Carel Jansen en
Nynke Boersma

Meten is weten? Over de waarde van de leesbaarheidsvoorspellingen van drie geautomatiseerde Nederlandse meetinstrumenten

Samenvatting

Uit eerder onderzoek van Kraf, Lentz & Pander Maat (2011) is bekend dat de leesbaarheidsvoorspellingen van drie geautomatiseerde instrumenten voor Nederlandse teksten aanzienlijke verschillen kunnen vertonen. Nog niet helder was echter hoe de uitslagen van deze drie instrumenten (*Texamen*, *Klinkende Taal* en *de Accessibility Leesniveau Tool*) zich verhouden tot de feitelijke begrijpelijkheid van de teksten waarover een voorspelling wordt gedaan. Om daarover uitspraken te kunnen doen, is een onderzoek uitgevoerd met dezelfde negentien teksten die Kraf et al. gebruikten voor de onderlinge vergelijking van de drie instrumenten. Van elk van deze teksten werden vijf cloze test-varianten gemaakt. Van de 95 resulterende cloze tests werden er steeds drie of vier voorgelegd (over verschillende onderwerpen) aan 125 deelnemers aan dit onderzoek. Zo werden in totaal 475 cloze scores verzameld, die vervolgens in verband konden worden gebracht met de voorspellingen van de drie instrumenten. In twee gevallen (de Accessibility Leesniveau Tool en Klinkende Taal) werd een significante maar niet erg hoge correlatie gevonden. Met de uitslagen van Texamen bleken de gemiddelde cloze scores voor de negentien teksten niet significant te correleren. Ook werd nagegaan hoe het tekstbegrip samenhing met een aantal tekstkenmerken. Gevonden werd dat twee tekstkenmerken samen een groot deel van de variantie in de cloze scores verklaarden: gemiddelde woordlengte en gemiddelde zinslengte. Bij gebruik van deze twee klassieke variabelen uit het leesbaarheidsonderzoek bleken andere tekstkenmerken, zoals, de proportie frequente woorden en de type-token ratio, geen significante bijdrage aan de verklaarde variantie meer te leveren.

1 Inleiding

In een themanummer van het *Tijdschrift voor Taalbeheersing* uit 2011, getiteld *Begrijpelijke Taal*, bespreken Kraf, Lentz en Pander Maat (2011) de ontwikkeling van drie geautomatiseerde Nederlandse meetinstrumenten voor de leesbaarheid van teksten: *Texamen* van BureauTaal, *Klinkende Taal* van Gridline en de *Accessibility Leesniveau Tool* van de Stichting Accessibility. Ook doen Kraf et al. verslag van een experiment waarin ze de uitkomsten van deze instrumenten, alle drie beschikbaar op het internet,[1] met elkaar vergeleken. Daartoe bepaalden ze de onderlinge correlaties van de leesbaarheidsscores van deze drie instrumenten bij negentien teksten over vier verschillende onderwerpen. Verder gingen ze na hoe die leesbaarheidsscores samenhingen met enkele tekstkenmerken waarvan ze verwachtten dat die daar verband mee zouden houden. Voor wie op een snelle manier wil bepalen hoe leesbaar (of begrijpelijk; we gebruiken deze twee termen verder door elkaar) een bepaalde tekst is, leveren de resultaten van Kraf et al. al een goed beeld van de verschillen en overeenkomsten tussen deze instrumenten. Ook biedt het

werk van Kraf et al. nuttige informatie over de vraag door welke tekstkenmerken de scores van de vergeleken drie instrumenten meer en minder bepaald worden. Maar aan het eind van hun artikel wijzen de auteurs er met recht op dat daarmee nog geen onderzoek beschikbaar is waaruit blijkt wat de relatie is tussen begripsprestaties van volwassen Nederlanders enerzijds en de scores van de instrumenten anderzijds: "We kunnen daarom niets zeggen over hun validiteit." (p. 264).

In dit artikel bespreken we een onderzoek dat we hebben uitgevoerd met als belangrijkste doel wél uitspraken over de validiteit van de drie instrumenten te kunnen doen. We hebben daartoe dezelfde negentien teksten die gebruikt werden door Kraf et al. (2011) voorgelegd aan een groep van 125 lezers en we hebben bij hen nagegaan in welke mate ze de teksten begrepen. Tweede doel van ons onderzoek was na te gaan in hoeverre de feitelijke begrijpelijkheid van de teksten voorspeld kon worden uit een combinatie van een beperkt aantal tekstkenmerken. Daartoe hebben we de begripsscores van de lezers gerelateerd aan zeven tekstkenmerken, waaronder ook de kenmerken die Kraf et al. in hun onderzoek betrokken. Hoe ons onderzoek is uitgevoerd en wat de resultaten waren, komt aan de orde in paragraaf 3 en 4. Eerst gaan we kort in op de geschiedenis van het leesbaarheidsonderzoek, op de beslissingen die er bij de ontwikkeling van de nieuwe instrumenten zijn genomen, en op de uitkomsten van het vergelijkende onderzoek van Kraf et al. (2011).

2 Voorgeschiedenis

Zoals onder meer blijkt uit een overzichtsartikel van Jansen en Lentz (2008) wordt er inmiddels al zo'n negentig jaar gewerkt aan de ontwikkeling van leesbaarheidsformules: instrumenten die op basis van een aantal eenvoudig uit te voeren tellingen van kenmerken van een tekst een deugdelijke voorspelling over de begrijpelijkheid van die tekst opleveren. De leesbaarheidsformule die internationaal verreweg de grootste faam heeft verworven, heeft intussen de pensioengerechtigde leeftijd bereikt: de 'Reading Ease'-formule van Flesch uit 1948. Voor toepassing van deze formule was niet veel meer nodig dan het tellen van de zinnen, woorden en lettergrepen in de tekst, of in enkele fragmenten daarvan. De gevonden getallen moesten in deze formule worden ingevoerd: RE = 206,84 -(1,02 * SL) -(0,85 * WL). "RE" stond voor 'Reading Ease' (leesgemak), "SL" voor het gemiddelde aantal woorden per zin, en "WL" voor het gemiddelde aantal lettergrepen per woord. Hoe hoger de uitkomst van de formule, hoe gemakkelijker de tekst. Een kind kon de was doen, in het Engelse taalgebied maar ook in Nederland en Vlaanderen – zeker toen Douma hier zijn variant van de formule van Flesch had geïntroduceerd (Jansen & Lentz, 2008, p. 4).

In de jaren zeventig van de vorige eeuw kwam er van de kant van Nederlandse en Vlaamse taal- en tekstwetenschappers voor het eerst serieuze belangstelling voor leesbaarheidsformules. Zondervan en Van Caldenborgh (1971-1972) belichtten het fenomeen in een kritisch artikel; Van Hauwermeiren (1974-1975) ontwikkelde zijn eigen formules op basis van de scores van respondenten op *cloze tests* (zie hieronder meer over deze toetsvorm), en Jansen en Woudstra (1979) en Palm (1979) uitten stevige bezwaren tegen de manier waarop eerdere formules, maar ook die van Van Hauwermeiren tot stand gekomen waren. De kritiek uit wetenschappelijke hoek op de leesbaarheidsformules leidde tot een steeds negatievere

houding in de schrijfadviesliteratuur en daarmee tot een sterk teruglopende populariteit. Dat veranderde toen de personal computer gemeengoed werd, en het niet meer nodig was om zelf uit te rekenen wat de gemiddelde woord- en zinslengte was. Zeker nadat Microsoft besloten had om voor een aantal talen (niet voor het Nederlands) in het tekstverwerkingsprogramma Word leesbaarheidsformules in te bouwen als optie in de spelling- en grammaticacontrole, werd het een koud kunstje om de leesbaarheid van een tekst 'objectief' vast te (laten) stellen. De berekeningen die deze toen modern aandoende software automatisch uitvoerde, berustten echter nog steeds op de klassieke leesbaarheidsformules zoals die van Flesch uit 1948. De fundamentele bezwaren tegen die formules golden en gelden onverkort voor de software-varianten in de tekstverwerkingsprogramma's. Zoals gesteld in Jansen en Lentz (2008): het zag er indrukwekkend uit, maar het was oude wijn in nieuwe zakken. De wetenschappelijke basis was, ondanks de moderne technologie, nog net zo pover als in het midden van de vorige eeuw.[2]

Te vrezen valt dat er ook bij de nieuwste loten aan de tak van de Nederlandse leesbaarheidsformules, *Texamen*, *Klinkende Taal* en de *Accessibility Leesniveau Tool*, geen sprake is van een degelijk fundament. Zoals Kraf et al. (2011) vaststellen, is geen van deze drie meetinstrumenten gebaseerd op onderzoek dat is uitgevoerd bij echte lezers. Bij de ontwikkeling van *Texamen* werd een aantal kenmerken van tweehonderd teksten gekoppeld aan inschattingen van het begrijpelijkheidsniveau van die teksten door een onbekend aantal MBO-docenten. Bij *Klinkende Taal* werd eenzelfde procedure gevolgd maar kwamen de inschattingen van de kant van communicatie-experts met ervaring als trainer. En ook bij de *Accessibility Leesniveau Tool* hebben echte lezers geen rol gespeeld; voor de ontwikkeling van dit instrument is per niveau door uitgeverij *Eenvoudig Communiceren* ingeschat wat voor teksten daarbij zouden passen (zie Kraf et al., 2011, p. 253; 257; 259; 264).

Een tweede probleem rond de drie instrumenten betreft de vorm waarin de uitkomsten worden gegoten. In afwijking van de eerdere leesbaarheidsformules, waarvan de uitslagen werden uitgedrukt in getallen die onderling vergeleken en eventueel met gewenste opleidingsniveaus in verband gebracht konden worden, worden de scores van *Texamen*, *Klinkende Taal* en de *Accessibility Leesniveau Tool* uitgedrukt in zogenoemde 'taalniveaus'. Het gemakkelijkst zouden teksten zijn op A1-niveau, en dan loopt de moeilijkheidsgraad op van A2, B1, B2 en C1 tot en met C2. Het idee dat elke tekst een taalniveau zou hebben, is afkomstig van BureauTaal, dat die taalniveaus heeft afgeleid uit het *Common European Framework of Reference* (Engels: CEFR; Nederlands: ERK) voor de talen, dat in 2011 werd vastgesteld door de Raad van Europa. Tegen het concept taalniveau als maat voor tekstbegrijpelijkheid kunnen diverse bezwaren worden ingebracht. Om te beginnen heeft het CEFR helemaal geen betrekking op het moeilijkheidsniveau van teksten. Het gaat in het CEFR enkel en alleen over vaardigheden van mensen om mondeling en schriftelijk met elkaar te communiceren, en dan nadrukkelijk in een Europese taal die niet hun moedertaal is. Sinds enkele jaren kan dankzij de zes niveaus van het CEFR overal in Europa een eenduidige betekenis worden gegeven aan een uitslag van een test die bijvoorbeeld luidt dat een Nederlander een spreekvaardigheid in het Portugees heeft op niveau A1 ('beginner') of A2 ('beginner plus'), dat een Griek Italiaans spreekt op niveau of B1 ('halfgevorderde') of B2 ('gevorderde'), of dat een Duitser een leesvaardigheid heeft in het Frans op niveau C1 ('vergevorderde') of C2 ('bijna moedertaal'). Verder gaat het CEFR niet. Beweren dat een tekst een taalniveau heeft van

bijvoorbeeld B1 of C2, is niet meer dan een slag in de lucht. Zo'n uitspraak kan pas gedaan worden als er onderzoek is uitgevoerd waarin kenmerken van teksten gekoppeld zijn aan de vaardigheidsniveaus van lezers in termen van het CEFR. Zulk onderzoek is echter nog niet beschikbaar (zie ook Jansen, 2013).

Een derde probleem blijkt uit het vergelijkende onderzoek van Kraf et al. (2011). Daarin bleken de scores van de drie instrumenten flink van elkaar te verschillen. Er bleken maar vier van de zes theoretisch beschikbare taalniveaus voor te komen: geen enkele tekst scoorde op een of meer van de instrumenten taalniveau A1 of A2. Desondanks kwamen de uitslagen van *Texamen* en van *Klinkende Taal* in slechts tien van de negentien gevallen overeen, waren de scores van *Texamen* en de *Accessibility Leesniveau Tool* in slechts vijf van de negentien gevallen hetzelfde, en gold dat in dertien gevallen voor de uitslagen van *Klinkende Taal* en de *Accessibility Leesniveau Tool*. Tabel 1 (overgenomen uit Kraf et al., 2011) bevat de rangordecorrelaties tussen de scores van de drie instrumenten.

Tabel 1: Rangordecorrelaties tussen de taalniveaus toegekend door de drie instrumenten (uit Kraf et al., 2011, p. 261).

	Texamen	***Klinkende Taal***	***Accessibility***
Texamen	1	.64⋆	.60⋆
Klinkende Taal		1	.75⋆
Accessibility			1

⋆ Statistisch significant ($p<.01$)

Met Kraf et al. (2011, p. 261) constateren we dat hoewel alle gevonden correlaties significant zijn op 1%-niveau, een aanzienlijk deel van de variantie ($1-r^2$ namelijk) onverklaard blijft. Dat geldt met name bij de confrontatie van de uitslagen van *Texamen* en de andere twee instrumenten; in die gevallen zou de onverklaarde variantie bij intervaldata 59%, respectievelijk 64% zijn.

Als de uitslagen van drie instrumenten, zoals hier het geval is, onderling flink verschillen terwijl ze geacht worden hetzelfde te meten, dan kan dat verschillende oorzaken hebben. Denkbaar is dat een van de instrumenten in de meeste gevallen correcte voorspellingen oplevert, terwijl de uitkomsten van de andere instrumenten vaak incorrect zijn. Mogelijk is ook dat twee van de instrumenten veel correcte scores opleveren, en dat de uitkomsten van het derde instrument vaak incorrect zijn. Ten slotte kan het zo zijn dat geen van de instrumenten adequaat functioneert, en dat ze dus alle drie feitelijk incorrecte voorspellingen opleveren. Welke van de genoemde opties aan de orde is, valt alleen vast te stellen door de voorspellingen te toetsen aan de werkelijkheid, in dit geval door de relaties te bepalen tussen de scores van de instrumenten en de begripsprestaties van echte lezers.

Zoals hierboven gezegd moesten Kraf et al. (2011) de vraag naar de validiteit van de drie nieuwe instrumenten onbeantwoord laten. Hun onderzoek liet geen conclusies toe over de relatie tussen feitelijk tekstbegrip en scores van de instrumenten. Met het onderzoek dat we hieronder bespreken wilden we wel tot zulke conclusies kunnen komen, en wel door leesbaarheidsvoorspellingen van de drie instrumenten over een aantal teksten te confronteren met begripsprestaties bij echte lezers van die teksten. Daarnaast gingen we na hoe de begripsscores van die lezers samenhingen met een beperkt aantal tekstkenmerken.

3 Methode

3.1 Materiaal Gebruik is gemaakt van dezelfde teksten als die waarmee Kraf et al. (2011) nagingen in hoeverre de meetresultaten van *Texamen*, *Klinkende Taal* en de *Accessibility Leesniveau Tool* met elkaar overeenkwamen. Dat materiaal bestond oorspronkelijk uit twintig teksten, maar voor slechts negentien van die teksten konden Kraf et al. beschikken over de uitslagen van alle drie instrumenten. Van één tekst ontbrak in de informatie die BureauTaal ter beschikking stelde het meetresultaat van *Texamen*. Deze tekst is ook in ons onderzoek niet meegenomen. De negentien teksten die ook wij gebruikten, bestonden steeds uit ongeveer 270 woorden (M=270.74; SD=10.41). De teksten waren verdeeld over vier categorieën. Vijf teksten waren afkomstig van een woningbouwvereniging en waren gericht aan huurders, in vijf gevallen ging het om krantenberichten over uiteenlopende binnenlandse onderwerpen, er waren vier fragmenten uit polisvoorwaarden van verzekeringen, en vijf teksten kwamen uit de roddelrubriek van een landelijk dagblad. Zo wilden Kraf et al. (2011) een zekere spreiding in zowel onderwerp als complexiteit bereiken; datzelfde gold dus voor ons onderzoek.

3.2 Toets: de cloze test Als instrument om de begrijpelijkheid van de negentien teksten bij werkelijke lezers te meten, werd de *cloze test* gebruikt, oorspronkelijk ontwikkeld door Taylor (1953). In een *cloze test* worden systematisch woorden weggelaten uit de tekst en vervangen door een streepje, en wordt aan de deelnemers aan het onderzoek gevraagd om op de streepjes precies die woorden te noteren waarvan zij vermoeden dat die in de eigenlijke tekst gestaan hebben. Het aantal correcte antwoorden wordt uitgedrukt in een percentage van het totaal aantal woorden dat moest worden ingevuld, en daaruit wordt afgeleid hoe goed de deelnemer de eigenlijke tekst begrepen zou hebben. Gaat het erom de leesvaardigheid van individuele deelnemers vast te stellen, dan worden aan hen diverse teksten met verschillende moeilijkheidsgraad voorgelegd die volgens de *cloze procedure* bewerkt zijn, en wordt voor iedere deelnemer afzonderlijk bepaald wat diens gemiddelde *cloze score* bij de verschillende teksten is. Gaat het om de begrijpelijkheid van een bepaalde tekst voor een gegeven doelgroep, dan wordt aan een aantal lezers uit die doelgroep een *cloze*-versie van die tekst voorgelegd, en wordt vervolgens bepaald wat de gemiddelde *cloze score* van die lezers is.

De ratio achter het gebruik van een *cloze test* als meetinstrument voor tekstbegrip en tekstbegrijpelijkheid is de volgende. Wie een *cloze test* uitvoert, moet een beroep doen op kennis en vaardigheden die ook nodig zijn in het leesproces. Goodman (1967) karakteriseerde het leesproces bij gevorderde lezers al treffend als een *psycholinguistic guessing game*, een selectief proces waarin niet aan elke letter en aan elk woord even veel aandacht wordt besteed. Onderzoek in de achterliggende decennia naar het proces van begrijpend lezen houdt dat beeld in grote lijnen in stand. Lezers nemen in een tekstpassage waar ze hun aandacht op richten heel snel een aantal vormen waar (zoals beginkapitalen, letterstokken en –staarten, eerste en laatste letters van de woorden, en spaties tussen woorden), en die waarnemingen combineren ze razendsnel met wat ze al gelezen hebben in de tekst, met wat ze al weten over het onderwerp van de tekst (en ruimer: met hun kennis van de wereld), en met hun kennis van de taal waarin de tekst geschreven is. Op basis van de combinatie van waarnemingen en inhoudelijke en talige kennis op grafemisch, syntactisch, semantisch en pragmatisch niveau komen lezers, zonder dat ze zich daar bewust van zijn, tot voorspellingen over wat er in het desbetreffende deel van de tekst zal staan. Die voorspellingen controleren ze vervolgens weer

razendsnel. Door de resultaten van die controle wordt hun verdere leesgedrag gestuurd: succesvolle voorspellingen leiden tot versnelling van het leesproces; ontkrachte voorspellingen leiden tot vertraging. Hoe leesbaarder een tekst geschreven is, en hoe groter iemands inhoudelijke en talige kennis, des te vaker diens voorspellingen correct zullen blijken, en met des te meer vaart de tekst gelezen en geïnterpreteerd kan worden.[3]

De taken waarvoor een deelnemer aan een *cloze test* zich gesteld ziet, hebben sterke overeenkomsten met de hierboven beschreven taken van de lezer. Om met kans op succes te kunnen voorspellen welke woorden zijn weggelaten, moet iemand die een *cloze test* invult een beroep doen op dezelfde soorten kennis als die de lezer nodig heeft om een tekst vlot te kunnen lezen: kennis van wat er in de tekst tot nu toe aan de orde is geweest, kennis over het onderwerp van de tekst (en ruimer: kennis van de wereld), en kennis van de taal waarin de tekst geschreven is, op grafemisch, syntactisch, semantisch en pragmatisch niveau.

Het is dan ook niet verrassend dat *cloze scores* hoog blijken te correleren met uitslagen van traditionele tekstbegriptoetsen. Uit studies die in 1967 al beschikbaar waren, concludeert Bormuth (1967) dat "the use of the cloze readability procedure seems to result in valid measurements of the comprehension difficulty of written instructional material. The correlations between cloze readability and conventional comprehension test scores are high, and none of the research has presented evidence that the processes employed in responding to cloze readability tests are in any major sense distinguishable from those employed in responding to conventional comprehension tests." (p. 20-21). Hij verwijst daarbij onder meer naar onderzoek van Taylor (1953): r=.76; onderzoek van Jenkinson (1957): r=.82; en onderzoek van Bormuth zelf (Bormuth, 1962): r=.73 en r=.84. Een factoranalyse met negen *cloze tests* en zeven traditionele meerkeuze-begripstoetsen bevestigde Bormuth (1969) in zijn stellingname dat er geen aanleiding is om te menen dat de twee soorten toetsen iets anders zouden meten.

In haar proefschrift uit 2007 constateert ook Kamalski dat in een validatie-experiment de interne betrouwbaarheid van de *cloze test* in orde bleek en dat er sprake was van convergente validiteit van de *cloze scores* en de scores op twee andere begripsmaten ('sorting tasks' en 'mental model tasks'); op grond daarvan komt ze tot de conclusie dat de "cloze tasks' can certainly be useful and valid" (Kamalski, 2007, p. 103; 105). Van nog recenter datum is de studie van Gellert en Elbro (2013). Ook zij vonden een sterke samenhang (r=.84) tussen een *cloze test* en een traditionele tekstbegriptest. Empirisch onderzoek van Horton (1974-1975) bij 112 leerlingen van ongeveer 15 jaar oud naar de constructvaliditeit van de *cloze test* onderstreept de congruentie met traditionele tekstbegriptoetsen. Dezelfde vier *Structure-of-Intellect*-vaardigheden (cognitie van semantische eenheden, evaluatie van semantische relaties, divergente productie van semantische implicaties, en divergente productie van semantische klassen) die gerelateerd waren aan de scores op traditionele tekstbegriptoetsen bleken ook gerelateerd te zijn aan de scores op een *cloze test*. Horton's conclusie is helder: "The construct validity of cloze procedure has been established. The construct was defined as the ability to deal with the linguistic structure of the language; it is related to the ability of the subject to deal with the relationships among words and ideas." (p. 250).

De *cloze test* kan op verschillende manieren uitgevoerd worden. Zo moet er een keuze gemaakt worden tussen twee manieren waarop woorden in een tekst weggelaten kunnen worden. Dat kan op een willekeurige manier (*fixed ratio*) of op een selectieve manier (*rational fill in*). Wordt er gekozen voor de *fixed-ratio* aanpak dan laat de onderzoeker steeds het n^{e} woord weg na het vorige woord dat is weggelaten (bijvoorbeeld het eerste, zesde, elfde, enzovoort woord). Bij de *rational fill-in* aanpak bepaalt de onderzoeker zelf welke woorden worden weggelaten, bijvoorbeeld door te besluiten alleen inhoudswoorden door streepjes te vervangen (zie onder meer Gellert & Elbro, 2013).

Als voordeel van de *rational fill-in* aanpak wordt wel gezien dat de onderzoeker ervoor kan zorgen dat er geen woorden buiten beschouwing blijven (inhoudswoorden bijvoorbeeld) waarvan hij meent dat de voorspelbaarheid in elk geval getoetst moeten worden. Als voordeel van de *fixed-ratio* aanpak wordt vaak beschouwd dat de resulterende *cloze test* ongevoelig is voor wellicht inadequate keuzes van de onderzoeker. Daarnaast is de *fixed ratio* aanpak erg gemakkelijk toe te passen. Bachman (1985) komt op basis van een vergelijking van de *cloze*-scores van in totaal 910 studenten tot de conclusie dat beide vormen goed vergelijkbaar zijn; in zijn onderzoek correleerden de scores bij beide vormen ook hoog met andere taalvaardigheidsmaten (p. 546; 550). In zijn grootschalige onderzoek gericht op de ontwikkeling van een meetinstrument voor leesbaarheid en leesvaardigheid in het basisonderwijs koos Staphorsius (1994) voor de *fixed ratio* aanpak (bij hem: 'mechanische deletie'). In geen van de pleidooien voor de *rational fill-in* aanpak (bij hem 'selectieve deletie') vond hij een aanvullende bijdrage aan de theoretische verankering van de *cloze procedure*; ook vond hij geen eenduidige aanwijzingen voor een grotere betrouwbaarheid van de *cloze test* bij een *rational fill-in* aanpak dan bij een *fixed ratio* aanpak (p. 134-135).

Een tweede besluit dat moet worden genomen nadat ervoor gekozen is een *cloze test* te gebruiken, betreft de wijze van scoring. Wordt er besloten tot exacte scoring dan wordt een woord dat de deelnemer op een bepaald streepje heeft ingevuld alleen dan als correct beschouwd als het - afgezien van duidelijke spelfouten - precies overeenkomt met het woord dat op die plaats in de tekst stond. Wordt semantische scoring toegepast, dan wordt een antwoord ook als correct beschouwd wanneer het ingevulde woord weliswaar niet precies overeenkomt met het woord dat er oorspronkelijk stond, maar wel goed zou passen op de lege plaats waar het wordt ingevuld.

Op het eerste gezicht lijkt de semantische scoringsmethode de beste. Wanneer een lezer een goed gekozen synoniem invult voor het woord dat er eigenlijk had moeten staan, is dat immers een aanwijzing dat die lezer de betreffende passage goed begrepen heeft. Uit vergelijkend onderzoek echter blijkt een sterk verband tussen de uitkomsten van de exacte en de semantische scoringsmethode. Zo vond Laing (1988) bij toepassing van een *cloze test* op twee passages met elk vijftig weggelaten woorden correlaties van .83 en .87. Werden de antwoorden van de deelnemers bij de honderd weggelaten woorden samengenomen, dan werd een correlatie van .92 gevonden (p. 58-59). In het Nederlandse taalgebied gaf Staphorsius (1994) uit praktische en empirische overwegingen de voorkeur aan exacte scoring. Hij baseerde zijn keuze op vier studies naar de onderlinge relaties tussen de uitkomsten van *cloze tests* waarin de twee scoringsmethodes werden gebruikt; daarbij werden steeds correlaties ge-

vonden die hoger waren dan .95 (p. 149-150). O'Toole en King (2011) vergeleken de resultaten bij exacte en de semantische scoring van 447 Australische middelbare-schoolleerlingen bij drie verschillende *cloze tests*. Hun conclusie is dat het best semantische scoring (bij hen 'conceptual scoring') kan worden toegepast als het erom gaat een rangorde te bepalen van individuele leerlingen op basis van hun leesvaardigheid. Docenten/onderzoekers kunnen immers van geval tot geval bepalen in welke mate ze een bepaald antwoord als indicator voor een meer of minder goed ontwikkelde leesvaardigheid beschouwen. Is het doel van de test echter de leesbaarheid van een bepaalde tekst voor verschillende groepen vast te stellen of van verschillende teksten voor een bepaalde groep, dan verdient volgens O'Toole en King de exacte scoring de voorkeur. Blijkens hun onderzoek worden bij semantische scoring sterke lezers bevoordeeld en zwakke lezers benadeeld, wat bij een ongelukkige keuze van de *cloze test*-deelnemers kan leiden tot een te positief beeld van de leesbaarheid van een tekst voor die tweede groep. Bij exacte scoring geldt dat bezwaar volgens O'Toole en King (2011) niet. Daarnaast noemen ze de exacte scoring "easier, less subjective, sufficiently reliable [and] highly correlated with conceptual scoring" (p. 140).

In ons onderzoek werd het keuzeprobleem tussen de *fixed ratio*-aanpak en de *rational fill-in* aanpak opgelost door van elke tekst vijf *fixed ratio cloze test* varianten te maken: een waarin het eerste, zesde, elfde, enzovoort woord werden weggelaten, een waarin het tweede, zevende, twaalfde, enzovoort woord werden weggelaten, een waarin het derde, achtste, dertiende, enzovoort woord werden weggelaten, enzovoort. Zo konden we er zeker van zijn dat de voorspelbaarheid van alle woorden uit alle teksten een even belangrijke rol in de uiteindelijke tekstbegripscores zou spelen, en dat er geen woorden buiten de test zouden vallen waarvan de voorspelbaarheid van belang zou kunnen zijn voor het tekstbegrip. Daarmee werd tegemoet gekomen aan het bezwaar van onder meer Kobayashi (2002) dat bij toepassing van de *fixed ratio* aanpak de toevallige keuze van het eerste woord de cloze scores bij een gegeven tekst in belangrijke mate zou kunnen beïnvloeden. Elk woord in de teksten die wij gebruikten, werd immers precies even vaak weggelaten - een aanpak die ook wordt aanbevolen in O'Toole en King (2010).

Ook besloten we de exacte scoringsmethode toe te passen. Daarmee was zowel de snelheid als de betrouwbaarheid van de scoring gediend en gezien de hierboven gemelde hoge correlaties met de resultaten van de semantische scoringsmethode werd het risico van een verminderde validiteit gering geacht.

3.3 Deelnemers Het onderzoek werd afgenomen bij 125 volwassenen (bekenden van de tweede auteur, dan wel bekenden van die bekenden), die allen het Nederlands als moedertaal hadden. De leesvaardigheid van de deelnemers kon om praktische redenen niet op een directe wijze worden gemeten. Dat zou een te groot beslag gelegd hebben op de tijd die zij voor dit onderzoek ter beschikking hadden. Wel kon het door de deelnemers zelf ingeschatte leesvaardigheidsniveau worden bepaald door aan hen te vragen om te reageren op achttien zogenaamde *can do statements* afkomstig uit *DIALANG* (Alderson, 2005; Alderson & Huhta, 2005).

DIALANG is een taalvaardigheidstoetsprogramma op het internet, ontwikkeld op basis van het CEFR, met subsidies van de Europese Commissie binnen de context van het Socrates

programma (*Dialang*, z.j.). Hoofddoel van DIALANG is vreemdetaalgebruikers en -leerders diagnostische informatie te geven over hun taalvaardigheid in een van de veertien Europese talen, met betrekking tot schrijven, lezen, luisteren, grammatica en vocabulaire. DIALANG-gebruikers krijgen niet alleen per vaardigheid een groot aantal toetsitems voorgelegd waarmee hun prestaties kunnen worden bepaald, maar ze moeten ook reageren op een aantal items (de *can do statements*) waarmee kan worden nagegaan op welk CEFR-niveau ze hun eigen vaardigheden inschatten. Voorbeelden van *can do statements* zijn stellingen als 'Ik kan korte, eenvoudige persoonlijke brieven begrijpen' (A2-niveau), 'Ik kan artikelen en rapporten begrijpen over eigentijdse problemen waarin de schrijvers een bepaalde houding of een bepaald standpunt innemen' (B2-niveau) en 'Ik kan een breed scala van lange en complexe teksten begrijpen en daarbij de expliciete en impliciete nuances van stijl en betekenis volledig begrijpen' (C2-niveau).

In Alderson (2005) worden voor een groep van in totaal 1803 respondenten uit elf Europese landen die niet het Engels als moedertaal hadden, rangordecorrelaties gepresenteerd tussen feitelijke toetsscores (in termen van CEFR-niveaus) voor lezen, luisteren en schrijven in het Engels en zelfbeoordelingen (ook in termen van CEFR-niveaus) over diezelfde deelvaardigheden, gebaseerd op reacties op *can do statements*. De samenhang tussen feitelijke toetsscores en zelfbeoordelingen bleek hoog te zijn: de rangcorrelatiecoëfficiënten waren respectievelijk .91, .87 en .84.

Dat er sprake is van sterke samenhang tussen zelfbeoordelingen en toetsscores in een vreemde taal, impliceert nog niet dat oordelen en feitelijke prestaties ook dicht bij elkaar liggen - noch voor die vreemde taal, noch voor de moedertaal (zie voor opvallende discrepanties in dit opzicht bij Nederlandse taalgebruikers Van Onna & Jansen, 2008). Wel gingen we er, gegeven de hoge correlaties in Alderson (2005), vanuit dat de zelf ingeschatte leesvaardigheidsniveaus die we af konden leiden uit de reacties van de deelnemers aan ons onderzoek op de *can do statements*, bruikbare informatie opleverden om na te gaan of de deelnemers goed verdeeld waren over de onderzoeksgroepen (zie hieronder). De uitkomst van een chi-kwadraattoets (vier zelf toegekende leesvaardigheidsniveaus: B1, B2, C1 en C2, verdeeld over vijf groepen deelnemers die verschillende combinaties van teksten voorgelegd kregen; zie hieronder) bleek in dit opzicht geruststellend: $\chi^2(12)=5.18$; $p=.95$.

3.4 Design en procedure Van elk van de negentien teksten werden zoals gezegd in paragraaf 3.2 vijf verschillende *cloze test*-varianten gemaakt: een variant waarbij om te beginnen het eerste woord van de lopende tekst werd weggelaten, een tweede variant waarbij het tweede woord als eerste werd weggelaten, een derde variant waarbij werd begonnen met het derde woord, enzovoort. In alle *cloze tests* werd vervolgens steeds het vijfde woord weggelaten na het vorige woord dat was weggelaten.[4] Bij de verdeling van de 95 *cloze test*-varianten die zo ontstonden, met steeds ongeveer 54 woorden die moesten worden ingevuld, werden vijf groepen met elk 25 deelnemers onderscheiden. De eerste groep kreeg *cloze tests* voorgelegd over de eerste huurtekst, de eerste krantentekst, de eerste polistekst en de eerste roddeltekst. De *cloze tests* voor de tweede groep betroffen de tweede huurtekst, de tweede krantentekst, de tweede polistekst en de tweede roddeltekst. De derde groep kreeg *cloze tests* voorgelegd over de derde huurtekst, de derde krantentekst, de derde polistekst en de derde roddeltekst; de *cloze tests* voor de vierde groep betroffen de vierde huurtekst, de

vierde krantentekst, de vierde polistekst en de vierde roddeltekst. De laatste groep ging aan het werk met drie in plaats van vier *cloze tests*: de vijfde huurtekst, de vijfde krantentekst en de vijfde roddeltekst. Een tekst over polisvoorwaarden kreeg deze deze groep dus niet voorgelegd; over dat onderwerp waren slechts van de eerste vier teksten de uitslagen van alle drie meetinstrumenten beschikbaar. Dit alles resulteerde in 475 ingevulde *cloze tests*, gelijkelijk verdeeld over de negentien teksten. Om ongewenste volgorde-effecten te voorkomen, werd ervoor gezorgd dat de verschillende onderwerpen bij de deelnemers in een systematisch gevarieerde volgorde aan bod kwamen.

Het onderzoek werd uitgevoerd in sessies van de tweede auteur met iedere deelnemer afzonderlijk. Bij het begin daarvan ontving de deelnemer een boekje met de *cloze test*-varianten die moesten worden ingevuld, voorafgegaan door een korte schriftelijke instructie, die ook mondeling werd gegeven. Nadat de proefpersonen hun boekjes hadden ingeleverd, werden de ingevulde woorden gescoord (alleen wat exact overeenkwam met het oorspronkelijke woord werd als correct beschouwd), en werden per *cloze test*-variant en vervolgens per tekst de percentages correcte antwoorden berekend. Scores lager dan 10% werden als indicatie beschouwd dat de deelnemer bij de desbetreffende *cloze test* zijn taak niet serieus had genomen; die scores werden daarom buiten beschouwing gelaten. Dat betrof in totaal tien *cloze scores* bij acht verschillende teksten.[5] Alles bijeen werden dus (475-10=) 465 *cloze*-scores verzameld.

3.5 Tekstkenmerken Om na te gaan wat het verband was tussen de begripsscores van de deelnemers aan ons onderzoek en de vier tekstkenmerken die Kraf et al. (2011) in hun onderzoek gebruikten (proportie frequente woorden; type-token ratio; gemiddelde afstand onderwerp-persoonsvorm; gemiddelde afstand lijdend voorwerp-persoonsvorm), correleerden we de frequenties waarmee deze vier tekstkenmerken zich voordeden in de negentien onderzochte teksten met de gemiddelde *cloze*-scores voor deze teksten.[6] Daarnaast berekenden we de waarden van twee klassieke leesbaarheidsindicatoren: gemiddelde woordlengte (het gemiddeld aantal karakters per woord) en zinslengte (het gemiddeld aantal woorden per zin) van elk van de teksten. Ook bepaalden we, op suggestie van Hacquebord en Lenting-Haan (2012), de lengte van elke tekst (het totaal aantal woorden). Vervolgens berekenden we de correlaties van deze drie eenvoudig meetbare kenmerken met de gemiddelde *cloze*-scores. Ten slotte konden we met behulp van een regressieanalyse nagaan wat de door de zeven tekstkenmerken gezamenlijk verklaarde variantie in de *cloze*-scores zou zijn, en konden we bepalen welk aandeel elk van de tekstkenmerken afzonderlijk in die verklaarde variantie zou hebben.

4 Resultaten

Het eerste doel van dit onderzoek was de leesbaarheidsvoorspellingen van de drie instrumenten over een aantal teksten te confronteren met begripsprestaties bij echte lezers van die teksten. Daartoe werden rangordecorrelaties (Spearman's Rho) berekend tussen de uitslagen van de drie meetinstrumenten bij de negentien teksten en de gemiddelde *cloze*-scores van de deelnemers aan het onderzoek. Tabel 2 bevat de uitslagen van de meetinstrumenten zoals gemeld in Kraf et al. (2011), aangevuld met de door ons gevonden *cloze*-scores.

Tabel 2: Uitslagen van de meetinstrumenten zoals gemeld in Kraf et al. (2011, p. 260), aangevuld met cloze-scores (gemiddelden en standaarddeviaties).

	Texamen	*Klinkende Taal*	*Accessibility*	*Cloze*-score: M (SD)
Tekst 1 (huur)	C1	C1	B2	64.86 (11,93)
Tekst 2 (huur)	C1	C2	C2	37.91 (18.52)
Tekst 3 (huur)	C1	C1	B2	52.50 (12.35)
Tekst 4 (huur)	B2	B2	B1/B2	65.70 (13.13)
Tekst 5 (huur)	C1	C1	B2	60.96 (10.60)
Tekst 6 (krant)	C1	C1	B2	52.35 (14.81)
Tekst 7 (krant)	C1	B2	B2	56.36 (11.57)
Tekst 8 (krant)	C1	C1	B2/C1	51.88 (17.07)
Tekst 9 (krant)	C1	B2	C1	54.59 (14.74)
Tekst 10 (krant)	C1	C1	C1	61.34 (14.07)
Tekst 11 (polis)	C1	C2	C2	45.54 (18.53)
Tekst 12 (polis)	C1	C2	C2	49.50 (18.56)
Tekst 13 (polis)	C1	C2	C2	40.94 (13.39)
Tekst 14 (polis)	C1	B2	B2	59.68 (19.97)
Tekst 15 (roddel)	B2	B2	B2	55.94 (18.18)
Tekst 16 (roddel)	B1	B2	B1/B2	65.20 (12.60)
Tekst 17 (roddel)	C1	C1	B2/C1	65.69 (12.03)
Tekst 18 (roddel)	B2	B2	B2	59.30 (15.00)
Tekst 19 (roddel)	B1	B2	B2	57.04 (12.95)

Zoals Kraf et al. (2011, p. 259-260) opmerken, valt op in Tabel 2 dat de taalniveaus A1 en A2 niet voorkomen en dat *Texamen* aan veertien van de negentien teksten taalniveau C1 toekent. Opvallend aan de gemiddelde *cloze*-scores is dat soms bij teksten uit een en dezelfde categorie aanzienlijke verschillen werden gevonden. Zo leverde een van de teksten gericht aan huurders de laagste van alle scores op (37.91) terwijl voor een andere tekst uit deze categorie de hoogste score van alle teksten gevonden werd (65.70).

In Tabel 3 zijn de onderlinge rangordecorrelaties uit Kraf et al. (2011) te vinden, aangevuld met de rangordecorrelaties met de *cloze*-scores uit ons onderzoek.

Tabel 3: Rangordecorrelaties tussen de taalniveaus toegekend door de drie instrumenten zoals gemeld in Kraf et al. (2011, p. 261), aangevuld met rangordecorrelaties met de in dit onderzoek gevonden cloze-scores.

	Texamen	*Klinkende Taal*	*Accessibility*	*Cloze*-score
Texamen	1	.64*	.60*	-.39
Klinkende Taal		1	.75*	-.61*
Accessibility			1	-.69*

* Statistisch significant (p<.01)

Opvallend aan de uitkomsten in Tabel 3 is dat tussen de *Texamen*-uitslagen en de *cloze*-scores geen significant verband gevonden werd. Anders gezegd: de data geven geen aanwijzing dat de relatie tussen de *Texamen*-uitslagen en tekstbegrip zoals hier gemeten, op iets anders dan toeval berust. De correlaties tussen de uitslagen van de andere twee instrumenten en de *cloze*-scores bleken wel significant, en zoals verwacht ook negatief (hoe lager de *cloze test* score, hoe hoger het voor tekstbegrip benodigde taalniveau). Daarbij moet wel worden opgemerkt dat ook in het beste geval, dat van de *Accessibility Leesniveau Tool*, slechts 48% (r^2 namelijk) van de variantie in de *cloze*-scores zich bij intervaldata zou laten verklaren uit de uitslagen van dit meetinstrument.

Het tweede doel van ons onderzoek was na te gaan hoe de begripsscores van de lezers zouden samenhangen met een aantal kenmerken van de teksten. Daartoe werd een lineaire regressieanalyse uitgevoerd (stapsgewijs) met de gevonden gemiddelde *cloze*-score als afhankelijke variabele en de zeven gemeten tekstkenmerken als predictoren. In het licht van de soms relatief hoge correlaties tussen de tekstkenmerken (zie Tabel 4) werd eerst nagegaan of zich daarbij een collineariteitsprobleem kon voordoen.

Tabel 4: Correlaties (product-moment) tussen tekstkenmerken en cloze-scores.

	WL	***ZL***	***TL***	***PFW***	***TTR***	***O-PV***	***LV-PV***	***Cloze*-score**
WL	1	.57*	-.34	-.32	-.62**	.61**	.63**	-.69**
ZL		1	-.04	-.14	-.45	.51*	.78**	-.68**
TL			1	-.04	.32	.12	-.05	.13
PFW				1	.03	-.37	-.36	.24
TTR					1	-.21	-.44	.28
O-PV						1	.73**	-.53*
LV-PV							1	-.66**
Cloze-score								1

* Statistisch significant (p<.05)
** Statistisch significant (p<.01)

WL: aantal karakters per woord (M=5.31; SD=0.45)
ZL: aantal woorden per zin (SD=18.29; SD=4.82)
TL: totaal aantal woorden (M=270.74; SD=10.41)
PFW: proportie frequente woorden (M=.62; SD=.04)
TTR: type-token ratio (M=.55; SD=.06)
O-PV: afstand onderwerp-persoonsvorm (M=2.81; SD=0.96)
LV-PV: afstand lijdend voorwerp-persoonsvorm (M=3.73; SD=2.12)

Er bleek inderdaad sprake te zijn van een collineariteitsprobleem; weglating van de predictor LV-PV (afstand lijdend voorwerp-persoonsvorm) loste dat probleem op. De regressieanalyse die werd uitgevoerd met de zes overgebleven predictoren leverde een significant resultaat op voor gemiddelde woordlengte: R^2=.47; β=-.69; p=.001. Uitbreiding met de predictor gemiddelde zinslengte liet een significante verbetering van de voorspellingen zien: R^2=.60;

p(Fchange)=.04; bijdrage predictor gemiddelde woordlengte: β=-.44; p=.04; bijdrage predictor gemiddelde zinslengte: β=-.43; p=.04. Uitbreiding van deze twee predictoren met de vier andere tekstkenmerken leverde geen significante verbetering van de voorspellingen op.

5 Conclusies

De uitkomsten van ons onderzoek leiden tot gerede twijfel over de validiteit van de drie onderzochte meetinstrumenten. Kraf et al. (2011) lieten al zien dat de uitslagen van *Texamen*, *Klinkende Taal* en de *Accessibility Leesniveau Tool* bij toepassing op de negentien teksten die in hun onderzoek werden gebruikt, flinke verschillen vertoonden. Uit de confrontatie van de voorspellingen van de drie meetinstrumenten met de *cloze*-scores die wij in dit onderzoek verzamelden, bleek dat er tussen *Texamen* en die scores geen significant verband gevonden werd. Daarmee presteerde *Texamen* duidelijk het slechtst van de drie getoetste meetinstumenten. Wordt de tekstbegrijpelijkheid wel goed voorspeld door de twee andere instrumenten? Echt enthousiast kunnen we ook daar niet over worden. Ook in het beste geval, dat van de *Accessibility Leesniveau Tool*, achten we het gevonden verband (r=-.69) daarvoor niet sterk genoeg. Ook met dat instrument liet zich minder dan de helft van de variantie in de begripsscores verklaren. Anderson en Davison wezen bijna dertig jaar geleden al op het bestaan van leesbaarheidsformules waarmee 60% tot 80% van de variantie in gemiddelde tekstbegripscores kon worden verklaard (Anderson & Davison, 1984, p. 9; zie echter ook hieronder voor hun kritische commentaar bij de statistische analyses die tot dit soort variantie-informatie leiden).

De regressieanalyse die we uitvoerden met de tekstkenmerken die werden gemeten in Kraf et al. (2011) plus tekstlengte, gemiddelde woord- en gemiddelde zinslengte, liet zien dat bij gebruik van alleen gemiddelde woord- en gemiddelde zinslengte als predictoren, 60% van de variantie in de *cloze*-scores kon worden verklaard.[7] Dat suggereert dat met een geautomatiseerde klassieke leesbaarheidsformule die werkt met die twee variabelen, correlaties met de *cloze*-scores bereikt kunnen worden die die van de drie Nederlandse meetinstrumenten overtreffen. Dat blijkt in de praktijk ook zo te zijn. Wanneer de leesbaarheid van de negentien hier onderzochte teksten wordt gemeten met de geautomatiseerde *Coleman-Liau Index*, die gebaseerd is op het aantal karakters per woord en het aantal woorden per zin (*Readability Formulas*, z.j.), dan wordt een significante en relatief hoge rangordecorrelatie gevonden met de *cloze*-scores uit ons onderzoek: r=-.79; p<.001. En zelfs wanneer alleen het gemiddeld aantal karakters per woord als voorspeller wordt gebruikt, wordt een rangordecorrelatie bereikt met onze *cloze*-scores die die van elk van de drie Nederlandse meetinstrumenten overtreft (r=-.73; p<.001).[8]

Dit betekent beslist niet dat met de *Coleman-Liau Index* of met het gemiddeld aantal karakters per woord het ideale instrument gevonden zou zijn om de begrijpelijkheid van teksten mee te voorspellen. Daarvoor zijn de bezwaren nog te zeer van kracht die al sinds het begin van de jaren tachtig zijn ingebracht tegen instrumenten die met niet meer werken dan gemiddelde woord- en zinslengte. Zo wijzen Anderson en Davis (1984) erop dat bij de berekening van de verklaarde variantie bij leesbaarheidsformules maar zelden wordt verdisconteerd dat de gemiddelde begripsscores voor de teksten geaggregeerd zijn over individu-

ele lezers. Zou daar wel rekening mee worden gehouden, dan vallen er volgens hen voor de desbetreffende formules aanmerkelijk lagere percentages verklaarde variantie te verwachten (p. 9-10). Dat zou dan dus ook gelden voor de drie nieuwe instrumenten die in deze studie centraal stonden. Verder doen de leesbaarheidsformules geen recht aan de complexiteit en verwevenheid van allerlei mogelijke oorzaken van tekstmoeilijkheid en gaan ze voorbij aan de verschillende situaties waarin lezers zich meer en minder moeite willen getroosten om een tekst te doorzien. Misschien nog wel het hinderlijkst is het dat het gebruik van de formules al snel leidt tot het klassieke *correlatie=causatie* misverstand. Wie alleen iets aan de gemiddelde woord- en zinslengte verandert, heeft daarmee niet meteen ook de begrijpelijkheid beslissend beïnvloed, hoe aantrekkelijk die gedachte misschien ook moge zijn (zie ook Jansen, 1995; 2012).

Aan een meetinstrument voor de begrijpelijkheid van Nederlandse teksten dat beter voldoet, wordt thans gewerkt in *LIN* (LeesbaarheidIndex voor het Nederlands), een samenwerkingsproject van onderzoekers van de Universiteit Utrecht, de Radboud Universiteit Nijmegen, het CITO en de Nederlandse Taalunie. In Pander Maat (2012) en LIN (z.j.) wordt in grote lijnen de aanpak beschreven waarvoor gekozen is in dit project, dat wordt uitgevoerd in het kader van het NWO-programma Begrijpelijke Taal. Gebruik makend van recent computationeel-linguïstisch onderzoek wordt een negentigtal teksten automatisch geanalyseerd op een groot aantal tekstkenmerken; over alle teksten worden *cloze tests* afgenomen bij middelbare-schoolleerlingen, en met behulp van *eye tracking* wordt informatie verzameld over het leesproces bij een deel van die teksten. De LIN-onderzoekers spreken met recht van een lange weg die moet worden afgelegd om een serieuze leesbaarheidsindex te kunnen ontwikkelen (Pander Maat, 2012, p. 38). Uit ons relatief kleinschalige onderzoek, met een beperkt aantal deelnemers, teksten en tekstkenmerken, moge blijken dat bij de ontwikkelingen van *Klinkende Taal*, de *Accessibility Leesniveau Tool* en met name *Texamen* te veel bochten zijn afgesneden om een van deze instrumenten als zo'n serieuze leesbaarheidsindex te kunnen kwalificeren.

Noten

1. De *Accessibility Leesniveau Tool* is gratis (zie www.accessibility.nl/kennisbank/tools/leesniveau-tool). Voor gebruik van de andere twee instrumenten moet worden betaald; zie www.klinkendetaal.nl/introductie en www.texamen.nl/over-texamen-4.html (sites geraadpleegd op 14 maart 2013).
2. Voor een overzicht van de problemen in het leesbaarheidsonderzoek, zie Kraf en Pander Maat (2009).
3. Over het proces van begrijpend lezen en het resulterende mentale model van de informatie in de tekst, zie onder meer Noordman en Maes (2000) en Van den Broek (2009). Oller (1979, 34) spreekt in een breder verband van een *pragmatic expectancy grammar*. Hij refereert ook aan de verwijzing van Taylor (1953) bij de introductie van de *cloze test* naar een notie uit de Gestaltpsychologie *closure*, die het mensen mogelijk maakt problemen op te lossen door ontbrekende elementen aan te vullen (p. 42). Volgens Oller proberen we altijd te voorspellen wat in een gegeven serie van elementen het volgende zal zijn, ook wanneer we communiceren met anderen. Hoe accurater onze voorspellingen zijn, hoe beter het proces van probleemoplossing verloopt, ook in de communicatie.

4. Conform de conventies bij de toepassing van de *cloze test* werden woorden uit de volgende categorieën niet weggelaten: eigennamen, tijd- en richtingaanduidingen, bedragen, woorden tussen haakjes, en woorden in titels en tussenkopjes.
5. Het betrof de teksten 2 (2 deelnemers), 5 , 6, 11, 12 (2 deelnemers), 13, 14 en 15.
6. We zijn Rogier Kraf, Leo Lentz en Henk Pander Maat zeer erkentelijk voor de data die ze ons in dit verband ter beschikking wilden stellen, en ook voor de teksten die we in dit onderzoek hebben kunnen gebruiken.
7. Ter vergelijking: onderzoek bij de ontwikkeling van leesbaarheidformules voor het Afrikaans waarbij zestien mogelijk relevante tekstkenmerken werden gebruikt, leidde tot twee formules die elk niet meer dan 40% van de variantie in begripsscores bij de onderzoeksdeelnemers verklaarden (Van Rooyen, 1986, p. 65).
8. Relatief hoge rangordecorrelaties met de *cloze test* uitslagen worden ook gevonden als voor de uitslagen van de *Coleman-Liau* index en voor de gemiddelde woordlengte ordinale schalen met slechts vier meetpunten worden gebruikt, zoals dat feitelijk ook het geval is met de taalniveaus die de hier besproken instrumenten bij de negentien onderzochte teksten opleverden: *Coleman-Liau* en *cloze*-scores: r= -.78 ($p<.001$); gemiddelde woordlengte en *cloze*-scores: r= -.65 ($p<.01$).

Literatuur

Alderson, J.C. (2005). *Diagnosing foreign language proficiency. The interface between learning and assessment*. London: Continuum.

Alderson, J.C., & Huhta, A. (2005). The development of a suite of computer-based diagnostic tests based on the Common European Framework. *Language Testing, 22 (3)*, 301-320.

Anderson, R.C., & Davison, A. (1984). *Conceptual and empirical bases of readability formulas*. Center for the study of reading, Technical report no. 392. University of Illinois at Urbana-Champaign.

Bachman, L.F. Performance on cloze tests with fixed-ratio and rational deletions. *TESOL Quarterly, 19 (3)*, 535-556.

Bormuth, J.R. (1962). *Cloze Tests as measures of readability*. Unpublished doctoral dissertation. Indiana University.

Bormuth J.R. (1969). Factor validity of cloze tests as measures of reading comprehension ability. *Reading Research Quarterly, 4 (3)*, 358-365.

Broek, P.W. van den (2009). *Cognitieve en neurologische processen tijdens begrijpend lezen. Fundamenteel onderzoek en onderwijskundige toepassing*. Oratie Universiteit Leiden.

Dialang (z.j.). www.lancs.ac.uk/researchenterprise/dialang/about (geraadpleegd op 14 maart 2013).

Gellert, A.S., & Elbro, C. (2013). Cloze tests may be quick, but are they dirty? Development and preliminary validation of a cloze test of reading comprehension. *Journal of Psychoeducational Assessment, 31 (1)*, 16-28.

Goodman, K. (1967). Reading: A psycholinguistic guessing game. *Journal of the Reading Specialist, 6 (4)*, 126-135.

Hacquebord, H., & Lenting-Haan, K. (2012). Kunnen we de moeilijkheid van teksten meten? Naar concrete maten voor de referentieniveaus. *Levende Talen Tijdschrift, 13 (2)*, 14–23.

Hauwermeiren, P. van (1974-1975). Leesbaarheidsformules voor informatieve Nederlandse teksten. *Spektator, 4 (3),* 499-520.

Horton, R.J. (1974-1975). The construct validity of cloze procedure. An exploratory factor analysis of cloze paragraph reading and Structure-of-Intellect tests. *Reading Research Quarterly, 10 (2)*, 248-251.

Jansen, C.J.M. (1995). *Rekenen met taal*. Oratie Technische Universiteit Eindhoven.

Jansen, C. (2012). *Voorkomen is beter. Op weg naar effectievere teksten in de gezondheidscommunicatie*. Oratie Rijksuniversiteit Groningen.

Jansen, C. (2013). De nieuwste kleren van de keizer. 'Teksten op B1-niveau' als leeg begrip. *Onze Taal, 82 (2)*, 56-57.

Jansen, C. & Lentz, L. (2008). Hoe begrijpelijk is mijn tekst? De opkomst, neergang en terugkeer van de leesbaarheidsformules. *Onze Taal, 77 (1)*, 4-7.

Jansen, C.J.M., & Woudstra, E.T. (1979). Theorie en praktijk van het Nederlandse leesbaarheidsonderzoek. Een analyse van twee formules. *Tijdschrift voor Taalbeheersing, 1 (1),* 43-60.

Jenkinson (1957). *Selected processes and difficulties in reading comprehension*. Unpublished doctoral dissertation. University of Chicago.

Kamalski, J.M.H. (2007). *Coherence marking, comprehension and persuasion. On the processing and representation of discourse*. Dissertatie Universiteit Utrecht.

Kobayashi, M. (2002). Cloze tests revisited. Exploring item characteristics with special attention to scoring methods. *The Modern Language Journal, 86 (4)*, 571-586.

Kraf, R., Lentz, L., & Pander Maat, H. (2011). Drie Nederlandse instrumenten voor het automatisch voorspellen van begrijpelijkheid. Een klein consumentenonderzoek. *Tijdschrift voor Taalbeheersing, 33 (3)*, 249-265.

Kraf, R., & Pander Maat, H. (2009). Leesbaarheidsonderzoek. Oude problemen, nieuwe kansen. *Tijdschrift voor Taalbeheersing 31 (2*), 97-123.

Laing, J.B. (1988). *Cloze procedure. A comparison of exact and acceptable scoring.* University of Victoria, Canada.

LIN: A validated reading level tool for Dutch (z.j.). www.nwo.nl/onderzoek-en-resultaten/onderzoeksprojecten/29/2300173129.html (geraadpleegd op 14 maart 2013).

Noordman, L.G.M., & Maes, A. (2000). Het verwerken van tekst. In A. Braet (Red.), *Taalbeheersing als Communicatiewetenschap* (pp. 29-60). Bussum: Coutinho.

O'Toole, J.M., & King, R.A.R. (2010). A matter of significance. Can sampling error invalidate cloze estimates of text readability? *Language Assessment Quarterly, 7(4)*, 303-316.

O'Toole, J.M., & King, R.A.R. (2011). The deceptive mean. Conceptual scoring of cloze entries differentially advantages more able readers. *Language Testing, 28 (1)*, 127-144.

Oller, J.W. (1979). *Language tests at school*. London: Longman.

Onna, B. van, & Jansen, C. (2008). Nederland talenland? Over de beheersing van Engels, Duits en Frans in Nederlandse organisaties. *Levende Talen Tijdschrift, 9 (1)*, 18-26.

Palm, H. (1979). Methodologische kanttekeningen bij de constructie en validering van leesbaarheidsformules. *Tijdschrift voor Taalbeheersing, 1 (2),* 180-192.

Pander Maat, H. (2012). Effectief Communiceren (2). De lange weg naar een serieuze leesbaarheidsindex. *Tekstblad, 18 (4)*, 36-38.

Readability Formulas (z.j.). www.readabilityformulas.com/coleman-liau-readability-formula.php (geraadpleegd op 14 maart 2013).

Rooyen, R. van (1986). Eerste Afrikaanse leesbaarheidsformules. *Communicatio, 12 (1)*, 59-69.

Staphorsius, G. (1994). *Leesbaarheid en leesvaardigheid. De ontwikkeling van een domeingericht meetinstrument*. Proefschrift Universiteit Twente.

Taylor W. L. (1953). 'Cloze Procedure'. A new tool for measuring readability. *Journalism Quarterly, 30 (3)*, 415- 433.

Zondervan, F., & Caldenborgh, P. van (1971-1972). Leesbaarheidsformules, constructie en betrouwbaarheid. *Spektator, 1 (3)*, 341-351.