Kennis van woorden: de toetsing van diepe woordkennis

R. Schoonen en M. Verhallen

Samenvatting

Kennis van woorden is meer dan het correct benoemen of het aanwijzen van het juiste plaatje in een woordenschattoets. In dit artikel wordt ingegaan op 'diepte van woordkennis' en op het belang van diepere woordkennis voor de (taal)ontwikkeling van kinderen in het onderwijs. Een nieuwe toetsvorm van gestructureerde woordassociaties wordt gepresenteerd als operationalisatie van diepe woordkennis. In een empirisch onderzoek naar de betrouwbaarheid en validiteit van deze toetsvorm voor vijfde- en zevende-groepers uit het basisonderwijs blijkt dat de nieuwe toetsvorm in belangrijke mate aan de verwachtingen voldoet. Te verwachten verschillen in diepe woordkennis tussen groepen kinderen kunnen met de toets aangetoond worden en de toetsscores correleren hoog met de scores op een vergelijkbare toets.

1 Inleiding

Woordenschat is de afgelopen jaren steeds meer in de belangstelling komen te staan, zowel in onderzoek als in de praktijk van het onderwijs. Deze aandacht is geheel in overeenstemming met het belang dat aan woordkennis moet worden toegekend: bij taalverwerving én bij alledaags taalgebruik nemen woorden en hun betekenis een sleutelpositie in.

Met betrekking tot taal*verwerving* heeft onderzoek laten zien dat woordenschat nauw gerelateerd is aan andere aspecten van taalontwikkeling: lexicale maten zijn goede voorspellers van andere linguïstische vaardigheden van eerste- en tweede-taalleerders (Anderson & Freebody, 1981; Verhoeven & Vermeer, 1989). Ook in het taal*gebruik* blijken de woorden de centrale dragers van betekenis. Communicatie stagneert al snel als enkele woorden niet gekend worden of niet in hun onderling verband begrepen worden; omgekeerd kan men met enkele (inhouds)woorden al redelijk duidelijk boodschappen overbrengen, zo blijkt uit de stijl van telegrammen. In het onderwijs is het nauwelijks anders dan in het dagelijks leven daarbuiten.

Op school staan woorden centraal bij de overdracht van kennis; een minimale en vaak veel meer dan minimale, woordkennis is voorwaarde om met succes de dagelijkse lessen te kunnen volgen. Als kinderen de woorden die mondeling of schriftelijk in de klas gebezigd worden, niet voldoende kennen, zullen ze het leergesprek, de uitleg of de schoolboekteksten niet goed bevatten. Daarbij, en misschien wel daarom, is de uitbreiding van de woordenschat met de bijbehorende begrippen een doelstelling van de meeste vormen van onderwijs, zeker van het basisonderwijs. Zowel bij taal als bij andere vak- en vormingsgebieden moeten leerlingen nieuwe woorden, nieuwe betekenissen en nieuwe relaties tussen woorden leren.

De relatie tussen woordkennis en schoolsucces is in diverse studies aangetoond en in het licht van onderwijsprestaties is het verband tussen lexicale kennis en begrijpend lezen uitgebreid bestudeerd. Zowel bij eerste- als bij tweede-taalverwervende kinderen blijkt er een sterke samenhang tussen de omvang van de woordenschat en de vaardigheid in het begrijpend lezen (zie voor een overzicht o.a. Anderson & Freebody, 1981; Coady, 1995; Appel & Vermeer, 1996). In die onderzoeken blijft echter onderbelicht welke facetten van woordkennis nuttig zijn voor welke facetten van tekstbegrip, terwijl het zowel voor de theorievorming als voor de praktijk van het onderwijs belangrijk is om inzicht te hebben in de precieze aard van de relatie tussen facetten van woordkennis en begrijpend lezen. Als het om tekstbegrip of leesvaardigheid gaat, dan is de complexiteit van het construct algemeen geaccepteerd; er zijn vele studies naar een componentiële analyse van leesvaardigheid (zie onder meer Barr, Kamil, Mosenthal & Pearson, 1991; Schoonen

& Wolf, 1985; Singer & Ruddell, 1985). Woordenschat wordt daarentegen vaak expliciet of impliciet als een eenduidig begrip gezien, dat qua operationalisatie ook nauwelijks problemen lijkt te kennen. In dit artikel willen we dat beeld van woordkennis bijstellen en uitbreiden. We zullen een aantal onderbelichte facetten van woordkennis voor het voetlicht halen en daarbij aansluitend ingaan op enkele operationalisaties van woordkennis.

1.1 De woordenschat: een schat aan woorden?

Het begrip 'woordenschat' verwijst naar een reservoir, een schat, aan woorden die de taalgebruiker tot zijn beschikking heeft. Als men het in het onderwijs of in onderzoek over de woordenschat heeft, dan gaat het vooral om de omvang van dat reservoir. Het beeld van woordenschat als een losse verzameling van gekende woorden schiet echter op twee punten tekort. Ten eerste impliceert het dat een woord alleen wel of niet gekend zou kunnen zijn. Woorden kunnen echter ook gedeeltelijk gekend worden of juist heel goed in al zijn gebruiksmogelijkheden begrepen worden; deze variatie in de kwaliteit van woordkennis komt in het genoemde beeld van woordenschat niet tot uiting, terwijl we toch moeten aannemen dat voor de meeste woorden zal gelden dat individuen enorm verschillen in hun kennis van en hun ervaring met die woorden. Een indeling van woorden in wel of niet gekend doet in dat opzicht geen recht aan de psychologische werkelijkheid.

Het tweede punt is dat woorden, zowel in onderzoek als in het onderwijs, veelal als losse items beschouwd worden. Ook dit is niet conform de (psychologische) realiteit; op basis van psycholinguïstisch onderzoek moeten we aannemen dat woorden juist hun betekenis ontlenen aan de relaties die ze hebben met andere woorden in ons mentaal lexicon. Woorden figureren in een semantisch netwerk van gerelateerde woorden. Woordenschatuitbreiding is dan ook meer dan het verwerven van geïsoleerde, semantische 'units': nieuwe woorden worden ingebed in het netwerk van de woordenschat en dat betekent dat er allerlei verbindingen met andere woorden tot stand moeten komen (Aitchison, 1994; Verhallen & Verhallen, 1994). De bij het eerste punt genoemde variatie in woordkennis bestaat voor een belangrijk deel uit verschillen in het aantal relaties dat bij een woord is 'aangehecht' en de aard van deze relaties. Op basis van bovengenoemde punten kunnen we stellen dat het beeld zoals dat vaak geïmpliceerd is in woordenschattoetsing en woordenschatonderwijs sterk gesimplificeerd is. Met woordenschat gaat het niet alleen om het aantal woorden (d.i. de *breedte* van de woordenschat), maar ook om de hoeveelheid en kwaliteit van de kennis van woorden (d.i. *diepe* woordkennis).

1.2 Breedte en diepte van de woordenschat

Op school breiden de kinderen hun woordkennis in twee richtingen uit: in de breedte en in de diepte (Anderson & Freebody, 1981). Aan de ene kant leren kinderen steeds meer nieuwe woorden bij en neemt de omvang van de woordenschat toe (breedte). Aan de andere kant worden kinderen in het onderwijs geconfronteerd met nieuwe betekenissystemen en betekenisrelaties waarmee de kennis van reeds bekende woorden wordt verrijkt (diepte). Durking, Crowther en Shire (1986) karakteriseren de ontwikkeling van de woordenschat in de schooljaren als volgt:

> *(...) the processes of vocabulary development during this period are clearly more than additive, and involve both the enrichment of the children's knowledge of the organisational structures relating items in the lexicon, and refinement of knowledge of the meanings of individual words* (o.c.: 77).

Nieuwe betekenisstructuren worden in het onderwijs geleerd door een steeds verdergaand proces van generaliseren, categoriseren en abstraheren.

Een kind kent van het woord 'vogel' bijvoorbeeld eerst concrete kenmerken op grond van eerste ervaringen: de vogels in de tuin, op de TV of in het plaatjesboek. Later wordt, vooral op school, de betekenistoekenning uitgebreid en meer gedecontextualiseerd. Kinderen leren abstraheren van toevallige kenmerken ('alle vogels hebben veren, maar niet alle vogels zijn klein') en bouwen kennis op over de algemene categorie VOGEL: de vogel bouwt nesten, legt eieren enz. Allerlei verschillende aspecten worden aan het begrip of concept VOGEL toegevoegd. Het woord 'vogel' krijgt een uitgebreidere betekenislading, terwijl steeds meer relaties met andere woorden wor-

den begrepen. Kinderen leren op school echter niet alleen meer, maar ook andersoortige betekenisrelaties: naast syntagmatische betekenisrelaties (zoals vogel-vliegen, vogel-nest, vogel-ei) ontwikkelt zich gaandeweg het begrip van paradigmatische betekenisrelaties (vogel-dier, vogel-mus). In tegenstelling tot syntagmatische relaties zijn paradigmatische betekenisrelaties hiërarchisch. De hiërarchische classificatie kenmerkt zich door klasse-inclusie (een zwaluw is een trekvogel, is een vogel, is een dier) en veel schoolse kennis is op deze manier 'logisch' geordend en voor te stellen als een boomstructuur. Paradigmatische relaties zijn voor het onderwijs essentieel omdat ze door coördinatie, subordinatie en superordinatie generalisaties mogelijk maken: de kenmerken van een hoger liggende categorie (bijv. VOGEL) gelden voor alle daaraan ondergeschikte categorieën (TREKVOGEL, ZWALUW etc.). Met andere woorden: de betekenisaspecten van het woord 'vogel' zijn per definitie overdraagbaar op de woorden 'trekvogel' en 'zwaluw'. Bij de opbouw van het semantisch netwerk en de verdieping van woordkennis staan paradigmatische relaties centraal: zowel generalisatie als abstractie/categorisatie wordt binnen bereik van het kind gebracht (vgl. Cruse, 1986; Kuczaj, 1982; Vygotsky, 1962).

Kwaliteit van woordkennis en het kunnen hanteren van woordrelaties is belangrijk voor de schoolprestaties. In formele onderwijssituaties is het niet voldoende als woorden oppervlakkig en in beperkte contexten begrepen of geleerd worden. De vraag 'hoe goed kent een kind de woorden?' zou meer aandacht moeten krijgen. Het gaat bij deze vraag om achterliggende aspecten van woordkennis die niet altijd direct waarneembaar zijn, omdat de gangbare toetsen en woordenschatlessen appelleren aan minder vérgaande woordkennis. Het overige (zaakvak)onderwijs verwacht daarentegen vaak wel uitgebreidere woordkennis, die niet alle kinderen zullen bezitten. Zo kunnen verschillen in woordkennis tussen kinderen makkelijk over het hoofd gezien worden en toch de bron zijn van taal- en onderwijsproblemen of -achterstanden. Het volgende voorbeeld is illustratief (De Haan, 1988). De Haan vroeg kinderen de betekenis van het woord 'industrie' te omschrijven nadat ze een tekst uit een aardrijkskundeboek gelezen hadden waarin dit woord centraal stond. Ze noteerde de volgende antwoorden: 'waar planten groeien', 'een gebied waar je iets kunt doen', 'met andere landen importeren en exporteren' of 'dat is eh, landbouw, eh, nou alles bij elkaar eigenlijk we, wat je kunt bedenken, landbouw, visserij'. Leerkrachten zullen in dergelijke gevallen bij vervolglessen ernstig rekening moeten houden met grote verschillen in betekenistoekenning aan het woord 'industrie'.

Recent onderzoek van Verhallen (1994) naar betekenistoekenning aan eenvoudige woorden door kinderen heeft uitgewezen dat er op het gebied van diepe woordkennis belangrijke verschillen tussen kinderen zijn. In de studie werd aan negen- en elfjarige Turkse en Nederlandse kinderen een 'interview' afgenomen waarin de kinderen gestimuleerd werden om al hun kennis van de verschillende betekenisaspecten van een woord te expliciteren. De bevraagde woorden konden als bekend voor de kinderen verondersteld worden: ze zouden bij oppervlakkige beschouwing als gekend gelden. Bij een vergelijking van de betekenisomschrijvingen bleken er op twee dimensies systematische verschillen tussen de Nederlandse en Turkse kinderen te zijn: a) Turkse kinderen kennen aan Nederlandse woorden die ze verworven lijken te hebben, *minder* verschillende betekenisaspecten toe dan hun Nederlandse leeftijdsgenoten, en b) de betekenisaspecten die de Turkse kinderen noemen, zijn in vergelijking met die van de Nederlandse kinderen relatief *minder paradigmatisch* van aard. Deze kwalitatieve verschillen in woordkennis tussen de Turkse en Nederlandse kinderen bleken vergelijkbaar met die tussen negen- en elfjarigen, zowel binnen de Nederlandse als de Turkse groep. Waar oppervlakkig bezien geen verschillen zijn (de kinderen 'kenden' alle woorden), komen bij doorvragen wel degelijk belangrijke verschillen aan het licht.

Voor het onderwijs is het belangrijk dat leerkrachten goed inzicht krijgen in de aard en uitgebreidheid van de woordkennis van hun leerlingen. Voor onderzoek is het evenzeer van belang een genuanceerd beeld te hebben van de 'woordenschat' van proefpersonen om theoretisch veronderstelde relaties tussen woorden-

schat en andere (taal)variabelen te kunnen exploreren of toetsen. De vraag die nu rijst, is: 'hoe operationaliseer je diepe woordkennis?'.

2 De operationalisatie van diepe woordkennis

Er zijn ten behoeve van onderzoek en onderwijs verschillende standaardinstrumenten ontwikkeld waarmee de breedte of omvang van de woordenschat kan worden onderzocht, b.v. de Cito-woordenschattoetsen (Cito, 1992) en de actieve en passieve woordenschattoetsen van de Taaltoets Allochtone Kinderen (TAK) (Verhoeven, Vermeer & Van der Guchte, 1986). Het ontbreekt nog aan vergelijkbare instrumenten waarmee diepe woordkennis op eenvoudige wijze kan worden gepeild. Kwaliteit van woordkennis komt aan de orde in een definitietaak, zoals b.v. in de TAK-bovenbouw (Verhoeven & Vermeer, 1993). De definitietaak is een relatief complexe taak waarbij (formele) definitievaardigheid én (diepe) woordkennis samen in het geding zijn.

Het gebruik van de interviewmethode, zoals in Verhallen (1994), levert weliswaar inzicht in diepe woordkennis op, maar is zeer arbeidsintensief en laat geen of weinig experimentele manipulatie toe: de onderzoeker is afhankelijk van wat in de proefpersoon opkomt tijdens het interview (de afname) en hetgeen de proefpersoon kan en wil verwoorden. Voor onderwijs en onderzoek is het wenselijk om over eenvoudige, betrouwbare en valide operationalisaties van diepe woordkennis te beschikken.

Wesche en Paribakht (1996) bespreken in een overzichtsartikel verschillende methoden om woordkennis te toetsen met voor- en nadelen van verschillende procedures. Het gaat hierbij steeds om een afweging tussen eenvoud van de procedure, zodat voldoende woorden aan bod kunnen komen, en diepgang (validiteit) van de meting. Ook voor de operationalisatie van diepe woordkennis is gezocht naar een dergelijke balans. In de praktijk komt het er op neer dat de toets behalve klassikaal afneembaar, bij voorkeur ook objectief – zonder tussenkomst van een beoordelaar – scoorbaar moet zijn. Een eenvoudige toetsvorm en objectieve scoring komen de betrouwbaarheid van een toets ten goede. Eenvoud en betrouwbaarheid staan daarentegen vaak op gespannen voet met validiteit (cf. Wesdorp, 1981). We zullen dus een evenwicht moeten vinden tussen validiteit enerzijds en eenvoud en betrouwbaarheid anderzijds.

Read (1993) heeft een toetsvorm voor woordkennis ontwikkeld 'that would involve a simple response task and allow broad coverage of a set of words while, at the same time, probing depth of knowledge of words in some meaningful way' (p. 358). In Reads format worden woorden niet gepresenteerd als losse items, maar als onderdeel van een (mogelijk) semantisch netwerk. Een stimuluswoord is gecombineerd met acht andere woorden, waarvan vier gerelateerd zijn aan het stimuluswoord en vier andere niet. De taak voor de respondent is om de vier gerelateerde woorden te identificeren. Read (Read, 1993, p. 359) geeft als voorbeeld het stimuluswoord 'edit':

edit			
arithmetic	film	pole	publishing
revise	risk	surface	text

De respondent moet in bovenstaand voorbeeld niet alleen een (bijna) synoniem, 'revise', herkennen, maar ook twee woorden die vaak voorkomen met het stimuluswoord (collocaties), 'film' en 'text', en een woord dat een deel-geheelrelatie heeft met het stimuluswoord, namelijk 'publishing'.

De aanname bij deze toetsvorm is dat een taalleerder met diepere woordkennis beter in staat is om de geassocieerde woorden te identificeren (die verschillende aspecten van betekenis van het stimuluswoord representeren) dan degene wiens woordkennis 'minder diep' is. Op een eenvoudige manier kan van een betrekkelijk groot aantal stimuluswoorden de (diepere) woordkennis vastgesteld worden. Bovendien maakt deze toetsvorm door een goede keuze van de antwoordalternatieven (d.i. de acht keuzewoorden) het mogelijk om specifieke betekenisrelaties te manipuleren en te bevragen, zodat men meer inzicht kan krijgen in het soort kennis dat proefpersonen van een woord hebben en in de ontwikkeling van die kennis. Tenslotte doet de toetsvorm – hoewel die schriftelijk is – slechts een gering beroep op de (m.n. technische) leesvaardigheid, hetgeen

zowel voor onderzoek als voor onderwijs voordelen biedt.

2.1 De constructie van de WoordAssociatieTaak (WAT)

Uitgaande van het 'woordassociatie'-format van Read (1993) is de WoordAssociatieTaak (WAT) ontwikkeld voor kinderen in de bovenbouw van het basisonderwijs. In de WAT wordt van de kinderen verwacht dat ze bij een stimuluswoord drie uit zes associaties kiezen. Bij elk item komt het stimuluswoord in het midden te staan van zes meer of minder geassocieerde woorden. Het kind wordt gevraagd steeds drie verbindingslijnen te trekken (zie Figuur 1).

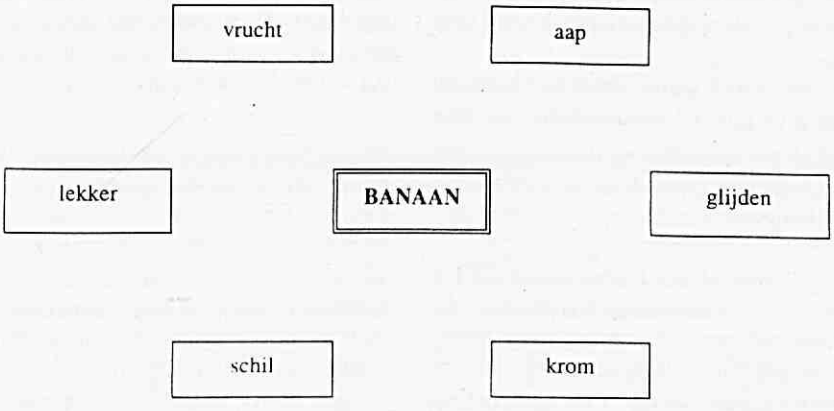
De beoogde goede antwoorden representeren verschillende soorten relaties die in semantisch netwerk verondersteld worden, zoals paradigmatische relaties (te weten superordinatie, subordinatie, synonymie), partonomische relaties (constituenten) en gedecontextualiseerde syntagmatische relaties (definiërende perceptuele kenmerken, inherente kenmerken en/of middel-doel relaties) (zie ook Verhallen, 1994). In Figuur 1 zijn de drie te selecteren woorden de superordinaat 'vrucht', de partonomische associatie 'schil' en het definiërende perceptuele kenmerk 'krom'.

Er is er voor gekozen om de verschillen tussen 'goede' en 'foute' antwoorden gradueel te maken. Bij de ontwikkeling van woordkennis speelt niet alleen categorisering, maar ook generalisatie en abstractie een belangrijke rol. Op grond van deze processen wordt betekenistoekenning geleidelijk meer gedecontextualiseerd. Om de diepte van woordkennis (met name de graad van decontextualisatie van de betekenistoekenning) genuanceerder te bepalen wordt kinderen gevraagd om onderscheid te maken tussen woorden die *altijd* bij het stimuluswoord horen en woorden die op een meer contextgebonden wijze met het stimuluswoord geassocieerd zijn. De taak is om bewust betekenisrelaties af te wegen en te selecteren. Met de keuze van de afleiders is gespeculeerd op contextuele of incidentele relaties tussen woorden die geen of weinig verband houden met de betekenis van het stimuluswoord. Dit betekent dat er geen absolute grens tussen goede of foute antwoorden is. Als een kind associatieve relaties (zoals in het voorbeeld BANAAN-aap) kiest, wordt dit wel opgevat als een indicatie voor de contextgebondenheid van de betekenistoekenning aan 'banaan' en het relatieve belang van dergelijke syntagmatische verbindingen in het lexicon.

Uitgangspunt bij het construeren van de items was dat alle woorden bekend verondersteld moesten kunnen worden bij negenjarige kinderen uit groep 5 van de basisschool (de jongste doelgroep), hoewel men daar uiteraard nooit zeker van kan zijn.

Op basis van verschillende streefwoordenlijsten en frequentielijsten (Coenen & Vermeer, 1988; Schrooten & Vermeer, 1994; Van Gelderen, 1994) zijn de stimuluswoorden geselecteerd. Voor de pilotversies zijn alle woorden gecontroleerd in de beheersingslijst van Coenen en Vermeer (1988); bij de bewerking van deze pilotversies tot de hier beproefde versies is soms van deze lijst afgeweken[1].

Alleen inhoudswoorden (zelfstandige naam-



Figuur 1 Een voorbeelditem uit de woordassociatietaak (WAT)

woorden, werkwoorden en bijvoeglijke naamwoorden) als de belangrijkste dragers van betekenis zijn in de toets opgenomen. Deze woorden kunnen verwijzen naar zowel concrete als abstracte zaken en er is gestreefd naar een spreiding over verschillende semantische domeinen.

Na een pilot met een proeftoets van 27 items bij 133 kinderen uit groep 5 en 7, en enkele hardop-denkprotocolanalyses (Bekius, 1995), zijn enkele items inhoudelijk bijgesteld en is de verzameling items uitgebreid tot 50. Bij de uitbreiding zijn met het oog op het verdere validatie-onderzoek 20 woorden uit de definitie-taak van de TAK-bovenbouw gebruikt als stimuluswoord. Deze woorden zijn ook in de vorm van de WAT omgezet door er steeds zes geassocieerde woorden aan toe te voegen.

Uit de aldus verkregen verzameling van items zijn aselect twee toetsversies van 30 items samengesteld; de twee versies hebben een overlap van 10 items. In het vervolg verwijzen we naar deze versies met WAT-A en WAT-B.

3 Onderzoeksopzet[2]

3.1 Proefpersonen

In het onderzoek richtten we ons op kinderen in de bovenbouw van het basisonderwijs. Met het oog op een analyse van leeftijdsverschillen is ervoor gekozen om leerlingen uit groep 5 en 7 te werven. Daarbij streefden we naar spreiding in te verwachten taalvaardigheidsniveau en naar voldoende deelname van anderstalige kinderen. Bij de keuze van de scholen is er sprake van een 'gelegenheidssteekproef' die in niet-statistische zin als willekeurig beschouwd mag worden.

In het onderzoek zijn in totaal 822 kinderen uit groep 5 en groep 7 van het basisonderwijs betrokken geweest. De kinderen zijn afkomstig van 19 verschillende scholen in Noord- en Zuid-Holland, Utrecht en Friesland. Enkele kinderen zijn om verschillende redenen (zie verderop) buiten beschouwing gelaten, zodat voor de analyse de gegevens van 795 kinderen beschikbaar waren. Een aantal kinderen uit de steekproef is benaderd voor een extra toetsafname met de eerdergenoemde definitietaak uit de TAK-bovenbouw (Verhoeven & Vermeer, 1993), zodat een vorm van soortgenootvaliditeit geëvalueerd kan worden (zie verderop). Op aanwijzing van de leerkracht zijn hiervoor een zwakke, twee gemiddelde en een goede leerling gevraagd. Informatie over de taalachtergrond van de kinderen is eveneens bij de leerkracht ingewonnen.

Tabel 1 geeft een overzicht van hun verdeling over taalachtergrond en jaargroep.

3.2 Instrumenten

Voor het onderzoek hebben de leerlingen een van de twee toetsversies (WAT-A of WAT-B) voorgelegd gekregen. Om te voorkomen dat door vermoeidheid of tijdgebrek systematisch de laatste items van beide versies niet beproefd zouden worden, is van elke versie een variant gemaakt met de items in de omgekeerde volgorde (*), zodat er feitelijk vier toetsvarianten waren: A, A*, B en B*.

Voor elk WAT-item is gescoord welke verbindingslijnen er door de leerling getrokken zijn. De toetsscore is het aantal items waarvoor de drie beoogde verbindingslijnen getrokken zijn.

Een kleine groep leerlingen werd de definitietaak uit de TAK-bovenbouw afgenomen (zie Verhoeven & Vermeer, 1993), waarbij de leerlingen van 25 woorden een definitie moeten geven, die vervolgens op een driepuntsschaal (0,1,2) beoordeeld wordt.

Tabel 1

Aantal deelnemende kinderen aan de WAT en de definitietaak (TAK), totaal en uitgesplitst naar taalachtergrond en jaargroep

leerjaar		Taalachtergrond							Totaal
		Nederlands	Marokkaans	Turks	Surinaams	Fries	overig	onbekend	
groep 5	WAT	231	59	25	33	22	40	1	411
	TAK	23	9	1	3	2	4	-	42
groep 7	WAT	199	50	31	31	31	42	-	384
	TAK	19	13	4	5	2	1	-	44

3.3 Procedure

De dataverzameling heeft plaatsgevonden in de maand februari 1996 en is uitgevoerd door proefleiders (studenten Algemene Taalwetenschap), die hiervoor over een uitgebreid protocol beschikten om de afnamen zoveel mogelijk te uniformeren. In een instructie van ongeveer 10 minuten werd aan de hand van twee voorbeelditems en via het principe van de geleide instructie de procedure met de kinderen doorgenomen. De kinderen hebben de WAT klassikaal gemaakt. De vier varianten zijn aselect aan kinderen toegewezen.

De voor de definitietaak geselecteerde kinderen hebben deze taak in een individuele sessie met de proefleider gemaakt volgens de instructie van de TAK. Ook de scoring van de definities is volgens de voorgeschreven procedure uitgevoerd.

3.4 Analyses

De betrouwbaarheid van een toets kan op verschillende manieren gedefinieerd worden, zoals test-hertestbetrouwbaarheid of interne consistentie. In ons onderzoek richten we ons op de interne consistentie van de toets als indicatie van de betrouwbaarheid. Daarnaast zal een één-factormodel voor de toets gepast worden. De scoring van de antwoorden van de leerlingen kan in principe 'mechanisch' uitgevoerd worden, zodat de scoringsbetrouwbaarheid hier geen punt van discussie is.

De validiteit van een toets is niet eenduidig vast te stellen. Validiteitsbepaling moet men zien als een continu proces waarin men evidentie verzamelt voor het idee dat men (alleen) meet wat men beoogt te meten (Messick, 1989; Shepard, 1993). Het begrip 'diepe woordkennis' is relatief nieuw en de validatie van operationalisaties van diepe woordkennis staat dan ook nog aan het begin van dat 'continue proces'. Om te beginnen gaan we na of de toets verschillen kan aantonen tussen groepen kinderen die naar bekend is ook daadwerkelijk verschillen in hun diepe woordkennis ('known-group validity', Kerlinger, 1973). Jongere kinderen hebben minder diepe woordkennis dan oudere kinderen, en Nederlandstalige kinderen hebben over het algemeen meer diepe woordkennis van het Nederlands dan anderstalige kinderen (zie Verhallen, 1994). Een valide toets moet in ieder geval deze verschillen kunnen laten zien.

Verder inzicht in de validiteit van de toets krijgen we door na te gaan of de toetsscores sterk samenhangen met scores op toetsen die hetzelfde of een sterk verwant construct meten ('soortgenootvaliditeit', De Groot, 1961). De nieuwe toets zou ten minste substantieel moeten correleren met andere toetsen die (aspecten van) diepe woordkennis meten: de definitietaak uit de TAK is zo'n andere toets.

De efficiëntie of bruikbaarheid van de toets wordt globaal geanalyseerd door na te gaan hoe de afname-omstandigheden waren, zoals afnametijd, aantal uitvallers, aantal ontbrekende of 'ongeldige' (onbruikbare) scores e.d.

4 Resultaten

Omdat de gang van zaken bij de afnamen van een nieuwe toets(vorm) onderdeel is van de evaluatie van de kwaliteit van de toets(vorm), namelijk als aspect van de utiliteit van de toets, zullen we deze paragraaf beginnen met enkele gegevens over de afnamen. We gaan eerst in op de uitval en 'missings' tijdens de dataverzameling. Vervolgens komen de beschrijvende statistieken aan de orde (gemiddelden, scoreverdelingen e.d.). Hieruit kunnen we afleiden of de moeilijkheidsgraad van de toets goed is en of de toets voldoende differentieert tussen leerlingen. Voor de verdere analyses is het handig als we kunnen abstraheren van de versies en volgorde-varianten van de toets die we gemaakt hebben. Op basis van de beschrijvende statistieken wordt nagegaan of de beproefde versies en varianten vergelijkbaar zijn. Als de toetsversies en -varianten niet verschillen, dan kan in volgende analyses desgewenst het onderscheid achterwege blijven. Vervolgens wordt de betrouwbaarheid (interne consistentie) van de toetsversies op twee manieren geschat. Ten slotte zullen de resultaten gepresenteerd worden die inzicht geven in de validiteit van de toets, aangenomen dat de betrouwbaarheid van voldoende niveau is om de validiteit verder te onderzoeken.

4.1 De bruikbaarheid van de toets

De toetsafname en de voorafgaande instructie gaven geen problemen voor de proefleiders. Ook als we de frequentie van 'ongeldige' scores bekijken, blijkt dat er weinig leerlingen zijn die kennelijk de bedoeling niet begrepen hebben of zich vergissen.

Tabel 2

Beschrijvende statistieken voor WAT-A en WAT-B (30 items): gemiddelde (M), standaarddeviatie (SD), laagst en hoogst behaalde score (min-max), scheefheid en gepiektheid van de verdeling (skew. en kurt.) en steekproefgrootte (n)

		M	SD	min-max	skew.	kurt.	n
Groep 5							
	WAT-A	15.47	5.67	3-27	-.01	-.74*	204
	WAT-B	15.22	5.49	3-28	.13	-.54	207
Groep 7							
	WAT-A	20.48	4.57	5-29	-.49*	.13	201
	WAT-B	21.22	5.00	5-30	-.86*	.51	183

* Skewness, respectievelijk kurtosis is groter dan 2 maal de bijbehorende standaardfout en wijkt daarmee significant af van nul.

Van de oorspronkelijke 822 leerlingen die elk 30 items maakten, zijn er slechts 15 kinderen die weleens *meer* dan de gevraagde drie strepen getrokken hebben. In dertien gevallen gaat het om een eenmalige 'vergissing', één kind begaat drie keer deze fout en één kind heeft kennelijk echt moeite om zich aan de instructie te houden: het streept bij veertien items meer dan drie alternatieven aan.

Dat kinderen *minder* dan drie alternatieven aanstrepen komt duidelijk frequenter voor. Dit is op zich begrijpelijk. Als een kind niet alle drie de beoogde associaties herkent, zal het misschien niet verder komen dan het aanstrepen van een of twee betekenisrelaties. Ruim 83% van de kinderen streept steeds drie alternatieven aan, slechts 1% van de kinderen streept bij meer dan vijf items minder dan drie alternatieven aan. Hierbij inbegrepen zijn kinderen die items overgeslagen hebben en dus geen enkele streep getrokken hebben en kinderen die onduidelijke (ambigue) strepen getrokken hebben.

De leerlingen die bij een stimuluswoord absoluut geen associaties hebben, hebben weinig kans om het item op basis van raden goed te scoren. Uit de zes alternatieven zijn twintig combinaties van drie (strepen) te maken. Bij een toets van 30 items is een verwachte score bij consequent raden 1.5, afgerond 2. Het blijkt dat zes kinderen (0.7%) een score van slechts 2 behalen. Het is niet duidelijk of dit wijst op een (volledig) gebrek aan diepe woordkennis, gebrek aan motivatie of onduidelijkheid over wat de bedoelde antwoorden zijn. De kinderen met een score op kansniveau (d.i. een totaalscore van maximaal 2 goed) zijn bij de overige analyses buiten beschouwing gelaten.

Samenvattend kan op basis van bovenstaande informatie gesteld worden dat het toetsformat in ieder geval geschikt is voor de beoogde leeftijdsgroep en zich leent voor klassikale afname binnen een redelijk tijdbestek (20 à 30 minuten, exclusief instructie), en in dat opzicht bruikbaar en 'intern efficiënt' is.

4.2 Moeilijkheidsgraad en differentiatie

Uit het databestand zijn de gegevens van de zes kinderen verwijderd die op kansniveau scoorden (0.7%). Bovendien zijn kinderen die meer dan 10% van de toets (d.w.z. meer dan drie items) gemist hebben buiten beschouwing gelaten; het gaat in totaal om nog eens 21 kinderen (2.6%). De uiteindelijke steekproefgrootte komt hiermee op 795: 405 voor WAT-A en 390 voor WAT-B.

Om inzicht te geven in de moeilijkheidsgraad van de toets en in de mate waarin hij individuele verschillen te zien geeft, worden in Tabel 2 enkele beschrijvende statistieken voor de beide versies[3] gerapporteerd.

Uit Tabel 2 kunnen we opmaken dat de kinderen in *groep 5* gemiddeld ongeveer 50% van de items goed scoren. Er is daarbij een behoorlijke spreiding in scores (zie SD en min-max). Deze spreiding wordt uiteraard in de hand gewerkt door de heterogene samenstelling van de steekproef. De scores benaderen redelijk een normaalverdeling. Uitgaande van de vergelijkbaarheid van de subgroepen die versie A of B gemaakt hebben vanwege de aselecte toewijzing (zie ook verderop), zijn de scores op beide versies ook goed vergelijkbaar.

Voor *groep 7* kan men constateren dat gemiddeld genomen de kinderen tweederde

van de toets goed maken. Ook hier is nog steeds een behoorlijke spreiding in scores, zij het dat nu meer kinderen het absolute maximum van 30 benaderen. De spreiding is daardoor iets kleiner dan in groep 5 en de verdelingen zijn iets scheef naar links. Overigens gaat het om kleine afwijkingen van normaliteit. Ook nu zijn de scoreverdelingen van beide versies goed vergelijkbaar.

Hoewel de verdeling van de versies (WAT-A en WAT-B) over de leerlingen aselect was, is nagegaan in hoeverre de betreffende groepen leerlingen vergelijkbaar zijn. Vergelijkbaarheid van deze groepen leerlingen is voorwaarde voor een zinvolle vergelijking van WAT-A en -B. De versies hadden samen een overlap van tien items waarop de leerlingen vergeleken kunnen worden. Noch in groep 5 noch in groep 7 is er een noemenswaardig verschil tussen de makers van versie A of B: de verschillen zijn niet significant en verklaren niet of nauwelijks variantie ($F(1,409)=2.28$, $p=.13$, $\eta^2=.01$ voor groep 5 en $F(1,382)=.68$, $p=.41$, $\eta^2=.00$ voor groep 7).

Samenvattend kunnen we stellen dat de toetsen qua moeilijkheid goed geschikt zijn voor groep 5 en (in iets mindere mate) groep 7. De verdelingen wijken niet sterk af van normaliteit en geven veel individuele verschillen te zien. Als de versies A en B niet veel in betrouwbaarheid verschillen, kan in het validatie-onderzoek van het onderscheid geabstraheerd worden, omdat we mogen aannemen dat de (willekeurig samengestelde) versies niet of nauwelijks verschillen.

4.3 Betrouwbaarheids- en itemanalyse

De betrouwbaarheid wordt geëvalueerd door de interne consistentie van de toetsversies vast te stellen. Daarbij kan nagegaan worden of er individuele items zijn die wellicht minder goed functioneren in die zin dat ze een lage of negatieve itemrestcorrelatie vertonen en iets anders lijken te meten dan de rest van de items. Daarnaast is een één-factormodel gepast volgens de instrumentele variabele-methode van Hägglund (Hägglund, 1982; Eiting, 1992). De passing wordt uitgedrukt in een 'adjusted goodness-of-fit'-index (agfi). De resultaten van de betrouwbaarheidsanalyse staan in Tabel 3.

Tabel 3

Betrouwbaarheidsschattingen voor WAT-A en WAT-B (Cronbachs α), de passing van een één-factormodel (agfi) en extreme en gemiddelde itemrestcorrelatie ($\leq r_{ir} \leq$, $\bar{r}_{ir}$). Aantal items is 30

		α	agfi	$\leq r_{ir} \leq$	$\bar{r}_{ir}$
Groep 5					
	WAT-A	.83	.97	.18-.53	.34
	WAT-B	.82	.97	.01-.52	.32
Groep 7					
	WAT-A	.75	.95	.09-.41	.26
	WAT-B	.80	.97	.04-.51	.31

Over het algemeen mag men de toetsen redelijk (intern) consistent noemen; alleen bij de A-versie in groep 7 valt de betrouwbaarheid onder de vaak gehanteerde norm van .80. De passing van een één-factormodel is voor beide versies en beide leeftijdsgroepen goed te noemen (agfi ≥.95). In geen van de gevallen levert een twee-factormodel een betere passing op.

De individuele items functioneren over het algemeen goed: alle itemrestcorrelaties zijn positief en gemiddeld genomen redelijk. Verwijdering van het slechtste item uit een toets levert geen of nauwelijks verbetering op van de interne consistentie van de betreffende toets als geheel. Daarom is voor de volgende analyses afgezien van itemselectie.

4.4 Validiteitsanalyse

Zoals in de vorige paragraaf al aangegeven is, is de validiteit van een toets niet zo eenvoudig of eenduidig in een index uit te drukken als de betrouwbaarheid. De hieronder te presenteren resultaten kunnen dan ook niet meer zijn dan eerste indicaties van de validiteit van de toets. Als vorm van 'known group' validiteit worden eerst de verschillende leeftijd- en taalgroepen met elkaar vergeleken. Vervolgens beschouwen we een vorm van soortgenootvaliditeit ('congruent validity').

Known-group validiteit. Bij de beschrijvende statistieken (Tabel 2) bleek al dat negen- en elfjarigen zich op de toetsen goed onderscheiden. Hoewel beide leeftijdsgroepen heterogeen zijn qua samenstelling in taalachtergrond, is ook het leeftijdsverschil nog duidelijk aantoonbaar.

In een tweewegsvariantie-analyse[4] (zie Tabel 4) blijkt dat het leeftijdseffect significant en in termen van proportie verklaarde variantie

Tabel 4

Gemiddelde per taalachtergrond en leerjaar met tussen haakjes de steekproefgrootte (links) en de bijbehorende uitkomsten van een tweewegsvariantie-analyse (rechts)

Taal	Leerjaar 5	Leerjaar 7		Variantie-analyse *MS*	*df*	*F*	*p*
Nederlands	17.3 (253)	22.1 (230)	Taal	3362	1	148	.000
anders	12.2 (157)	18.9 (154)	Leerjaar	6167	1	271	.000
			Interactie	166	1	7	.007
			Residu	23	790		

(η^2) groot is[5] ($F(1,790)=271$, $p=.000$, $\eta^2=.22$). Circa 22 procent van de variantie in de scores op de WAT kan op conto van leerjaar (groep 5 versus groep 7) geschreven worden.

Op niveau van de individuele items blijkt dat alle items door groep 7 leerlingen beter gemaakt worden dan door groep 5 leerlingen. Voor WAT-A geldt dat het bij 20 van de 30 items om een significant verschil gaat (χ^2-toets met Yates' correctie) met een gemiddelde effectgrootte (ϕ) voor de 30 van .18[6]. Voor WAT-B is dit aantal zelfs 24 (gemiddelde $\phi=.21$).

Als we een indeling van 'bekende groepen' maken naar taalachtergrond van de leerlingen, kunnen we op twee manieren te werk gaan, namelijk Nederlandstalig versus anderstalig, of een uitgebreidere indeling zoals in Tabel 1.

Bij de eerste indeling beschouwen we de Friezen als autochtone Nederlanders als Nederlandstalig (vgl. De Jong & Riemersma, 1994) en de Marokkaanse, Turkse, Surinaamse en 'overige' leerlingen als anderstalig. De toets moet bestaande verschillen in diepe woordkennis tussen Nederlands- en anderstaligen kunnen reproduceren in termen van duidelijke scoreverschillen tussen de twee groepen.

In de eerdergenoemde tweewegsvariantie-analyse (zie Tabel 4) met taalachtergrond en leerjaar als factoren blijkt er ook een significant effect van taalachtergrond te zijn ($F(1,790)=148$, $p=.000$, $\eta^2=.12$). Bovendien blijkt deze variabele (enigszins) te interageren met leeftijd (zie Tabel 4).

De (kleine) interactie wordt veroorzaakt doordat de anderstalige kinderen relatief iets meer vooruitgaan van leerjaar 5 naar leerjaar 7 dan de Nederlandstalige kinderen[7]; een interactie die Verhallen (1994) ook met de interviewtaak vond. Overigens is het interactie-effect zeer klein, zeker ook ten opzichte van de twee hoofdeffecten ($\eta^2=.006$ versus .12 en .22) en mogelijk mede veroorzaakt door de benadering van het 'plafond van de toets' door de Nederlandstalige kinderen in groep 7.

Ook bij de specifiekere indeling naar taalachtergrond: Nederlands, Fries, Surinaams, Marokkaans, Turks en 'overig' kunnen voor de taalgroepen voorspellingen geformuleerd worden wat betreft hun prestaties op een Nederlandse taaltoets, zoals de WAT. Men mag aannemen dat de Nederlandstalige kinderen de best ontwikkelde diepe woordkennis hebben (zie Verhallen, 1994; Verhallen & Schoonen, 1993). De Friese kinderen zullen naar verwachting niet of nauwelijks onderdoen voor de eentalige Nederlandse kinderen (vgl. De Jong & Riemersma, 1994). Vervolgens zijn er twee taalgroepen, namelijk Turkse en Marokkaanse kinderen, voor welke in peilingsonderzoek in het onderwijs bij herhaling (taal)achterstanden geconstateerd zijn. Onderling verschillen de beide groepen weinig; als er verschillen zijn dan zijn die meestal in het voordeel van de Marokkaanse kinderen (vgl. Verhoeven & Vermeer, 1989). De Surinaamse groep neemt vaak een middenpositie in tussen het niveau van de Nederlandstalige kinderen en de Turks/Marokkaanse kinderen (Verhoeven & Vermeer, 1989; Driessen, Jungbluth, Van Langen & Vierke, 1996). Het is een misvatting dat Surinaamse kinderen (als groep) Nederlandstalig zijn; hun problemen worden nogal eens onderschat (De Haan, 1994). De verwachte prestatievolgorde is dus: {NL/FR}, SU, {MA/TU}. De categorie 'overig' is heterogeen, omdat het om kinderen uit nieuwe immigratielanden gaat, maar ook om kinderen uit gemengde huwelijken, bijv. Nederlands/Turks, maar ook Nederlands/Engels. Naar verwachting zal deze groep ook een tussenpositie innemen en bovendien betrekkelijk heterogeen zijn (d.i. een grote

Tabel 5
Gemiddelde (M), standaarddeviatie (SD) en steekproefgrootte (n) per taalachtergrond en per leerjaar

		Nederlands	Fries	Surinaams	Marokkaans	Turks	overig
Leerjaar 5							
	M	17.35	16.77	14.55	11.03	9.60	13.48
	SD	*5.22*	*5.09*	*3.98*	*4.40*	*3.94*	*4.62*
	(*n*)	(231)	(22)	(33)	(59)	(25)	(40)
Leerjaar 7							
	M	22.10	22.39	20.06	18.00	18.52	19.33
	SD	*4.24*	*4.58*	*4.36*	*4.77*	*4.77*	*5.33*
	(*n*)	(199)	(31)	(31)	(50)	(31)	(42)

standaarddeviatie te zien geven). In Tabel 5 staan de gemiddelde prestaties van de onderscheiden taalgroepen, uitgesplitst naar leerjaar.

De verwachte prestatievolgorde vinden we terug in Tabel 5, zowel voor groep 5 als groep 7. De Nederlands- en Friestalige kinderen scoren het hoogst en onderscheiden zich nauwelijks van elkaar. De Marokkaanse en Turkse kinderen presteren het zwakst; in groep 5 is er nog wel een verschil tussen de beide groepen in het voordeel van de Marokkaanse kinderen. Deze tendens vindt men ook elders in de literatuur (Verhoeven & Vermeer, 1989), in groep 7 lijkt dat verschil verdwenen. De Surinaamse kinderen en de 'overigen' nemen zoals verwacht een tussenpositie in. Van een grotere heterogeniteit van de laatste groep ten opzichte van de andere groepen lijkt alleen sprake in groep 7, niet in groep 5. Overigens dient men de gerapporteerde statistieken (varianties en gemiddelden) niet al te absoluut te interpreteren, omdat er geen sprake is van een representatieve steekproef. Bovendien moet aangetekend worden dat de kinderen van de verschillende taalachtergronden niet evenredig verdeeld zijn over alle deelnemende scholen, zodat schooleffecten e.d. niet uitgesloten kunnen worden.

Soortgenootvaliditeit. Om een indicatie te krijgen van de soortgenootvaliditeit is, zoals in de *Analyse*-sectie beschreven, bij een subgroep van leerlingen (individueel) de definitietaak van de TAK afgenomen. Analyse van de correlatie van de WAT met de definitietaak kan verdere indicaties over de validiteit van de WAT geven.

In totaal zijn 91 kinderen getoetst. Op basis van de eerder beschreven uitval op de WAT beschikken we over de scores van 86 kinderen op de definitietaak én op de WAT. De 42 leerlingen uit groep 5 scoorden gemiddeld 18.19 (SD=10.01) van de 50 maximaal te behalen punten (25 items à twee punten); de 44 leerlingen uit groep 7 scoorden met 27.91 (SD=10.59) duidelijk hoger. Voor beide groepen was de interne consistentie van de toets goed: Cronbachs α= .89 respectievelijk .91.

De vraag is nu in hoeverre de definitietaak en de WAT vergelijkbare informatie opleveren. In Tabel 6 wordt de correlatie tussen de twee typen scores gerapporteerd.

Tabel 6
Correlatie (pmc) tussen de WAT en de definitietaak uit de TAK met bijbehorende steekproefgrootte (n); tussen haakjes de correlatie gecorrigeerd voor attenuatie

	leerjaar 5	leerjaar 7
pmc	.69 (.80)	.71 (.84)
n	42	44

Dat beide correlaties significant zijn is in dit verband niet zo interessant, belangrijker is dat ze substantieel zijn. Ongeveer 50% van de variantie in de prestaties op de definitietaak en op de WAT is gemeenschappelijk.

Omdat beide variabelen niet perfect betrouwbaar zijn, wordt de samenhang tussen beide enigszins 'afgezwakt' (attenuatie). Correctie voor attenuatie laat zien dat rekeninghoudend met de onbetrouwbaarheid in beide metingen de correlatie maximaal .80, respectievelijk .84 kan bedragen. In beide gevallen ligt r=1.0 niet in het 95%-betrouwbaarheidsinterval, waarmee aangenomen mag worden dat de variabelen niet volledig samenvallen. Beide toetsen vertonen dus een sterke 'ware' samenhang, maar meten niet exact hetzelfde.

5 Conclusie en discussie

Onze primaire doelstelling was de ontwikkeling en validering van een woordenschattoets die verder reikt dan het benoemen of aanwijzen van plaatjes. De beoogde toets zou moeten appelleren aan wat we noemen 'diepe woordkennis'. Dat diepe woordkennis van belang is voor het (taalgebruik in het) onderwijs en dat m.n. anderstalige kinderen daar op uitvallen hebben we elders betoogd en gedemonstreerd (Verhallen, 1994; Verhallen & Schoonen, 1993).

Hoewel we niet de pretentie willen hebben dat de toets niet voor verbetering vatbaar is, mogen we toch concluderen dat deze eerste (grootschalige) resultaten bemoedigend zijn. Een toets met 30 items blijkt voldoende intern consistent en de items mogen als één factor beschouwd worden. Uitgaande van een gemiddelde α van .80 voor de versies met 30 items, mag men een interne consistentie van .87 verwachten bij afname van 50 items (Spearman-Browns formule voor homogene toetsverlenging).

Met de toets kan men eveneens goed verschillende 'known' groepen onderscheiden, groepen waarvan men mag aannemen dat ze verschillen in diepe woordkennis. Overigens bleek hierbij een (zeer) kleine interactie tussen taalachtergrond en jaargroep; de achterstanden van de anderstaligen zouden iets kleiner zijn in groep 7 dan in groep 5. Dit lijkt in strijd met bevindingen van Verhoeven en Vermeer (1996). Zij vinden voor de leeswoordenschat een interactie-effect dat er op wijst dat de achterstand van mediterrane anderstalige kinderen groter wordt in de bovenbouw van het basisonderwijs. Hoewel de Nederlandstalige kinderen in groep 7 relatief hoog scoren op de WAT, kan een plafondeffect niet de (volledige) verklaring zijn; de genoemde kinderen scoren gemiddeld 22 van de maximaal 30. Het feit dat we Surinaamse en Arubaanse/Antilliaanse kinderen bij de anderstalige kinderen gerekend hebben, waar Verhoeven en Vermeer uitsluitend voor de mediterrane kinderen een interactie-effect vinden, kan evenmin de (volledige) verklaring zijn. Als men namelijk de scores van de Turkse en Marokkaanse kinderen apart bekijkt (Tabel 5), ziet men nog steeds dat de achterstand in groep 7 kleiner is dan in groep 5. Kennelijk zijn de ontwikkelingspatronen voor diepe woordkennis en de leeswoordenschat niet gelijk, want Verhallen (Verhallen, 1994; Verhallen & Schoonen, 1993) vindt met een andere operationalisatie van diepe woordkennis eenzelfde interactie als hier gerapporteerd. Het kan zijn dat anderstalige kinderen op iets oudere leeftijd pas de zgn. 'paradigmatic shift' doormaken waarin het semantisch netwerk gereorganiseerd wordt. Als zij deze 'shift' eenmaal doorgemaakt hebben, kan de achterstand op hun Nederlandstalige leeftijdsgenoten in *diepe* woordkennis kleiner worden. Deze mogelijke verklaring voor de hier gerapporteerde interactie laat onverlet dat anderstalige kinderen waarschijnlijk een minder gevarieerd Nederlands taalaanbod genieten dan hun Nederlandstalige leeftijdsgenoten met als gevolg een steeds verder achterblijvende breedte van de woordenschat (d.i. de interactie zoals gerapporteerd door Verhoeven en Vermeer (1996)).

Ten slotte blijken de scores op de woordassociatietoets sterk samen te hangen met de scores op een definitietaak (r=.80, resp. .84, gecorrigeerd voor attenuatie). De substantiële correlatie tussen de WAT en de TAK-definitietaak is van belang omdat definitievaardigheid algemeen beschouwd wordt als een vorm van diepe woordkennis (Snow, Cancino, De Temple & Schley, 1991; Verhallen, 1994). Een kanttekening die evenwel bij de correlatieschatting gemaakt moet worden is dat de ongeveer de helft van de woorden van de betreffende versie van de WAT die de kinderen maakten, terugkeerden in de definitie-taak van de TAK. Deze overlap in woorden leidt uiteraard eveneens tot hogere correlatieschattingen dan wanneer men met twee duidelijk verschillende tests te maken heeft.[8]

De beide taken meten weliswaar in hoge mate dezelfde vaardigheid of kennis, maar zijn niet zonder meer inwisselbaar. Enerzijds pretenderen ze niet volledig hetzelfde te meten: de definitietaak heeft een produktief karakter waarbij ook de formulering een rol speelt en de WAT is receptief waarbij men relaties binnen het semantisch netwerk moet afwegen. Anderzijds zijn er praktische verschillen: de hier gepresenteerde WAT heeft de prettige eigenschap dat hij klassikaal en vlot af te nemen is en niet de beoordelingsproblemen van de definitietaak kent.

In onze scoring van de items hebben we ervoor gekozen om een antwoord pas 'goed' te rekenen als alle drie de beoogde relaties aangestreept worden door de leerling. Men zou kunnen overwegen om elk item op een vierpuntsschaal van 0 tot 3 te scoren, waarbij elke terecht aangestreepte relatie een punt oplevert, of om – zoals Read gedaan heeft – elke relatie als een item op te vatten. In het laatste geval neemt het aantal 'items' ogenschijnlijk toe, maar er ontstaat een ongewenste onderlinge afhankelijkheid van 'items' rond één stimuluswoord die allerlei ongewenste psychometrische consequenties heeft. Bovendien gaat de raadkans een grotere rol spelen. Dit laatste geldt ook bij de scoring volgens een vierpuntsschaal. Door te raden zal men niet zo snel een item volledig goed scoren, maar één of twee punten zijn toch al gauw in de wacht te slepen. Overigens blijkt het gebruik van een vierpuntsschaal niet veel extra informatie op te leveren. Een WAT-score zoals hier in de analyses gebruikt, correleert .96 (n=795) met een WAT-score gebaseerd op vierpuntsschalen per item.

Een andere manier van scoring ontstaat als men zich beperkt tot uitsluitend de paradigmatisch gerelateerde woorden (super- en subordinaten en synoniemen). Uiteraard heeft niet elk stimuluswoord evenveel van deze gerelateerde woorden om zich heen staan. Bij de meeste stimuluswoorden gaat het om een of twee (hiërarchisch) paradigmatisch gerelateerde woorden, maar bij enkele stimuluswoorden om drie of geen. Nu was de instructie zo dat er overal drie woorden verbonden moesten worden met het stimuluswoord. Daarmee speelt de kans dat men toevallig het paradigmatisch gerelateerde woord aanstreept weer een grotere rol. Als men alleen de paradigmatische relaties scoort, wordt de toets iets te makkelijk (plafondeffect) en de verschillen tussen bijvoorbeeld de zes taalgroepen (iets) kleiner. Bovendien correleert een score voor alleen hiërarchische paradigmatische relaties nog steeds sterk met de 'algemene' toetsscore (r=.83). Het lijkt dus te gaan om diepe woordkennis als geheel en niet om een specifiek type relaties uit het semantisch netwerk van de kinderen, zoals hiërarchisch paradigmatische relaties.

Gegeven deze operationalisatie van diepe woordkennis wordt het interessant de relatie tussen diepe woordkennis en andere taalvaardigheden nader te bestuderen. Hoe is diepe woordkennis gerelateerd aan 'oppervlakkige' woordkennis? En welke rol speelt diepe woordkennis in bijvoorbeeld begrijpend lezen? Gaat het bij de diepe woordkennis om de beschikbaarheid van deze kennis (na b.v. enig nadenken) of kan men (b.v. tijdens het lezen) alleen van diepe woordkennis profiteren als deze kennis min of meer automatisch geactiveerd wordt. Onderzoek naar antwoorden op deze vragen zal niet alleen het begrip diepe woordkennis nader preciseren in het geheel van taal- en andere cognitieve vaardigheden, maar het zal ook aanwijzingen kunnen geven voor wenselijk (woordenschat- en zaakvak)onderwijs. Niet alleen bij de taallessen, maar ook binnen het zaakvakonderwijs moet diepe woordkennis aandacht krijgen en moet aandacht besteed worden aan het decontextualiseren en verdiepen van woordkennis, zowel voor Nederlands- als anderstalige kinderen. Inzicht in het begrip 'diepe woordkennis' is van groot belang omdat achter de gelijke oppervlakkige woordkennis een wereld van verschillen schuil kan gaan.

Noten

1. Overigens is de bruikbaarheid van frequentielijsten betrekkelijk, omdat ze sterk afhankelijk zijn van de kwaliteit van het corpus waarop ze gebaseerd zijn, veelal geschreven taal van volwassenen. Zo is een woord als 'voedsel' met een frequentie van 66/mln veel frequenter dan het voor kinderen ongetwijfeld vertrouwdere 'wekker' of 'banaan', die beide een frequentie van minder dan 10/mln hebben. Het belang van een begrip voor een kind of de vertrouwdheid met het specifieke domein zijn evenzeer van invloed op de kennis van woorden.
2. Het empirisch gedeelte van het onderzoek is uitgevoerd in samenwerking met een groep studenten in kader van een module Tweede-Taalverwerving. Verslag van die module wordt gedaan in Birkhoff & Boeve (1996).
3. Er is geabstraheerd van volgorde-varianten; deze varianten hebben geen significant effect gehad op de prestaties op de toets. Per leerjaar is zowel voor WAT-A als voor de WAT-B nagegaan

of de toetsprestaties van de leerlingen verschilden afhankelijk van de volgorde-variant die men maakte (A vs. A* en B vs. B*). Volgorde-variant kan in die analyses niet meer dan 0.5% (groep 5) en 1.0% (groep 7) van de variantie verklaren.

4. Gezien de grote overeenkomst tussen de twee versies van de WAT is bij deze en volgende analyses geen onderscheid meer gemaakt naar versie (A of B).

   Achteraf zijn wij gewezen op de mogelijkheid – en misschien wel wenselijkheid – van het equivaleren van de toetsen (zie b.v. Engelen & Eggen, 1993), alvorens van één WAT-score uit te gaan. Bij equivalering zou gebruik gemaakt kunnen worden van het 'ankertoets-design' waarbij de overlap in de twee versies als anker kan dienen. De scores op het niet-overlappende deel van WAT-B zouden volgens de functie $e_A(B)$= .96*B* + .08 getransformeerd moeten worden om beter op dezelfde schaal van het niet-overlappende deel van A te komen. Uit de functie blijkt al dat het effect gering zal zijn en de totale geëquivaleerde toetsscore voor WAT-B correleert .9993 met de 'gewone' toetsscore en het gecorrigeerde gemiddelde is 0.2 lager dan het ongecorrigeerde. Correctie ten behoeve van equivalering is in dit artikel achterwege gebleven; de gemiddelden in Tabel 4 en 5 zouden maximaal .27 veranderen en de correlaties in Tabel 6 helemaal niet.
5. Volgens de vuistregels van Cohen (1988) zijn effecten vanaf η^2=.14 (d.i. f=.40) groot te noemen.
6. In termen van Cohen (1988) is .10 een klein en .30 een middelmatig verschil.
7. Het design is niet gebalanceerd in de zin dat het aantal observaties in alle cellen even groot is, hetgeen betekent dat de factoren gecorreleerd zijn. Als men door weging de grootte van de subgroepen even groot maakt, veranderen de uitkomsten nauwelijks.
8. In de totale groep van 86 kinderen die beide toetsonderdelen, WAT en TAK, gemaakt hebben, is de geobserveerde correlatie .75. Als men de scores van de WAT uitsplitst naar TAK- en niet-TAK-woorden dan bedraagt de correlatie met de definitietaak .79 respectievelijk .62, waarbij aangetekend moet worden dat de deelscores (voor TAK- en niet-TAK-woorden in de WAT) uiteraard onbetrouwbaarder zijn dan de WAT-totaalscore.

Literatuur

Aitchison, J. (1994). *Words in the mind. An introduction to the mental lexicon* (2nd ed.). Oxford, UK & Cambridge, USA: Blackwell Publishers.

Anderson, R.C., & Freebody, P. (1981). Vocabulary knowledge. In J. Guthrie (Ed.), *Comprehension and teaching: research reviews* (pp. 77-117). Newark, DE: International Reading Association.

Appel, R., & Vermeer, A. (1996). Uitbreiding van de Nederlandse woordenschat van allochtone leerlingen in het basisonderwijs. *Pedagogische Studiën, 73*, 82-92.

Barr, R., Kamil, M.L., Mosenthal P., & Pearson, P.D. (Eds.) (1991). *Handbook of reading research* (Volume II). New York: Longman.

Bekius, A. (1995). *De operationalisatie van diepe woordkennis. Tussentijds onderzoeksverslag.* Amsterdam: vakgroep ATW, UvA (Intern rapport).

Birkhoff, M., & Boeve, L. (Red.) (1996). *Operationalisatie van woordkennis. Onderzoeksverslag in het kader van de vervolgmodule Tweede-Taalverwerving II.* Amsterdam: Instituut voor Algemene Taalwetenschap/Universiteit van Amsterdam (Intern rapport).

Cito (1992). *Woordenschattoets 1 (groep 3/4).* Arnhem: Cito.

Coady, J. (1995). Research on ESL/EFL vocabulary acquisition: putting it in context. In Th. Huckin, M. Haynes & J. Coady (Eds.), *Second language reading and vocabulary learning* (pp. 3-23). Norwood, NJ: Ablex Publishing Corporation.

Coenen, M., & Vermeer, A. (1988). *Nederlandse woordenschat allochtone kinderen.* Tilburg: Zwijsen.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cruse, D.A. (1986). *Lexical semantics.* Cambridge: Cambridge University Press.

Driessen, G., Jungbluth, P., Langen, A. van, & Vierke, H. (1996). *PRIMA-cohortonderzoek. Technische rapportage ITS-deel.* Nijmegen: ITS.

Durkin K., Crowther, R.D., & Schire, B. (1986). Children's processing of polysemous vocabulary in school. In K. Durkin (Ed.), *Language development in the school years* (pp. 172-202). London: Croom Helm.

Engelen, R.J.H., & Eggen, T.J.H.M. (1993). Equivaleren. In T.J.H.M. Eggen & P.F. Sanders (Red.), *Psychometrie in de praktijk* (pp. 309-348). Arnhem: Cito Instituut voor Toetsontwikkeling.

Eiting, M.H. (1992). *Reliability. A program for reliability analysis (update).* Amsterdam: SCO-Kohnstamm Instituut.

Gelderen, A. van (1994). *Taalvaardigheidseisen in het zaakvakonderwijs voor eentalige en meertalige kinderen: moeilijkheden in de instructietaal en leerdoelen voor het taalonderwijs op de basisschool.* Amsterdam: SCO Kohnstamm Instituut (rapport 362).

Groot, A.D. de (1961). *Methodologie. Grondslagen van onderzoek en denken in de gedragswetenschappen.* 's-Gravenhage: Mouton & Co (8e druk, 1975).

Haan, D. de (1988). Leren in je eigen taal. *Jeugd in School en Wereld,* nr. 3, 10-15.

Haan, D. de (1994). *Deep Dutch. Towards an operationalization of school language skills* (Academisch proefschrift). Amsterdam: Universiteit van Amsterdam/IFOTT.

Hägglund, G. (1982). Factor analysis by instrumental variables methods. *Psychometrika, 47,* 209-222.

Jong, S. de, & Riemersma, A.M.J. (1994). *Taalpeiling yn Fryslân. Onderzoek naar de beheersing van het Fries en het Nederlands aan het einde van de basisschool.* Ljouwert/Leeuwarden: Fryske Akademy, 1994, nr 780 (Academisch proefschrift Katholieke Universiteit Brabant).

Kerlinger, F.N. (1973). *Foundations of behavioral research* (2nd ed.). New York etc.: Holt, Rinehart and Winston, Inc.

Kuczaj, S.A. (1982). Acquisition of word meaning in the context of the development of the semantic system. In C.J. Brainerd & M. Pressley (Eds.), *Verbal processes in children* (pp. 95-123). New York: Springer Verlag.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (pp. 13-103) (3rd ed.). New York: American Council on Education/MacMillan Publishing Co.

Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing, 10,* 355-371.

Schoonen, R., & Wolf, T. (1985). Lezen: vaardigheid en proces. Een empirisch onderzoek naar lezen als taalvaardigheid en als cognitief proces. In W.K.B. Koning (Red.), *Taalbeheersing in theorie en praktijk* (pp. 326-334). Dordrecht: Foris.

Schrooten, W., & Vermeer, A. (1994). *Woorden in het basisonderwijs. 15000 woorden aangeboden aan leerlingen.* Tilburg: Tilburg University Press (Studies in meertaligheid 6).

Shepard, L.A. (1993). Evaluating test validity. In L. Darling-Hammond (Ed.), *Review of research in education* (Vol. 19) (pp. 405-450). Washington: AERA.

Singer, H., & Ruddell, R.B. (Eds.) (1985). *Theoretical models and processes of reading* (3rd ed.). Newark, DE: International Reading Association.

Snow, C.E., Cancino, H., De Temple, J., & Schley, S. (1991). Giving formal definitions: A linguistic or metalinguistic skill. In E. Bialystok (Ed.), *Language processing in bilingual children* (pp. 90-112). Cambridge: Cambridge University Press.

Verhallen, M. (1994). *Lexicale vaardigheid van Turkse en Nederlandse kinderen. Een vergelijkend onderzoek naar betekenistoekenning* (Academisch proefschrift). Amsterdam: Universiteit van Amsterdam/IFOTT.

Verhallen, M., & Schoonen, R. (1993). Word definitions of monolingual and bilingual children. *Applied Linguistics, 14,* 344-365.

Verhallen, M., & Verhallen, S. (1994). *Woorden leren, woorden onderwijzen. Handreiking voor leraren in het basis- en voortgezet onderwijs.* Hoevelaken: CPS.

Verhoeven, L., & Vermeer, A. (1989). *Diagnose van kindertaal. Nederlandse taalvaardigheid van autochtone en allochtone kinderen.* Tilburg: Zwijsen.

Verhoeven, L., & Vermeer, A. (1993). *Taaltoets allochtone kinderen: bovenbouw. Diagnostische toetsen voor de vaardigheid Nederlands bij allochtone en autochtone kinderen in de bovenbouw van het basisonderwijs.* Tilburg: Zwijsen.

Verhoeven, L., & Vermeer, A. (1996). *Taalvaardigheid in de bovenbouw. Nederlands van autochtone en allochtone leerlingen in het basis- en mlkonderwijs.* Tilburg: Tilburg University Press.

Verhoeven, L., Vermeer, A., & Guchte, C. van de (1986). *Taaltoets allochtone kinderen; toetspakket.* Tilburg: Zwijsen.

Vygotsky, L. (1962). *Thought and Language.* Cambridge, MA: MIT Press.

Wesche, M., & Paribakht, T.S. (1996). Assessing second language vocabulary knowledge: Depth versus breadth. *Canadian Modern Language Review, 53,* 13-40.

Wesdorp, H. (1981). *Evaluatietechnieken voor het moedertaalonderwijs. Een inventarisatie van beoordelingsmethoden voor de stelvaardigheid,*

het begrijpend lezen, de spreek-, luister- en discussievaardigheid. 's-Gravenhage: SVO/Staatsuitgeverij.

Manuscript aanvaard 29-4-1998

Auteurs

R. Schoonen: Leerstoelgroep tweede-taalverwerving, Faculteit Geesteswetenschappen, en SCO-Kohnstamm Instituut, Faculteit Pedagogische en Onderwijskundige Wetenschappen van de Universiteit van Amsterdam

M. Verhallen: Leerstoelgroep tweede-taalverwerving, Faculteit Geesteswetenschappen van Universiteit van Amsterdam en Expertisecentrum Nederlands als tweede taal van de Hogeschool Haarlem

Correspondentie-adres: R. Schoonen, Leerstoelgroep Tweede-taalverwerving, Faculteit Geesteswetenschappen, Universiteit van Amsterdam, Spuistraat 210, 1012 VT Amsterdam. e-mail: rob.schoonen@hum.uva.nl

Abstract

Knowledge of words: Testing 'deep lexical knowledge'.

R. Schoonen & M. Verhallen. *Pedagogische Studiën*, 1998, *75*, 153-168.

Vocabulary knowledge implies more than simply naming stimulus pictures correctly or matching pictures to an auditory stimulus as most vocabulary tests require. This article discusses the concept of 'deep lexical knowledge' and its importance in the development of children's language, particulary in the context of education. A special testing format of structured word associations for measuring deep lexical knowledge is presented. The reliability and validity of this new format is tested empirically in a study of grade 3 and 5 primary students ('groep 5' and 'groep 7'). Both the reliability and the known-group and concurrent validity turn out to be satisfactory.