het gebruik van booleaanse operatoren. De proefpersoon weet niet hoe data in het 'jaar'-venster moeten worden ingevoerd.
Volledigheid	De namen van auteurs ontbreken in de zoekresultaten. De helpfunctie geeft alleen informatie in het Engels, niet in het Nederlands.
Feedback	De catalogus geeft geen foutmelding als de proefpersoon een fout maakt. De catalogus geeft niet aan hoe de zoekresultaten geordend zijn (op jaar, op eerste auteur, etc.)

Tabel 3 geeft een overzicht van de typen problemen die in de beide condities zijn ontdekt. Er waren op dit punt geen significante verschillen: alle vijf probleemtypen werden in vergelijkbare aantallen in beide condities gevonden. Terminologie en data-invoer werden door de proefpersonen in beide condities als meest problematisch ervaren.

Tabel 3. *Aantal problemen uit verschillende categorieën per proefpersoon in de synchrone en retrospectieve hardopdenkconditie*

	Synchroon		Retrospectief		
	Gemiddeld	**SD**	**Gemiddeld**	**SD**	**Significantie**
Lay-out	2.9	1.2	2.6	1.3	n.s.
Terminologie	4.1	1.5	4.1	2.0	n.s.
Data-invoer	4.9	1.2	4.9	1.2	n.s.
Volledigheid	1.1	0.9	1.2	0.6	n.s.
Feedback	1.0	1.0	0.9	0.6	n.s.
Totaal	13.9	3.3	13.6	4.1	n.s.

De analyses tot nu toe waren gericht op de algemene trends in de resultaten, waarbij individuele problemen buiten beschouwing bleven. Een vergelijking van de lijsten met problemen die in beide condities naar voren kwamen, geeft een indruk van de mate van overlap tussen de synchrone en de retrospectieve hardopdenkconditie. Van de in totaal 72 verschillende problemen werd 47% gevonden in beide condities, 31% alleen in de synchrone hardopdenkconditie, en 22% alleen in de retrospectieve hardopdenkconditie. Meer overlap bestaat wanneer gekeken wordt naar de overlap in probleemdetecties. We kijken dan steeds per probleemdetectie of die ook in de andere conditie naar voren is gekomen. Tabel 4 laat zien dat 89% van alle probleemdetecties betrekking heeft op problemen die door proefpersonen in beide condities werden gevonden.

Tabel 4. *Percentage problemen die uniek zijn voor één van beide hardopdenkcondities*

	Uniek synchroon	**Uniek retrospectief**	**Ontdekt in beide**
Lay-out	10	12	78
Terminologie	1	6	93
Data-invoer	6	2	92
Volledigheid	11	4	84
Feedback	8	2	91
Totaal	6	5	89

Het algemene beeld dat ontstaat, is dat synchrone en retrospectieve hardopdenkprotocollen vergelijkbaar zijn wat betreft het aantal en de typen gevonden problemen. De twee methoden verschillen echter met betrekking tot de wijze waarop deze problemen naar voren komen: synchrone hardopdenkprotocollen brengen meer problemen aan het licht die tijdens de taakuitvoering te observeren zijn; retrospectieve hardopdenkprotocollen resulteren in meer problemen aan de hand van de verbalisaties van proefpersonen. De verbalisaties spelen een ondergeschikte rol in de synchrone hardopdenkprotocollen. Dit resultaat is opmerkelijk, omdat men er bij de hardopdenkmethode als usability test juist van uitgaat dat de verbale protocollen cruciaal zijn voor het ontdekken van problemen. Uit het huidige experiment blijkt echter dat de verbale protocollen niet zozeer nieuwe problemen aan het licht brengen, maar voornamelijk dienen ter ondersteuning van de problemen die ook waarneembaar zijn. Het feit dat observeerbare problemen significant meer voorkomen in de synchrone hardopdenkconditie kan, zoals eerder beschreven, worden verklaard door de zwaardere werkbelasting van de proefpersonen. Dat is een eerste aanwijzing voor de reactiviteit van de (synchrone) hardopdenkmethode in dit experiment. Om deze reden is het interessant om te kijken of de dubbele werkbelasting ook invloed heeft gehad op de taakuitvoering van de proefpersonen.

3.2 Taakuitvoering. De taakuitvoering van de proefpersonen in beide condities is bekeken aan de hand van twee indicatoren: de succesvolle voltooiing van de zeven opdrachten en de tijd die nodig was om deze opdrachten te voltooien. Tabel 5 geeft een overzicht van de resultaten van beide indicatoren. Met betrekking tot de benodigde tijd zijn er geen significante verschillen gevonden, noch per taak noch voor de gehele set van zeven taken. Blijkbaar had het hardop denken in de synchrone conditie geen vertragend effect op de taakuitvoering van de proefpersonen. De dubbele werkbelasting in de synchrone hardopdenkconditie had echter wel invloed op de mate van succes bij het voltooien van de taken. De proefpersonen in deze conditie waren significant minder succesvol in het correct voltooien van de complete takenset dan de proefpersonen in de retrospectieve conditie (t-test, $t=2.252$, $df=38$, $p<0.05$, Cohen's $d=0.71$). Dit resultaat correspondeert met de eerdere bevinding dat de synchrone hardopdenkprotocollen ook meer waarneembare problemen bevatten dan de retrospectieve protocollen. Dit bevestigt het eerdere vermoeden van reactiviteit van de synchrone hardopdenkprotocollen. Hierbij moet worden opgemerkt dat de proefpersonen in het algemeen moeite hadden met het uitvoeren van de opdrachten: gemiddeld genomen werd slechts 40% van de taken met succes voltooid. De lastigste taak (taak 7) werd door slechts één van de 40 proefpersonen met succes volbracht; de eenvoudigste taak (taak 4) door 38 van de 40 proefpersonen. In de discussieparagraaf zullen we hier nader op ingaan.

Tabel 5. *Taakuitvoering in de synchrone en retrospectieve hardopdenkcondities*

	Synchroon		Retrospectief		
	Gemiddeld	**SD**	**Gemiddeld**	**SD**	**Significantie**
Aantal taken dat succesvol is afgerond	2.6	1.0	3.3	1.0	$p<0.05$
Totale tijd benodigd voor de zeven taken	21.1	5.7	19.6	5.0	n.s.

3.3 Ervaringen van de proefpersonen. De vragenlijst met betrekking tot de ervaringen van de proefpersonen tijdens het onderzoek richtte zich op drie aspecten: de ervaringen met synchroon of retrospectief hardopdenken, de werkwijze bij de taakuitvoering en de aanwezigheid van onderzoeker en opnameapparatuur.

Over hun ervaringen met (synchroon of restrospectief) hardopdenken moesten de proefpersonen op vijfpuntsschalen aangeven of zij deze activiteit moeilijk, onaangenaam, vermoeiend, onnatuurlijk of tijdrovend vonden. Omdat deze vragen samen geen betrouwbare schaal vormden, werd elke vraag individueel geanalyseerd. Deze individuele analyses (zie tabel 6) laten zien dat er geen significante verschillen waren in de oordelen van de proefpersonen over het hardopdenken (t-test). Over het algemeen oordeelden de proefpersonen redelijk neutraal, met scores rond het midden van de vijfpuntsschaal. Voor de synchrone hardopdenkconditie betekent dit dat de reactiviteit van de methode, zoals die in de vorige paragrafen naar voren kwam, niet als zodanig door de proefpersonen zelf werd ervaren.

Tabel 6. *Proefpersoonervaringen met betrekking tot het hardopdenken*

	Synchroon		Retrospectief		
	Gemiddeld	**SD**	**Gemiddeld**	**SD**	**Significantie**
Moeilijk – gemakkelijk	2.4	0.8	2.7	1.2	n.s.
Onprettig – prettig	2.7	0.8	2.9	1.0	n.s.
Wel – niet vermoeiend	3.4	1.0	3.8	1.4	n.s.
Onnatuurlijk – natuurlijk	3.4	0.9	3.0	1.5	n.s.
Wel – niet tijdrovend	3.2	1.2	3.2	1.1	n.s.

NB: Scores op een vijfpuntsschaal (1 = negatief, 5 = positief)

De proefpersonen werd vervolgens gevraagd om in te schatten in hoeverre hun werkwijze verschilde van een normale werkwijze: sneller of langzamer, gerichter of minder gericht, etc. De resultaten, die in tabel 7 zijn weergegeven, bevatten wederom geen significante verschillen tussen de synchrone en de retrospectieve hardopdenkconditie. In beide condities waren de proefpersonen van mening dat hun werkwijze slechts weinig verschilde van hun normale manier van werken. Na een hercodering van de antwoorden om elke afwijking van de normale werkwijze (naar beide kanten van de schaal) vast te stellen (waarbij de middelste score als 0 werd gecodeerd en de extremen als 2), bleek dat de acht variabelen een betrouwbare schaal vormden (Cronbach's alpha = 0.84). Op deze schaal bleken de proefpersonen in de retrospectieve hardopdenkconditie, naar hun eigen oordeel, significant meer afwijkend te hebben gewerkt tijdens dan de proefpersonen in de synchrone hardopdenkconditie (met een gemiddelde afwijking van 0.33 versus 0.29; t-test, $t=2.242$, $df=38$, $p<.0.05$, Cohen's $d=0.72$).

Dit betekent dat, in tegenstelling tot de eerdere bevindingen met betrekking tot probleemdetecties en taakuitvoering, de proefpersonen in de retrospectieve hardopdenkconditie meer reactiviteit ervoeren dan de proefpersonen in de synchrone conditie. Dit kan echter te maken hebben met het tijdstip waarop de proefpersonen in de retrospectieve hardopdenkconditie de vragenlijst invulden. Zij deden dit na afloop van het bekijken en becommentariëren van hun video-opname, en het is goed voorstelbaar dat de per definitie onnatuurlijke taak van het achteraf commentaar geven in hun beoordeling van de taakuitvoering is mee genomen.

Tabel 7. *Proefpersoonervaringen met betrekking tot hun werkwijze in de test, vergeleken met hun normale werkwijze*

	Synchroon		Retrospectief		
	Gemiddeld	**SD**	**Gemiddeld**	**SD**	**Significantie**
Sneller – langzamer	2.7	0.7	2.3	0.8	n.s.
Meer – minder gericht	2.6	0.6	2.1	0.9	n.s.
Meer – minder geconcentreerd	3.3	0.6	3.5	0.9	n.s.
Meer – minder vasthoudend	2.6	0.9	2.7	0.9	n.s.
Meer – minder succesvol	3.0	0.5	2.9	0.7	n.s.
Meer – minder prettig	3.2	0.5	3.4	0.6	n.s.
Meer – minder oog voor fouten	2.6	0.7	2.2	0.7	n.s.
Meer onstpannen – meer gespannen	3.4	0.6	3.7	0.5	n.s.

NB: Scores op een vijfpuntsschaal (3 = geen verschil met de normale werkwijze)

Het laatste deel van de vragenlijst betrof de aanwezigheid van de onderzoeker en het gebruik van opnameapparatuur. De proefpersonen werd eerst gevraagd aan te geven in hoeverre ze het onplezierig, onnatuurlijk of storend hadden gevonden dat de onderzoeker tijdens het experiment aanwezig was. Vervolgens werden dezelfde vragen nog een keer gesteld, maar dan met betrekking tot het gebruik van de opnameapparatuur. Voor alledrie de aspecten van de testsituatie kon een voldoende betrouwbare schaal (op basis van twee vragen) worden gevormd (Cronbach's alpha = 0.66 voor 'onplezierig', 0.81 voor 'onnatuurlijk' en 0.62 voor 'storend'). De resultaten staan in tabel 8. De scores met betrekking tot '(on)plezierig' en '(on)natuurlijk' waren noch negatief noch positief, en verschilden niet significant tussen de twee condities. De scores met betrekking tot '(niet) storend' waren tamelijk positief in beide condities, maar de proefpersonen in de synchrone hardopdenkconditie vonden de testsituatie nog minder storend dan de proefpersonen in de retrospectieve hardopdenkconditie (t-test, t=2.368, df=33.4, p<.05, Cohen's d=0.75).

Dit laatste verschil kan wederom worden verklaard door het tijdstip waarop de proefpersonen in de retrospectieve hardopdenkconditie de vragenlijst invulden. Een tweede verklaring ligt in het feit dat de aanwezigheid van de onderzoeker in het eerste gedeelte van de retrospectieve conditie minder functioneel was dan tijdens de synchrone conditie. Ook zou het achteraf bekijken van hun eigen proces confronterend kunnen zijn voor de proefpersonen. Tot slot zou ook de werkdruk van de proefpersonen een verklaring kunnen zijn. De proefpersonen in de synchrone hardopdenkconditie moesten gelijktijdig taken uitvoeren en hardopdenken, waardoor ze minder aandacht konden besteden aan de onderzoeker en de opnameapparatuur.

Tabel 8. *Proefpersoonervaringen met betrekking tot de testsituatie: de aanwezigheid van de onderzoeker en opnameapparatuur*

	Synchroon		Retrospectief		
	Gemiddeld	**SD**	**Gemiddeld**	**SD**	**Significantie**
Onplezierig	2.8	0.3	2.7	0.8	n.s.
Onnatuurlijk	2.9	0.7	3.1	1.3	n.s.
Storend	4.3	0.6	3.7	0.9	p<.05.

NB: Scores op een vijfpuntsschaal (1 = negatief, 5 = positief)

Al met al ondersteunen de oordelen van de proefpersonen de validiteit van zowel de synchrone als de retrospectieve hardopdenkprotocollen. De vragen resulteerden in overwegend positieve oordelen voor beide evaluatiemethoden. Uit de vragenlijst kwam echter een aantal verschillen naar voren tussen beide condities die slecht corresponderen met de gegevens betreffende de probleemdetecties en de taakuitvoering uit paragraaf 3.1 en 3.2. De proefpersonen in de retrospectieve hardopdenkconditie ondervonden meer reactiviteit als gevolg van de testsituatie, en vonden deze situatie storender dan de proefpersonen in de synchrone hardopdenkconditie. Dit kan op een reëel verschil tussen beide methoden duiden, maar het is ook aannemelijk dat dit verschil is veroorzaakt door een achteraf minder gelukkige keuze in de onderzoeksprocedure (waarbij de proefpersonen in de retrospectieve hardopdenkconditie de vragenlijst niet direct na de taakuitvoering, maar pas na het becommentariëren van de video-opname invulden).

4. Discussie

Het hier gerapporteerde onderzoek toont aan dat er zowel overeenkomsten als verschillen zijn tussen synchrone en retrospectieve hardopdenkprotocollen. De verschillen die tussen beide condities gevonden zijn, geven een nieuwe kijk op de validiteit van hardopdenkprotocollen voor usability testing. Hoewel beide methoden vergelijkbaar waren wat betreft de aantallen en de typen probleemdetecties, verschilden ze significant in de wijze waarop deze probleemdetecties tot stand kwamen. De synchrone hardopdenkconditie leverde significant meer problemen op die puur op basis van observatie naar voren kwamen. De retrospectieve hardopdenkconditie resulteerde in significant meer problemen die niet te observeren waren, maar alleen door de verbalisaties van proefpersonen aan het licht kwamen. Deze resultaten maken duidelijk dat synchroon hardopdenkonderzoek een meer getrouwe representant is van een strikt taakgerelateerde usability test, terwijl retrospectief hardopdenkonderzoek een breder scala aan gebruikersreacties lijkt op te leveren. Dit komt overeen met de bevindingen uit het onderzoek van Bowers & Snyder (1990), waarbij proefpersonen in de retrospectieve hardopdenkconditie allerlei verklaringen en suggesties verwoordden, terwijl de proefpersonen in de synchrone hardopdenkconditie zich vaak beperkten tot het beschrijven van hun handelingen. Om de waarde van de feedback van beide methodes te bepalen is verder onderzoek nodig naar de predictieve validiteit van synchrone en retrospectieve hardopdenkprotocollen: Hoe belangrijk zijn de gevonden problemen? Zijn er veel false alarms, met name in de geobserveerde problemen bij synchroon hardopdenkende proefpersonen en in de geverbaliseerde problemen in retrospectieve hardopdenkprotocollen?

Met betrekking tot het gebruik van synchrone hardopdenkprotocollen brengen de resultaten van dit onderzoek twee belangrijke zaken aan het licht. De eerste is de bijzonder beperkte bijdrage van de proefpersoonverbalisaties aan de opbrengst (in aantal gevonden gebruikersproblemen) van de usability test. De verbalisaties van proefpersonen resulteerden nauwelijks in de detectie van nieuwe problemen, en dienden voornamelijk ter ondersteuning of verklaring van problemen die ook te observeren waren in de proefpersoonhandelingen. Dit kan op zich ook een belangrijke bijdrage zijn, zeker voor de diagnose en het op waarde schatten van de gevonden problemen. Maar toch moet worden vastgesteld dat de verbalisaties in de synchrone hardopdenkconditie een minder belangrijk onderdeel van de usability test waren dan doorgaans in handboeken over usability testing wordt gesuggereerd.

Een tweede en nog belangrijker observatie is dat de synchrone hardopdenkprotocollen reactiviteit veroorzaakten in de usability test. Dit komt overeen met eerdere bevindingen van Russo, Johnson & Stephens (1989), die de validiteit van hardopdenkprotocollen bestudeerden voor verschillende cognitieve taken en concludeerden dat het hardopdenken de taakuitvoering zowel kon hinderen als bevorderen. Deze observatie weerspreekt de resultaten van Bowers & Snyder (1990), die geen verschil vonden in de taakuitvoering van synchroon en retrospectief hardopdenkende proefpersonen.

In het hier gerapporteerde onderzoek had het hardop denken een consistent en plausibel negatief effect op de taakuitvoering. De extra taak om tijdens de taakuitvoering gedachten te verbaliseren zorgde ervoor dat de proefpersonen meer fouten maakten en minder succesvol waren in het verrichten van de zeven taken. Aan de hand van dit resultaat lijkt de twijfel gerechtvaardigd of de taakuitkomst in een synchrone hardopdenktest een correcte

indicatie geeft van de gebruikersvriendelijkheid van een communicatiemiddel, en of de problemen die in hardopdenkonderzoek worden gevonden per definitie ook echte gebruikersproblemen zijn. Onderzoek naar de predictieve validiteit, zoals omschreven door De Jong & Schellens (2000, 2002), van hardopdenkresultaten is dus geen overbodige poging om vast te stellen wat feitelijk al bekend is, maar een belangrijke stap in verder onderzoek naar de reactiviteit van de methode. Er bestaat ten slotte een reële mogelijkheid dat een probleem gevonden in een synchrone hardopdenktest (gedeeltelijk) is veroorzaakt door de gehanteerde methode zelf. Vergeleken met stilwerkende proefpersonen levert de hardopdenkmethode weliswaar meer probleemdetecties op, maar deze zijn grotendeels toe te schrijven aan extra problemen in de taakuitvoering, en niet zozeer aan de verbalisaties zelf.

Of deze constatering al dan niet schadelijk is, staat overigens nog ter discussie. De meeste usability tests hebben tot doel gebruikersproblemen te identificeren en beoordelen, en het is vol te houden dat het een voordeel van synchroon hardopdenken is dat dergelijke problemen kennelijk gemakkelijker aan het licht komen. In die interpretatie is de drempel om taken succesvol te verrichten alleen wat hoger. Natuurlijk is zo'n positieve draai aan de resultaten alleen te geven als uit onderzoek blijkt dat deze extra geobserveerde problemen in de praktijk corresponderen met reële gebruikersproblemen.

De meest plausibele verklaring voor de reactiviteit van synchrone hardopdenkprotocollen ligt in de werkbelasting van de proefpersonen: de moeilijkheidsgraad van de taken kan een cruciale factor in dit onderzoek zijn geweest. De gegevens met betrekking tot de taakuitvoering maken duidelijk dat de zeven taken die de proefpersonen voorgelegd kregen, erg moeilijk waren. De cognitieve belasting van de taken in combinatie met de extra belasting van het hardopdenken lijkt een negatief effect te hebben gehad op zowel de verbalisaties als de taakuitvoering van proefpersonen. De gaten in de verbalisaties worden onderschreven door Ericsson & Simon (1993, p.91), die stellen dat proefpersonen mogelijk stoppen met verbaliseren wanneer ze te zwaar cognitief belast zijn. Het negatieve effect op de taakuitvoering wordt echter niet eenduidig verklaard door de bestaande literatuur (Russo, Johnson & Stephens, 1989; Ericsson & Simon, 1993). In tegendeel, er zijn ook studies die aantonen dat het synchroon hardopdenken juist een positief effect op taakuitvoering heeft (Loxterman, Beck & McKeown 1994). Het zou dan ook interessant zijn om de relatie tussen taakcomplexiteit, verbalisatie en taakuitvoering van proefpersonen in een synchrone hardopdenktest nader te onderzoeken.

Een laatste opmerking betreft de generaliseerbaarheid van het huidige onderzoek. Het gaat hier om een eerste vergelijkende studie, waarbij slechts één object was betrokken. Een belangrijk kenmerk van de UBVU-catalogus en de taken gebruikt in dit onderzoek is dat er in de wijze waarop de proefpersonen met de computer werkten veel te observeren viel. Het zou interessant zijn om te kijken of vergelijkbare resultaten naar voren komen bij applicaties met een minder duidelijk waarneembaar gebruikersproces. Een replicatie van dit onderzoek met documentatie, websites of interfaces met een meer open taakdomein zou een nuttige follow-up zijn om synchrone en retrospectieve hardopdenkprotocollen verder te onderzoeken.

Al met al duiden de resultaten van dit onderzoek erop dat synchrone en retrospectieve hardopdenkprotocollen kunnen worden beschouwd als gelijkwaardige maar duidelijk verschillende evaluatiemethoden. Een sterk, en nieuw argument dat voor het gebruik van retrospectieve hardopdenkprotocollen pleit, is dat ze minder gevoelig voor de invloed van taakcomplexiteit zouden kunnen zijn, zowel met het oog op reactiviteit als met betrekking

tot de volledigheid van de verbalisaties. In richtlijnen voor hardopdenkonderzoek wordt vaak gesteld dat de onderzoeker taken dient te formuleren met een gemiddelde moeilijkheidsgraad, zodat de deelnemers noch in een automatisch werkproces vervallen noch belast worden met een te zware cognitieve belasting. Bij usability testing zijn deze richtlijnen echter niet altijd haalbaar. Tenslotte liggen de kwaliteit van het testobject en de selectie van realistische taken doorgaans niet in de handen van het usability test team.

Noten

⋆ Dit artikel is een Nederlandse bewerking van Van den Haak, De Jong & Schellens (2003).

Bibliografie

Barnum, C.M. (2002). *Usability testing and research*. New York: Longman.

Battleson, B., A. Booth & J. Weintrop (2001). Usability testing of an academic library web site: a case study. *Journal of Academic Librarianship, 237*, 188-198.

Boren, M.T. & J. Ramey (2000). Thinking aloud: reconciling theory and practice. *IEEE Transactions on Professional Communication, 43*,261-278.

Bowers, V.A. & H.L. Snyder (1990). Concurrent versus retrospective verbal protocols for comparing window usability. *Proceedings of the Human Factors Society 34th Meeting, 8-12 October 1990* (pp.1270-1274). Santa Monica: HFES.

Branch, J.L. (2000). Investigating the information-seeking processes of adolescents: The value of using think alouds and think afters. *Library & Information Science Research, 22*, 371-392.

Campbell, N. (ed.) (2001). *Assessment of library-related Web sites: Methods and case studies*. Chicago: LITA.

Dumas, J.S. & J.C. Redish (1999). *A practical guide to usability testing*. Revised edition. Exeter: Intellect.

Ericsson, K.A. & H.A. Simon (1993). *Protocol analysis: Verbal reports as data*. Revised edition. Cambridge, MA: MIT Press.

Haak, M.J. van, M.D.T. de Jong & P.J. Schellens (2003). Retrospective vs. concurrent think-aloud protocols: testing the usability of an onlike library catalogue. *Behaviour & Information Technology, 22,* 339-351.

Henderson, R.D., M.C. Smith, J. Podd & H. Varela-Alvarez (1995). A comparison of the four prominent user-based methods for evaluating the usability of computer software. *Ergonomics, 38*, 2030-2044.

Hoc, J.M. & J Leplat (1983). Evaluation of different modalities of verbalization in a sorting task. *International Journal of Man-Machine Studies, 18*, 283-306.

Jansen, C.J.M. & M.F. Steehouder (1989). *Taalverkeersproblemen tussen overheid en burger. Een onderzoek naar verbeteringsmogelijkheden van voorlichtingsteksten en formulieren*. Dissertatie Universiteit Twente, Enschede. 's-Gravenhage: Sdu.

Janssen, D., L. van Waes & H. van den Bergh (1996). Effects of thinking aloud on writing processes. In C.M. Levy & S. Randell (eds.), *The science of writing. Theories, models, individual differences, and applications* (pp.233-250). New Jersey: Lawrence Erlbaum.

Jong. M. de & P.J. Schellens (1995). *Met het oog op de lezer. Pretestmethoden voor schriftelijk voorlichtingsmateriaal.* Amsterdam: Thesis

Jong, M. de & P.J. Schellens (2000). Toward a document evaluation methodology: What does research tell us about the validity and reliability of methods? *IEEE Transactions on Professional Communication, 43*, 242-260.

Jong, M. de & P.J. Schellens (2002). Tekstevaluatie. Onderzoek naar de validiteit van probleemopsporende methoden. *Tijdschrift voor Taalbeheersing, 24*, 146-166.

Kuusela, H. & P. Paul (2000). A comparison of concurrent and retrospective verbal protocol analysis. *American Journal of Psychology, 113*, 387-404.

Loxterman, J.A., I.L. Beck & M.G. McKeown (1994). The effects of thinking aloud during reading on students' comprehension of more or less coherent text. *Reading Research Quarterly, 29*, 353-367.

Nielsen, J. (1993). *Usability engineering*. Boston, MA: Academic Press.

Norlin, E. & C.M.I.Winters (2002). *Usability testing for library websites: a hands-on guide*. Chicago: American Library Association.

Rubin, J. (1994). *Handbook of usability testing: How to plan, design, and conduct effective tests*. New York: John Wiley.

Russo, J.E., E.J. Johnson & D.L. Stephens (1989). The validity of verbal protocols. Memory & *Cognition, 17*, 759-769.

Short, E.J., S.W. Evans, S.E. Friebert & C.W. Schatschneider (1991). Thinking aloud during problem solving: Facilitation effects. *Learning and Individual Differences, 3* (2), 109-122.

Taylor, K.L. & J.P. Dionne (2000). Accessing problem-solving strategy knowledge: The complementary use of concurrent verbal protocols and retrospective debriefing. *Journal of Educational Psychology, 29*, 413-425

Teague, R., K. De Jesus & M. Nunes-Ueno (2001). Concurrent vs. post-task usability test ratings. *Proceedings of the Conference on Human Factors and Computing Systems, 31 March – 5 April 2001* (pp.289-290). Seattle, WA: ACM SIGCHI.

Marinda Hall
Menno de Jong en
Michaël Steehouder

Multiculturele website-evaluatie

Verschillen tussen individualistische en collectivistische proefpersonen

Samenvatting

Een website is geëvalueerd met proefpersonen afkomstig uit twee culturen: een collectivistische, contextgevoelige cultuur en een individualistische, weinig contextgevoelige cultuur. Voor de test werd gebruik gemaakt van retrospectieve hardopdenkprotocollen en de plus-minmethode. Uit het onderzoek blijkt dat de plus-minmethode aanzienlijk minder problemen aan het licht brengt bij proefpersonen uit een collectivistische cultuur. Retrospectieve hardopdenkprotocollen blijken minder gevoelig voor culturele invloeden, maar er zijn wel verschillen: proefpersonen uit een collectivistisch cultuur vallen minder vaak uit de rol van gebruiker die ze in de test moeten aannemen en hun commentaar is vaak indirecter geformuleerd dan dat van de proefpersonen uit een individualistische cultuur.

1. Inleiding

Ontwerpers van websites gericht op internationale doelgroepen moeten rekening houden met cultuurverschillen, dat spreekt vanzelf. Er is dan ook een groeiende interesse in de vakliteratuur voor kwesties die met internationale en interculturele aspecten van websites te maken hebben. Zo is er onderzoek gedaan naar de relatie tussen cultuurkenmerken en internetgebruik (bijvoorbeeld La Ferle, Edwards & Mizuno 2002), de mate waarin culturele achtergronden van invloed zijn op de waardering van websites (bijvoorbeeld O'Keefe e.a. 2000; Simon 2001) en de vraag hoe internationale organisaties op hun websites omgaan met culturele kwesties (bijvoorbeeld Marcus & Gould 2000; Becker 2002; Okazaki & Rivas 2002). Arnold (1998) geeft een overzicht van linguïstische, culturele, juridische en technische complicaties die zich voordoen bij het ontwerpen van websites voor internationale doelgroepen.

Een voor de hand liggende aanpak om een website geschikt te maken voor bezoekers met verschillende culturele achtergronden is die verscheidenheid ook tot zijn recht te laten komen bij het evalueren of testen van de site (Hoft 1995; Nielsen 2000). Voor zo'n evaluatie zijn verschillende methoden voorhanden (zie Schriver 1989; De Jong & Schellens 1995, 1997), en er zijn goede handleidingen voor usability testing beschikbaar (zoals Dumas & Redish 1993; Rubin 1994; Barnum 2002; Schweibenz & Thissen 2003). De Jong & Schellens (2000, 2002) concluderen uit een overzicht van de beschikbare methodologische

literatuur dat een evaluatie met beoogde gebruikers een effectieve manier is om de bruikbaarheid van documenten en interfaces te beoordelen en te verbeteren. Zowel *in-use* evaluatiemethoden, zoals hardopdenk usability testing, als *non-use* methoden, zoals de plus-en-min-methode, blijken geschikt. Er is echter nooit nagegaan wat de invloed is van de culturele achtergrond van proefpersonen op het verloop en de resultaten van evaluatieonderzoek. Werken hardopdenkprotocollen hetzelfde bij proefpersonen uit Europa en Azië, en zijn de resultaten ook vergelijkbaar? En hoe gedragen beide groepen proefpersonen zich in een plus-en-minonderzoek? Tot dusver is het onderzoek naar de validiteit en de bruikbaarheid van evaluatiemethoden vrijwel uitsluitend gedaan met proefpersonen uit Noord-Amerika en West-Europa. Het is niet duidelijk of de conclusies van dat onderzoek ook opgaan voor proefpersonen uit andere culturen.

Er is wel enig onderzoek beschikbaar naar de invloed van andere persoonsvariabelen op de feedback die verkregen wordt in evaluatieonderzoek. Uit onderzoek van De Jong & Schellens (2001) bleek dat mannelijke en vrouwelijke proefpersonen verschillende commentaren gaven op brochures; ook bleek dat hoger opgeleiden meer problemen in brochures signaleerden dan lager opgeleiden, en meer aandacht hadden voor problemen met de tekststructuur. Van Versseveld (1995) toonde aan dat de betrokkenheid van proefpersonen bij het onderwerp invloed kan hebben op het commentaar dat wordt verkregen op een brochure. In een plus-en-mintest noemden hoog-betrokkenen meer problemen in een brochure, en richtten ze zich vooral meer op volledigheidsproblemen (behoefte aan meer informatie), terwijl de laag-betrokkenen juist meer relevantieproblemen noemden. In een experiment met het testen van vragenlijsten vonden Diamantopoulos, Reynolds & Schlegelmilch (1994) dat deelnemers met inhoudelijke voorkennis en kennis van vragenlijsten beter in staat waren om allerlei gebreken in de vraagformulering te ontdekken (zoals ambigue vragen of ontbrekende antwoordmogelijkheden). Blijkbaar maakt het uit wat voor proefpersonen aan een doelgroepgerichte pretest deelnemen.

Het internationale karakter van het World Wide Web roept de vraag op naar het effect van nationale cultuur als relevant achtergrondkenmerk van proefpersonen in een pretest. Het is immers genoegzaam bekend dat deze variabele invloed heeft op allerlei andere vormen van gedrag (Hall 1977; Hofstede 1980, 1994; Smith & Bond 1998; Trompenaars & Hampden-Turner 1998). De hoofdvraag van ons onderzoek is dan ook: *Beïnvloeden cultuurverschillen tussen proefpersonen de aard van de feedback die verkregen wordt bij de evaluatie van een website?* Om deze vraag te beantwoorden, voerden we een webevaluatie uit met proefpersonen uit West-Europa en uit Azië en Afrika en vergeleken we zowel de resultaten als de ervaringen van de proefpersonen.

Uit het spectrum van mogelijke evaluatiemethoden kozen we er twee, namelijk retrospectieve hardopdenkprotocollen en de plus-en-minmethode. De eerste staat model voor een *in-use* benadering. De proefpersonen voeren taken uit met behulp van een website en hun handelingen worden opgenomen op video. Na afloop bekijken ze de video en proberen ze onder woorden te brengen wat ze gedacht hebben tijdens de uitvoering van de taken (cf. Nielsen 1993). We kozen voor een retrospectieve in plaats van een synchrone hardopdenksessie omdat het onderzoek werd gehouden in het Engels, wat niet de moedertaal was van de proefpersonen. We veronderstellen dat de taak om hardop te denken in een vreemde taal een te zware cognitieve belasting voor de proefpersonen is, waaronder zowel de taakuitvoering als het hardopdenken zou kunnen lijden. Uit eerder onderzoek blijkt dat synchroon en retrospectief hardopdenken goed vergelijkbare resultaten opleveren (Hoc &

Leplat 1983; Van den Haak, De Jong & Schellens 2003).

De plus-en-minmethode is een typische *non-use* methode (De Jong 1998). De proefpersonen wordt gevraagd om een document te lezen en plussen en minnen in de marge te zetten op plaatsen waar ze positieve of negatieve leeservaringen hebben. In het tweede deel van een plus-en-minsessie worden de proefpersonen geïnterviewd over de redenen waarom ze plussen en minnen noteerden. Hoewel er enkele pogingen zijn gedaan om deze methode toe te passen voor de evaluatie van websites, hebben we er in dit onderzoek voor gekozen te werken met geprinte versies van enkele webpagina's.

Beide methoden zijn gangbaar in de communicatiepraktijk. Met het oog op mogelijke cultuurverschillen is de combinatie van methoden interessant omdat ze beide een hoge mate van interactie met de proefpersoon met zich meebrengen, maar verschillende eisen stellen aan de proefpersonen. Bij de hardopdenkmethode moeten de proefpersonen zich gedragen als echte gebruikers van de website en inzicht geven in de fouten die ze maken en de twijfels die ze hebben bij het uitvoeren van hun taken. Bij de plus-minmethode vervullen de proefpersonen de rol van beoordelaar en moeten ze hun oordelen over de website geven aan de proefleider. Beide methoden kunnen in bepaalde opzichten bedreigend zijn voor een proefpersoon.

2. Dimensies van cultuurverschillen

Cultuurverschillen worden vaak gekarakteriseerd aan de hand van zogenaamde cultuurdimensies. Dat zijn aspecten van een cultuur die gemeten kunnen worden in relatie tot andere culturen. Een toonaangevende reeks dimensies werd ontwikkeld door Hofstede (1994, 2001). Op basis van een vragenlijst die werd ingevuld door 116.000 personeelsleden van IBM in 50 landen en 20 talen onderscheidde hij de volgende dimensies:

- masculiniteit versus femininiteit,
- hoge versus lage onzekerheidsvermijding,
- grote versus kleine machtsafstand,
- individualisme versus collectivisme,
- lange- versus korte-termijnoriëntatie.

Vergelijkend onderzoek heeft aangetoond dat de dimensie *individualisme-collectivisme* de belangrijkste is als het gaat om verschillen tussen culturen (Ting-Toomey 1998). In individualistische culturen zijn de banden tussen mensen relatief los: mensen worden geacht vooral voor zichzelf en de meest directe familieleden te zorgen. In collectivistische culturen leven mensen van geboorte tot dood in hechte groepen, die een levenslange beschermde omgeving vormen en daarvoor in ruil een groot beroep doen op loyaliteit. De Verenigde Staten en Zweden zijn sterk individualistische landen; de Arabische landen en Indonesië scoren juist laag op de individualisme-index (IDV).

Verschillende auteurs stellen dat deze dimensie sterk is gerelateerd aan de manier waarop mensen met elkaar communiceren. Dit sluit aan op een ander onderscheid van Hall (1977), tussen contextgevoelige (high-context) en weinig contextgevoelige (low-context) culturen (Hofstede 2001; Ting-Toomey 1998). In weinig contextgevoelige culturen dient communicatie expliciet, direct en eenduidig te zijn. In contextgevoelige culturen ligt veel informatie besloten in de context, of is deze geïnternaliseerd in de personen die met elkaar

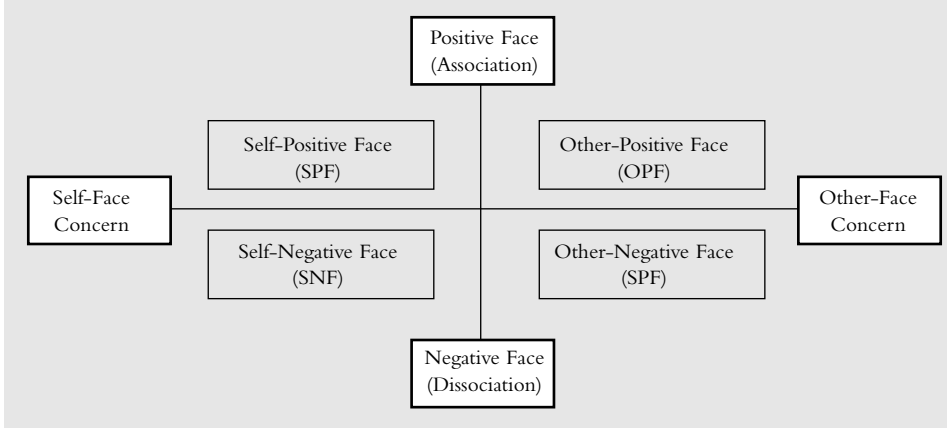
communiceren; er wordt weinig expliciet gemaakt in communicatieve boodschappen. Duitsland, Scandinavië, de Verenigde Staten en Zwitserland bezitten een relatief weinig contextgevoelige cultuur, terwijl Aziatische en Mediterrane landen juist gekenmerkt worden door een contextgevoelige cultuur. Contextgevoelige culturen corresponderen met collectivistische culturen, terwijl weinig contextgevoelige culturen doorgaans individualistisch zijn. In individualistische culturen moeten veel vanzelfsprekende zaken expliciet gezegd worden (Hofstede, 2001 p. 212).

Het verschil tussen de West-Europese en de Aziatisch en Afrikaanse proefpersonen in ons onderzoek kan goed gekarakteriseerd worden door een combinatie van Hofstedes dimensie individualisme-collectivisme en Halls onderscheid naar contextgevoeligheid. Aziatische en Afrikaanse proefpersonen bevinden zich dan aan de collectivistische en contextgevoelige zijde van het spectrum, West-Europese proefpersonen aan de individualistische en weinig contextgevoelige zijde. Vanzelfsprekend gaat het hierbij om verschillen tussen groepen, en doen we geen voorspellingen over het gedrag van individuen. Bovendien moeten we in het oog houden dat culturen meestal verschillen op meer dimensies tegelijk. Onze karakterisering van de twee groepen is dus een (bedoelde) simplificatie van de werkelijkheid, die, zoals we hieronder zullen betogen, een vruchtbare basis is om hypothesen te ontwikkelen voor ons onderzoek.

3. Cultuur en beleefdheid: hypothesen voor het onderzoek

Ting-Toomey (1998) verbond theorieën over cultuurverschillen met de beleefdheidstheorie van Brown & Levinson (1990). Centraal in deze beleefdheidstheorie staat de behoefte van mensen om hun 'gezicht te behouden'. Mensen willen door anderen gewaardeerd worden (positive face) en ze willen niet door anderen gedwongen worden tot gedrag dat ze niet wensen (negative face). Brown & Levinson hebben een typologie van gezichtsbedreigende handelingen ontwikkeld en onderzoek gedaan naar de wijze waarop die voorkomen. Voorbeelden van positief-gezichtsbedreigende handelingen zijn afkeuren en kritiek geven: de spreker (S) laat merken dat hij of zij de verlangens, eigendommen of persoonlijke eigenschappen van de hoorder (H) niet op prijs stelt. Negatief-gezichtsbedreigende handelingen zijn bijvoorbeeld opdrachten: S geeft te kennen dat hij wil dat H iets doet. Maar ook bijvoorbeeld een aanbod kan een negatief-gezichtsbedreigende handeling zijn: S verplicht H om het aanbod te accepteren of af te slaan, en wellicht om in de toekomst een wederdienst te bewijzen. Hoewel er culturele verschillen zijn in de manier waarop inhoud wordt gegeven aan het begrip 'face', is de erkenning van het belang ervan voor mensen universeel.

Aan de hand van een onderscheid tussen 'self-face concern' en 'other-face concern' en van het hierboven genoemde onderscheid tussen positive en negative face ontwikkelde Ting-Toomey (1998) een model met vier kwadranten (zie figuur 1). Mensen die gericht zijn op 'self-positive face' (SPF), zoeken in hun communicatie aansluiting bij en waardering van anderen. Mensen die gericht zijn op 'other-positive face' (OPF), houden in hun communicatie vooral rekening met de behoefte aan aansluiting en waardering van anderen. Mensen die gericht zijn op 'self-negative face' (SNF) proberen in hun communicatie vooral hun eigen vrijheid van handelen en autonomie te beschermen. Mensen die gericht zijn op 'other-negative face' (ONF) tonen in hun communicatie veel respect voor de handelingsvrijheid van anderen.



Figuur 1. *Vier kwadranten van facework (Ting-Toomey 1998, p.218)*

Volgens Ting-Toomey verschillen individualistische, weinig contextgevoelige culturen in veel opzichten van collectivistische, contextgevoelige culturen. Voortbouwend op figuur 1 ontwikkelde ze een set theoretische aannames over verschillen in 'facework' tussen de twee culturen (zie tabel 1). Deze aannames vormden de basis voor de hypothesen die we voor ons onderzoek hebben geformuleerd. Hieronder zullen we de zeven hypothesen beschrijven en toelichten.

Tabel 1. *Facework in individualistische, weinig contextgevoelige en in collectivistische, contextgevoelige culturen (gebaseerd op Ting-Toomey, 1998, p.230)*

Elementen van 'face'	Individualistische culturen	Collectivistische culturen
Identiteit	Nadruk op 'ik'	Nadruk op 'wij
Primaire zorg voor	Self-face	Other-face
Behoefte aan	Negative face	Positive face
Strategieën gericht op	Self-positive en self-negative face	Other-positive en other-negative face
Stijl	Controlerend, confronterend en oplossingsgericht	Voorkomend, vermijdend en vriendelijk
Taalhandelingen	Directe taalhandelingen	Indirecte taalhandelingen
Nonverbaal gedrag	Directe emotionele uitdrukkingen	Indirecte emotionele uitdrukkingen

De eerste vraag die we ons stellen, is in hoeverre culturele achtergrond van invloed is op de resultaten van de pretest. Bij de plus-en-minmethode is deze invloed te verwachten. Proefpersonen die in een plus-en-mintest problemen naar voren willen brengen, moeten openlijk en direct kritiek op de website geven. Dit past niet bij de neiging tot indirecte communicatie en bij de 'other-face concern' die in collectivistische culturen dominant zijn. Bij de hardopdenkprotocollen is voorhands geen verschil te verwachten. Veel van de problemen die in hardopdenkonderzoek aan het licht komen, zijn direct gerelateerd aan knelpunten die zich voordoen tijdens de taakuitvoering. Er is geen reden om te veronderstellen dat één van beide groepen proefpersonen meer knelpunten zal ervaren dan de andere. We hebben voor ons onderzoek namelijk twee groepen proefpersonen geworven met een gelijkwaardig opleidingsniveau (zie paragraaf 4.2).

H1 De plus-en-minmethode zal minder problemen aan het licht brengen bij proefpersonen uit collectivistische culturen dan bij proefpersonen uit individualistische culturen.

H2 De retrospectieve hardopdenkprotocollen zullen evenveel problemen aan het licht brengen bij proefpersonen uit collectivisitische culturen als bij proefpersonen uit individualistische culturen.

De culturele oriëntatie kan daarnaast een verschil in de appreciatie van de testmethode met zich meebrengen. Beide onderzochte methoden hebben consequenties voor de 'face' van de proefpersoon. We veronderstellen daarbij dat er een relatie is tussen het type 'facework' dat dominant is in beide culturen en de ervaringen van de proefpersonen tijdens het evaluatieonderzoek. In collectivistische culturen is er met name sprake van 'other-face concern' en een behoefte aan 'positive face'. Beide worden bedreigd door een opzet met de plus-en-minmethode, waarbij proefpersonen immers kritiek moeten uiten op een website. Individualistische culturen zijn met name gericht op een 'self-face concern' en op 'negative face'. De 'self-face concern' wordt bedreigd door een test met retrospectieve hardopdenkprotocollen, omdat daarin duidelijk wordt wat de proefpersonen allemaal fout doen tijdens de taakuitvoering (hoewel ze natuurlijk kunnen besluiten om de schuld van deze problemen te zoeken bij de website en niet bij zichzelf; zie hypothese 5). De behoefte aan 'negative face' wordt bedreigd doordat de proefpersonen in de onderzoekssetting worden gedwongen om de website op een bepaalde manier, aan de hand van een set taken, te gebruiken (hoewel ze natuurlijk kunnen besluiten om hun rol als proefpersoon ruimer op te vatten; zie hypothese 6).

H3 De proefpersonen uit collectivistische culturen zullen positiever oordelen over de retrospectieve hardopdenkprotocollen dan de proefpersonen uit individualistische culturen.

H4 De proefpersonen uit individualistische culturen zullen positiever oordelen over de plus-en-minmethode dan de proefpersonen uit collectivistische culturen.

Een derde aspect dat vanuit cultureel oogpunt interessant is, betreft het verschijnsel 'blaming' (Schriver, 1997): de mate waarin proefpersonen zichzelf, het product of de testsituatie de schuld geven wanneer zij tegen problemen aanlopen tijdens een usability test. De vraag of proefpersonen in ons onderzoek zichzelf (intern) dan wel de website of de testsituatie (extern) de schuld geven van problemen kan cultuurafhankelijk zijn. Vanwege hun 'other-face concern' kunnen proefpersonen uit collectivistische culturen meer geneigd zijn om zelf de schuld op zich te nemen; vanwege hun 'self-face concern' zouden proefpersonen uit individualistische culturen juist meer geneigd kunnen zijn om de schuld bij externe factoren te zoeken.

H5 Proefpersonen uit collectivistische culturen zullen in het hardopdenkonderzoek meer geneigd zijn om de schuld voor problemen tijdens de taakuitvoering bij zichzelf te zoeken dan proefpersonen uit individualistische culturen.

Een vierde mogelijke invloed van cultuurverschillen betreft de rol van proefpersonen in de test. In hardopdenkonderzoek moeten proefpersonen zich gedragen als gebruikers die een beperkt aantal, niet door henzelf gekozen, taken uitvoeren en met name aandacht besteden aan direct aan die taken gerelateerde problemen. Deze rol kan in strijd zijn met de behoefte aan 'negative face' die bij individualistisch proefpersonen sterker aanwezig is dan bij collectivistische proefpersonen. Als reactie op deze beperkingen zouden ze gedurende het proces andere rollen kunnen aannemen dan de strikte gebruikersrol die in het onderzoek opgesloten ligt. Deze rollen kunnen in hun hardopdenkcommentaar naar voren komen. We onderscheiden de volgende rollen: de rol van proefpersoon (ingaan op de eigen ervaringen tijdens de test en de eigen prestaties), de rol van internetgebruiker (vertellen over het eigen gebruik van het World Wide Web in normale omstandigheden) en de rol van reviewer (oordelen geven over de website).

H6 In het retrospectieve hardopdenkonderzoek zullen proefpersonen uit individualistische culturen minder geneigd zijn om zich te houden aan de door de taken geïmpliceerde gebruikersrol dan proefpersonen uit collectivistische culturen.

De laatste hypothese heeft betrekking op de manier waarop de proefpersonen zich in de test uitdrukken. Mensen uit collectivistische culturen hebben naar verwachting een voorkeur voor indirecte taalhandelingen bij het leveren van commentaar, omdat ze meer gericht zijn op 'other-face concern' en 'positive face'. We hebben dit overigens alleen onderzocht in de retrospectieve hardopdenkprotocollen. De plus-en-minresultaten bevatten zoveel commentaar en zoveel combinaties van directe en indirecte uitingen dat een betrouwbare scoring van de taalhandelingen niet goed mogelijk was.

H7 Proefpersonen uit collectivistische culturen zullen meer geneigd zijn om indirecte en eufemistische formuleringen te kiezen voor de kritiek die ze hebben op de website dan proefpersonen uit individualistische culturen.

4. Methode

Om de bovenstaande hypothesen te toetsen, hebben we een website geëvalueerd met een combinatie van retrospectieve hardopdenkprotocollen en de plus-en-minmethode. In het onderzoek werden twee groepen proefpersonen betrokken: assistenten in opleiding (aio's) van Aziatische of Afrikaanse afkomst en West-Europese aio's. Hieronder bespreken we de website die we hebben gebruikt, de proefpersonen en de procedure. Vervolgens gaan we in op de afhankelijke variabelen in het onderzoek.

4.1 Onderzochte website: Web of Science. Voor ons onderzoek waren we op zoek naar een website die voldeed aan de volgende criteria:

- Een instructieve functie: voor het retrospectieve hardopdenkonderzoek was het wenselijk dat proefpersonen een aantal realistische taken aan de hand van de website konden uitvoeren.
- Substantiële tekstuele inhoud: voor de plus-en-minmethode was het wenselijk dat de website tekstuele informatie bevatte die de proefpersonen konden lezen en becommentariëren.

- Geen cultuurspecifieke inhoud: de website moest gericht zijn op gebruikersgroepen uit verschillende culturen. Het zou even waarschijnlijk moeten zijn dat Aziatische/Afrikaanse en West-Europese proefpersonen de website zouden bezoeken en gebruiken.

Op grond van deze criteria hebben we de *Web of Science* database gekozen als onderzoeksobject. *Web of Science* wordt gepubliceerd en bijgehouden door het Institute for Scientific Information (ISI). De kern van de database bestaat uit citatie-indexen. Wetenschappers kunnen er opzoeken hoe vaak en in welke artikelen er naar een bepaald artikel is verwezen. *Web of Science* omvat een groot aantal wetenschappelijke tijdschriften in de alfa-, bèta- en gammawetenschappen. Voor aio's geldt *Web of Science* als een belangrijke informatiebron: het is één van de manieren om systematisch wetenschappelijke literatuur te zoeken. Ter illustratie bevat figuur 2 het 'full search' scherm van de website.

Figuur 2. *De 'full search' pagina van het Web of Science*

4.2 Proefpersonen. We hebben voor het onderzoek twee homogene groepen proefpersonen geworven, die alleen verschilden op hun nationale herkomst. We vroegen mannelijke aio's in de technische wetenschappen aan de Universiteit Twente om deel te nemen. Deze steekproef bleek om meerdere redenen geschikt voor ons onderzoek:

- De proefpersonen behoorden allemaal tot de doelgroep van *Web of Science*.
- Het opleidingsniveau van de proefpersonen was in beide groepen gelijk. Alle proefpersonen hadden hun mastersdiploma gehaald en werkten aan hun promotieonderzoek. Een gelijkwaardig opleidingsniveau is belangrijk, omdat eerder onderzoek van De Jong & Schellens (2001) uitwees dat er een relatie is tussen opleiding en de hoeveelheid en de aard van de feedback in een pretest.
- De proefpersonen uit Azië en Afrika waren op het moment van onderzoek slechts één tot twee jaar in Nederland, en kunnen daarom nog worden beschouwd als representanten van de cultuur van hun land van herkomst.

- Voor alle proefpersonen was Engels niet hun moedertaal. Vanwege de sterk uiteenlopende nationaliteiten van de proefpersonen hadden we besloten om het gehele onderzoek, in beide groepen in het Engels af te nemen. De taalvaardigheid van de proefpersonen kan immers invloed hebben op hun verbalisaties in de retrospectieve hardopdenkprotocollen en op het soort en de hoeveelheid feedback die ze geven bij de plus-en-minmethode (de website zelf was ook in het Engels). Gezien de vooropleiding van de deelnemers en de Engelstalige praktijk aan de Universiteit Twente, mogen we aannemen dat de beheersing van het Engels in beide groepen min of meer gelijk was.

Deze selectiecriteria maken het goed mogelijk om eventuele culturele verschillen op het spoor te komen, maar er is natuurlijk ook een keerzijde, in termen van generaliseerbaarheid: het onderzoek beperkt zich tot hoger opgeleide en mannelijke proefpersonen. Toekomstig onderzoek zou zich mede moeten richten op proefpersonen met lagere opleidingsniveaus en op vrouwelijke proefpersonen.

Proefpersonen die voldeden aan de eerdergenoemde criteria werd mondeling gevraagd of ze bereid waren om mee te werken aan ons onderzoek. In totaal wilden 38 aio's meedoen: 20 proefpersonen uit een individualistische cultuur (allen uit Nederland) en 18 uit collectivistische culturen (uit India, Indonesië, China, Turkije en Soedan).

4.3 Procedure. Het onderzoek vond plaats in individuele sessies in een onderzoeksruimte aan de Universiteit Twente. Het onderzoek bestond uit vier delen. Eerst werd de proefpersonen gevraagd om zeven taken met *Web of Science* uit te voeren. Hun taakuitvoering werd opgenomen met behulp van het programma HyperCam (http://www.hyperionics. com). Omdat het soort taken mogelijk van invloed is op de resultaten van een usability test (zie Sienot 1997; Van Waes 2000), kregen de proefpersonen in totaal zeven taken, verdeeld over twee categorieën: zoektaken en toepassingstaken. Om de eventuele invloed van voorkennis op de taakuitvoering uit te sluiten, besloten we om geen taken op te nemen die aansloten op de technisch-wetenschappelijke literatuur, maar in plaats daarvan dicht bij ons eigen vakgebied te blijven. Hieronder staat een overzicht van alle (in het Nederlands vertaalde) taken.

1. Hoeveel artikelen zijn er beschikbaar over het onderwerp 'communication theory'? Zorg ervoor dat u artikelen vindt waarin beide woorden opeenvolgend worden gebruikt.
2. Zoek op hoeveel artikelen er beschikbaar zijn over 'web evaluation'. Deze keer hoeven de twee woorden niet opeenvolgend te zijn.
3. Bewaar het laatste zoekresultaat op de A-drive van deze computer.
4. Hoeveel artikelen geschreven door Jan H. Spyridakis zijn er beschikbaar in de database?
5. Staat er in de database een tijdschrift met de titel *IEEE Transactions on Knowledge and Data Engineering*?
6. Hoeveel tijdschriften staan er in de Science Citation Index Expanded?
7. Hoe vaak werd het artikel van P.J. Schellens in *Technical Communication* geciteerd in andere artikelen?

Na de taakuitvoering werd de proefpersonen gevraagd om de schermopnamen te bekijken en hardop te denken in het Engels over de manier waarop de taken zijn uitgevoerd. De proefpersonen mochten het afspelen even stoppen wanneer ze dat wilden. Hun verbalisaties werden op cassette opgenomen. Van Someren, Barnard & Sandberg (1994) stellen dat retrospectieve data niet altijd waarheidsgetrouw zijn, vooral als er de nodige tijd zit tussen de taakuitvoering en de uitleg achteraf. Om dergelijke vertekeningen tegen te gaan, moesten de proefpersonen meteen na de taakuitvoering doorgaan met het bekijken en becommentariëren van de schermopnamen.

Vervolgens werd de proefpersonen gevraagd om enkele uitgeprinte helppagina's van *Web of Science* met behulp van de plus-en-minmethode te evalueren. In de onderzochte pagina's werd uitleg gegeven over het doel van *Web of Science*, over de tijdschriften die in de database zijn opgenomen, en over zoekmogelijkheden op de site. Eerst zetten de proefpersonen plussen en minnen in de kantlijn voor passages die ze positief dan wel negatief waardeerden. Benadrukt werd dat de proefpersonen zelf mochten uitmaken om welke redenen ze plussen en minnen plaatsten en dat ze de teksteenheid voor een plus of min ook zelf konden kiezen (variërend van individuele woorden tot een hele pagina). Nadat de proefpersonen klaar waren met het lezen en plussen en minnen zetten, werd in individuele interviews getracht om de redenen voor elke plus en min te achterhalen. Dit interview werd wederom op cassette opgenomen.

Tot slot vulden de proefpersonen een vragenlijst in, waarmee we extra gegevens verzamelden over (1) hun oordelen over de twee evaluatiemethoden, (2) de mate waarin zij zichzelf, de website of de testsituatie de schuld gaven van de problemen die ze in het retrospectieve hardopdenkonderzoek waren tegengekomen, en (3) hun plaats op het continuüm collectivistisme-individualisme. Voor dat laatste gebruikten we, bij gebrek aan beter, Hofstedes IDV-index, hoewel er, mede door Hofstede zelf, vraagtekens gezet zijn bij de validiteit van dit instrument als middel om individuele culturele oriëntaties te meten (Hofstede, 2001, p.497).

4.4 Afhankelijke variabelen in het onderzoek. Om de opbrengst van de beide evaluatiemethoden te onderzoeken (H1 en H2), werd het gemiddelde aantal geconstateerde hardopdenk- en plus-en-minproblemen per proefpersoon berekend. Bij de plus-en-minmethode werd iedere negatieve opmerking van de proefpersonen gecodeerd als probleem. Bij de retrospectieve hardopdenkprotocollen werden alle afwijkingen van het optimale handelingsverloop als probleem gecodeerd, evenals de opmerkingen die proefpersonen maakten om hun twijfel, verrassing, afkeuring en dergelijke kenbaar te maken.

De oordelen van de proefpersonen over de twee methoden (H3 en H4) werden onderzocht door drie sets vragen, alle op vijfpunts Likert schalen. Een eerste set van vijf vragen richtte zich op de ervaringen van de proefpersonen in het retrospectieve hardopdenkonderzoek (bijvoorbeeld: 'Ik voelde me ongemakkelijk bij het uitvoeren van de taken'). Een tweede set van vier vragen had betrekking op de ervaringen in het plus-en-minonderzoek (bijvoorbeeld: 'Ik vond het niet prettig om de website met deze methode te evalueren'). Een derde set van vier vragen betrof een vergelijking van de twee evaluatiemethoden (bijvoorbeeld: 'Door *Web of Science* op papier te evalueren kon ik betere aanwijzingen voor verbeteringen geven dan door de taken uit te voeren').

De mate waarin de proefpersonen zichzelf, de website of de testsituatie de schuld gaven van de problemen die ze waren tegengekomen (H5), werd onderzocht met een set van

negen vragen (weer op vijfpunts Likert schalen). De proefpersonen moesten aangeven in hoeverre ze mogelijke verklaringen voor hun gebruikersproblemen onderschreven. Drie vragen hadden betrekking op de eigen vaardigheden (bijvoorbeeld: 'doordat ik niet goed heb gelezen'), drie op de kwaliteit van *Web of Science* (bijvoorbeeld: 'door een gebrekkige gebruikersvriendelijkheid van de website') en drie op de testsituatie (bijvoorbeeld: 'doordat de onderzoeker over mijn schouder meekeek').

De vraag in hoeverre de proefpersonen zich in het retrospectieve hardopdenkonderzoek hielden aan de opgelegde gebruikersrol (H6) werd onderzocht door middel van een analyse van de hardopdenkprotocollen. Iedere afwijking van de typische gebruikersrol werd gemarkeerd en geduid in termen van de drie alternatieve rollen die we eerder onderscheidden (proefpersoon, internetgebruiker en reviewer). In de analyse keken we naar het totale aantal afwijkingen van de opgelegde gebruikersrol, maar ook naar de rollen afzonderlijk.

De directheid in de formuleringen van de commentaren (H7) werd onderzocht aan de hand van een lijst met letterlijke uitingen van proefpersonen uit de retrospectieve hardopdenkprotocollen, die vervolgens werd voorgelegd aan 12 studenten Toegepaste Communicatiewetenschap, met de opdracht om aan elke uiting een directheidsscore op een vijfpuntsschaal te geven. De twaalf studenten vormden samen een betrouwbaar beoordelingsinstrument om directe en indirecte uitingen van elkaar te onderscheiden (Cronbach's alfa = .82). De analyse beperkte zich tot de proefpersonen die in hun hardopdenkprotocollen commentaar hadden gegeven op de website. De directheidsscores per uiting werden vergeleken met de culturele achtergrond van de proefpersonen als onafhankelijke variabele.

5. Resultaten

De retrospectieve hardopdenkprotocollen en de plus-en-minmethode hebben beide veel relevante problemen in de *Web of Science* website aan het licht gebracht. Zo werd in het hardopdenkonderzoek gemiddeld 33% van de taken zonder succes uitgevoerd. Per taak varieerde dit percentage van 3% (voor de derde opdracht: het bewaren van zoekresultaten op diskette) tot 54% (voor de tweede opdracht: het uitvoeren van een zoekopdracht). Veel problemen hadden te maken met het invoeren van onderwerps- en auteursgegevens op het zoekscherm. Daarnaast bleken twee eigenaardigheden van de interface zeer contra-intuïtief voor vrijwel alle gebruikers: (1) na het gebruik van de "back"-toets van de browser moet de nieuwe webpagina vaak opnieuw geladen worden, en (2) wie na het invoeren van zoektermen de enter-toets op het toetsenbord indrukt, wist daarmee alle net ingevoerde zoekgegevens en belandt vervolgens op de homepage van *Web of Science*. Met name dat laatste zal regelmatige *Web of Science* bezoekers vermoedelijk bekend voorkomen.

Bij de verwerking van de data is één van de collectivistische proefpersonen uiteindelijk buiten de analyse gehouden, omdat hij zich bij beide methoden onttrok aan de rol die hem was toebedeeld. Ondanks de instructies weigerde hij de zeven taken uit het hardopdenkonderzoek uit te voeren en gaf hij bij de plus-en-minmethode geen enkel specifiek commentaar op de voorgelegde helppagina's. In plaats daarvan verkende hij de website en de helppagina en uitte hij zijn bewondering ervoor ("It's an exhaustive list and it's definitely very handy to the layman who does not know how and what kinds of words could be en-

tered for a particular search. Indeed very good"). Deze handelwijze kan eventueel worden gezien als een extreme variant op een collectivistisch houding.

Hieronder zullen we de resultaten van de overige 37 proefpersonen bespreken aan de hand van de zeven hypothesen die we hadden opgesteld. Voordat we daarmee beginnen, gaan we in op twee relevante achtergrondkenmerken van de proefpersonen: hun scores op de IDV-index en hun eerdere ervaringen met *Web of Science*.

5.1 Achtergrondkenmerken van de proefpersonen. Een eerste vraag met betrekking tot de achtergrondkenmerken van de proefpersonen is of er, naast hun land van herkomst, een onafhankelijke bevestiging kan worden gevonden voor de veronderstelde verschillen in culturele oriëntatie. Aan het einde van de onderzoekssessies kregen de proefpersonen daartoe de vier vragen uit Hofstedes IDV-index voorgelegd. De IDV-index van de twee groepen bleek niet significant te verschillen (t-toets, t=.554, df=35, p=.583). De Nederlandse proefpersonen scoorden precies zoals verwacht, maar de Aziatische en Afrikaanse proefpersonen scoorden individualistischer dan op grond van hun nationaliteit werd verwacht.

Een mogelijke verklaring betreft het type proefpersonen dat aan ons onderzoek heeft deelgenomen. Door selectie- en/of assimilatieprocessen zouden onze proefpersonen uit Azië en Afrika minder collectivistische kenmerken kunnen hebben dan we op voorhand hadden aangenomen. De invloed van selectieprocessen is waarschijnlijk, omdat al deze proefpersonen de ingrijpende beslissing hadden genomen om voor een wetenschappelijke carrière tijdelijk te emigreren naar een ver en onbekend land. Dergelijk avontuurlijk gedrag sluit beter aan op een individualistische oriëntatie dan op een collectivistische. Assimilatieprocessen zijn ook mogelijk, omdat alle proefpersonen inmiddels enige tijd in Nederland woonden en zich wellicht hebben aangepast aan Nederlandse normen.

Wellicht belangrijker is dat de IDV-index geen bewezen valide maat is om cultuurverschillen te meten. Het gaat om vier items die vreemd genoeg geen inhoudelijke relatie hebben met typisch individualistische of collectivistische kenmerken, en dus hoogstens als predictor kunnen fungeren. Hofstede (2001, p.497) is zelf ook niet onverdeeld optimistisch over de betrouwbaarheid en de validiteit van de index, met name voor het gebruik ervan om individuele verschillen in kaart te brengen. En in het algemeen is er twijfel aan de mogelijkheden om met behulp van vragenlijsten culturele verschillen aan te tonen (zie Peng, Nisbett & Wong 1997). Op grond van deze overwegingen blijven we de proefpersonen beschouwen als representanten van hun nationale culturen, ondanks de uitslagen van de IDV-scores. Een aanvullend argument-achteraf zullen we nog geven in paragraaf 6.2.

Een tweede relevante vraag over de twee groepen proefpersonen betreft hun eerdere ervaringen met *Web of Science*: in hoeverre zijn de beide groepen proefpersonen vergelijkbaar op dit punt? De meerderheid van de proefpersonen had *Web of Science* al eens eerder gebruikt. Er bleek hierbij echter een bijna-significant verschil te zijn tussen de individualistische en de collectivistische proefpersonen: er waren relatief minder collectivistische proefpersonen met ervaring met *Web of Science* (58% tegenover 90%, Fisher's exact test, p=.052). De intensiteit waarmee *Web of Science* in de laatste drie maanden was gebruikt, bleek vrijwel gelijk bij de twee groepen proefpersonen (1,8 tegenover 1,9 keer, t-toets, t=.103, df=26, p=.918). Omdat voorkennis van invloed zou kunnen zijn op de feedback die de proefpersonen geven (Diamantopoulos, Reynolds & Schlegelmilch, 1994), hebben we besloten om de eerdere ervaring van de proefpersonen als extra (dichotome) variabele mee te nemen in onze analyses van de ontdekte problemen.

5.2 Aantal ontdekte problemen. Bij de plus-en-minmethode verwachtten we dat de collectivistische proefpersonen minder problemen zouden noemen dan de individualistische proefpersonen (H1). Bij de retrospectieve hardopdenkprotocollen verwachtten we geen verschillen tussen de beide groepen (H2). Zoals te zien is in tabel 2, werden beide hypothesen bevestigd in dit onderzoek. De eta^2 bij de plus-en-minresultaten duidt op een substantieel verschil tussen de twee groepen proefpersonen. De eerdere ervaring met *Web of Science*, die we als extra variabele in de analyses hadden meegenomen, bleek bij beide methoden geen effect te hebben op het aantal ontdekte problemen. Er was ook geen sprake van een interactie-effect.

Tabel 2. *Gemiddeld aantal problemen per proefpersoon in Web of Science*

	Individualistisch	**Collectivistisch**	**Significantie**
Lezersproblemen in het plus-en-minonderzoek	4.8	1.8	$F(1,33)=8.97$, $p<.01$, $eta^2=.21$
Gebruikersproblemen in de de retrospectieve hardopdenkprotocollen	7.2	9.5	n.s.

Naast de specifieke problemen leverden de retrospectieve hardopdenkprotocollen ook een indicatie op van het overall succes van de proefpersonen. We gaven al eerder aan dat er relatief veel taken waren die zonder succes werden uitgevoerd. Bij een vergelijking van het aantal met succes afgeronde taken hebben we geen significante verschillen gevonden. Zowel de culturele achtergrond van de proefpersonen als hun eerdere ervaring met de website had geen invloed op het succes in de taakuitvoering.

5.3 Oordelen over de twee evaluatiemethoden. We verwachtten dat de collectivistische proefpersonen positiever zouden oordelen over de retrospectieve hardopdenkprotocollen dan de individualistische proefpersonen (H3), en dat de individualistische proefpersonen juist positiever zouden zijn over de plus-en-minmethode (H4). Beide hypothesen werden in het onderzoek niet bevestigd. De vijf vragen over de retrospectieve hardopdenkprotocollen en de vier vragen over de plus-en-minmethode vormden geen betrouwbare schalen en moesten dus apart worden geanalyseerd. Op alle vragen vonden we geen significante verschillen tussen de twee groepen proefpersonen. Over het algemeen oordeelden de proefpersonen positief over beide evaluatiemethoden.

De vier vragen waarin de proefpersonen de retrospectieve hardopdenkprotocollen en de plus-en-minmethode moesten vergelijken, vormden twee adequate schalen. De ene was gericht op de ervaringen als proefpersoon (Cronbach's alfa = .60); de andere op de inschattingen van het belang van de zelf gegeven feedback (Cronbach's alfa = .67. Ook voor deze schalen vonden we geen significante verschillen tussen de twee groepen proefpersonen, maar bij de inschatting van het belang van de gegeven feedback vonden we wel een opmerkelijke, niet verwachte tendens: de collectivistische proefpersonen neigden naar een positiever oordeel over de plus-en-minmethode dan de individualistische proefpersonen (t-toets, $t=1.975$, $df=35$, $p=.056$).

5.4 Schuldvraag bij gebruikersproblemen. We verwachtten dat collectivistische proefpersonen meer dan de individualistische proefpersonen geneigd zouden zijn om de schuld van

gebruikersproblemen bij zichzelf te zoeken (H5). Dit is onderzocht aan de hand van negen vragen in de vragenlijst aan het einde van elke sessie. Drie vragen stonden voor interne schuldtoekenning, zes vragen voor externe schuldtoekenning (waarvan drie betrekking hadden op de kwaliteit van de website en drie op de testsituatie). De vragen over externe schuldtoekenning vormden twee adequate schalen (Cronbach's alfa = .64 voor de kwaliteit van de website, en .61 voor de testsituatie); de vragen over interne schuldtoekenning niet. Zoals te zien is in tabel 3, hebben we geen significante verschillen gevonden tussen de beide groepen proefpersonen over de schuldvraag bij gebruikersproblemen. Anders dan in het onderzoek van Schriver (1997) bleken de proefpersonen in ons onderzoek niet geneigd om vooral zichzelf de schuld te geven van de problemen waar ze tegenaan liepen. De gunstige oordelen van de proefpersonen over de testsituatie maakt duidelijk dat de opzet van het retrospectieve hardopdenkonderzoek in hun ogen geslaagd was: de taken waren realistisch en de testsituatie was niet storend.

Tabel 3. *Waar ligt de schuld voor de gebruikersproblemen in het hardopdenkonderzoek?*

	Individualistisch	Collectivistisch	Significantie
Extern: de onnatuurlijke testsituatie (gemiddelde van drie items)	4.4	4.4	n.s.
Extern: de kwaliteit van het *Web of Science* (gemiddelde van drie items)	3.2	3.3	n.s.
Intern: gebrek aan ervaring met databases	4.4	4.5	n.s.
Intern: gebrek aan ervaring met het *Web of Science*	2.7	2.5	n.s.
Intern: niet goed gelezen	3.0	3.3	n.s.

N.B: Scores op een vijfpuntsschaal (1 = mee eens – 5 = mee oneens)

5.5 Proefpersoonrollen in het hardopdenkonderzoek. De volgende hypothese richtte zich op het gedrag van de proefpersonen tijdens het retrospectieve hardopdenkonderzoek. We verwachtten dat de collectivistische proefpersonen minder geneigd zouden zijn om af te wijken van de (door de methode veronderstelde) typische gebruikersrol dan de individualistische proefpersonen (H6). Daarbij onderscheidden we drie mogelijke andere rollen die de proefpersonen zouden kunnen aannemen: proefpersoon, internetgebruiker en reviewer. Hieronder hebben we uit de protocollen een aantal voorbeelden van elke rol opgenomen.

Uitingen vanuit de rol van proefpersoon:
I was doing a few checks. That's why I didn't complete all the questions.
I learned something from this today. I should pay more attention.
It's a stupid way of doing it. But it's the only way I know. I'm getting annoyed with myself.
Don't look at this. This is so stupid. I even typed it wrong!

Uitingen vanuit de rol van internetgebruiker:
Normally, I never use save.
Normally, I would ask someone else whether they know how to do it.
I always press back on the browser, not on the page.
That's what I usually do. In my field of research …

Uitingen vanuit de rol van reviewer:
I think this shouldn't be so strict.
That's not very good that you have to use a list. How difficult can it be to implement…
That was the main problem: enter.
It's not one of my favorite databases.

De resultaten zijn te vinden in tabel 4. Wanneer gekeken wordt naar alle afwijkingen van de standaard-gebruikersrol, wordt de hypothese bevestigd: de collectivistische proefpersonen hielden zich meer aan de gebruikersrol dan de individualistische proefpersonen. De gevonden waarde voor Cohen's d duidt op een gemiddeld tot groot effect. Kijken we naar de afzonderlijke niet-gebruikersrollen, dan is er geen sprake van significante verschillen. Op grond van dit resultaat mag verwacht worden dat individualistische proefpersonen aan een usability test een grotere diversiteit aan problemen naar voren zullen brengen dan collectivistische proefpersonen. Collectivistische proefpersonen houden zich meer aan de gebruikersrol die voortvloeit uit de taken die ze krijgen.

Table 4. *Gemiddeld aantal afwijkingen van de gebruikersrol in de hardopdenkprotocollen*

	Individualistisch	**Collectivistisch**	**Significantie**
Proefpersoon	3.1	2.2	n.s.
Internetgebruiker	2.1	1.1	n.s.
Reviewer	1.4	0.7	n.s.
Totaal	6.6	4.0	t-toets (eenzijdig), t=1.818, df=35, p<.05, Cohen's d=0.61

5.6 Directe en indirecte formuleringen bij het geven van commentaar. We verwachtten dat proefpersonen uit collectivistische landen vaker dan de individualistische proefpersonen zouden kiezen voor indirecte en eufemistische formuleringen (H7). Een eerste opmerkelijke bevinding betreft het aantal commentaren dat door de twee groepen proefpersonen is gemaakt. Geheel in lijn met hypothese 6 was het overgrote deel van de commentaren in de retrospectieve hardopdenkprotocollen afkomstig van individualistische proefpersonen (70% tegenover 30%).

Hieronder zijn voorbeelden van directe en indirecte commentaren opgenomen (samen met de gemiddelde scores op de schaal direct-indirect (1-5), die door twaalf studenten Toegepaste Communicatiewetenschap zijn toegekend). Er bleek sprake van een significant verschil tussen de beide groepen, in de richting van onze hypothese. Commentaren afkomstig van collectivistische proefpersonen hadden een gemiddelde score op de schaal direct-indirect van 2,8; commentaren van individualistische proefpersonen eindigden met een gemid-

delde score van 2,2. Het verschil is statistisch significant (t-toets, t=2.507, df=37, p<.05) en correspondeert met een groot effect (Cohen's d = .89).

Directe formuleringen van commentaar:
Stupid thing (score 1,1).
That's very annoying (score 1,2).
Doesn't work properly. Annoying (score 1,3).
Of course, it still didn't work (score 1,4)

Indirecte formuleringen van commentaar:
There's something strange … I guess it's not important (3,5)
It's not the most convenient way … (3,3)
I think this shouldn't be so strict (3,2)
It's not one of my favorite databases (3,2)

6. Conclusies en discussie

In deze paragraaf zullen we eerst conclusies trekken over de effecten die culturele verschillen op de dimensie individualisme-collectivisme hebben op de feedback die verzameld wordt in een pretest. Daarna bespreken we het opvallende verschil in ons onderzoek tussen het feitelijke gedrag van proefpersonen en hun antwoorden op opiniërende vragen. Tot slot gaan we, op grond van onze ervaringen, in op de problematiek van intercultureel onderzoek.

6.1 Culturele invloeden op website-evaluatie. De belangrijkste conclusie uit ons onderzoek is dat de culturele achtergrond van de proefpersonen inderdaad een factor is die van invloed kan zijn op de feedback die in een website-evaluatie wordt verzameld. Hoewel de IDV-index geen verschillen liet zien tussen de twee groepen in ons onderzoek, bleken de proefpersonen die op grond van hun nationaliteit een collectivistische respectievelijk individualistische cultuur vertegenwoordigden, zich in een aantal gevallen overeenkomstig de verwachtingen te gedragen. Het advies dat in veel webdesign-literatuur wordt gegeven om internationale websites te evalueren met gebruikersgroepen uit verschillende landen blijft waardevol, maar daarbij moet de kanttekening worden gemaakt dat de waarde van de evaluatie*methoden* die in dergelijk onderzoek gebruikt worden, eveneens cultuurafhankelijk kan zijn.

Dat geldt met name voor de plus-en-minmethode, die bij proefpersonen uit individualistische, weinig contextgevoelige culturen meer commentaren uitlokt dan bij proefpersonen uit collectivistische, contextgevoelige culturen. Opvallend genoeg kwam dit duidelijke verschil niet tot uiting in de oordelen die de collectivistische proefpersonen gaven over de plus-en-minmethode: ze waren zelf verhoudingsgewijs optimistisch over de waarde van hun plus-en-mincommentaar. In het tot dusverre beschikbare onderzoek, waarin de plus-en-minmethode steeds als waardevol pretestinstrument naar voren komt, werd altijd gewerkt met West-Europese of Amerikaanse proefpersonen (zie De Jong 1998). Op grond van de

resultaten van het hier gepresenteerde onderzoek lijkt de conclusie gerechtvaardigd dat de methode wellicht minder geschikt is in collectivistische, contextgevoelige culturen. Overigens is in ons onderzoek alleen de kwantiteit van het commentaar gemeten; het is niet uit te sluiten dat de vergelijking anders uitvalt als ook de kwaliteit van het plus-en-mincommentaar wordt meegenomen.

De retrospectieve hardopdenkprotocollen lijken minder te worden beïnvloed door culturele verschillen tussen proefpersonen, doordat de fouten die proefpersonen maken de ruggengraat van de methode vormen en er op voorhand geen redenen zijn om te veronderstellen dat proefpersonen uit de ene cultuur meer fouten maken dan proefpersonen uit de andere cultuur. Toch vonden we ook hier twee effecten van de culturele achtergrond van de proefpersonen. Ten eerste waren de proefpersonen uit de collectivistische, contextgevoelige culturen meer geneigd om dicht bij de in het onderzoek opgelegde gebruikersrol te blijven. Waar de proefpersonen uit de individualistische, weinig contextgevoelige culturen regelmatig commentaren gaven vanuit andere perspectieven (kritiek geven op de website vanuit de rol van reviewer, commentaar geven op de testsituatie of reflecteren op de dingen die ze normaal gesproken met het internet of met databases zouden doen), waren de proefpersonen uit de collectivistische, contextgevoelige culturen meer geneigd om in het hele hardopdenkonderzoek de aangereikte gebruikersrol aan te houden. Men zou kunnen zeggen dat ze zich als de ideale proefpersonen voor een usability test gedroegen, maar ook het commentaar van proefpersonen die even buiten de gebruikersrol traden, bevatte vaak waardevolle suggesties voor de webdesigner. Ten tweede bleken de proefpersonen uit de collectivistische, contextgevoelige culturen geneigd om hun commentaren minder direct te formuleren dan de individualistische proefpersonen. Dit zou in de praktijk consequenties kunnen hebben in de revisiefase, waarin de onderzoeker(s) en webdesigner(s) moeten inschatten hoe ernstig de ontdekte problemen zijn.

Op een wat abstracter niveau blijkt het onderscheid tussen collectivistische, contextgevoelige en individualistische, weinig contextgevoelige culturen, zoals geoperationaliseerd door Ting-Toomey (1998), een vruchtbare benadering om voorspellingen te doen over verschillen tussen West-Europese en Aziatische of Afrikaanse proefpersonen. Niet alle hypothesen werden in ons onderzoek bevestigd, maar er was duidelijke ondersteuning voor vier van de hypothesen die we op grond van Ting-Toomey hadden opgesteld.

6.2 Verschillen tussen gedrag en zelfrapportage. Een opmerkelijke discrepantie in onze data betreft het verschil tussen gedrag en zelfrapportage. De drie hypothesen die in ons onderzoek niet werden bevestigd (namelijk over de waardering van de twee gebruikte evaluatiemethoden en de schuldvraag bij gebruikersproblemen) zijn getoetst door middel van zelfrapportage. De drie significante verschillen die we wél hebben gevonden (het aantal plus-en-minproblemen, de afwijkingen van de opgelegde gebruikersol en de directheid van de commentaren) waren gebaseerd op het feitelijke gedrag van de proefpersonen. Dit verschil kan verklaard worden aan de hand van Hofstedes (1994) 'ui-diagram' van cultuurkenmerken, dat veronderstelt dat de culturele oriëntatie van mensen gelaagd is (met waarden als de kern en symbolen als de buitenste laag). Het gedrag van de proefpersonen lijkt een meer fundamentele laag van hun culturele oriëntatie te vertegenwoordigen dan de antwoorden die ze geven in een vragenlijst. De antwoorden in de vragenlijst kunnen vertekend zijn door sociale wenselijkheid. En sociale wenselijkheid blijkt ook weer een factor die gerelateerd is aan culturele kenmerken: onderzoek van Middleton & Jones (2000) maakt

duidelijk dat de invloed van sociale wenselijkheid groter is in collectivistische culturen dan in individualistische culturen. Zo geredeneerd is te verwachten dat de invloed van de culturele achtergrond op het gedrag en met name de attitude van proefpersonen in een pretest groter is wanneer die proefpersonen worden onderzocht in hun land van herkomst, zonder dat selectie- en assimilatieprocessen invloed hebben gehad.

Deze invloed van sociale wenselijkheid zou ook kunnen gelden voor de scores op de IDV-index, die immers ook berusten op zelfrapportage. Dit is een reden te meer om aan het ontbreken van verschillen op deze index niet te veel consequenties te verbinden.

6.3 Intercultureel communicatieonderzoek. Onze ervaringen in dit onderzoek roepen twijfel op over twee populaire benaderingen in intercultureel onderzoek. De eerste betreft het inschakelen van immigranten als getrouwe representanten van de cultuur van hun land van herkomst. Als gevolg van selectie- en assimilatieprocessen kunnen deze immigranten behoorlijk afwijken van de culturele oriëntatie in hun oorspronkelijke land. Deze afwijkingen zullen het eerst merkbaar zijn in vragenlijstonderzoek. Natuurlijk hangt veel af van het doel van het onderzoek: als het niet gaat om wereldwijde culturele verschillen maar om verschillen binnen de nationale context, zoals in het onderzoek van Lentz & Hulst (2000), is er natuurlijk geen probleem. De tweede twijfel betreft het gebruik van vragenlijsten, zoals de IDV-index, om culturele verschillen te meten. Dergelijke instrumenten richten zich primair op de buitenste lagen van cultuur en dringen niet gemakkelijk door tot de waarden en gewoonten die voor een belangrijk deel bepalend zijn voor het gedrag.

Hoewel we op grond van ons onderzoek een aantal duidelijke conclusies hebben kunnen trekken over de invloed van culturele verschillen op website-evaluatie, hebben onze ervaringen ons ook bewust gemaakt van de geweldige complexiteit van intercultureel onderzoek. Culturele verschillen kunnen op veel verschillende manieren invloed hebben, zowel op de communicatie als op het onderzoek zelf. Een vergelijking met de bekende Russische matruschkapoppen ligt voor de hand: als de eerste pop geopend wordt, verschijnt er een tweede, identiek maar een beetje kleiner; als de tweede pop geopend wordt, komt de derde te voorschijn, etcetera. Ons interculturele onderzoek behelsde het openen van de tweede matruschkapop, maar terwijl we daarmee bezig waren, werden we ons al bewust van de derde en de vierde. Zoals we eerder stelden, spelen interculturele verschillen mogelijk ook weer een rol bij de manier waarop de proefpersonen vragenlijsten invullen (sociale wenselijkheid). En weer een ander onderzoek dat we recentelijk op het spoor kwamen, laat zien dat de culturele overeenkomst tussen proefleider en proefpersonen mogelijk ook van belang is. In het onderzoek van Vatrapu (2002) bleken usability interviews vruchtbaarder als de onderzoeker en de proefpersoon een gedeelde culturele achtergrond hadden. In ons onderzoek werden alle sessies 'gewoon' door dezelfde Nederlandse onderzoekster geleid.

We hopen met dit onderzoek te hebben aangetoond dat multiculturele aspecten relevant zijn voor de opzet en uitvoering van een website-evaluatie. Naast de bovengenoemde methodologische complicaties schuilt er in dit type onderzoek nog een ander probleem: het gevaar dat de resultaten leiden tot een stereotiep beeld van de groepen proefpersonen en hun culturele oriëntatie. We benadrukken dat we het hier hebben over verschillen tussen groepen, niet tussen individuen. Ondanks alle mogelijke bezwaren zijn we van mening dat intercultureel onderzoek naar website-evaluatie in de huidige context van internationalisering essentieel is voor een beter begrip van de waarde en beperkingen van gangbare evaluatiemethoden.

Bibliografie

Arnold, M. (1998). Building a truly World Wide Web: A review of the essentials of international communication. *Technical Communication, 45*, 197-206.

Barnum, C.M. (2002). *Usability testing and research.* New York: Longman.

Becker, S.A. (2002). An exploratory study on Web usability and the internationalization of US e-business. *Journal of Electronic Commerce Research, 3*, 265-278.

Brown, P. & S.C. Levinson (1990). *Politeness: Some universals in language usage*. Cambridge: Cambridge University Press.

Diamantopoulos, A., N. Reynolds & B. Schlegelmilch (1994). Pretesting in questionnaire design: The impact of participant characteristics on error detection. *Journal of Marketing Research, 36*, 295-311.

Dumas, J.C. & J.S. Redish (1993). *A practical guide to usability testing*. Norwood, NJ: Ablex.

Haak, M. van den, M. de Jong & P.J. Schellens (2003). Hardopdenkprotocollen als pretestmethode: Synchroon en retrospectief hardopdenken vergeleken. *Tijdschrift voor Taalbeheersing, 25*, 236-252

Hall, E. T. (1977). *Beyond culture*. Garden City, NY: Anchor Press/Doubleday.

Hoc, J.M. & J. Leplat (1983). Evaluation of different modalities of verbalisation in a sorting task. *International Journal of Man-Machine Studies, 18*, 283-306.

Hofstede, G. (1994). *Culture and organizations: Software of the mind*. London: Harper Collins.

Hofstede, G. (2001). *Culture's consequences: Comparing values, behaviours, institutions, and organizations across nations*. Second edition. Beverly Hills, CA: Sage.

Hoft, N. (1995). *International technical communication: How to export information about high technology*. New York: John Wiley.

Jong, M. de (1998). *Reader feedback in text design. Validity of the plus-minus method for the pretesting of public information brochures*. Amsterdam: Rodopi.

Jong, M. de & P.J. Schellens (1995). *Met het oog op de lezer. Pretestmethoden voor schriftelijk voorlichtingsmateriaal.* Amsterdam: Thesis.

Jong, M. de & P.J. Schellens (1997). Reader-focused text evaluation: An overview of goals and methods. *Journal of Business and Technical Communication, 11*, 402-432.

Jong, M. de & P.J. Schellens (2000). Toward a document evaluation methodology: What does research tell us about the validity and reliability of methods? *IEEE Transactions on Professional Communication, 43*, 242-260.

Jong, M. de & P.J. Schellens (2001). Readers' background characteristics and their feedback on documents: The influence of gender and educational level on evaluation results. *Journal of Technical Writing & Communication, 31*, 267-281.

Jong, M. de & P.J. Schellens (2002). Tekstevaluatie. Onderzoek naar de validiteit van probleemopsporende methoden. *Tijdschrift voor Taalbeheersing, 24*, 146-166.

La Ferle, C., S.M. Edwards & Y. Mizuno (2002). Internet diffusion in Japan: Cultural considerations. *Journal of Advertising Research, 42*, 2, 65-79.

Lentz, L. & J. Hulst (2000). Babel in document design: the evaluation of multilingual texts. *IEEE Transactions on Professional Communication, 43*, 313-322.

Marcus, A. & E.W. Gould (2000). Cultural dimensions and global web user-interface design: What? So what? Now what? *Proceedings of the 6th Conference on Human Factors and the Web*. http://www.tri.sbc.com/hfweb/marcus/hfweb00_marcus.html.

Middleton, K.L., & J.L. Jones (2000). Socially desirable response sets: the impact of country culture. *Psychology & Marketing, 17*, 149-163.

Nielsen, J. (1993). *Usability engineering*. Boston, MA: Academic Press.

Nielsen, J. (2000). *Designing Web usability: The practice of simplicity*. Indianapolis, IN: New Riders.

Okazaki, S. & J.A. Rivas (2002). A content analysis of multinationals' Web communication strategies: Cross-cultural research framework and pre-testing. *Internet Research, 12*, 380-390.

O'Keefe, R.M., e.a. (2000). From the user interface to the consumer interface: Results from a global experiment. *International Journal of Human-Computer Studies, 53*, 611-628.

Peng, K., R.E. Nisbett & N.Y.C. Wong (1997). Validity problems comparing values across cultures and possible solutions. *Psychological Methods, 2*, 329-344.

Rubin, J. (1994). *Handbook of usability testing. How to plan, design, and conduct effective tests*. New York: John Wiley.

Schriver, K.A. (1989). Evaluating text quality: The continuum from text-focused to reader-focused methods. *IEEE Transactions on Professional Communication, 32*, 238-255.

Schriver, K.A. (1997). *Dynamics in document design. Creating text for readers*. New York: John Wiley.

Schweibenz, W. & F. Thissen (2003). *Qualität im Web. Benutzer-freundliche Webseiten durch Usability Evaluation*. Berlin: Springer.

Sienot, M. (1997). Pretesting Web sites. A comparison between the plus-minus method and the think aloud method for the World Wide Web. *Journal of Business and Technical Communication, 11*, 469-482.

Simon, J.S. (2001). The impact of culture and gender on Web sites. *Data Base for Advances in Information Systems, 32*, 18-37.

Smith, P.B. & M.H. Bond (1998). *Social psychology across cultures*. Second edition. London: Prentice Hall.

Someren, M.W van, Y.F. Barnard & J.A.C. Sandberg (1994). *The think aloud method. A practical guide to modelling cognitive processes*. London: Academic Press.

Ting-Toomey, S. (1998). Intercultural conflicts styles. A face-negotiation theory. In: Y.Y. Kim & W.B. Gudykunst (eds.), *Theories in intercultural communication* (pp. 213-235). Newbury Park, CA: Sage.

Trompenaars, F. & C. Hampden-Turner (1998). *Riding the waves of culture. Understanding cultural diversity in global business*. New York: McGraw-Hill.

Vatrapu, R. (2002). Culture and international usability testing: The effects of culture in interviews. Master's thesis, Virginia Polytechnic Institute and State University. http://scholar.lib.vt.edu/theses/available/etd-09132002-083026/unrestricted/Vatrapu_Thesis.pdf.

Versseveld, C. van (1995). Betrokkenheid bij voorlichting. De invloed van betrokkenheid op pretest-commentaar. *Tekst[blad], 1* (3), 40-44.

Waes, L. van (2000). Thinking aloud as a method for testing the usability of Web sites: The influence of task variation on the evaluation of hypertext. *IEEE Transactions on Professional Communication, 43*, 279-291.

Zahedi, F., W.V. van Pelt & J. Song (2001). A conceptual framework for international Web design. *IEEE Transactions on Professional Communication, 44*, 83-103.

Leo Lentz en
Menno de Jong

In welke termen denken lezers over tekstproblemen?

Samenvatting

In diverse methoden voor tekstevaluatie leggen we aan proefpersonen categorieën voor van lezersproblemen, zoals *stijl, structuur* en *begrip*. Soms in de vorm van schaaltjes waarop zij een score moeten aangeven of in open vragen zoals: *"Wat vindt u van de stijl van deze tekst?"* Soms ook in de vorm van aandachtspunten waarop lezers hun commentaar moeten richten. Onduidelijk is echter welke betekenis lezers die niet elke dag met tekstkwaliteit bezig zijn, aan deze categorieën toekennen. In welke termen formuleren zij zelf probleemtypen? En welke betekenis kennen zij toe aan die termen? In dit artikel rapporteren wij over een onderzoek waarin aan vijftig proefpersonen gevraagd is een ordening aan te brengen in een reeks problemen die andere lezers bij eerder onderzoek spontaan gerapporteerd hebben. De resultaten maken duidelijk dat men gemiddeld vijf tot negen probleemtypen onderscheidde. Deze verschillende probleemtypen worden besproken en met voorbeelden geïllustreerd.

1. Inleiding

In welke termen denken leken over lezersproblemen? Wanneer zij moeite hebben met de interpretatie van een passage in een brochure, zien zij dat dan als een *begripsprobleem*? En beschouwen zij een betuttelende toon als een kwestie van *stijl*? Of zijn dat termen voor categorieën die wij als experts in tekstkwaliteit de afgelopen decennia hebben ontwikkeld om typen lezersproblemen aan te duiden, maar die niet aansluiten op de denkwereld van gewone lezers? In de literatuur zijn verschillende overzichten van categorieën gepresenteerd. De Jong en Schellens (2000) noemen in een recent overzicht bijvoorbeeld de volgende categorieën:

- begrip: hieronder vallen problemen op het niveau van een woord, zin of tekstfragment met het begrijpen en toepassen van de informatie;
- acceptatie: lezers betwisten als feiten gepresenteerde eenheden in de tekst, zijn het oneens met waardeoordelen of wijzen adviezen af;
- waardering: lezers geven de voorkeur aan een aantrekkelijker formulering zonder een probleem te ervaren met begrip of acceptatie;

- structuur: lezers hebben een probleem met de volgorde waarin eenheden gepresenteerd worden of met de markering van de structuur;
- relevantie: lezers vinden bepaalde informatie overbodig of een onderdeel te ver uitgewerkt;
- volledigheid: lezers vragen om meer informatie over een specifiek punt of een verdere uitwerking daarvan;
- vormgeving: lezers geven kritiek op de uiterlijke vorm van de tekst of op afbeeldingen in het document;
- correctheid: lezers geven kritiek op een passage wat betreft de spelling, interpunctie, zinsbouw of andere tekstconventies.

Dit overzicht is gepresenteerd met de bedoeling tot een categorisering te komen van allerlei mogelijke problemen die lezers in teksten kunnen ervaren. Andere overzichten, zoals in Maes, Ummelen en Hoeken (1996) zijn bedoeld om oordelen van lezers over teksten te genereren. Ook daar vinden we een categorisering van dimensies:

begrijpelijkheid	makkelijk	–	moeilijk
	eenvoudig	–	ingewikkeld
	duidelijk	–	onduidelijk
	overzichtelijk	–	onoverzichtelijk
	logisch opgebouwd	–	onlogisch opgebouwd
	bondig	–	omslachtig
aantrekkelijkheid	interessant	–	oninteressant
	afstandelijk	–	aansprekend
	afhoudend	–	uitnodigend
	saai	–	boeiend
	onpersoonlijk	–	persoonlijk
	eentonig	–	afwisselend

Blijkbaar vallen *opbouw* en *overzichtelijkheid* hier onder een categorie van een hogere orde zoals *begrijpelijkheid*, terwijl het verwante *structuur* bij Schellens en De Jong een zelfstandige categorie is. Wat zijn zinvolle rubrieken voor tekstevaluatie? Hoe ordenen en benoemen we die? En hoe werken we ze uit in afzonderlijke items voor vragenlijsten die getest zijn op betrouwbaarheid?

Een antwoord op deze vragen is niet goed mogelijk zonder een grondige bezinning op de vraag hoe leken die niet elke dag met tekstevaluatie bezig zijn, zelf denken over problemen met teksten. Welke categorieën onderscheiden zij? Hoe ordenen en benoemen zij die? Met iedere vragenlijst, met elke schaal en met elk lijstje met aandachtspunten doen we immers een beroep op hun interpretatie van onze termen. Ruim twintig jaar geleden vroeg Cicourel (1982) in een klassiek geworden artikel al aandacht voor het probleem van de ecologische validiteit in allerlei onderzoeksmethoden waarvoor geldt dat zij

> "[...] *presuppose something about the way people are able to analyze textual materials. For example, in responding to questions in a reading test, the respondent must utilize several sources of knowledge that the researcher used in interviews and surveys, therefore, presumes a theory of communication and comprehension that is seldom addressed […].*"

In het onderzoek[1] waarover we hier rapporteren, staat dat perspectief van validiteit centraal. Ons gaat het daarbij met name om de begripsvaliditeit van de instrumenten die in evaluatieonderzoek gangbaar zijn: vragenlijsten die met behulp van schaaltjes de waardering van teksten meten of met open vragen een oordeel ontlokken over bijvoorbeeld de *stijl* of *geloofwaardigheid* van een document. Maar ook voor een evaluatiemethode als Focus is de vraag relevant, omdat proefpersonen met dat software-programma bij elk commentaar zelf een passende probleemcategorie moeten kiezen.

2. Opzet van het onderzoek

In essentie hebben de geformuleerde onderzoeksvragen betrekking op de categorisering van problemen die de lezer zelf ervaart bij het lezen van een tekst. Deze gedachte leidt tot een mogelijke opzet waarbij elke proefpersoon zijn of haar eigen commentaren op een tekst in probleemtypen onderverdeelt. Een nadeel van die werkwijze is dat elke proefpersoon met specifieke eigen problemen komt. De ervaring leert ons dat in een pretest met bijvoorbeeld dertig proefpersonen hoogstens een vijftal problemen door meer dan tien proefpersonen gezamenlijk genoemd zou worden, en geen enkel probleem door meer dan twintig. Over de meeste problemen zou derhalve in zo'n opzet niet meer dan één uitspraak gedaan worden.

Wij kozen daarom voor een andere opzet, waarbij de lezersproblemen vooraf gegeven zijn en de proefpersonen dus een categorisering maken van *andermans* commentaren. We hebben twee groepen van 25 proefpersonen geconfronteerd met een deel van de feedback die in eerder evaluatieonderzoek verzameld was, verdeeld over telkens 40 kaartjes met op elk kaartje een nummer van de regel van de tekst en het probleem zoals lezers dat eerder verwoord hadden.

De proefpersonen lazen allereerst de tekst rustig door en namen vervolgens het eerste kaartje van de stapel en legden dat voor zich neer. Daarna pakten ze het tweede kaartje en stelden zichzelf de vraag of dat een vergelijkbaar soort probleem weergaf of dat het echt iets anders was. In het laatste geval legden ze het ernaast. Bij het derde kaartje gingen zij na of dat probleem beter bij het eerste of het tweede paste of dat het weer een nieuw soort probleem was. Zo vormden zij gaandeweg stapeltjes van lezersproblemen die naar hun idee bij elkaar hoorden. Tussentijds mochten ze zo vaak als ze wilden kaartjes veranderen van stapel. Ze mochten ook stapels samenvoegen en splitsen. Als alle kaartjes geordend waren, vroeg de proefleider of ze nog iets wilden veranderen. Zo nee, dan startte de tweede fase van de opdracht.

Er waren nu een aantal stapeltjes gevormd met lezersproblemen. Voor elk stapeltje werden de volgende vragen gesteld:

- Waarom horen deze kaartjes bij elkaar? Wat hebben ze gemeenschappelijk?
- Hoe zou u dit groepje problemen noemen, als u er een woord voor zou moeten kiezen?
- Is dat een term die u vaker gebruikt in het dagelijks leven?
- Welk kaartje vindt u het beste voorbeeld van die groep? En welke het slechtste?
- Kunt u een omschrijving geven van deze term? Hoe zou een definitie luiden?

Op deze manier werden de proefpersonen in feite gestuurd om in eigen termen een semantische analyse te maken van de categorieën die ze zojuist zelf gecreëerd hadden. Aldus ontstond voor elke proefpersoon een dataset met in groepjes geordende kaartjes, en voor elke categorie een term plus een karakterisering van de kenmerkende essentie.

2.1 Het materiaal. In twee sessies is gewerkt met verschillend materiaal. De eerste keer is gewerkt met een persuasieve tekst (*Kijk uit voor je huid*), waarop feedback was verkregen met behulp van het softwareprogramma Focus. De tweede keer is gewerkt met een meer informatieve en instructieve tekst (*Je eerste baan*), waarop feedback was verzameld met de plus-en-minmethode. De spreiding over de genres is nagestreefd teneinde te voorkomen dat we uitsluitend met categorieën te maken zouden krijgen die sterk aan een bepaald domein gebonden zouden zijn. De eerste tekst komt uit de gezondheidsvoorlichting en waarschuwt voor huidkanker als gevolg van zonnebrand. De tweede tekst is afkomstig van de Belastingdienst en geeft informatie aan jongeren die voor het eerst een baan krijgen en daardoor in fiscaal vaarwater terechtkomen.

Beide teksten zijn in eerder onderzoek met verschillende methoden geëvalueerd. De belastingtekst is onderworpen aan de plus-en-minmethode in het promotieonderzoek van De Jong (1998). De andere tekst is in eerder Utrechts onderzoek[2] geëvalueerd met behulp van het softwareprogramma Focus[3]. Proefpersonen die met Focus werkten, kregen de tekst op scherm aangeboden en konden in een aparte kolom hun commentaar noteren.

De pretest met Focus leverde 148 verschillende problemen op; de plus-en-minmethode leverde 157 problemen op. Omdat het niet realistisch is proefpersonen ongeveer 150 kaartjes voor te leggen, is voor beide teksten een selectie van veertig commentaren gemaakt. Daarbij zijn de volgende criteria gehanteerd.

- Het commentaar moest helder geformuleerd zijn, omdat een onbegrijpelijke reactie nu eenmaal niet goed te categoriseren valt.
- We hebben vermeden dat twee kaartjes naar dezelfde passage en een soortgelijk probleem verwijzen, omdat de toekenning van het ene kaartje aan een categorie dan van invloed zou kunnen zijn op de categorisering van het tweede kaartje; elk probleem moest dus duidelijk herkenbaar en uniek zijn.
- Problemen die door meer dan één proefpersoon genoemd zijn, hadden de voorkeur boven unieke probleemdetecties.
- Er is alleen gebruik gemaakt van commentaar dat betrekking had op tekst; reacties op illustraties en vormgeving zijn niet meegenomen.
- Het probleem moest duidelijk gelokaliseerd kunnen worden in de tekst, zodat de proefpersonen de betreffende passage (via de regelnummering) snel zouden kunnen vinden.

- De problemen moesten betrekking hebben op redelijk over de tekst verspreide passages en een grote variëteit aan problemen weerspiegelen.

De laatste twee criteria verdienen enige toelichting. Het vijfde criterium is gekozen om te vermijden dat proefpersonen langdurig op zoek zouden gaan naar de passage waar het commentaar betrekking op heeft. Dit leidt echter tot het gevaar dat op die manier de meer globale reacties buiten de categorisering blijven. En in het laatste criterium schuilt een groot gevaar van circulariteit. Wie alleen problemen met geloofwaardigheid en begrip selecteert, moet niet verbaasd opkijken als veel proefpersonen de kaartjes keurig in die beide categorieën sorteren, ofschoon menigeen een geheel eigen categorisering zal kiezen. Hier is onzes inziens echter sprake van een dilemma: wie geen aandacht schenkt aan spreiding van commentaren, loopt het gevaar van een toevallige eenzijdige selectie; en wie wel vooraf de commentaren wil spreiden over een brede waaier aan probleemtypen loopt het gevaar van een cirkelgang. Wij hebben in dit onderzoek voor beide opties gekozen: de feedback op de gezondheidstekst is mede geselecteerd op het criterium van spreiding volgens de categorieën van Focus; de feedback op de belastingtekst is vooral geselecteerd op grond van het derde criterium - reacties die door minimaal twee respondenten zijn geformuleerd, zonder rekening te houden met spreiding.

2.2 Proefpersonen en afname. Aan elk van beide sessies namen 25 proefpersonen deel. Alle proefpersonen waren hoger opgeleid (vwo-, hbo- of universitair niveau), maar hadden geen speciale expertise op het gebied van tekstschrijven. De leeftijd van de proefpersonen varieerde van 22 tot 66 jaar. In totaal namen 22 mannen en 28 vrouwen aan de sessies deel.

We hebben de procedure bij de afname hierboven al kort geschetst:

- eerst de tekst rustig lezen,
- dan de kaartjes een voor een op stapeltjes leggen,
- vervolgens elk stapeltje omschrijven met behulp van een term, een goed en slecht voorbeeld en een definitie.

Dit gebeurde in een rustige één-op-één situatie bij de proefleiders thuis, bij proefpersonen thuis en soms ook op een rustig moment op het werk van de proefpersoon. Elk kaartje was voorzien van een nummer. De proefleider noteerde de nummers waaruit de stapeltjes uiteindelijk bestonden op een scoreformulier en noteerde daar tevens de antwoorden op de vragen. Gemiddeld duurde de afname ongeveer een uur per proefpersoon.

2.3 De analyse. Van elk genummerd kaartje is in een tabel aangegeven hoe vaak het samengevoegd is met elk van de overige 39 kaartjes. Aldus ontstaat voor elk kaartje een rangorde met de meest frequente combinaties. In tabel 1 ziet u als voorbeeld vier rangordes met de hoogste frequenties voor de kaartjes 9, 24, 25 en 6.

Tabel 1. *Rangordes van combinaties van vier kaartjes met tussen haakjes de frequentie waarmee die combinatie is gemaakt door proefpersonen (N=25)*

9	**24**	**25**	**6**
25 (23)	25 (20)	9 (23)	24 (18)
24 (19)	9 (19)	24 (20)	9 (17)
6 (17)	6 (18)	6 (17)	25 (17)
14 (9)	33 (8)	14 (9)	33 (9)
36 (9)	14 (7)		2 (8)
33 (8)	36 (7)		4 (8)
26 (7)			14 (7)

Kaartje 9 is dus 23 keer gecombineerd met kaartje 25 en is 19 keer met kaartje 24 gecombineerd. In de volgende kolommen zien we dat die beide kaartjes ook weer vaak met elkaar zijn gecombineerd. Aldus ontstaat in deze tabel een beeld waarin de kaartjes 9, 24, 25 en 6 een kerngroep vormen. Het kaartje met nummer 14 wordt daar niet meer toe gerekend, omdat het verval in de frequentie tussen kaartje 6 (minimaal 17 keer gecombineerd in deze groep) en kaartje 14 (maximaal 9 keer) precies op deze grens intreedt. Een kerngroep van combinaties is vastgesteld telkens wanneer zo'n verval zichtbaar werd. Soms was echter geen plotseling verval zichtbaar omdat de frequenties geleidelijk terugliepen; in die gevallen is afhankelijk van de hoogste frequentie een streep getrokken ergens tussen een frequentie van 6 tot 10.

Nadat zulke kerngroepen waren gevormd, is vervolgens nagegaan welke proefpersonen die combinatie in een stapeltje gemaakt hadden. In het bovenstaande voorbeeld zijn dus de gegevens verzameld van alle proefpersonen die de combinatie 9, 24, 25 en 6 gemaakt hebben. Van die proefpersonen zijn alle namen genoteerd die aan de combinatie zijn gegeven en de omschrijvingen die daarvoor zijn geformuleerd. Bovendien is genoteerd welk kaartje als beste vertegenwoordiger van die categorie genoemd is. Aldus ontstond voor elke kerngroep een reeks benamingen en omschrijvingen. De analyse van die gegevens leidde tot de resultaten die in het vervolg van dit artikel gerapporteerd worden.

3. Resultaten

Hoeveel groepjes creëerden de proefpersonen? De gemiddelden verschilden in de beide settings. De kaartjes met reacties op de belastingtekst werden gemiddeld in 5.4 categorieën verdeeld, terwijl de feedback op de gezondheidstekst in gemiddeld 8 stapeltjes werd verdeeld. Het kleinst aantal categorieën was 3 (door vier proefpersonen) en het grootste aantal was 11 (door twee proefpersonen). De meeste proefpersonen maakten tussen de vijf en negen groepjes.

Wat zijn dat voor groepjes? Hieronder vatten we voor de beide settings de resultaten samen per categorie. Niet elke categorie was echter even duidelijk. Sommige probleemcategorieën vertoonden duidelijk minder samenhang dan andere. We beginnen met de vijf categorieën waarover grote duidelijkheid bestond.

3.1 Overbodige Informatie. In de teksten over huidkanker en belasting zaten de volgende kaartjes die tot de kerngroep gerekend werden.

Kijk uit voor je huid	**Je eerste baan**
Leuk om te weten, maar je hebt er niks aan.	*De mededeling dat de belasting wordt afgedragen staat er twee keer in.*
Het lijkt me dat iedereen dit wel weet.	*De informatie over terugvragen staat er twee keer in.*
Dit is overbodig als het niet wordt toegelicht.	*In dit tekstblok staat twee keer dezelfde zin over aanvragen.*
Dit is al eerder gezegd, kan dus weg.	

Er waren in totaal 26 proefpersonen die de kern van deze categorie in een groep bij elkaar brachten. Zij gaven er de volgende namen aan:

- overbodige informatie/overbodigheid
- onnodige informatie
- dubbelop/doublures/dubbelingen
- overlap

De eerste naam krijgt de voorkeur van een meerderheid van achttien proefpersonen. In de definities van de proefpersonen komen twee elementen vaak terug: het herhalen van informatie en het geven van informatie die niet ter zake doet, niet relevant is, niet van toepassing is, of onnodig is. Wij komen tot de volgende omschrijving:

> *Er is sprake van overbodige informatie als een passage voor de lezer niet ter zake doet of onnodig wordt herhaald.*

Overigens waren er negen proefpersonen die bovenstaande kerngroep gevormd hadden, maar toch niet bij de nadere analyse betrokken zijn. Bij nader inzien bleek dat zij de bovenstaande groep hadden samengevoegd met de onderstaande groep, en aldus een ruimere overkoepelende categorie hadden gevormd.

3.2 Ontbrekende informatie. Ook de tweede categorie kwam als een duidelijke groep naar voren. Hiervoor werden kaartjes samengevoegd met de volgende commentaren van lezers.

Kijk uit voor je huid	**Je eerste baan**
Er staat niet bij dat je je hiertegen ook moet beschermen.	*Waar vind je het dichtstbijzijnde belastingkantoor?*
Ik mis voorbeelden van huidziektes waar het hier om gaat.	*Er moet direct gezegd worden wat je moet doen als je je sofi-nummer kwijt bent.*
Er staat "hoofdzakelijk" maar waar UV nog meer van komt, staat er niet bij.	*Wat is de sanctie als je geen formulier aanvraagt?*

Er waren in totaal 31 proefpersonen die de kerngroep van deze categorie gevormd hadden. Zij gaven daarvoor de volgende benamingen:

- ontbrekende informatie/ontbreking
- onvolledigheid
- gemiste informatie/informatieleemte
- vragen, vragen n.a.v. de tekst, onbeantwoorde vragen.

In de omschrijvingen kwamen de woorden "ontbreken" en "onvolledig" erg vaak terug. Bovendien gaven proefpersonen vaak het gevolg van die onvolledigheid aan: lezers begrijpen de tekst niet goed meer, omdat er iets ontbreekt. Daarom luidt voor deze categorie de omschrijving:

> *Er is sprake van ontbrekende informatie als de lezer meer informatie nodig heeft om de tekst goed te begrijpen.*

3.3 Geloofwaardigheid. De derde categorie kwam als een iets minder duidelijke groep uit de data naar voren. We vinden in deze categorie de volgende commentaren als kerngroep.

Kijk uit voor je huid	**Je eerste baan**
Ik vraag me af of het waar is dat de meeste mensen verstandig zonnen.	*Het is onterecht dat je geen studiefinanciering meer krijgt.*
Er staat dat afwijken van de instructies de gezondheid nauwelijks kan schaden.	*Soms moet je premies betalen, maar kun je er niet van profiteren.*
Volgens mij klopt dat niet.	*Dit feit klopt niet: het kost nooit geld,*
"Die mensen" verwijst naar een groep voor wie de folder toch juist niet bedoeld is?	*je verdient altijd wel iets.*
	De opmerking dat je ook over vakantiegeld moet betalen is onterecht.

Er waren 19 proefpersonen die de kern van deze categorie in één stapeltje vormden. De terminologie voor deze groep lezerreacties was niet erg eenduidig. We kwamen de volgende benamingen tegen:

- geloofwaardigheid
- oordeel/persoonlijke mening/waardeoordeel/opinie
- waarheid/betrouwbaarheid/correctheid/ervaringen/incorrecte details
- inhoudelijke discussie.

De term "geloofwaardigheid" is slechts vier keer genoemd door alle vijftig proefpersonen, hetgeen betekent dat we hier allerminst met een voor de handliggende benaming te maken hebben. De alternatieve namen werden echter nog minder vaak genoemd. Sommige proefpersonen kozen namen waarmee ook voorgaande categorieën meegenomen werden, zoals bijvoorbeeld "inhoud". In de omschrijvingen vonden we eveneens een grote verscheidenheid. Maar er waren wel gemeenschappelijke elementen, zoals: een inhoudelijk oordeel over de tekst, de lezer gaat in discussie met standpunten in de tekst, betwijfelt of de tekst inhoudelijk juist is, de lezer gelooft niet wat er in de tekst beweerd wordt, vraagt zich af of de tekst wel klopt. Wij maken een onderscheid tussen twee aspecten in deze omschrijvingen:

de twijfel over de waarheid van een feitelijke mededeling en het betwisten van een standpunt in de tekst over een kwestie. Daarmee komen we tot de volgende omschrijving:

> *Er is sprake van een probleem met de geloofwaardigheid van de tekst als de lezer betwijfelt of de tekst inhoudelijk klopt of als de lezer het niet eens is met standpunten in de tekst.*

3.4 Formulering. De voorgaande categorieën hebben alle met de inhoud van de tekst te maken. In de volgende stapeltjes die proefpersonen maakten, gaat het meer om tekstkenmerken. De volgende commentaren hadden met name betrekking op formuleringskwesties.

Kijk uit voor je huid	**Je eerste baan**
*Ik zou eerder kiezen voor "Kijk uit **met** je huid" dan voor "Kijk uit **voor** je huid".*	*De zin is kinderachtig: natuurlijk krijg je loon als je gaat werken.*
"Er zijn natuurlijk zonaanbidders" lijkt mij mooier.	*De eerste zin is te ingewikkeld.*
Als je kiest voor "slecht", zou ik "gunstig" door "goed" vervangen.	*De tekst is over het algemeen te simpel geschreven.*
	De mededeling bruto-netto is kinderachtig.

Deze groepering van lezersfeedback is door 36 van de vijftig lezers aangebracht, hetgeen een hoge score is. Zij gaven zeer uiteenlopende benamingen aan deze groep:

- formulering/woordformulering
- stijl/taalgebruik
- schrijfwijze
- woordkeuze
- redactioneel
- tekstoordeel
- doelgroep

De term "doelgroep" wekt misschien enige verbazing, maar de omschrijving hierbij luidde: "de wijze waarop in de richting van de doelgroep wordt geschreven". De term "formulering" werd het meest gekozen, gevolgd door "stijl". In de omschrijvingen komt duidelijk naar voren dat het niet om de inhoud van de tekst gaat, maar om nuances in de woordkeus of in de zinsbouw, waarbij overigens woordkeus veel vaker genoemd wordt. Een proefpersoon gaf als omschrijving: "de manier waarop zinnen en woorden zijn geschreven". Veel proefpersonen gaven in hun formulering een stilistisch verfraaiingsperspectief aan en wezen tegelijkertijd op een aspect van duidelijkheid: "Door de keuze van een ander woord wordt de tekst duidelijker of mooier." Dat aspect van verduidelijking verklaart waarom sommige kaartjes niet alleen bij deze categorie terechtkwamen, maar ook in een ander groepje functioneerden dat meer betrekking had op het begrip van de tekst. Dit roept de vraag op of verduidelijking een onderdeel moet zijn van de omschrijving. Wij kiezen ervoor die beide aspecten mee te nemen in de omschrijving:

Er is sprake van een formuleringsprobleem als de lezer moeite heeft met bepaalde woorden in de tekst of met de manier waarop de zin is geschreven, bijvoorbeeld omdat iets te ingewikkeld of juist te kinderlijk gezegd wordt of omdat de stijl lelijk gevonden wordt.

3.5. Structuur. In de vijfde categorie brachten de proefpersonen de volgende kaartjes met lezerreacties in één groep bij elkaar.

Kijk uit voor je huid	**Je eerste baan**
Deze toelichting komt te laat. Ik zou de voor- en nadelen eerder noemen.	*De functie van de twee subparagrafen is niet helder.*
Ik vind deze informatie over ziektes niet goed geordend. Ik zou een duidelijker onderscheid aanbrengen.	*De structuur van het onderdeel is niet helder. Vreemd dat er halverwege opeens subparagrafen zijn.*
Het is logischer om eerst UV-A en dan UV-B te noemen	*De structuur van het onderdeel is vreemd: begonnen moet worden met de loonbelasting-verklaring.*

Er zijn 31 proefpersonen die deze groep gevormd hebben. Zij geven daarvoor redelijk eenduidige termen:

- (tekst)structuur
- (tekst)opbouw/ordening/(tekst)indeling
- informatievolgorde

Opvallend is dat in de feedback op de belastingtekst het woord "structuur" zelf al twee keer voorkomt in de probleemomschrijving. Deze term is door onze proefpersonen het meest genoemd en lijkt daarmee de beste kandidaat om de categorie mee aan te duiden. In de omschrijvingen komen telkens de woorden indeling, opbouw, volgorde en structuur weer terug, soms ook vergezeld van vormgeving en lay-out. Wij kiezen op grond van de gepresenteerde formuleringen voor de volgende omschrijving:

Er is sprake van een structuurprobleem wanneer de volgorde van de informatie in de tekst niet goed is, waardoor de lezer geen helder beeld krijgt van de opbouw van de tekst en het overzicht verliest.

Tot zover de vijf categorieën waarvoor de proefpersonen duidelijk herkenbare groepen gevormd hadden. Een zesde categorie staat in evaluatieonderzoek vaak centraal, maar is in onze data minder duidelijk zichtbaar naar voren gekomen. Het gaat hier om een type problemen dat wij doorgaans met de term "begripsproblemen" aanduiden.

3.6 Onduidelijkheid. In de analyse van de data die verkregen zijn met de reacties op de gezondheidstekst is wel een herkenbare categorie zichtbaar geworden die op begripsproblemen betrekking lijkt te hebben, maar in het onderzoek met de belastingtekst is een dergelijke groep niet duidelijk naar voren gekomen. Het gaat om de volgende reacties.

Kijk uit voor je huid

Wordt hier met "zonnebrand" zonnebrandcrème bedoeld of verbranden door de zon?
Wat is "buiten bescherming"? Is dat altijd wanneer je buiten bent, ook in de regen?
Betekent dit dat je je dus ook onder de zonnebank moet beschermen?
Gaat het hier nu over de zon of over zonnebanken?
Ik snap niet wat hiermee bedoeld wordt.

Deze groepering van reacties is dus slechts door één van beide groepen gemaakt, en wel door 17 van de 25 proefpersonen. Zij gaven aan deze categorie de volgende namen:

- onduidelijkheid
- begrip
- betekenisprobleem
- onduidelijke formulering
- boodschapprobleem

Op grond van deze resultaten lijkt duidelijk sprake te zijn van een categorie die betrekking heeft op problemen met het vormen van een mentale representatie van een deel van de tekst. Maar hoe komt het dan dat bij de feedback op de belastingtekst deze categorie niet naar voren is gekomen? Een eerste verklaring zou kunnen zijn dat in de kaartjes onvoldoende reacties verwoord werden die op deze categorie duiden. Daar is echter geen sprake van. Die kaartjes waren er in voldoende mate, maar ze zijn op een andere manier gegroepeerd. Een deel van die kaartjes kwam terecht in de categorie "ontbrekende informatie". Sommige proefpersonen gaven die categorie de naam "vragen" of "vraagstellingen". Het ging om bijvoorbeeld onderstaande reacties.

Je eerste baan

Waar vind je het dichtstbijzijnde belastingkantoor?
Er moet direct gezegd worden wat je moet doen als je je sofi kwijt bent.
Wat is de sanctie als je geen formulier aanvraagt?

Anderen brachten die kaartjes onder in de categorie "formulering". Een gebrekkige formulering kan immers een begripsprobleem tot gevolg hebben, zoals misschien het geval is bij de reactie *De eerste zin is te ingewikkeld*. Deze tweede verklaring lijkt onaannemelijk omdat de kern van de groep "formulering" in de belastinggroep eerder bestaat uit kritiek op een te kinderachtige stijl. Een nadere analyse van de feedback (op de belastingtekst) die door de proefpersonen is ondergebracht bij "ontbrekende informatie" leidde tot twee sterk verwante clusters van kaartjes die beide betrekking lijken te hebben op dezelfde categorie. Hieronder geven we de kerngroepen van die beide stapeltjes weer.

Je eerste baan	
Waar vind je het dichtstbijzijnde belasting-kantoor?	*De beschrijving van de tariefgroep roept vragen en onzekerheid op.*
Hier staat dat je je paspoort mee moet nemen, maar is er geen ander identiteitsbewijs mogelijk?	*De concrete gevolgen voor partner/ouder zijn onduidelijk.*
Wat is de sanctie als je geen formulier aanvraagt?	*De relatie tussen loonheffing en andere premies is niet duidelijk.*

Een mogelijke verklaring voor de ordening van deze reacties in twee groepen is dat in de linkerkolom vragen geformuleerd zijn, die vooral klinken als een roep van de lezer om meer informatie, terwijl in de rechterkolom meer de stem van een beschouwende criticus doorklinkt die een diagnose formuleert van een mankement in de tekst. Uiteraard waren er ook proefpersonen (6) die dit onderscheid niet maakten. In de keuzes voor benamingen voor beide subgroepen of voor de overkoepelende groep komt "begrip" niet één keer voor. Wel komen we de term "onduidelijk(heid)" negen keer tegen. Wanneer we de kaartjes op een rij zetten die door deze negen proefpersonen als beste vertegenwoordiger van de groep "onduidelijkheid" worden genoemd, ontstaat het volgende overzicht.

Je eerste baan
In de folder staat niet wat er wordt gedaan met de belasting.
De beschrijving van de tariefgroep roept vragen en onzekerheid op. De zin over reiskosten is niet duidelijk.
Er moet direct gezegd worden wat te doen als je je sofi kwijt bent.
De relatie tussen premies en loonheffing is niet helder.
De structuur van het onderdeel is vreemd: begonnen moet worden met de loonbelastingverklaring.
Vreemd dat er halverwege opeens subparagrafen zijn.
De functie van de twee subparagrafen is niet helder.

Dit overzicht wijst niet onmiddellijk in de richting van een verborgen groep begripsproblemen, alhoewel die wel vertegenwoordigd zijn. De laatste drie lezerreacties zijn echter van een andere orde, en de eerste reactie heeft toch meer betrekking op een wens om meer informatie dan op een echt begripsprobleem.

In de omschrijvingen die proefpersonen geven voor de term "onduidelijkheid" komen we de volgende definities tegen:

- je weet niet wat wordt bedoeld met iets dat in de tekst wordt gezegd;
- de lezer snapt niet wat met de tekst wordt bedoeld;
- wat met de tekst bedoeld wordt, komt niet duidelijk over op de lezer;
- als je een tekst leest, dan begrijp je die niet zomaar, want de tekst roept vragen op over de bedoeling of betekenis van woorden en zinnen.

En voor "begrip" komen we (in de data van de gezondheidstekst) de volgende omschrijvingen tegen:

- onbegrip voor tekstinhoud ontstaat, de tekst roept vragen op;
- de lezer begrijpt niet wat er in de tekst staat;
- het is niet duidelijk wat ergens mee bedoeld wordt, de lezer snapt het niet;
- een term, zin of uitdrukking is niet duidelijk, de lezer begrijpt niet wat de schrijver bedoelt.

Uit deze omschrijvingen komt een duidelijk onderscheid naar voren met de categorie "ontbrekende informatie". Het gaat hier immers om informatie die wel in de tekst te vinden is, maar die niet goed begrepen wordt. In de laatst geciteerde omschrijving wordt een relatie zichtbaar met "formulering", omdat uiteraard een gebrekkige formulering daar de oorzaak van kan zijn.

Wij concluderen dat er toch wel reden is om een categorie begripsproblemen te onderscheiden, die we echter op grond van de data liever met de term "onduidelijkheid" aanduiden. Deze term is immers door respectievelijk zeven en negen proefpersonen naar voren gebracht en in de eerste groep is er ook daadwerkelijk een als zodanig herkenbare categorie gevormd van lezerreacties. De resultaten van de tweede groep leiden echter tot voorzichtigheid, omdat de term mogelijkerwijs door proefpersonen als overlappend wordt ervaren met de categorie "ontbrekende informatie". In de omschrijving van de term zou dit onderscheid dan ook duidelijk geformuleerd moeten worden, bijvoorbeeld op de volgende manier.

> *Er is sprake van onduidelijkheid wanneer de tekst ergens wel informatie over geeft, maar als die passage vragen oproept, omdat de lezer de inhoud ervan niet goed begrijpt.*

3.7 Leestekens en spelling. In de gezondheidstekst waren in het eerdere evaluatieonderzoek bewust een aantal fouten aangebracht in de spelling en interpunctie. Dit was gedaan met het doel om na te gaan of hoger en lager opgeleide proefpersonen hier verschillend op zouden reageren (zie Lentz en De Jong, 2000). Op een aantal kaartjes waren lezerreacties zichtbaar die hierop betrekking hadden. Deze kaartjes zijn door sommige proefpersonen in twee groepen ondergebracht en door andere in één gezamenlijke groep. Het gaat om de volgende reacties.

Interpunctie	**Spelling / Correctheid**
Hier hoort ook een puntkomma.	*Volgens mij moet er* spelbreker *staan in plaats van* speldbreker.
Deze komma is overbodig.	*Er moet* verhoogd *staan in plaats van* verhoogt.
De dubbele aanhalingstekens maken het erg onduidelijk.	Daarmee *moet* Daarvan *zijn.*
Ik zou dat kan niet gezond zijn *tussen aanhalingstekens zetten.*	

Het is de vraag of dit een zinvolle categorie is, omdat we in evaluatieonderzoek immers geen lezers inschakelen teneinde na te gaan of de tekst goed gespeld is. In principe leggen we een incorrect gespelde tekst niet aan lezers voor. Toch weten we uit ervaring dat lezers vaak de behoefte hebben commentaar te leveren op komma's en punten, en op de spelling

van woorden. Ook onder leken bevindt zich menige meester Pennewip, die het niet kan laten zijn deskundigheid op dit punt tentoon te stellen.

Hoe noemen de proefpersonen deze categorie? Wij kregen de volgende benamingen:

- leestekens / interpunctie
- spelfout / spelling
- taalfout / taal- en spelfout.

In de omschrijvingen werden deze termen vaak herhaald, maar er was ook vaak sprake van "punten en komma's" en één creatieve proefpersoon schonk ons het prachtige "taalelastiekjes". Wie in evaluatieonderzoek deze categorie aan proeflezers aan wil bieden, zou volgens ons het best interpunctie en spelling samen kunnen nemen onder de naam *Leestekens en Spelling*. Daarbij past de volgende omschrijving.

> *Er is sprake van een probleem met de leestekens als die op een verkeerde plaats in de tekst staan of als ze ontbreken.*
> *En er is sprake van een probleem met de spelling als de woorden niet gespeld zijn volgens de regels van het Nederlands.*

4. Conclusie en discussie

Leken die niet beroepsmatig met taal en tekst bezig zijn, kunnen in evaluatieonderzoek geconfronteerd worden met de volgende categorieën en de daarbij passende omschrijvingen. Er is sprake van:

overbodige informatie	als informatie voor u niet ter zake doet of onnodig wordt herhaald.
ontbrekende informatie	als u meer informatie nodig heeft om de tekst goed te begrijpen.
onduidelijkheid	wanneer de tekst ergens wel informatie over geeft, maar als die passage vragen oproept, omdat u de inhoud ervan niet goed begrijpt.
formuleringsprobleem	als u moeite heeft met bepaalde woorden in de tekst of met de manier waarop de zin is geschreven, bijvoorbeeld omdat iets te ingewikkeld of juist te kinderlijk gezegd wordt of omdat u de stijl lelijk vindt.
structuurprobleem	wanneer de volgorde van de informatie in de tekst niet goed is, waardoor u geen helder beeld krijgt van de opbouw van de tekst en het overzicht verliest.
geloofwaardigheidsprobleem	als u betwijfelt of de tekst inhoudelijk klopt of als u het niet eens bent met bepaalde standpunten in de tekst.
foutieve leestekens	als die op een verkeerde plaats in de tekst staan of ontbreken.
spellingsfouten	als de woorden niet gespeld zijn volgens de regels van het Nederlands.

Twijfel is mogelijk over de groep "onduidelijkheid", omdat die in één van de beide sessies niet als zodanig duidelijk naar voren kwam, hetgeen verklaard kan worden door de overlap met "ontbrekende informatie" en in mindere mate met de categorie "formuleringsprobleem". Bovendien is twijfel mogelijk over de zinvolheid van een categorie die op spelling en leestekens is gericht, omdat dat een type reacties is waar we in evaluatieonderzoek niet direct behoefte aan hebben.

Uit bovenstaand overzicht moet niet geconcludeerd worden dat elke proefpersoon deze categorieën spontaan creëerde tijdens het sorteren van kaartjes met lezerreacties. Het overzicht is tot stand gekomen door een analyse van de meest sterke verbanden tussen de verschillende kaartjes. Dat er ook volkomen andersoortige categorieën zijn genoemd, blijkt uit onderstaand overzicht.

- wijzigingen
- algemene informatie
- verdieping
- eerlijkheid
- redactioneel
- belangrijke opmerking
- onbelangrijke opmerking
- persoonlijk
- arbeidsproces
- belastingontduiking
- aanspreekniveau
- persoonlijke gevolgen
- flauwekul
- opmaak
- kader
- omschrijvingsfouten
- tekstueel
- inhoudelijk
- inhoudelijke details

Sommige proefpersonen sorteerden volgens een dimensie van relevantie van het commentaar met als polen belangrijk - onbelangrijk. Anderen sorteerden op een dimensie als inhoud - tekst - vorm. En nog weer anderen kozen inhoudelijke rubrieken die uit het thema van de tekst voortkwamen; zij bedachten groepen met titels als *arbeidsproces* en *persoonlijke gevolgen*. Misschien had voor die proefpersonen de instructie duidelijker geformuleerd moeten worden.

Het onderzoek levert een grootste gemene deler op voor de categorisering van problemen. Toch is het duidelijk dat ook met deze categorieën lezers soms zullen twijfelen in welke rubriek ze een probleem moeten indelen. Bijvoorbeeld omdat zij een ander indelingssysteem hanteren, of omdat zij voor andere labels zouden kiezen. Het onderzoek bevestigt echter in grote lijnen de bruikbaarheid van de lijst met probleemcategorieën, zoals die aan het begin van dit artikel gepresenteerd is, en waar we in Focus al jaren mee werken. Het helpt ons daarnaast om de probleemomschrijvingen en -afbakeningen meer op de kennis van leken af te stemmen en daarmee de ecologische validiteit van ons instrumentarium te vergroten.

Meer onderzoek is nodig: (1) naar de mate waarin mensen met de hier gepresenteerde categorieën uit de voeten kunnen en de problemen die ze daarbij ervaren; (2) naar de precieze afbakening tussen probleemcategorieën; (3) naar de mate waarin de resultaten afhankelijk zijn van het gekozen teksttype (brochure), omdat denkbaar is dat bij een handleiding of een website een gedetailleerder arsenaal aan probleemtypen nodig is, dat specifiek gericht is op problemen met de toepassing van die informatie; (4) naar de mate waarin de resultaten afhankelijk zijn van het opleidingsniveau van de lezers, dat in dit onderzoek relatief hoog was.

Noten

1 Het betreft onderzoek dat aan de Universiteit Utrecht is uitgevoerd door Viola Teepe en Aart van Kapel in het kader van een scriptieproject van Lentz en De Jong.

2 Meer over dit onderzoek is te vinden in Lentz en De Jong (2000), waarin de verschillen worden besproken tussen de feedback op deze tekst van hoger en lager opgeleide proefpersonen.

3 Over Focus als instrument voor tekstevaluatie is meer te vinden in De Jong en Lentz (2001), waar o.a. een vergelijking wordt gemaakt met feedback die verkregen is met de plus-en-minmethode.

Bibliografie

Cicourel, A.V. (1982). Interviews, surveys, and the problem of ecological validity. *The American Sociologist, 17* (2), 11-20.

Jong, M. de (1998). *Reader feedback in text design. Validity of the plus-minus method for the pretesting of public information brochures*. Amsterdam: Rodopi.

Jong, M. de & L. Lentz (2001). Focus: Design and evaluation of a software tool for collecting reader feedback. *Technical Communication Quarterly, 10*, 387-401.

Jong, M. de & P.J. Schellens (2000). Toward a document evaluation methodology: what does research tell us about the validity and reliability of evaluation methods? *IEEE Transactions on Professional Communication, 43*, 242-260.

Kapel, A. van (2000). *Taalgebruikers, teksten en probleemcategorieën. Een onderzoek naar het lexicon van taalgebruikers voor categorieën van lezersproblemen aan de hand van bestaande feedback op een tekst*. Doctoraalscriptie Universiteit Utrecht, afdeling Taalbeheersing.

Lentz, L. & M. de Jong (2000). Hoger en lager opgeleide proefpersonen en hun feedback op een voorlichtingstekst. In: R. Neutelings, N. Ummelen & A. Maes (red.) *Over de grenzen van de taalbeheersing* (pp. 317-326). Den Haag: SDU.

Maes, A., N. Ummelen & H. Hoeken (1996). *Instructieve teksten. Analyse, ontwerp en evaluatie*. Bussum: Coutinho.

Teepe, V. (1999). *Wat is dat voor probleem? Een onderzoek naar de categorieën waarin tekstproblemen ingedeeld worden en de termen die aan die categorieën worden gegeven*. Doctoraalscriptie Universiteit Utrecht, afdeling Taalbeheersing.

Boek beoordelingen

Gemert-Pijnen, J. van (2003). *Het totstandkomen en functioneren van infectiepreventieprotocollen. Een onderzoek naar communicatie gestuurd door wet- en regelgeving.*
Dissertatie Universiteit Twente. Enschede: Thesis. Promotor: prof. dr. P.J. Schellens.

Het onderzoek van Van Gemert richtte zich op protocollen in Nederlandse ziekenhuizen en andere zorginstellingen. Protocollen zijn voorbeelden van *regelgestuurde communicatie-uitingen*: door de overheid verplicht gestelde documenten die dienen als leidraad voor het personeel om onveilige situaties te voorkomen c.q. te beëindigen. Net als bijvoorbeeld veiligheidsvoorschriften in industriële bedrijven en schoolwerkplannen in het basisonderwijs vervullen protocollen in zorginstellingen meerdere functies tegelijk. De instelling kan laten zien dat er aan de verplichtingen krachtens de wet- en regelgeving is voldaan, de veiligheid op organisatieniveau wordt als het goed is door de protocollen vergroot en individuele werknemers kunnen eruit afleiden wat hun in een gegeven situatie te doen staat. Waar in eerdere proefschriften over regelgestuurde communicatie-uitingen – dat van Lentz & Van Tuijl (1989) over schoolwerkplannen en dat van Elling (1991) over industriële veiligheidsvoorschriften – veel werk werd gemaakt van de functionele analyse van de onderzochte tekstsoort, is Van Gemert in haar inleidende hoofdstuk aanmerkelijk beknopter. Onder verwijzing naar het werk van Elling (en opvallend genoeg niet naar dat van Lentz & Van Tuijl of naar later werk van taalbeheersingsonderzoekers als Pander Maat op dit gebied), staat ze kort stil bij de principiële problematiek van de afstemming van regelgestuurde communicatie in organisaties op de onderscheiden doelen en doelgroepen. Duidelijk wordt al snel dat de nadruk in dit proefschrift ligt op empirisch onderzoek naar het ontwerp, de implementatie en de evaluatie van protocollen, en naar het functioneren van protocollen in de praktijk.

Het onderzoek spitst zich toe op twee voorbeelden van *infectiepreventieprotocollen*: protocollen waarmee besmetting (HIV of hepatitis bijvoorbeeld) moet worden voorkomen als gevolg van accidenteel bloedcontact (ook wel *prikaccidentprotocollen*) en protocollen waarmee MRSA (methicilline-resistente staphylococcus aureus) – infecties voorkomen moeten worden (aangeduid als *MRSA-protocol*). MRSA-infecties zijn gevreesd omdat ze zich snel in ziekenhuizen verspreiden en aldus tot grote epidemieën kunnen leiden; MRSA-stammen zijn resistent voor veel soorten antibiotica, wat bestrijding van MRSA in ziekenhuizen urgent maakt. Alleen zo kan worden voorkomen dat behandeling van de desbetreffende infecties op een bepaald moment helemaal niet meer mogelijk is.

Dat Van Gemert juist deze protocollen, die

beide deel uitmaken van de arbo-zorg, gekozen heeft, berust er onder meer op dat ze om een communicatieve aanpak vragen die deels verschillend is. De *prikaccidentprotocollen* zijn primair gericht op de individuele werknemer, die om accidenteel bloedcontact te voorkomen een aantal routinehandelingen moet internaliseren en daarnaast moet weten wat te doen als zich toch een incidenteel accident voordoet. Bij de *MRSA-protocollen* staat de samenwerking tussen diverse werknemers voorop. Wanneer bij een patiënt die bijvoorbeeld in een buitenlands ziekenhuis opgenomen is geweest MRSA wordt geconstateerd, moeten er onmiddellijk grootschalige maatregelen worden genomen om verspreiding tegen te gaan: de patiënt gaat in isolatie, medewerkers die met de patiënt in contact zijn geweest worden naar huis gestuurd, en de afdeling waar de patiënt gelegen heeft wordt gesloten. Om dat allemaal goed te laten verlopen zijn heldere afspraken nodig, die worden vastgelegd in een protocol dat, zoals Van Gemert duidelijk maakt, in dit geval moet functioneren als gemeenschappelijk interpretatiekader voor de uit te voeren handelingen.

De opzet van het onderzoek dat Van Gemert heeft uitgevoerd met voorbeelden van deze twee soorten protocollen typeert ze in haar tweede hoofdstuk als een exploratieve case studie, waarin ze drie soorten activiteiten ondernam. In vijf ziekenhuizen liet ze *vragenlijsten* beantwoorden om na te gaan in hoeverre de doelgroepen bekend waren met de protocollen van het eigen ziekenhuis en de daarin vereiste maatregelen, of ze in staat en bereid waren die maatregelen daadwerkelijk uit te voeren en of ze zich daarin ook voldoende gestimuleerd voelden. Aan de hand van een *gebruiksonderzoek* in de vorm van een praktijktoets probeerde ze vast te stellen in welke mate de protocollen in de praktijk bruikbaar waren, en via interviews met schrijvers en andere actoren die voor de protocollen verantwoordelijk waren, probeerde ze het ontwerp- en implementatieproces te reconstrueren.

De gestandaardiseerde en ge-preteste vragenlijsten waarmee informatie werd verzameld over het functioneren van de protocollen werden voorgelegd aan in totaal 133 respondenten. Die respondenten waren afkomstig uit vijf in de verslaglegging geanonimiseerde algemene ziekenhuizen, die onder meer werden gekozen op basis van regionale spreiding. De respondenten kwamen uit de beroepsgroepen verpleging, OK-assistenten, specialisten/arts-assistenten, analisten en huishoudelijk personeel.

Voor de praktijktoets gebruikte Van Gemert zoals ze zelf schrijft een 'aangepaste versie van de hardop-werkmethode'. Aan in totaal 92 proefpersonen (allen ook respondent in het vragenlijstonderzoek) werden praktijkscenario's voorgelegd (*hardop-zoektaken*) waarbij aan de hand van de protocollen (die per ziekenhuis in een aantal opzichten enigermate verschilden) moest worden bepaald welke handelingen in welke situatie relevant zouden zijn. Anders dan in hardopwerk-onderzoek gebruikelijk is, werd van de proefpersonen niet gevraagd om de desbetreffende handelingen ook daadwerkelijk hardop denkend uit te voeren. Van Gemert meldt dat ze als proefleider tijdens de hardopwerksessies een 'actievere inbreng [heeft] gehad dan strikt genomen volgens de hardopwerk-methode toelaatbaar is' (p. 38). Refererend aan opmerkingen van Ericsson & Simon (1984), de grondleggers van het moderne hardopdenkonderzoek, en meer nog aan parafrases van die opmerkingen bij document design onderzoekers Boren &

Ramey (2000), stelt Van Gemert dat het soms gerechtvaardigd is om af te wijken van de strikte afnameprocedures 'for example if the participant is stuck or the system crashes'.

De interviews (gedeeltelijk gestructureerd) die dienden ter reconstructie van het ontwerp- en implementatieproces werden afgenomen bij in totaal 48 informanten: beleidsvertegenwoordigers, penvoerders, actoren die betrokken waren bij het implementatieproces en actoren die betrokken waren bij de evaluatie. De vragen die de informanten werden gesteld, waren deels voor allen dezelfde, en deels actorspecifiek. Centraal stonden documenten (protocollen, concepten, brondocumenten) waaruit fragmenten werden gekozen die aan de informanten werden voorgelegd met de vraag te vertellen hoe bepaalde beslissingen tot stand gekomen waren.

Door de combinatie van een relatief grootschalige en met zorg gespreide vragenlijstafname, hardopwerk-sessies en interviews kreeg Van Gemert de beschikking over een grote en rijk geschakeerde dataverzameling. Daarbij moet wel worden opgemerkt dat de informatie die deze data konden opleveren over de totstandkoming en het feitelijke gebruik van protocollen, nogal indirect van aard was. De interviews lieten niet meer toe dan een procesbeschrijving *achteraf*, met de vragenlijsten kon niet meer in beeld worden gebracht dan *gerapporteerd* gedrag van de betrokkenen, en in de praktijktoetsen was de taakuitvoering beperkt tot *opzoekwerk* en bleven daadwerkelijke handelingen buiten beeld. Ook al doordat de onderzoeker zelf vaker ingreep dan Ericsson & Simon ooit bedoeld kunnen hebben (van *system crashes* of vergelijkbare calamiteiten kan bij opzoektaken nauwelijks sprake zijn geweest) is het de vraag of dit deel van het onderzoek de aanduiding hardopwerk-aanpak wel verdient. In elk geval hadden de beperkingen die de indirectheid van de verworven informatie opleverde voor het beeld dat van de ontwerp- en gebruikspraktijk van ziekenhuisprotocollen geschetst kon worden, in het discussiehoofdstuk meer aandacht mogen krijgen.

In hoofdstuk 3 worden de resultaten gepresenteerd van het vragenlijstonderzoek. Het beeld dat zich uit de antwoorden van de respondenten af laat leiden, is overwegend positief. De doelgroep is op de hoogte van het bestaan van de protocollen en laat weten bekend te zijn met de gevraagde maatregelen. Ook zegt men in meerderheid de protocollen zelf te hebben geraadpleegd. De attitude tegenover de protocollen is positief en men is in meerderheid bereid de gewenste maatregelen te treffen. Ook is de eigen-effectiviteitsverwachting bij de meeste respondenten groot: men verwacht dat het lukt om de nodige maatregelen te treffen. Over de institutionele stimulans is men minder positief: veel respondenten laten weten behoefte te hebben aan meer aandacht voor de protocollen binnen de organisatie.

Uit de analyse bleken, aldus Van Gemert op p. 100, extreme verschillen tussen de beroepsgroepen in kennis, attitude en gerapporteerd gedrag. Huishoudelijk personeel bleek het minst op de hoogte, en schatte de uitvoerbaarheid van de voorzorgsmaatregelen rond accidenteel bloedcontact lager in dan de andere beroepsgroepen. Verpleegkundig personeel bleek het meest gemotiveerd om te doen wat er volgens de protocollen verlangd wordt, artsen stonden het meest gereserveerd tegenover de preventiemaatregelen. Ook waren er verschillen waarneembaar tussen

de onderzochte ziekenhuizen, en dan met name waar het preventieproblemen op organisatieniveau MRSA) betreft.

Daarbij dient echter te worden aangetekend dat Van Gemert haar uitkomsten niet gecorrigeerd heeft voor alfa-inflatie. Ze rapporteert steeds de uitkomsten van een non-parametrische vorm van variantie-analyse, waarbij met behulp van de Kruskal-Wallis test werd nagegaan of de relatie tussen de onderzochte variabelen in een situatie met meer dan twee onafhankelijke steekproeven significant was. Waar dat het geval bleek, werden steeds post-hoc analyses uitgevoerd om de vraag preciezer te kunnen beantwoorden waar het gevonden verband aan toe te schrijven was. Daarbij werd steeds de Mann-Whitney U-test gehanteerd. Voorbijgegaan werd daarbij aan het risico dat wanneer er zoals hier een grote hoeveelheid vergelijkingen tussen groepen onderling gemaakt worden, alleen al op basis van toeval een aantal significante verschillen gevonden zullen worden die er feitelijk niet zijn. Verstandig is het om in zo'n geval niet een alfa van .05 maar een kleinere alfa te kiezen, en aldus een soort Bonferroni-correctie toe te passen zoals die bekend is uit de parametrische variantie-analyse. Had Van Gemert dat gedaan, dan had dat geleid tot een lager aantal significante verschillen tussen de vergeleken groepen.

Het relatief zonnige beeld dat de antwoorden bij de vragenlijsten opleverden, wordt door de resultaten van het gebruiksonderzoek in hoofdstuk 4 niet bevestigd. Het bleek dat respondenten minder bekend waren met de protocollen dan ze zelf dachten, en dat ze de maatregelen minder goed uit konden voeren dan ze zelf inschatten. De MRSA-protocollen blijken daarbij het minst slecht te functioneren. Van Gemert schrijft dat toe aan de grote impact van een MRSA-probleem op de bedrijfsvoering, waardoor de aandacht voor maatregelen om zo'n probleem te voorkomen of zo nodig te bestrijden, bij de betrokkenen mogelijkerwijs wordt verhoogd.

In hoofdstuk 5 en 6 rapporteert Van Gemert de resultaten die de interviews opleverden. Daarbij blijkt een discrepantie tussen het gewenste en het feitelijk gevoerde protocolbeleid: door langdurige discussie over de uitvoering van het beleid is er met de beleidsontwikkeling veel tijd gemoeid; van integrale kwaliteitszorg is nauwelijks sprake, het zicht op het gehele proces van ontwerp, implementatie en evaluatie ontbreekt, en – misschien wel het belangrijkst – ontwerpers blijken door hun positie in de organisatie weinig zicht te hebben op het functioneren van de protocollen in de praktijk. Protocollen worden voornamelijk geschreven door samenwerkende materie-experts zoals ziekenhuishygiënisten en arts-biologen, die in een aantal gevallen door een commissie worden geadviseerd. Op organisatieniveau wordt het protocol gezien als een verantwoordingsdocument, als norm voor geschillen en als toets of er correct gehandeld is. Op individueel niveau worden er vooral kennisdoelen nagestreefd (waar is het protocol te vinden; wat moet er gebeuren?); de algemene verwachting is dat de werknemers wel in staat en gemotiveerd zullen zijn om de nodige maatregelen te treffen. Uitgangspunt voor het ontwerp van de protocollen vormen landelijk opgestelde richtlijnen. Daarin vinden de auteurs in de ziekenhuizen echter onvoldoende steun, zo blijkt uit de interviews. Ook ondervindt men in het eigen ziekenhuis in termen van tijd, personeel en middelen, slechts beperkte ondersteuning. Het management geeft aan de totstandkoming en ook de implementatie van ade-

quate protocollen slechts een lage prioriteit. Evaluatie vindt dan ook alleen plaats als zich incidenten voordoen die om maatregelen vragen die in de protocollen terug te vinden moeten zijn. Al met al, zo concludeert Van Gemert, verlopen ontwerp, implementatie en evaluatie weinig bevredigend, noch voor de betrokken actoren, noch voor de gebruikers. Net zoals gevonden werd in ander onderzoek naar regelgestuurde communicatie via teksten, zoals dat van Elling en van Lentz & Van Tuijl, blijkt dat ook bij de ontwikkeling van ziekenhuisprotocollen meestal geen bevredigende oplossing gevonden wordt voor het probleem van de onverenigbaarheid van de verantwoordingsplicht voor de organisatie enerzijds en de afstemming op de praktijk van de individuele gebruiker anderzijds. De vraag dringt zich dan ook op of de nagestreefde polyfunctionaliteit van dit soort documenten wel haalbaar is. Wellicht verdient het de voorkeur om voor gebruik in de dagelijkse praktijk protocollen een lage prioriteit te geven, en in plaats daarvan een voor de medewerkers continu bereikbare helpservice in te richten. Blijkens uitlatingen van proefpersonen in Van Gemert's praktijktoets gaat daar in twijfelsituaties zoals die zich door de aard van de problematiek van infectiepreventieprotocollen nu eenmaal vaak voordoen, ook de voorkeur van de betrokkenen naar uit.

In haar slothoofdstuk besteedt Van Gemert terecht veel aandacht aan de discrepantie die ze waarneemt tussen de positieve uitkomsten van haar vragenlijstonderzoek en de aanmerkelijk negatievere bevindingen die zowel de praktijktest als de interviews opleverden. De belangrijkste verklaringen die daarvoor ze naar voren brengt, liggen in sociaal wenselijke antwoorden, zelfoverschatting en gedragsrationalisatie bij de respondenten die antwoorden op de vragen gaven. Pas in de praktijktoets bleek hoe het er werkelijk voorstaat als men de protocollen probeert te lezen en erin op te zoeken wat er gedaan moet worden. Optimistische verwachtingen van schrijvers en lezers worden dan al snel gelogenstraft. Een aanbeveling die Van Gemert als reactie daarop doet, is voortaan twee soorten protocollen op te stellen: *organisatieprotocollen* met een verantwoordingsfunctie en *doelgroepspecifieke protocollen* met een gebruiksfunctie. De organisatieprotocollen dienen dan ter legitimering van het preventiebeleid en als norm bij geschillen; deze protocollen lijken daarmee vooral goed te moeten worden opgeborgen in een archiefkast waar ze alleen uit hoeven te worden gehaald als er toezichthoudende instanties tevreden moeten worden gesteld. De doelgroepspecifieke protocollen die Van Gemert bepleit, moeten het echte werk doen; die moeten in de dagelijkse praktijk op de werkvloer worden gebruikt.

De vraag is hoe realistisch deze oplossing is. Het lijkt onwaarschijnlijk dat toezichthouders bereid zouden zijn hun formele oordeel over het infectiepreventiebeleid in een ziekenhuis te baseren op protocollen die alleen uit de kast worden gehaald als de inspectie op bezoek komt, terwijl in de praktijk een afwijkend protocol het richtsnoer voor het handelen van de medewerkers zou zijn. Hoe in zo'n geval zou kunnen worden aangetoond dat beide soorten protocollen functioneel equivalent zijn, laat Van Gemert in het midden. Het is ook lastig voor te stellen hoe dat zou moeten. Een betere oplossing is daarom wellicht de eerder genoemde inrichting van een helpservice, die dan liefst permanent bereikbaar moet zijn voor vragen vanuit de diverse beroepsgroepen. De medewerkers van zo'n helpservice zouden dan de enigen zijn die de achterliggende protocollen nog per se goed zouden hoeven kennen. Op basis daarvan kunnen ze dan, in inter-

actie met de vragenstellers, op de juiste momenten de adequate antwoorden op vragen over infecties of andere onderwerpen van protocollen formuleren. Een dergelijke helpservice zou ook als taak kunnen krijgen om regelmatig reminders voor preventief gedrag rond te sturen en om bij incidenten zelf communicatie-initiatieven te nemen – taken die Van Gemert in haar aanbevelingen weggelegd ziet voor 'voorbeeldstellende personen', zoals direct leidinggevenden en sleutel-actoren (p. 246). Het kan zijn dat Van Gemert een permanente helpservice onhaalbaar acht omdat de kosten daarvan te hoog zouden uitvallen. Maar misschien zou een samenwerkingverband tussen verschillende ziekenhuizen dat probleem kunnen verkleinen.

Uit het onderzoek van Van Gemert is duidelijk geworden dat er in elk geval iets moet gebeuren: de bestaande praktijk van infectiepreventieprotocollen voldoet niet. Daarmee is een belangrijke bijdrage geleverd aan de kennis van actuele communicatieproblemen in Nederlandse ziekenhuizen. Daarnaast heeft Van Gemert het taalbeheersingsonderzoek naar teksten in regelgestuurde communicatie verrijkt met een grote hoeveelheid zorgvuldig verzamelde gegevens van actoren die bij de totstandkoming en het gebruik van dit type teksten betrokken zijn. Voeg daarbij een uiterst helder gestructureerd verslag van activiteiten en bevindingen, en het moge duidelijk zijn dat het vele werk dat Van Gemert aan haar onderzoek heeft besteed een mooie bijdrage aan het vakgebied heeft opgeleverd.

Carel Jansen

Signaleringen

Claes, Marie-Thérèse & Gerritsen, Marinel (2002). *Culturele waarden en communicatie in internationaal perspectief*. Bussum: Coutinho. ISBN 9062833047. Prijs: € XXX (276 pp.)

'Cultuur is communicatie en communicatie is cultuur'. Met dit citaat van Edward T. Hall, de grondlegger van het denken over cultuurverschillen in relatie tot communicatie, opent *Culturele waarden en communicatie in internationaal perspectief*. De auteurs van dit boek beogen een overzicht te geven van theorieën over cultuurverschillen, van de consequenties die deze verschillen kunnen hebben voor de communicatie en van de verbale en non-verbale communicatie-elementen waarvan de betekenis van cultuur tot cultuur kan variëren. Met name wordt daarbij aandacht besteed aan verschillen en overeenkomsten tussen de Nederlandse en de Belgisch-Vlaamse cultuur. Een interessante vergelijking, want de taal van Nederlanders en Vlamingen is nagenoeg identiek, maar de cultuur geenszins.

In een inleidend hoofdstuk worden de begrippen cultuur, communicatie en interculturele communicatie gedefinieerd en toegelicht. Voor cultuur achten de auteurs gedeelde waarden bepalend. Communicatie wordt toegelicht aan de hand van een model van Targowski en Bowman, met tien aspecten van communicatie, variërend van de 'fysische link' tot en met de 'opslaan/terughalenlink'. En in de slotparagraaf wordt onderzoek naar interculturele communicatie afgezet tegen crosscultureel onderzoek.

Het tweede hoofdstuk is gewijd aan de diverse theorieën die de dimensies specificeren waarop de waarden die bepalend zijn voor cultuur kunnen verschillen. Kluckholm & Strodtbecks indeling in fijnmazige en grofmazige culturen (hier te lande overgenomen door Pinto), Hofstedes vijf dimensies – machtsafstand, individualisme/collectivisme, masculiniteit/ femininiteit, onzekerheidsvermijding en lange-/kortetermijngerichtheid – de zeven categorieën van Trompenaars en de evenzovele van Schwartz passeren de revue. Dit alles voegen de auteurs tezamen tot een breed amalgaam, met als kern Hofstedes dimensies.

In het hoofdstuk over communicatie worden cultuurbepaalde verschillen in het gebruik van communicatie-elementen beschreven. Bij verbale communicatie worden de verschillende manieren genoemd waarop de woordenschat van verschillende talen de wereld indeelt, en de verschillende waarden en betekenissen die begrippen en woorden in verschillende talen kunnen hebben. Bij non-verbale communicatie gaat het om verschillen in gebruik en betekenis van prosodische en paralinguïstische middelen en van afstand, mimiek, gebaar en zelfs geur. Verder wordt Halls onderscheid tussen culturen waarin veel, en culturen waarin weinig context

wordt gebruikt bij de interpretatie van communicatie aangehaald, en worden, naast andere verschillen, variaties in de mate waarin en de wijze waarop culturen beleefdheid uitdrukken genoemd.

Het vierde hoofdstuk inventariseert de overeenkomsten en verschillen tussen Nederland en Vlaanderen. Onze Zuiderburen blijken onder meer hoger te scoren dan wij op Hofstedes dimensies machtsafstand, masculiniteit en onzekerheidsvermijding. Maar hoewel in een feminiene cultuur, zoals de Nederlandse, persoonlijke verhoudingen belangrijker worden gevonden dan geld, en in een cultuur met hoge onzekerheidsvermijding, zoals de Vlaamse, regels nu juist weer erg belangrijk worden gevonden, zijn voor de Belg relaties toch weer belangrijker dan regels. Hoe dan ook, we blijken beide (althans volgens sommige onderzoekingen) erg individualistisch.

In het vijfde hoofdstuk worden verschillen en overeenkomsten geschetst tussen de twee in het vorige hoofdstuk besproken culturen en andere Europese culturen (binnen en buiten Europa). De Angelsaksische ('de Angelsaksische culturen zijn lagecontext-maatschappijen, op Groot-Brittanië na dat een wat hogere context heeft'), de Germaanse, de Noordse, de Romaanse culturen en die van Turkije worden aan de cultuurproef onderworpen. In een volgend hoofdstuk ondergaan de Aziatische, Arabische en Afrikaanse culturen hetzelfde lot.

Het slothoofdstuk, Omgaan met cultuurverschillen, beschrijft de stadia waarin interculturele communicatieve competentie wordt verworven, van ontkenning tot integratie, en de cultuurschok die mensen die voor langere tijd in het buitenland verblijven, ervaren.

Zoals uit bovenstaande samenvatting moge blijken, schrikken de auteurs niet terug voor enige generalisatie en simplificatie. Hoewel ze hier en daar plichtsgetrouw memoreren dat men 'behoedzaam moet zijn met stereotypering' en dat er niet altijd controleerbaar empirisch onderzoek ten grondslag ligt aan de typeringen, lijken ze zich zelf over het algemeen aan deze waarschuwingen weinig gelegen te laten liggen.

De hoofdmoot van het boek vormt de beschrijving van cultuurverschillen. Over communicatie of over interculturele communicatie is er weinig te vinden. Het blijft vaak bij algemeenheden en anekdotiek: in Duitsland hecht men veel waarde aan academische titulatuur, in Zuid-Europese culturen schudt men de handen twee tot drie keer stevig, maar in Noord-Europa blijft het bij één vrij slappe korte handdruk, de kinnewip betekent in Italië iets anders dan in Frankrijk. Er wordt nauwelijks verwezen naar de inmiddels toch aardig gegroeide stapel literatuur over wat er interactioneel en communicatief gebeurt wanneer mensen uit verschillende culturen met elkaar communiceren. En de lezer die zijn interculturele communicatieve competentie hoopte te vergroten, komt met dit boek ook niet echt verder. Het slothoofdstuk doet niet wat de titel belooft, en de vele, op zich aardige, voorbeelden en oefeningen waarmee het boek is doorspekt, zijn vooral gericht op het vergroten van begrip voor en bewustwording van cultuurverschillen. Beter intercultureel communiceren leer je er niet van. Daarvoor is het verband tussen cultuur en communicatie, het openingscitaat ten spijt, te weinig doorzichtig gemaakt.

M.A. van Rees

Dikstaal, Nico (2002). *Daar worden wij niet vrolijk van. Business- en kantoorjargon van A tot Z.* Utrecht: het Spectrum.ISBN 9027476233. Prijs: € 9,-. (131 pp.)

Mensen zijn niet altijd even origineel, zeker niet in hun taalgebruik. Het is niet iedereen gegeven zinnen te produceren als "De Keizer was sigarenfabrikant" en dat is ook helemaal niet erg, want als iedereen steeds maar origineel moest zijn zou het deelnemen aan de dagelijkse gespreks- en schrijfpraktijk wel erg veel voorbereiding kosten – als je tenminste aanneemt dat Elschot lang en goed over zijn beginzin van *Een ontgoocheling* heeft nagedacht. Originaliteit is zoals bekend een overspannen eis uit de Romantiek die jammer genoeg al snel nadien bezonken cultuurgoed is geworden, terwijl toch niet zo lang voor de Romantiek ieder ontwikkeld mens nog wist dat goed naäpen – mits men er tevens naar streefde het nageaapte te overtreffen – een verdienste was. Een gedachte die weer teruggaat op de klassieke retorica. En in één interpretatie is de retorica dan ook wel degelijk een leer die de tot publieke communicatie geneigde – of door beschuldigingen aan zijn adres gedwongen – spreker of schrijver een verzameling vaste patronen, figuren en clichés biedt waaruit hij kan putten om een stilistisch of argumentatief overtuigende tekst in elkaar te zetten. Gebruik wat er al bedacht is, was het idee, want het heeft zijn waarde bewezen. Toch werd in de klassieke retorica ook al gewaarschuwd tegen het klakkeloos opvolgen van wat anderen voorschreven. In de regel boden de retorische handboeken de gebruikers dan ook geen pasklare formuleringen, maar wezen ze op effectieve stramienen, waar sprekers en schrijvers zelf op konden variëren en in het beste geval een onverwachte wending aan konden geven. De klassiek-retorische stijl is daardoor nooit echt uit de mode geraakt. En als dat dreigt te gebeuren, zijn er altijd wel sprekers als De Hoop Scheffer die de retorische traditie in leven trachten te houden. Wat helaas niet altíjd tot verheffend taalgebruik leidt. Maar hoe clichématig 'retorisch' taalgebruik soms ook moge zijn, dit taalgebruik staat nog altijd heel ver af van het taalgebruik dat het onderwerp vormt van het boekje waar deze signalering over gaat.

Nico Dikstaal, journalist, heeft in *Daar worden wij niet vrolijk van* om en nabij de zevenhonderd woorden en uitdrukkingen bij elkaar gezet die volgens hem worden gebruikt om niets te zeggen, maar dan wel op een zo interessant mogelijke manier. Dat wil zeggen, op een manier die geacht wordt interessant te zijn. Het afschuwwekkende jargon dat Dikstaal beschrijft, ontleent zijn aantrekkingskracht voor de gebruiker in de eerste plaats aan het feit dat het in de betreffende omgeving door *iedereen* wordt gebruikt – niet in de laatste plaats door degenen aan wier gezag men in die omgeving is onderworpen. Vandaar de passendheid van Dikstaals benaming van dit jargon: kantoortaal. Het grote verschil met 'retorisch' taalgebruik is dat kantoortaal niet zómaar het overnemen van clichés behelst maar de totaal gedachteloze, gezagsgetrouwe en machinale vorm daarvan.

De biotoop van Dikstaals kantoortaal is, het zal geen verbazing wekken, het kantoor. Het aardige van de voorbeelden die Dikstaal geeft is echter dat een deel ervan zonder meer ook van toepassing is op het jargon dat een aantal jaar geleden zijn intrede heeft gedaan in de academische wereld. 'Flexibiliteit', bijvoorbeeld, is op de universiteit al enige jaren een term waarmee bij vooronderstelling een hoog goed wordt aangeduid. Om tot een 'maximaal' 'rendement' van het beschikbare aantal 'fte' te komen moet de 'hokjesgeest' bestreden worden en dient een zodanige 'cultuurom-

slag' plaats te vinden dat personeelsleden (i.c. universitaire (hoofd)docenten en hoogleraren) overal 'inzetbaar' zijn. Ook 'profileren' gooit in het hedendaagse universitaire milieu hoge ogen, met name in de onderlinge strijd tussen de universiteiten om tot een efficiënte 'werkverdeling' te komen in de aanstaande 'kenniseconomie'. Om duidelijk te maken wat voor 'club' je precies bent moeten er 'keuzes worden gemaakt'. Deze laatste toverformule wordt overigens ook vaak gebruikt als de argumenten op zijn: 'Maar waarom hebt u dat nu precies gedaan?" "Welnu, er moeten keuzes gemaakt worden". Einde keten.

Het zou ook niet realistisch zijn om te veronderstellen dat Dikstaals kantoortaal beperkt zou zijn tot het kantoor. Dat dit inderdaad niet zo is, maakt zijn boekje des te aardiger.

Peter Houtlosser

Uit de tijdschriften

Levende Talen Tijdschrift, jrg. 4, nr. 2.
Britta Bendieck en Marta Stehr stellen in het openingsartikel van dit nummer vast dat de Nederlandse schooljeugd helemaal niet meer zo onberedeneerd slecht over Duitsers denkt, maar geven voor de zekerheid aan hoe leraren er in hun lessen voor kunnen zorgen dat het Duitslandbeeld verbetert. Nanette Bienfait houdt naar aanleiding van haar onderzoek naar het effect van expliciet grammaticaonderwijs aan allochtone leerlingen in een internationale 'schakelklas' een pleidooi voor een meer functioneel-communicatief gerichte vorm van taalonderwijs. Hilde Hacquebord licht de theoretische achtergronden toe van het zogenoemde ontwikkelingsonderzoek. Ze geeft aan wat voor werkwijze erin wordt gehanteerd, belicht de potentie die het heeft voor het talenonderwijs en laat zien in welke opzichten het verschilt van traditioneel wetenschappelijk onderzoek. Esther Hafkenscheid formuleert naar aanleiding van een dit jaar in Groningen begonnen project om de schrijfvaardigheid van bovenbouwleerlingen te verbeteren enkele concrete suggesties voor het schrijven van de verschillende soorten teksten die in de onderwijspraktijk aan de orde kunnen komen. In het kader van een afgerond dissertatieproject over de taal- en schiftkennis van analfabeten gaat Jeanne Kurvers in op de manier waarop heel jonge kinderen en volwassen allochtone analfabeten tegen syllogismen aankijken, hoe ze een verhaal vertellen bij een reeks afbeeldingen en hoe ze verhalen navertellen die hun eerst zijn voorgelezen.

Moer, (2002) nr. 4, (2003) nrs. 1, 2.
De laatste aflevering van de vorige jaargang (de redactie van *Moer* heeft besloten het aantal afleveringen met ingang van 2002 van zes terug te brengen naar vier) bestaat uit twee artikelen van Corrine Sebregts. In het eerste artikel legt zij uit wat de methode van het begrijpend lezen door rolwisselend leren inhoudt: de leerling neemt bij het bespreken van een zakelijke tekst geleidelijk de rol van de leraar over, waardoor de leerling beter leert begrijpen wat hij leest en de leraar gestimuleerd wordt om op een andere wijze met leesvaardigheid 'om te gaan'. In het tweede artikel legt zij uit wat de testmethode inhoudt voor het meten van de woordenschatbeheersing van leerlingen in de brugklas en laat zij zien dat de methode een goed middel vormt om tijdig op het spoor te komen van leerlingen met een zwakke woordenschatbeheersing. In de rubriek *Forum* pleiten Cristel van de Hoef en Robbert Roosenboom ervoor Marcel Reijmerinks in 1995 verschenen roman *Alles moet anders* uit de schoolbibliotheken te weren omdat de racistische uitspraken die in het boek worden gedaan niet in dienst van de personages, maar de personages in dienst van de uitspraken zouden staan. De redactie roept de lezers op om op deze vergaande stelling te reageren. Verder

is er een bespreking van de dissertatie van Mylène Hanson, *Klassengesprekken. Een interactieve benadering van onderwijs in multiculturele klassen*, van de hand van Herman Giesbers.

Het eerste nummer van de lopende jaargang is een themanummer over schrijfonderwijs. Anne-Marie van de Wiel en Piet-Hein van de Ven beschrijven een volgens hen geslaagd experiment dat is uitgevoerd aan het Nijmeegse Montessoricollege met een schrijflessenserie waarin aan de hand van een 'bouwplan' teksten geschreven, besproken, herschreven en gelezen worden. Piet-Hein van de Ven rapporteert over de geobserveerde attitude van vierde- tot zesdeklassers van het vwo ten aanzien van het schrijven. Ook doet hij verslag van interviews over schrijven met zesdeklassers van het vwo. Hans Wegman volgde een eerstejaars leerling van de lerarenopleiding van de Hogeschool Arnhem in zijn pogingen om volgens de op de hogeschool voorgeschreven methode in drie schrijfronden tot een artikel te komen voor het jeugdblad *Kijk* over een onderwerp waarover tevoren uitgebreid informatie was gegeven. Mieke Smits en Marieke Willemen geven een verslag van de schrijfactiviteiten op de Regenboogschool in Malden, die bekend staat om het sociale en educatieve belang dat men er aan het schrijven toekent. Ad van der Logt houdt een pleidooi voor het elektronische zoekprogramma Webquests dat in zijn ogen een vorm van actief leren inhoudt, direct toepasbaar is in de klas en een goed uitgangspunt biedt voor verdere educatieve activiteiten. Jeanne Kurvers bespreekt *Verder lezen. Leesteksten voor anderstaligen* van Marilene Gathier en Dorine de Kruyf.

In nummer 2 geven Ed Elbers en Mariëtte de Haan in een artikel over interculturele communicatie aan hoe leerlingen op een multiculturele basisschool in Utrecht tijdens de wiskundeles woorden behandelen waarvan de betekenis nog aan geen van hen duidelijk is. Ruth Rijks, student aan de pabo, houdt een pleidooi voor de onderwijsvorm Slash-21, die geheel zou voldoen aan de nieuwe normen van deze tijd. Ivar Gierveld, een leraar Nederlands die in deze onderwijsvorm les heeft gegeven, geeft commentaar. Linda Scheeres en Karel Stokking doen verslag van onderzoek waarin zij zijn nagegaan in hoeverre er in het voortgezet onderwijs consensus bestaat over de manier waarop het schrijfonderwijs en het daarop aansluitende examen moeten worden ingericht.

Nederlandse Taalkunde, jrg. 8, nr. 2.

Het eerste artikel in dit nummer is van Hans van de Velde en Roeland van Hout, die laten zien dat een experimentele aanpak in het bestuderen van de deletie van de slot-*n* een goed inzicht verschaft in de externe en interne factoren die de uitspraak van de slot-*n* beïnvloeden. Reinhild Vandekerckhove doet verslag van onderzoek naar de distributie van de betrekkelijke voornaamwoordsvormen *die* en *dat* in Westvlaamse dialecten en in de Westvlaamse standaardtaal. Fout gebruik van de beide vormen blijkt precies daar op te treden waar Westvlaamse dialecten afwijken of kunnen afwijken van het Standaardvlaams. Helena Taelman en Steven Gillis stellen vast dat de neiging van tweejarige kinderen om in hun prosodische structurering het ritme te volgen van de trochee niet afdoende verklaart waarom ze bij het uitspreken van woorden lettergrepen weglaten of inkorten en accenten van de ene naar de andere lettergreep verplaatsen.

In de rubriek *Digitaal* gaan Joep Kruijsen en Jos Swanenberg na wat de Limburgse en Brabantse dialectdatabanken op het internet te bieden hebben. Marc Oostendorp bespreekt het fonologieboek *Optimality theory* van René Kager, Jan Nijen Twilhaar de taalkundig-antropologische

vergelijking tussen Oost en West in Jenny van der Toorn-Schuttes *Cultuur en tweedetaalverwerving*, Johan De Caluwe *The morphology of Dutch* van Geert Booij en Folkert Kuiken de Groningse dissertatie *Grammaticaonderwijs aan jongeren* van Nanette Bienfait.

Onze Taal, jrg. 72, nrs. 5, 6, 7/8.
Bart Bossers vraagt zich in nummer 5 af of het einde van het onderwijs in de eigen taal al in zicht is – wat hier niet wil zeggen het einde van het Nederlands in het universitaire onderwijs (voor sommige zorgelijk ingestelden ook een punt), maar het einde van het middelbaar onderwijs in de eigen allochtoonse taal; dit in verband met de aangekondigde (maar ondertussen vertraagde) afschaffing van het Onderwijs in Allochtone Levende Talen. Ninke Stukker en Frank Jansen constateren dat de hedendaagse *direct mail* geen terughoudendheid meer kent. Hoe directer, hoe beter, is tegenwoordig blijkbaar het idee. Mark Traa legt uit waarom mensen tegen dieren en, in het ergste geval, tegen planten praten. Het is een beetje als tegen jezelf praten, met dit verschil dat er in het laatste geval voornamelijk misnoegen wordt geuit, terwijl de dieren en de planten doorgaans op schaamteloze liefdesbetuigingen kunnen rekenen. Christine Swankhuisen en Bert Pol, van het bureau 'Tabula Rasa strategische communicatie', bespreken allerlei psychologische onderzoeken die uitwijzen dat mensen onbewust beïnvloed worden door de woordkeus van teksten die ze lezen. Schrijvers die graag effectieve teksten schrijven kunnen er hun voordeel mee doen. In de inleiding staat een wel heel eigenaardige bewering: iemand die een tekst leest met veel woorden van het type *bloemetjesjurk*, *rollator* en *vergeetachtig* gaat daarna vanzelf wat langzamer lopen. Lia van Elk en Peter-Arno Coppen vermoeden dat de partikels de laatste tijd zo sterk terrein winnen omdat ze de zinsstructuur duidelijker en daardoor het leven gemakkelijker maken. René Appel interviewt de van oorsprong Marokkaanse schrijver Hafid Bouazza, die veel blijkt te hebben opgestoken van de poëzie van Herman Gorter. Liesbeth Koenen bespreekt de handleiding voor het verbeteren van het Engels *Eindelijk Engels!* van Kevin Cook en Daniel Gibb.

Nummer 6 begint met een artikel waarin Ton den Boon de nieuwe woorden inventariseert die de oorlog tegen Irak heeft opgeleverd. Naast de uit de Eerste Wereldoorlog stammende dikke Bertha blijkt er nu ook een domme bom te bestaan, die anders dan de naam doet vermoeden een eigen willetje heeft. Het Poldernederlands is nog steeds niet uit de mode en zal dat voorlopig ook wel niet raken, want Loulou Edelman heeft met behulp van het computerprogramma *Praat* aangetoond dat het soort vrouwen dat volgens Jan Stroop de *ei* op zijn Poldernederlands als *aai* uitspreekt dat ook inderdaad doet. Stroop doet er zelf nog een schepje bovenop door te stellen dat de *aai*-klank ook steeds vaker te horen is doordat woorden als *me* en *z'n* steeds vaker als *mij* (-> *maai*) en *zijn* (-> *zaain*) worden uitgesproken. Robert Beekes toont aan dat *pietlut* in overeenstemming met de bevinding van de negentiende-eeuwse spreekwoordenboekenmaker P.J. Harrebomée en anders dan Harrebomées critici altijd zeiden wel degelijk uit de Bijbel afkomstig is (via *Put* en *Lud*, en *putlut*). In een zoveelste poging om op grond van tendensen in het verleden een voorspelling te doen over toekomstige ontwikkelingen, werpt Joop van der Horst ditmaal een loodje over het heden heen om, gegeven de huidige ontwikkeling van sterke naar zwakke verledentijdsvormen in samenstellingen (*gezogen* tegenover *gestofzuigd*), de ontwikkeling te peilen van sterke naar zwakke verleden-

tijdsvormen in het algemeen. Een goede voorspelling doen is ditmaal moeilijker, omdat er in het verleden vaak heen en weer is gezwabberd tussen sterk en zwak. Marcel Lemmens laat zien dat de verschuiving van bijwoorden in een zin vaak onbedoelde betekenisveranderingen oplevert. Zo suggereert 'We gaan eens even kijken of we opnieuw contact met Parijs kunnen krijgen' dat men, mits de aangekondigde poging slaagt, voor de tweede keer 'contact' met Parijs zal krijgen, terwijl de zin juist wordt uitgesproken omdat men nog *geen* 'contact' met Parijs heeft kunnen krijgen. Marian van Eupen, bedenkster van quizvragen, benadrukt dat het zodanig 'dichttimmeren' van quizvragen dat meer dan één goed antwoord wordt uitgesloten een karweitje is waar ook een goede taalkennis voor nodig is. Gegeven dat er tegenwoordig nogal wat – geld – op het spel staat, kan een niet-weloverwogen formulering tot grote problemen leiden. Leonie Hagen geeft aan welke slinkse wegen bedrijven nu al volgen om het in aantocht zijnde verbod op het stellen van leeftijdseisen in personeelsadvertenties te ontlopen. Zo kan een 'jong en dynamisch' bedrijf vanzelf geen oude knar als portier gebruiken. (Zo'n bedrijf zou wel kunnen worden veroordeeld wegens het begaan van een divisiedrogreden.) Marc Oostendorp, in de vorige aflevering daarvan afgehouden door zijn rubriek over taal voor kinderen, bespreekt weer een proefschrift. Het werkstuk heet *Taal- en formuleringsproblemen in de regelgeving. De taalopmerkingen in de adviezen van de Raad van State* en is geschreven door Karl Hendrickx.

Tekst[blad], jrg. 9, nr. 2.

Alaude Jaasma probeert te doorgronden hoe de kantoormeubelen- en kantoorartikelenfirma Ahrend communicatief gezien gestalte geeft aan het beeld dat men volgens de firma van de firma zou moeten hebben. Anneke Wurth, Jaap de Jong en Bas Andeweg bieden de lezers een schriftelijke versie van hun lezing 'Ik vlei veilig of ik vlei niet', gehouden op het VIOT-congres in Antwerpen in december 2002, waarin zij laten zien dat Nederlandse toespraakadviseurs en Nederlandse toespraakschrijvers eensgezind zijn in hun oordeel over de exordiale *captatio benevolentiae*-toop: geen geslijm, s.v.p. Voor tekstschrijvers die graag zouden optreden tegen opdrachtgevers die de naam van de tekstschrijver uit het colofon weglaten, maar niet weten hoe, biedt het artikel van Christina Mercken uitkomst. Het geeft aan wat de rechten zijn die de tekstschrijver in dezen kan laten gelden en wat hij kan doen als die rechten geschonden worden. Rijk Willemse is van mening dat tekstschrijvers niet alleen over het noodzakelijke vakmanschap moeten beschikken, maar ook over de vaardigheden om 'in woord en daad' de opdrachtgever te begeleiden in het proces waarin de door de opdrachtgever gewenste tekst tot stand komt. Dat betekent onder meer dat hij al in het stadium van de conceptualisering in dat proces betrokken moet zien te worden en er ook tot het bittere einde bij moet blijven. Als oud-beoordelaar licht Peter Bardoel de potentiële deelnemers aan de Internationale Competitie voor Technische Publicaties in over de ins en outs van de competitie. Derk Eimers heeft een artikel geschreven over de verschillende betekenissen van wat een van de belangrijkste woorden in de professioneelschrijvenwereld blijkt te zijn, 'netwerken'. Ben Vroom geeft de tien belangrijkste obstakels weer die de zoekende websitebezoeker de zoektocht kunnen bemoeilijken. In een aansluitend artikel legt hij uit dat de oplossing van de problemen bij het zoeken niet verkregen wordt als men probeert alle problemen op te lossen maar dat de belangrijkste problemen kunnen worden opgelost als

men zich op het oplossen van de belangrijkste problemen concentreert. Gwennie Bosma en Ariane Volz moesten de voorlichting over de kinderbijslag verzorgen. In dit artikel leggen ze uit dat ze daarbij als 'communicatieconcept' het nobele uitgangspunt hebben gekozen dat in de voorlichting de taal van de klant moet worden gesproken en niet die van de regelgeving. Judith Mulder interviewt Carel Jansen, hoogleraar Bedrijfscommunicatie aan de Letterenfaculteit van de Katholieke Universiteit Nijmegen, over zijn recente onderzoek naar AIDS-voorlichting in Zuid-Afrika. Marcel Uljee en Edwin Lucas ondervragen copywriter Robin Kemme over het tot stand komen van de twee radioreclames en de printadvertentie die hij gefabriceerd heeft en over zijn werk als copywriter in het algemeen. Ferdinand Pronk interviewt Betteke van Ruler, bijzonder hoogleraar bij de afdeling Communicatiewetenschap van de Universiteit Twente, over de professionalisering van de communicatiewetenschap. Susanne de Joode recenseert *Medische publiekscommunicatie* onder redactie van Frans J. Meijman en Frans Meulenberg. De column is dit keer van Peter Jan Schellens, die mismoedig wordt van alle opgewekte reclame- en voorlichtingsteksten die hij zich in het hedendaagse dagelijks leven moet laten welgevallen.

Tijdschrift voor Communicatiewetenschap, jrg. 31, nrs. 1, 2.

Nr. 1 van deze jaargang opent met een bijdrage aan een nieuwe rubriek getiteld 'Actuele discussies in de communicatiewetenschap'. De bijdrage komt van Ruben Koning, Allerd L. Peeters en Hans Beentjes, die duidelijk maken dat de discussie over de relatie tussen agressief gedrag van jongeren en geweld op de televisie niet is afgesloten met het recente geruchtmakende artikel van Jeffrey Johnson en anderen in *Science*, waarin het uiteindelijke bewijs voor de invloed van het laatste op het eerste definitief zou zijn geleverd. Moniek Buijzen en Patti Valkenburg doen in een regulier artikel verslag van een onderzoek waarin in een zogenoemde *vote-counting*-analyse de validiteit werd getoetst van de hypothese dat televisiereclame kinderen – onbedoeld – materialistisch en ongelukkig maakt en ook nog eens tot conflicten met de ouders leidt. Hoewel in geen geval een causaal verband kon worden aangetoond, werden alle hypothesen valide bevonden. Marianne Simons, Melanie van Wijk en Jan de Ridder rapporteren over onderzoek waarin zij vaststelden dat elektronische media (ICT) op organisationeel niveau een versnellende en versterkende invloed hebben op de identificatie van mensen met het bedrijf of de organisatie waar ze werken. Leen d'Haenens, Nick Jankowski, Ard Heuvelman, Cindy van Summeren en Madelon Kokhuis vinden in een onderzoek onder lezers van elektronische en gedrukte versies van *De Telegraaf* en *de Gelderlander* geen evidentie voor de hypothese dat mensen het nieuws dat ze in een elektronische versie tot zich nemen anders 'consumeren' en onthouden dan het nieuws dat ze in gedrukte versie lezen. Dianne Alting en Paul Nelissen zijn met het oog op de ontwikkeling van doelmatige 'alcoholmatigingscampagnes' nagegaan wat de achtergronden zijn van het alcoholgebruik onder studenten. Naast 'onberedeneerde factoren' en groepsgeest bleken vooral de hooggespannen verwachtingen over het resultaat van het innemen het drankgebruik te bevorderen. Hans van Driel bespreekt *Consumer behavior and managerial decision making* (2002) van Frank R. Kardes.

Machteld Smid en Hans Beentjes rapporteren in aflevering 2 over onderzoek waarin werd nagegaan of kinderen (uit allerlei etnische groepen) personages in het

televisieprogramma *Sesamstraat* herkennen als behorend tot een bepaalde etnische groep en, als dat zo is, in hoeverre hun waardering van die personages op 'etniciteit' gebaseerd is. De kans op herkenning bleek groter naarmate de huidskleur donkerder was en de context exotischer. De waardering werd hoofdzakelijk beïnvloed door gedrag en kleding, niet door etniciteit. Steven Eggermont bevestigt grotendeels de resultaten van onderzoek waarin werd vastgesteld dat 15- en 16-jarigen hun romantische verwachtingen sterk laten beïnvloeden door het ideaalbeeld van de knappe en sociaal vaardige liefdesgenoot dat ze in televisieprogramma's wordt voorgeschoteld. Wel is het zo dat die programma's een minimaal gehalte aan realisme moeten hebben en dat ouders en vrienden nog enige herstellende invloed op het werkelijkheidsbesef van de puber kunnen hebben. Kees van Rees en Koen van Eijck analyseren de samenstelling van de 'mediarepertoires' van de Nederlandse bevolking. Ze laten zien dat het media-consumerende publiek inzichtelijk kan worden 'gesegmenteerd' in onderscheiden publieken naar voorkeur voor type inhoud en type medium en naar een bepaalde combinatie tussen deze twee en dat de verschillen tussen de publiekssegmenten kunnen worden geïnterpreteerd met behulp van aanvullende kenmerken zoals tijdsbesteding en politieke interesse. Mark Deuze bespreekt *Media/Society: Industries, images, and audiences* (2002) van Favid Croteau en William Hoynes, Ellen Hijmans *Dilemma's in menselijke interactie* (2001) van Erica Huls en Jan Steyaert *Interface en cyberspace: Inleiding in de nieuwe media* van Jan Simons.

Toegepaste Taalwetenschap in Artikelen, (2002) nrs. 67, 68, (2003) nr. 69.

Nummer 67 bevat bijdragen van Anne-Marie van Hoof over de gebaren die mensen bewust of onbewust onder het spreken maken, Claudia Boonstra en Mirjam Koop over taalstimulering op peuterspeelzalen in de gemeente Emmen, Irene Willems en Leo Noordman over het begrijpen van semantische, volitionele en epistemische causale relaties in de context waarin ze worden gebruikt, Maritta Moisio over het maken van een Nederlandse grammatica voor Finse leerlingen, Suzanne Adema over de verhouding tussen voor- en achtergrond in de Aeneïs van Vergilius, Floor Buschenhenke over de geleidelijke afname van de moedertaal in het lexicon en An Neven over de ontwikkeling van een dieptetoets voor het meten van woordenschatkennis op basis van een combinatie van zelfstandig naamwoorden en werkwoorden.

Nummer 68 is een themanummer over taalbewustzijn. Er staan artikelen in van Jeanne Kurvers, die de metatalige kennis van analfabeten vergelijkt met die van kinderen en laagopgeleiden, Ineke van de Craats, die onderzoekt in hoeverre taalbewustzijn kan bijdragen aan het gemakkelijker verwerven van een tweede taal, June Eyckmans, Frank Boers en Renaud Beeckmans, die het belang onderzoeken van taalbewustzijn bij het verwerven van idiomatische uitdrukkingen in een vreemde taal, Anne Baker en Beppie van den Bogaerde, die de rol van 'code-mixing' onderzoeken in moeder-kind-interacties in dove families, en Roel van Steensel, die het verband beproeft tussen het type gezin waarin een kind opgroeit, de mate waarin het in de voor- en vroegschoolse periode thuis tot leren wordt gestimuleerd en het succes dat het vervolgens op school heeft. Annie van der Beek heeft vastgesteld dat leerlingen in kleine-kringgesprekken met de leraar complexere cognitieve taalfuncties gebruiken dan in traditionele klassikale situaties, Anne Vermeer behandelt de vooroordelen van leraren over de taalvaardigheid Nederlands van autochtone en

allochtone jongens en meisjes in het speciale basisonderwijs en Arina Byrdina en Korrie van Helvert gaan na hoe met de brontaal en de doeltaal kan worden omgesprongen in het onderwijs in het schrijven in een vreemde taal.

In het laatstverschenen nummer 69 laten Carry van de Guchte en Anne Vermeer zien wat een passende woordkeus kan inhouden bij het kiezen van woorden voor woordenschatlessen, pleit Alied Blom voor een tekstgerichte benadering in het taalverwervingsonderwijs, zet Josje Verhagen uiteen welke rol typologische universalia spelen in het leren van Nederlands als tweede taal, bespreekt Yvonne Schotvanger exploratief corpusonderzoek naar het leren gebruiken van partikelwerkwoorden in normale en gestoorde taalontwikkeling, doet Annika Nonhebel verslag van een kwalitatieve studie naar de non-manuele markering van indirecte verzoeken in de Nederlandse gebarentaal, rapporteert Merel Keijzer over onderzoek naar de effecten van een remediërend softwarepakket op de Engelse spelvaardigheid van Nederlandse dyslectische en taalzwakke leerlingen in het voortgezet onderwijs, zet Petra Jongmans een linguïstische benadering uiteen van een hoortrainingsprogramma voor dragers van cochleaire implantaten, bespreekt Anke Hulsker de ontwikkeling van een diagnostische toets voor het meten van de leesvaardigheid Engels, doet Annemarieke Hoekstra verslag van onderzoek naar de interactie tussen leerling en leraar in individuele 'hulpmomenten' tijdens de wiskundeles, behandelt Gerard Doetjes de rol van taalvariatie en taalafstand in de communicatie tussen Zweden, Noren en Denen, gaat E.L. van den Broek in op het fenomeen van het leggen van nadruk in gesproken taal en geeft Mariken Bindels een analyse van zogenoemde *chat*gesprekken tussen basisschoolleerlingen en onderwijsassistenten.

Peter Houtlosser

Nieuws uit het vakgebied

Promoties
Henrike Jansen is op 18 september aan de Universiteit van Amsterdam gepromoveerd op het proefschrift *Van omgekeerde strekking. Een pragma-dialectische reconstructie van a contrario-argumentatie in het recht*. Promotor was prof.dr. F.H. van Eemeren, co-promotor dr. E.T. Feteris.

Benoemingen
Met ingang van 1 oktober 2003 is dr. Luuk Van Waes aangesteld als hoogleraar Zakelijke en Technische Communicatie aan de faculteit Toegepaste Economische Wetenschappen van de Universiteit Antwerpen.

Met ingang van 1 januari 2003 is prof. dr. mr. Paul van den Hoven benoemd tot hoogleraar Taal en Communicatie bij het Onderwijsinstituut Media en Representatie van de Faculteit der Letteren van de Universiteit Utrecht. Bij het Onderwijsinstituut Nederlandse taal en cultuur van de zelfde faculteit is dr. Ted Sanders met ingang van 1 juli 2003 benoemd tot hoogleraar Taalbeheersing van het Nederlands. De leerstoelgroepen Taal en Communicatie en Taalbeheersing van het Nederlands vormen samen de disciplinegroep Taalbeheersing, met een gemeenschappelijke onderwijs- en onderzoekstaak.

Met ingang van 1 maart is dr. Henrike Jansen vast aangesteld aan de Universiteit Leiden als universitair docent bij de sectie Taalbeheersing van de opleiding Nederlands.

Congressen
Aan de Universiteit Leiden wordt op zaterdag 15 november 2003 naar aanleiding van het twintigjarig bestaan van de Leidse Taalbeheersing als afstudeerrichting een Taalbeheersingsdag gehouden. Het thema van de dag is *Retorica in de beroepspraktijk*. Er is een programma met lezingen van Toine Braet, Agnes Verbiest en Jaap de Jong en al dan niet ludieke bijdragen van oud-studenten. Er verschijnt ook een boek, *Retorica in de beroepsprakrijk. Leidse taalbeheersers over hun professionele passies*. Nadere informatie bij Henrike Jansen 071-5272120 of h.jansen@let.leidenuniv.nl.

Onderwijsprojecten
In het kader van het Socratesprogramma Minerva Project 2003-2005 zijn de Universiteit Antwerpen, de Universiteit Nijmegen, de Kungl Tekniska Högskolan Stockholm en de Uniwersystet Warszawski gezamenlijk begonnen met de ontwikkeling en implementatie van een digitaal, modulair, multi-linguaal Europees Schrijfcentrum dat ten doel heeft studenten schrijf- en leesvaardigheden bij te brengen op het gebied van de technische, zakelijke en wetenschappelijke communicatie en de huidige schrijfdidactiek verder te ondersteunen en te complementeren. Meer concreet wordt gestreefd naar de ontwikkeling van een aangepaste probleemgerichte didactiek voor het onderwijs van schriftelijke vaardigheden in professionele communicatie, een Europees netwerk van gespecialiseerde instituten om deze digitale omgeving effectief te integreren in een brede onderwijssituatie en een programma dat erop gericht is de integratie van het digitale schrijfcentrum bij docenten en studenten zo vlot mogelijk te laten verlopen.