Indexering van de leestechniek

G. Staphorsius en N.D. Verhelst

Samenvatting

De CLIB (Cito leesindex voor het basis- en speciaal onderwijs) is een index voor zowel de leesvaardigheid van lezers als voor de leesbaarheid van teksten. Met behulp van de CLIB kunnen de leesvaardigheid van een lezer en de leesvaardigheid die een tekst vereist met elkaar vergeleken worden. Daardoor is het mogelijk om in de dagelijkse onderwijspraktijk de leesvaardigheid van leerlingen en de moeilijkheid van teksten op elkaar af te stemmen. Omdat de CLIB verwijst naar een domein van teksten die lezers met een bepaald vaardigheidsniveau met begrip kunnen lezen, noemen we de CLIB een domeingerichte index. De CLIB is in de eerste plaats een index voor de begrijpelijkheid en het begrijpen van teksten. Het onderzoek dat deze index opleverde past in een lange traditie. Veel minder traditioneel is het onderzoek naar een index voor de door teksten vereiste leestechniek.

Toch zou een index voor de vereiste leestechniek didactisch gezien van belang kunnen zijn. We denken dan vooral aan de leerlingen die aanzienlijk beter kunnen begrijpen dan decoderen. We hebben met het oog daarop een onderzoek uitgevoerd. Het leverde de CILT (Cito Index voor de leestechniek) op. Met behulp van een multipele regressievergelijking ($R = .90$) kan worden bepaald wat de CILT van een tekst is. Deze geeft aan hoe hoog de technische leesvaardigheid is, die een tekst vereist. De predictoren in de regressievergelijking zijn 'woordlengte' en 'woordfrequentie'. Het criterium voor leestechniek is het aantal woorden uit teksten dat in één minuut gemiddeld gelezen wordt door basisschoolleerlingen in groep 5 tot en met 8. Voor de domeingerichte bepaling van de leestechniek van leerlingen kan in principe elke valide toets voor de meting van de leestechniek worden gebruikt. Wij zijn er in geslaagd de scores op de toetsen Leestempo van het Cito te transformeren naar de CILT-schaal.

1 Inleiding: leesbaarheidspredictie

Het zoeken naar een index voor de leesbaarheid van teksten impliceert een zogenaamd leesbaarheidsonderzoek. Het eerste leesbaarheidsonderzoek is dat van Vogel en Washburne (1928). Zij slaagden erin een voorspeller van verschillen in begrijpelijkheid te ontwikkelen. Een leesbaarheidsonderzoek heeft in grote lijnen de volgende opzet. Van teksten in een steekproef uit een gegeven populatie bepaalt men de vereiste leesvaardigheid, ofwel de leesbaarheid. Vervolgens gaat men na door welke tekstkenmerken de vastgestelde verschillen in leesbaarheid tussen teksten verklaard kunnen worden. Men gebruikt daarbij de multipele regressie-analyse als techniek. Deze levert een multipele regressievergelijking of leesbaarheidsformule op. Daarmee kan de leesbaarheid van teksten worden voorspeld die behoren tot de populatie waaruit het onderzoeksmateriaal getrokken is. We zullen hier ter verduidelijking de grote lijnen van het onderzoek weergeven dat de CLIB opleverde (zie Staphorsius, 1994). In dit onderzoek bestond het onderzoeksmateriaal uit een steekproef uit de populatie non-fictie voor de jeugd. Het begrip 'leesbaarheid' werd nader bepaald met 'begrijpelijkheid', de mate waarin teksten een beroep doen op de vaardigheid begrijpend lezen. Om een maat voor de begrijpelijkheid te kunnen berekenen maakten we van elke tekst in de steekproef met behulp van de cloze-procedure (Taylor, 1953; zie ook Staphorsius, 1994) een cloze-toets. De variant van de procedure die werd gebruikt, schrijft de deletie van elk zevende woord voor. We vroegen aan leerlingen in groep 5 tot en met 8 te proberen de weggelaten woorden weer in te vullen. Het gemiddeld aantal juiste reconstructies, de gemiddelde cloze-score van de teksten in onze steekproef gebruikten we als criterium voor 'leesbaarheid'. Vervolgens gingen we na door welke onafhankelijke variabelen dit criterium het beste voorspeld werd. In verreweg het meeste leesbaarheidsonderzoek

zijn de onafhankelijke variabelen eenduidig te identificeren kenmerken van teksten. Veel gebruikte kenmerken zijn: de woordlengte, de bekendheid en frequentie van woorden, de lengte van zinnen, maar ook de frequentie van complexe syntactische structuren. In het leesbaarheidsonderzoek dat Staphorsius in 1994 rapporteert, vormen het criterium en (uiteindelijk) 4 onafhankelijke variabelen of predictoren de invoer van een multipele regressie-analyse. De predictoren zijn: het percentage frequente woorden (FREQ77; later in dit artikel zal nog blijken waarvoor deze afkorting precies staat), het percentage verschillende woorden (PTYPES), de gemiddelde woordlengte in letters (GWL) en het percentage zinnen (PZW; het aantal woorden is de percenteerbasis) als syntactische maat. De vergelijking ziet er als volgt uit:

$$\text{CLIB} = 46 - 6.603 \times \text{GWL} + 0.474 \times \text{FREQ77} - 0.365 \times \text{PTYPES} + 1.425 \times \text{PZW}$$

In de formule is de CLIB de voorspelde gemiddelde cloze-score (in de uitgaven voor het onderwijs werkt het CITO overigens met een transformatie van die score). Het praktische nut van deze regressievergelijking is, dat de leesbaarheid van andere teksten dan die in de steekproef, niet empirisch bepaald behoeft te worden, althans voorzover die teksten behoren tot de populatie waaruit het onderzoeksmateriaal afkomstig is. De vergelijking wordt in de literatuur ook vaak aangeduid met 'leesbaarheidsformule'.

In het onderzoek waarvan we hier verslag doen, is de leesbaarheid **niet als begrijpelijkheid**, maar als **veronderstelde leestechniek** gedefinieerd. Het heeft niettemin in grote lijnen dezelfde opzet als het onderzoek naar de CLIB. Daarmee is ook de route voor het verslag van ons onderzoek naar de voorspelbaarheid van de door teksten voor de jeugd vereiste leestechniek bepaald. In de volgende paragraaf gaan we eerst in op het criterium. We definiëren daar ook de populatie teksten en bespreken de selectie van ons onderzoeksmateriaal daaruit. In de derde paragraaf gaan we in op de predictoren. In paragraaf 4 rapporteren we de resultaten van de multipele regressie-analyse. In de vijfde paragraaf verkennen we een mogelijkheid om de uitkomsten van de regressievergelijking die we in paragraaf 4 presenteren van een domein- of taakgerichte interpretatie te voorzien. Een score die domeingericht geïnterpreteerd kan worden, geeft aan welke taken een leerling al wel en welke taken hij nog niet met succes kan uitvoeren, of andersom, welke teksten geschikt zijn voor welke leerlingen.

2 Criterium voor de vereiste leestechniek

Meting van de door teksten vereiste leestechniek

Technisch lezen is het omzetten van grafemen in fonemen. Het belangrijkste aspect van de leestechniek is – nadat de teken-klank-koppelingen geleerd zijn – de snelheid waarmee het leesproces geschreven woorden herkent ofwel de koppeling tussen woordvorm en de daarmee in het interne lexicon geassocieerde kennis tot stand brengt.

Een bekend en als betrouwbaar en valide geaccepteerd instrument voor de meting van de leestechniek is de zogenaamde één-minuuttoets of kortweg: de EMT (Ballard, 1920, was voor wij weten de eerste die een één-minuuttoets gebruikte). Dit instrument bestaat in het algemeen uit een reeks van 100 tot 200 woorden. De maat voor de leestechniek die het levert, is het aantal woorden dat in 60 seconden correct verklankt wordt. In ons land is het gebruik van de test van Brus en Voeten (1973) zeer verbreid, zowel in het basisonderwijs als in onderzoek. PPON gebruikt een één-minuuttoets (Zwarts, 1990) en het project Leerlingvolgsysteem van het Cito heeft nog onlangs onder de naam 'Drie-Minuten-Toets' drie versies van dit type instrument uitgegeven (Verhoeven, 1995). Deze bestaat uit drie in moeilijkheid verschillende één-minuut-toetsen. Wij maken in ons onderzoek gebruik van een variant van de EMT. De reeksen woorden waaruit ons onderzoeksmateriaal bestaat, zijn afkomstig uit bestaande teksten. We gaan ervan uit dat als van een reeks A in één minuut leestijd meer woorden verklankt worden dan van een reeks B, de gelezen woorden in reeks A 'technisch' leesbaarder zijn dan de gelezen woorden in een reeks B.

Tabel 1
Gemiddelde, SD en bereik van de gemiddelde DMT3-scores per reeks (N = 48)

Gem. LVS-EMT	SD	Laagste waarde	Hoogste waarde
59.421	0.4185	58.580	60.358

Constructie en selectie

In ons onderzoek hebben we gebruik gemaakt van 48 reeksen van 150 woorden. De reeksen zijn afkomstig uit een steekproef van teksten, die op hun beurt weer komen uit een corpus, dat gevormd is ten behoeve van het CLIB-onderzoek (Staphorsius, 1994). Dit corpus bestaat uit een steekproef van 240 teksten 'fictie' en 240 teksten 'non-fictie' (zie voor een uitvoerige verantwoording van deze steekproef Staphorsius, Krom & De Geus, 1988). Uit elk van de subcorpora hebben we voor het onderzoek naar de 'technische' leesbaarheid aselect 24 teksten getrokken. Per tekst hebben we drie steekproeven van 50 woorden getrokken. In elke trekkingsronde begonnen we met een aselect bepaald woord van de eerste zeven woorden (de fragmenten in het corpus hebben een omvang van ongeveer 350 woorden). Vervolgens selecteerden we elk zevende woord. En dat drie keer. De gevolgde procedure – we zullen verder om bondig te kunnen verwijzen ook wel spreken van ETP, een afkorting van EMT-tekst-procedure – leidde uiteindelijk dus tot 48 woordreeksen. Net als bij Ballard (1920), maar anders dan in de hierboven met name genoemde één-minuut-toetsen zijn in deze reeksen ook functiewoorden opgenomen. Een ander gevolg van de ETP is dat dezelfde woordvormen meerdere keren in een reeks kunnen voorkomen (in de ETP is een stap opgenomen die voorkomt dat dezelfde woordvormen direct na elkaar in een reeks voorkomen).

We gaan er op grond van de gevolgde procedure van uit dat het gemiddelde en de spreiding van de woordkenmerken in de 48 reeksen woorden een afspiegeling zijn van de verdeling van woordkenmerken in de populatie fictie en non-fictie voor de jeugd. Als ons onderzoek inderdaad resulteert in een voorspeller van de door teksten vereiste leestechniek, kunnen we de belasting van de woordherkenner in het leesproces door een gegeven tekst uit de populatie fictie en non-fictie voor de jeugd voorspellen op basis van kenmerken van de woorden in die tekst.

Afname van de 48 reeksen van 150 woorden

We hebben omdat we niet alle 48 woordreeksen aan één proefpersoon of aan één groep proefpersonen konden afnemen en toch de ingezette leestechniek per reeks moesten controleren, het 'randomized blocking' principe toegepast. Dat vereist een voormeting van de leestechniek van proefpersonen. Op 5 scholen in een steekproef die we gebruikten voor eerder onderzoek hebben we aan alle leerlingen in groep 4 tot en met 8 de kaart met de moeilijkste reeks woorden in de Cito-LVS Drie-Minuten-Toets afgenomen (omdat dit de derde kaart in de Drie-Minuten-Toets is, spreken we verder kortweg van DMT3). Op basis van de scores op deze toets hebben we 48 groepen van 36 leerlingen samengesteld met eenzelfde gemiddelde score en standaardafwijking op de DMT3. Elke groep van 36 kreeg een van de 48 reeksen toegewezen.

In Tabel 1 geven we het gemiddelde van de gemiddelde DMT3-scores per verzameling. De gemiddelde DMT3-score per groep van 36 leerlingen komt overeen met de gemiddelde leestechniek in het basisonderwijs aan het einde van groep 5.

Resultaten

In Tabel 2 geven we het gemiddelde en de spreiding van de criteriumwaarden (het aantal in 1 minuut gelezen woorden) van alle 48 reeksen woorden, maar we onderscheiden in de

Tabel 2
Gemiddelden en SD van de criteriumwaarden in fictie en non-fictie samen en op fictie en non-fictie afzonderlijk

Fictie en non-fictie		Fictie		Non-fictie	
Gemiddelde	*SD*	*Gemiddelde*	*SD*	*Gemiddelde*	*SD*
81.16	6.73	81.84	5.55	80.48	7.79

tabel ook tussen de reeksen die het resultaat zijn van de toepassing van de EMT-tekst-procedure op de twee in onze steekproef te onderscheiden genres, fictie en non-fictie.

3 De predictoren

Het woordlengte-, het woordfrequentie- en het contexteffect

Welke kenmerken komen als predictoren in aanmerking? Het antwoord op deze vraag is uiteraard afhankelijk van de specificatie van het begrip 'technisch lezen' of 'leestechniek'. Interindividuele verschillen in leestechniek worden, nadat lezers de bestaande 'grafeem-foneem'-koppelingen geleerd hebben, bepaald door het tempo waarin lezers in staat zijn tot het herkennen van woorden. Zijn er kenmerken van woorden die het woordherkenningstempo tijdens het lezen van 'lopende' teksten beïnvloeden en die voor de predictie van de door teksten vereiste leestechniek relevant kunnen zijn? In de literatuur wordt naar twee robuuste effecten verwezen: het woordlengte- en het woordfrequentie-effect. (zie Thomassen, Noordman & Eling, 1984, 1991; Rayner & Pollatsek, 1989; Just & Carpenter, 1987). Korte en frequente woorden worden sneller herkend. Behalve op het woordlengte- en frequentie-effect wijst de literatuur op het contexteffect. In de discussie over dit effect gaat het er niet in de eerste plaats om of de context het leestempo al dan niet beïnvloedt. Centraal staat de vraag of het contexteffect intra- dan wel postlexicaal is: faciliteert de context of faciliteren kenmerken van organisatieniveaus in de tekst hoger dan het woord (zin, alinea enzovoort) de leestaak tijdens of na het herkennen van woorden? Onderzoek wijst soms op een intra-lexicaal effect, maar vaker op een postlexicale invloed van de context (Rayner & Pollatsek, 1989). Voorzover het contexteffect inderdaad postlexicaal is, hoeft het bij de selectie van de predictoren, noch bij de keuze van het criterium een belangrijke rol te spelen, want het gaat ons in de eerste plaats om verschillen in tempo voor zover die afhankelijk zijn van variabelen die inwerken op het gemak en de snelheid waarmee lezers decoderen.

Een intra-lexicaal contexteffect is het zogenaamde repetitie-effect. Het blijkt dat een woord dat in een experiment herhaald wordt aangeboden, bij de herhaalde aanbiedingen sneller herkend wordt. Hudson, Bergman, Houtmans en Nas (1984) maken melding van onderzoek van Scarborough et al. (1977) en Scarborough et al. (1979) waaruit blijkt dat dit effect zelfs na een dag nog meetbaar is. Woordherkenningsmodellen verklaren het repetitie-effect uit een verlaging van de herkenningsdrempel. Dit effect wijst erop dat 'diversiteit van de woordenschat' als predictor in het leesbaarheidsonderzoek bruikbaar zou kunnen zijn. Operationalisering van deze variabele lijkt zinvol omdat door onze EMT-tekst-procedure hetzelfde woord meerdere keren in een reeks kan voorkomen.

Predictoren

De maten voor woordlengte in ons onderzoek duiden we aan met GWL en GWLG, terwijl we naar de maat voor woordfrequentie verwijzen met FREQ77. Als maat voor de diversiteit van de woordenschat in teksten gebruiken we PTYPES.

GWL is de gemiddelde woordlengte in letters, het quotiënt van het aantal letters en het aantal woorden. GWLG is de gemiddelde woordlengte uitgedrukt in lettergrepen. Deze twee predictoren zullen elkaar, zo leert ons het leesbaarheidsonderzoek, sterk overlappen. Dat neemt niet weg dat GWL het criterium voor veronderstelde leestechniek nauwkeuriger zou kunnen voorspellen dan GWLG, of andersom. Om te kunnen bepalen welke van de twee de beste voorspeller is, betrekken we beide predictoren in ons onderzoek.

FREQ77 is de door ons gehanteerde maat voor 'woordfrequentie'. Deze is gebaseerd op het onderzoek van Staphorsius, Krom en De Geus (1988). Zij voerden op het zogenaamde P335-corpus, de hierboven al eerder genoemde steekproef van 480 teksten uit de populatie jeugdlectuur, frequentietellingen uit. Het resultaat van de tellingen is onder meer een naar frequentie geordende lijst van in het corpus voorkomende woorden. Als frequent definiren we de 998 meest frequente woorden. Het zijn woorden die twintig keer of vaker in het corpus voorkomen. Ze maken bijna 77 procent van de tokens in het corpus (n = 202526) uit. Het totaal aantal types = 18210; de gemiddelde frequentie is ruim 11. Voor de deellijst van 998

Tabel 3
Gemiddelde en SD predictoren in de woordreeksen fictie, non-fictie en fictie en non-fictie samen

	Fictie en non-fictie		Fictie		Non-fictie	
	Gemiddelde	*SD*	*Gemiddelde*	*SD*	*Gemiddelde*	*SD*
GWL	4.45	0.38	4.36	0.27	4.55	0.46
GWLG	1.48	0.13	1.45	0.10	1.50	0.16
GFREQ77	78.18	6.63	79.32	6.35	77.03	6.84
PTYPES	75.32	6.04	75.10	6.06	75.52	5.99

woorden geldt een gemiddelde frequentie van ongeveer 156. We noemen deze lijst in het vervolg de FREQ77-lijst. FREQ77 is het percentage tokens in een tekst dat in de FREQ77-lijst voorkomt.

PTYPES staat voor het percentage verschillende woorden. Deze maat voor de type/token-ratio, dit wil zeggen de verhouding tussen het aantal verschillende woordvormen en het totaal aantal woorden, is onze maat voor de diversiteit van de woordenschat.

Meetresultaten

In Tabel 3 geven we per te onderscheiden steekproef (fictie, non-fictie en fictie en non-fictie samen) het gemiddelde en de SD van de waarden voor GWL, GWLG, FREQ77 en PTYPES.

De waarden zijn, met het oog op de vereiste homogeniteit, gebaseerd op tellingen in de eerste 80 woorden van de reeksen, het aantal dat overeenkomt met het gemiddeld aantal in de reeksen gelezen woorden.

Tussen haakjes: uit de literatuur en uit de praktijk van het (aanvankelijk) leesonderwijs is bekend dat woorden met bepaalde medeklinkercombinaties (bijvoorbeeld: sch..., schr...) moeilijker te verklanken zijn dan woorden zonder die combinaties. Toch hebben we 'lastige medeklinkercombinaties' niet als predictor in ons onderzoek betrokken. 'Gewone' teksten verschillen namelijk niet of nauwelijks als het om dit kenmerk gaat. Daarvoor hebben we een logische verklaring: 'lastige medeklinkercombinaties' is geen variabele die door schrijvers van 'gewone' teksten gemanipuleerd wordt. Voor deze verklaring vinden we steun in de empirie. Als maat voor 'lastige medeklinkercombinaties' kan de gemiddelde lettergreeplengte per letter worden gehanteerd. Immers: lastige medeklinkercombinaties zijn relatief lang, dus hoe groter de lettergreeplengte, hoe vaker lastig te verklanken medeklinkercombinaties in een tekst zullen voorkomen. In ons onderzoeksmateriaal blijkt de lettergreeplengte gemiddeld ± 3.1 te zijn. De standaardafwijking van dit gemiddelde is ± 0.09. Uit de correlatie met ons criterium wordt de betekenis van de verschillen tussen de teksten in ons corpus wat deze predictor betreft duidelijk: er is geen correlatie. Voor andere woordkenmerken die de verklanking bemoeilijken geldt hetzelfde als voor 'lastige medeklinkercombinaties'. Als er onder deze kenmerken al potentiële predictoren schuil gaan (bijvoorbeeld niet-klankzuivere leenwoorden), dan bestaat er een aanzienlijke kans dat we die 'gevangen' hebben met de predictor FREQ77.

4 Multipele regressie-analyse

In paragraaf 2 gaven we aan hoe de criteriumscores in ons onderzoek bepaald zijn. In paragraaf 3 presenteerden we vervolgens een overzicht van de onafhankelijke variabelen of predictoren, waarvan we op grond van gegevens uit de literatuur konden verwachten, dat ze een substantieel gedeelte van de variantie van het criterium zouden verklaren. In deze paragraaf rapporteren we de resultaten van de uitgevoerde multipele regressie-analyse. In deze analyse gaat het er om de optimale gewichtsverhouding te bepalen waarmee de predictoren het criterium voorspellen. In Tabel 4 rapporteren we per te onderscheiden steekproef de correlatiecoëfficiënten tussen de predictoren onderling en ook de resultaten van de multipele regressie-analyse. In de multipele regressie-analyse bleek dat PTYPES' bijdrage aan de voorspelling van het criterium niet significant

Tabel 4

Resultaten regressie-analyses in de steekproeven fictie, non-fictie en in fictie en non-fictie samen

	Fictie en non-fictie				Fictie				Non-fictie			
	1	2	3	4	1	2	3	4	1	2	3	4
1 GWL	1.00				1.00				1.00			
2 GWLG	.95	1.00			.90	1.00			.96	1.00		
3 FREQ77	-.65	-.54	1.00		-.55	-.46	1.00		-.70	-.59	100	
4 PTYPES	.37	.27	-.44	1.00	.26	.06	-.48	100	.46	.40	-.41	1.00
5 CRIT	-.88	-.85	.73	-.39	-.86	-.84	.66	-.27	-.90	-.85	.78	-.49
INTERCEPT	114.49				126.84				110.35			
FREQ77	0.28				0.24				0.32			
GWL	-12.33				-14.59				-12.01			
R (R^2)	.90 (.82)				.89 (.79)				.92 (.85)			

was. Vandaar dat deze predictor geen deel uit maakt van de multipele regressievergelijkingen. Zoals verwacht is de overlap tussen GWL en GWLG zeer groot. In de vergelijking is geen plaats voor beide predictoren. GWL doet het als predictor iets beter dan GWLG. Daarom is GWL in de definitieve multipele regressieanalyse als predictor ingevoerd en niet GWLG.

Kruisvalidering

Een multipele regressie-analyse schat regressiecoëfficiënten uit de correlatiematrix die afkomstig is uit een steekproef van observaties; de schattingen kunnen zeer wel beïnvloed zijn door de specifieke eigenaardigheden van die steekproef. Als we de schattingen willen beschouwen en gebruiken als benaderingen van de regressiecoëfficiënten die in de populatie (van teksten) gelden, dan moeten we aantonen dat de schattingen niet al te veel zijn 'aangetast' door de eigenaardigheden van de steekproef waaruit ze geschat zijn.

Een techniek die men daar vaak voor gebruikt is de kruisvalidatie. Deze komt er op neer dat men de regressiecoëfficiënten schat uit één steekproef (de screeningssteekproef; Lord & Novick, 1968) en deze schattingen gebruikt om het criterium in een onafhankelijke steekproef (de calibratiesteekproef) te voorspellen. In het algemeen vindt men dat de multipele correlatie in de calibratiesteekproef lager uitvalt dan in de screeningssteekproef; een eventueel verschil wordt 'krimp' genoemd. Een kleine krimp is een teken van stabiliteit van de schatters van de regressiecoëfficiënten.

In ons onderzoek hebben we de teksten behorend tot de subcorpora 'Fictie' en 'Non-fictie' beurtelings de rol van screeningssteekproef en calibratiesteekproef toebedeeld, en aldus twee kruisvalidaties uitgevoerd. De geschatte regressiecoëfficiënten zijn weergegeven in Tabel 4. In Tabel 5 geven we een overzicht van de resultaten.

Uit Tabel 5 blijkt dat de krimp van de R van 'Non-fictie' naar 'Fictie' nihil is en dat er van 'Fictie' naar 'Non-fictie' sprake is van geringe krimp. We gingen ook na of de regressievergelijkingen voor het corpus 'Fictie' en 'Non-fic-

Tabel 5

Resultaten kruisvalidatie

	screeningssteekproef			
	Fictie		Non-fictie	
calibratiesteekproef	Fictie	Non-fictie	Fictie	Non-fictie
R	0.89	0.88	0.92	0.92

tie' als aan elkaar gelijk kunnen worden beschouwd: de verschillen tussen beide vergelijkingen blijken niet significant te zijn ($\alpha = .05$). Dat wijst op de validiteit van de vergelijkingen in de afzonderlijke steekproeven.

Vanwege een grotere stabiliteit komt de regressievergelijking op basis van de data in de steekproeven fictie en non-fictie samen in aanmerking voor gebruik in de praktijk (zie Kerlinger & Pedhazur, 1973, blz. 284). We noemen deze vergelijking Cito Index Leestechniek, afgekort CILT.

$$\text{CILT} = 114.49 + 0.28 \times \text{FREQ77} - 12.33 \times \text{GWL} \quad (1)$$

5 Van een relatieve naar een domeingerichte index

De CLIB, we spraken over deze index in de inleiding, verwijst als index van de leesbaarheid van teksten naar een subpopulatie of domein van leerlingen met een leesvaardigheid die minimaal vereist is om een tekst te kunnen lezen en, omgekeerd, als een index voor de leesvaardigheid, gemeten met de Clib-toetsen, verwijst de CLIB naar een domein van teksten. Als een leerling op de Clib-toetsen een CLIB van 42 heeft behaald, dan kan deze teksten begrijpen met een CLIB van 42 en lager. Kan ook de CILT van een domeingerichte betekenis worden voorzien? In ieder geval niet zonder een aanvulling op het onderzoek waarvan we in de vorige paragrafen verslag deden en waarin het uitsluitend te doen was om de 'ontwikkeling' van een relatieve index voor de leestechniek. Een domeingerichte CILT vereist een ander (en 'duurder') criterium voor de leestechniek. Nadat we ons CILT-onderzoek hadden afgerond, ontstond in het project dat eenvoudig klassikaal afneembare toetsen voor de leessnelheid voor het Cito-Leerlingvolgsysteem ontwikkelt (Toetsen Leestempo, Krom, 1996), de behoefte aan een domeingerichte index en de vraag is nu of we, gebruik makend van de resultaten van het CILT-onderzoek, de CILT alsnog van een domeingerichte betekenis kunnen voorzien. Dat is de vraag waar het in deze paragraaf om gaat. Op het praktisch nut van een domeingerichte CILT komen we in de bespreking terug.

Als we de CILT willen voorzien van een domeingerichte betekenis dan moeten we in staat zijn vast te stellen hoe groot de technische leesvaardigheid van een lezer moet zijn om met succes een tekst met een gegeven CILT te kunnen decoderen. Minder cruciaal voor een absolute interpretatie, maar wel handig is het als behalve de door teksten veronderstelde leestechniek ook de leestechniek van lezers in de CILT kan worden uitgedrukt. Ter bepaling van de gedachten geven we een mogelijke opzet van een onderzoek naar een domeingerichte CILT met als uitgangspunt de relatieve CILT. In zo'n onderzoek zouden we eerst kunnen vaststellen wat voor elke leerling in een steekproef van leerlingen de moeilijkste tekst is in een steekproef van teksten die hij of zij met succes technisch kan lezen. Op deze manier kunnen we de leestechniek van de proefpersonen uitdrukken in een CILT, namelijk de CILT van de moeilijkste tekst die hij of zij kan decoderen. Daarnaast zouden we met een valide en betrouwbare toets (bijvoorbeeld met de bovengenoemde toetsen Leestempo) kunnen vaststellen wat de decodeervaardigheid van de proefpersonen is. We kennen dan van elke leerling de CILT en de score op de decodeertoets. Als de steekproeven van teksten en leerlingen aan het doel van het onderzoek beantwoorden, kan worden nagegaan of vanuit de CILT de scores op de gebruikte toets kunnen worden voorspeld (of andersom).

Wij hebben een variant van het onderzoek met hierboven geschetste opzet uitgevoerd. Wij hebben gebruik gemaakt van gegevens uit een onderzoek dat Staphorsius (1994) uitvoerde naar de bruikbaarheid van een tekstenschaal voor de beoordeling van de leesvaardigheid van leerlingen en naar de validiteit van de CLIB als 'absolute' leesindex. Staphorsius vroeg leerkrachten (aantal = 87; aantal verschillende scholen: 21; totaal aantal leerlingen in groep 5 tot en met 8: 1844) de vaardigheid 'begrijpend lezen' en de 'leestechniek' van hun leerlingen te bepalen met behulp van een tekstenschaal. Deze schaal bestond uit 12 teksten met ongeveer dezelfde gemiddelde CILT (en een iets grotere spreiding) als in de 48 woordreeksen in het 'CILT-onderzoek'. Aan de leerkrachten werd gevraagd de vaardigheid 'begrijpend lezen' van elke leerling in hun klas te schatten door de moeilijkste tekst op de

tekstenschaal aan te geven die de leerling naar zijn of haar oordeel met begrip zou kunnen lezen. Bovendien vroeg Staphorsius de leerkrachten van elke individuele leerling aan te geven wat naar schatting de moeilijkste tekst was die deze technisch 'voldoende goed' zou kunnen lezen. In het onderzoek waren representatieve steekproeven leerlingen uit groep 5 tot en met 8 betrokken (zie Staphorsius, 1994). Alle leerlingen hadden ook de Clib-toetsen gemaakt zodat Staphorsius behalve over het leerkrachtenoordeel ook over een 'directe' maat voor de leesvaardigheid van elke leerling (uitgedrukt op de CLIB-schaal) beschikte. Van deze dataset maken we nu gebruik bij onze pogingen om de relatieve CILT om te zetten in een absolute CILT. We beschikken over een maat voor de leestechniek van de leerlingen via de tekstenschaal. Van alle leerlingen weten we zo wat volgens hun leerkracht de moeilijkste tekst is die zij 'technisch' kunnen lezen. Van die tekst kennen we ook de CILT. We weten via de direct bepaalde CLIB voor welke 'gemiddelde leerling' deze CILT geldt: de gemiddelde leerling midden groep 4, eind groep 4, midden groep 5 enzovoort. Via het onderzoek dat is uitgevoerd ter normering van onder meer de toetsen Leestempo (Krom, 1996) en de DMT3 kennen we de waarden voor het met deze toetsen bepaalde leestempo van deze 'gemiddelde leerlingen'. We weten nu bijvoorbeeld van de derde tekst op de tekstenschaal wat, uitgedrukt op de Leestempo-schaal, het gemiddelde leestempo is van leerlingen die deze tekst wel, maar de volgende op de tekstenschaal nog niet voldoende goed kunnen lezen. We merkten net al op dat de derde tekst net als de elf andere teksten ook een CILT heeft. We kunnen nu een analyse uitvoeren waarin we deze CILT-waarden regresseren op bijvoorbeeld de Leestempo-waarden van de 12 teksten. We zoeken dus de coëfficinten A en B, zo dat

$$CILT_i = A + B \times LEESTEMPO_i + \varepsilon_i \qquad (2)$$

waarbij de variantie van de residuen zo klein mogelijk is. Deze coëfficiënten bepalen de regressievergelijking waarmee we de waarde van een tekst op de Leestempo-schaal kunnen omzetten in een waarde op de CILT-schaal. Deze analyse hebben we inderdaad uitgevoerd (r = .94). We hadden uiteraard ook de gemiddelde Leestempo-scores als (domeingericht) criterium kunnen hanteren en kunnen nagaan welke combinatie van predictoren dit criterium het best voorspelde. De stabiliteit van de CILT op basis van 48 woordreeksen echter is groter dan de stabiliteit van een index gebaseerd op 12 teksten.

Van een leerling met een bepaalde score op de toetsen Leestempo (Krom, 1996) kunnen we nu vaststellen voor welke teksten deze vermoedelijk al wel een voldoende leestechniek heeft en voor welke teksten dat naar alle waarschijnlijkheid nog niet het geval is. Immers de scores op de toetsen Leestempo kunnen worden omgezet naar een CILT, terwijl met formule 1 ook de CILT van teksten kan worden bepaald. Dat maakt het mogelijk teksten (boeken, artikelen, brochures enzovoort) te vinden met een CILT die ongeveer overeenkomt met de via de toetsen Leestempo bepaalde CILT van de leerling. Zo is de CILT van een domeingerichte interpretatie voorzien. We hadden uiteraard dezelfde oefening kunnen uitvoeren met de LVS-EMT of een andere voor ons doel valide toets. Ons onderzoek heeft met dit resultaat een voorlopig einde. Voorlopig, want we zullen onze verkenningen van domeingerichte criteria voor de leestechniek voortzetten.

6 Bespreking

Lezers verschillen in leestechniek en teksten verschillen in vereiste leestechniek. In dit verslag hebben we een voorspeller van relatieve verschillen tussen teksten in vereiste leestechniek, geïntroduceerd. Ons onderzoeksmateriaal bestond uit 48 reeksen van 150 woorden. Deze zijn met behulp van de EMT-tekst-procedure op basis van teksten in een steekproef uit de populatie jeugdlectuur samengesteld. Als criterium voor de vereiste leestechniek hanteerden we het aantal woorden dat door de proefpersonen gemiddeld correct gelezen werd. De proefpersonen waren leerlingen uit het basisonderwijs met een leestechniek die overeenkomt met de gemiddelde leestechniek van de leerlingen aan het einde van groep 5. Als predictoren hanteerden we het percentage frequente woorden, het percentage verschillende woorden, de gemiddelde woordlengte in letters en in letter-

grepen. De multipele regressie-analyse resulteerde in de CILT, een regressievergelijking met het percentage frequente woorden en de gemiddelde woordlengte in letters als predictoren (R = .90; R^2 = .82). De bijdrage van de gemiddelde woordlengte in lettergrepen en van het percentage verschillende woorden aan de voorspelling van het criterium naast de andere predictoren was niet significant. Van willekeurige teksten kan, nadat het percentage frequente woorden en de gemiddelde woordlengte in letters in die teksten bepaald is, de CILT berekend worden. We zijn er in geslaagd onder meer de scores op de toetsen Leestempo (Krom, 1996) om te zetten in een CILT die aangeeft wat de moeilijkste tekst is die een leerling kan verklanken. Daarmee heeft de CILT, naast een relatieve, een domeingerichte betekenis gekregen. We komen nu toe aan de vraag naar de praktische waarde van de CILT. De onderwijspraktijk kan al beschikken over de CLIB. Kan het onderwijs met indexen als de CILT iets dat het met de CLIB niet kan?

Voor de meeste leerlingen in de fase van het aanvankelijk lezen zijn de teksten die ze te lezen krijgen conceptueel gezien triviaal terwijl de decodeerproblemen in verhouding nog groot zijn. Voor een grote meerderheid van de leerlingen smaakt de overwinning van die problemen dan vermoedelijk nog zó goed, dat de inhoudelijke eenvoud van het aangereikte leesgoed ze niet al te zeer stoort. De decodeervaardigheid neemt al tijdens het aanvankelijk leesonderwijs snel zo toe dat zelfs conceptueel gezien nog te moeilijke teksten technisch binnen het bereik van die vaardigheid komen. Het lijkt er dan ook op dat de leestechniek (aspect: tempo) zich bij de meeste leerlingen ruwweg vanaf eind groep 4 kan ontwikkelen door het lezen van conceptueel gezien passende teksten. Voor zwakke decodeerders ligt dit vermoedelijk enigszins anders. Voor een aantal van deze lezers blijft de discrepantie bestaan, waarvan aan het begin van het leesonderwijs voor vrijwel alle leerlingen sprake is. Door hun matige leestechniek kan het gebeuren dat deze leerlingen niet of nauwelijks toekomen aan het begrijpen van teksten die conceptueel gezien geen enkel probleem vormen (Perfetti, 1985). Voor leerlingen met een grote discrepantie tussen leestechniek en conceptuele mogelijkheden, zo vermoedde men ook in het project 'LVS-Leestempo' (Krom, 1996), zouden indexen zoals de CILT van didactische betekenis kunnen zijn, vooral als zou blijken dat teksten met een ongeveer gelijke conceptuele lading verschillen in de leestechniek die ze veronderstellen. Tabel 6 geeft van deze verschillen een indruk, maar vereist eerst een nadere toelichting.

In Tabel 6 vergelijken we op basis van Staphorsius (1996) het Clib-niveau van teksten (totale aantal = 9991) met de index van het Katholiek Pedagogisch Centrum (Visser, Van Laarhoven en Ter Beek, 1994) voor de **leestechniek**, het zogenaamde AVI-niveau (in het onderwijs ook wel 'de AVI' genoemd) van dezelfde teksten. De teksten zijn afkomstig uit bijna 1500 verschillende bronnen (fictie en

Tabel 6
Gemiddelde, SD en bereik van AVI-niveau per Clib-niveau

		AVI-niveau**			
Clib-niveau		*Gem.*	*SD*	min.	max.
4	(266)*	5.56(B5)	0.99	4(E4)	10(>E6)
5	(1342)	7.14(E5)	0.97	5(E4)	10(>E6)
6	(2967)	8.56(E6)	0.88	5(B5)	10(>E6)
7	(3045)	9.63(>E6)	0.61	4(M4)	10(>E6)
8	(1477)	9.97(>E6)	0.18	8(B6)	10(>E6)

* Aantal teksten in het aangegeven Clib-niveau

** Bijvoorbeeld (B4) betekent: komt ongeveer overeen met het gemiddelde begin groep 4; (M4) = ongeveer gemiddelde midden groep 4; (E4) = ongeveer gemiddelde eind groep 4; >E6 = niveau dat na het moment E6 behaald wordt.

non-fictie) en hebben een gemiddelde lengte van ruim 420 woorden. De Clib-niveaus zijn niveau-aanduidingen voor het ***begrijpen*** van teksten. Per Clib-niveau kan het gemiddelde en de standaardafwijking van het AVI-niveau van de teksten worden afgelezen. Bijna 900 teksten behoren tot hogere of lagere niveaus dan de in de tabel opgenomen Clib-niveaus. Bijvoorbeeld Clib-niveau 4 verwijst naar teksten die geschikt zijn voor de gemiddelde leerling aan **het einde** van groep 4 in de basisschool. Voor de berekening van de gemiddelde AVI-index per Clibniveau hebben we alle teksten die boven AVI-niveau 9 uitkomen met een 10 gescoord en alle teksten lager dan niveau 5 met een 4. Op basis van de normgegevens van Visser, Van Laarhoven en Ter Beek (1994) hebben we achter de AVI-niveaus aangegeven welke 'gemiddelde leerling' op welk moment over een bepaald niveau beschikt. De momenten 'begin leerjaar' (september/oktober) en 'eind leerjaar' (maart/april) zijn gebaseerd op de gerapporteerde empirische gegevens. De waarden voor het moment 'midden leerjaar' zijn door ons op grond van deze empirische waarden geschat.

Uit Tabel 6 blijkt dat teksten met een vergelijkbare conceptuele moeilijkheid (uitgedrukt in Clib-niveau) onderling wat betreft de volgens het AVI-systeem vereiste leestechniek behoorlijk verschillen, soms zelfs 'jaren'. Een index voor de leestechniek maakt het mogelijk te midden van teksten met ongeveer dezelfde conceptuele moeilijkheid de teksten te identificeren met de laagste decodeermoeilijkheid. Met een domeingerichte index kan bovendien worden vastgesteld of zelfs de tekst met de laagste decodeerlading in vergelijking met de aanwezige leestechniek wel leesbaar is. Indexen voor de leestechniek kunnen zo bezien in het onderwijs aan leerlingen die specifiek op de leestechniek uitvallen een rol spelen. Uit onderzoek in de onderwijspraktijk echter moet blijken of dat ook een substantiële rol is. De uitkomst van dat onderzoek is in de eerste plaats afhankelijk van de vraag in hoeverre het domeingerichte criterium staat voor de minimaal nodige leestechniek en in de tweede plaats van de vraag hoe vaak de betreffende leerlingen inderdaad aan een bij hun leesdoel passende en conceptueel en technisch gezien leesbare tekst geholpen kunnen worden. Tussen haakjes: het is opvallend dat teksten die door gemiddelde leerling 'eind groep 4' met begrip gelezen kunnen worden, volgens de AVI-normering voor deze leerlingen zelfs gemiddeld genomen nog te grote decodeerproblemen bevatten.

We sluiten af met enkele concluderende opmerkingen over de CILT. Deze index lijkt in ieder geval bruikbaar voor een domeingerichte beschrijving van het niveau en het van de ontwikkeling van de leestechniek in het primair onderwijs. Didactische betekenis heeft de CILT naast de CLIB naar het zich laat aanzien vooral voor het onderwijs aan zeer zwakke decodeerders. Vermoedelijk kan de differentiatie van het leesonderwijs aan de andere leerlingen gebaseerd worden op de vaardigheid 'begrijpend lezen'. Een strikt uitgevoerde differentiatie op basis van de leestechniek van deze leerlingen zou er, gelet op de gegevens in Tabel 6, zelfs toe kunnen leiden dat ze teksten 'moeten' lezen die ze niet kunnen begrijpen en dat ze teksten niet 'mogen' lezen, terwijl ze die wel zouden kunnen begrijpen.

Literatuur

Ballard, P.B. (1920). *Mental Tests.* Londen: University of London Press.

Brus, B.T., & Voeten, M.J.M. (1973). *Een-Minuut-Test, vorm A en B.* Nijmegen: Berkhout.

Hudson, P.T.W., Bergman, M.W., Houtmans, M.J.M., & Nas, G.L.J. (1984). De bestudering van woordherkenning als basis voor het lezen. In A.J.W.M. Thomassen, L.G.M. Noordman & P.A.T.M. Eling (Red.), *Het leesproces* (pp. 27 – 52). Lisse: Swets & Zeitlinger.

Just, M. A., & Carpenter, P.A. (1987). *The psychology of reading and language comprehension.* Newton: Allyn and Bacon.

Kerlinger, F.N., & Pedhazur, E.J. (1973). *Multiple regression in behaviourial research.* New York: Holt, Rinehart and Winston.

Krom, R.S.H. (1996). *Toetsen Leestempo.* Arnhem: CITO.

Lord, F.M., & Novick M.R. (1968). *Statistical theories of mental test scores.* New York: Addison Wesley.

Perfetti, C.A. (1985). *Reading ability.* New York: Oxford University Press.

Rayner, K., & Pollatsek, A. (1989). *Psychology of reading.* Englewood Cliffs: Prentice Hall.

Scarborough, D.L., Cortese, C., & Scarborough, H.S. (1977). Frequency and repetition effects in lexical memory. *Journal of Experimental Psychology: Human Perception and Performance, 3,* 1-17.

Scarborough, D.L., Gerard, L., & Cortese, C. (1979). Accessing lexical memory: The transfer of word repetition effects across task and modality. *Memory & Cognition, 7,* 3-12.

Staphorsius, G. (1992). *Clib-toetsen.* Arnhem: CITO.

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

Staphorsius, G. (1996). *Overzicht van tekstkenmerken van bijna 10.000 teksten voor de jeugd uit bijna 1500 verschillende bronnen.* (P335-doc 601). Arnhem: CITO.

Staphorsius, G., & Krom, R.S.H. (1985). *Cito leesbaarheidsindex voor het basisonderwijs.* Arnhem: CITO.

Staphorsius, G., Krom, R.S.H., & De Geus, K. (1988). *Frequenties van woordvormen en letterposities in jeugdlectuur.* Arnhem: CITO.

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

Thomassen, A.J.W.M., Noordman, L.G.M., & Eling, P.A.T.M. (1984). *Het leesproces.* Lisse: Swets & Zeitlinger.

Thomassen, A.J.W.M., Noordman, L.G.M., & Eling, P.A.T.M. (1991). *Lezen en begrijpen.* Lisse: Swets & Zeitlinger.

Verhoeven, L. (1995). *Drie Minuten Toets.* Arnhem: CITO.

Visser, J., Laarhoven, A. van, & Beek, A. ter (1994). *AVI-Toetspakket.* Den Bosch: KPC.

Vogel, M., & Washburne, C.W. (1928). An objective method of determining grade placement of childrens reading material. *Elementary School Journal, 28,* 373-381.

Zwarts, M. (1990). *Balans van het taalonderwijs aan het einde van de basisschool.* Arnhem: CITO

Manuscript aanvaard 24 – 10 – 1996

Auteurs

G. Staphorsius (wetenschappelijk medewerker, Cito Instituut voor Toetsontwikkeling)

N.D. Verhelst (wetenschappelijk medewerker, Cito Instituut voor Toetsontwikkeling; Hoogleraar Psychometrie, faculteit der Toegepaste Onderwijskunde, Universiteit Twente)

Adres: Cito Instituut voor Toetsontwikkeling
Postbus 1034
6801 MG Arnhem

Abstract

The Development of a Domain Referenced Index of the Decoding Load of Texts

G. Staphorsius & N.D. Verhelst. Pedagogische Studiën, 1997, 74, 154-164.

The CRIE (**C**ITO ***R***eadability and Reading ability ***I***ndex for ***E***lementary education) is an index of both the reading ability of students in Dutch primary education and the readability of non-fiction prose. P-CRIE, a computer program, can indicate – in CRIE-values, ranging from 0 -100, – reading ability required by texts. The CRIE tests are an example of domain-referenced measurement of reading comprehension. The CRIE of a given reader refers to a set or domain of texts which can be comprehended by that individual with a given probability of success. The CRIE tests and P-CRIE can contribute to the individualization of reading instruction. The CRIE is primarily an index of the comprehensibility of texts and reading comprehension.

In addition to an index of readability as comprehensibility, an index of the decoding load of texts might be useful, especially for readers having a poor performance on decoding (aspect: speed), but whose comprehension of oral language is quite normal. Joining an index of the decoding skill required by texts to an index of the comprehensibility of texts, the matching of reading proficiency and readability can be optimized.

To develop such an index (of both the decoding skill and the 'decode-ability' of texts, we carried out a study, which will be reported on here. Our investigation has resulted in the CID (**C**ito **I**ndex of Decoding) a multiple regression equation with 'word length' and 'word frequency' as predictors ($R = .90$). The mean number of words decoded in one minute was selected as a measure of the decoding load of texts.