Ludo Beheydt

TAALVAARDIGHEID EN TAALMATEN

1. Kritiek op kwantitatieve taalmaten

De moderne toegepaste taalkunde schijnt de idee te voeden dat men op basis van kwantitatieve gegevens over taalgebruik geen conclusies kan trekken met betrekking tot de taalvaardigheid (Appel e.a. 1976, p. 155). De argumenten daarbij zijn dan dat syntactische complexiteitsscores onbevredigend zijn (Appel 1972), dat kwantitatieve maten als gemiddelde zinslengte nietszeggend zijn omdat een vaardige taalgebruiker toch ook alles in korte zinnen gezegd kan krijgen, dat vocabulariumdiversiteit te mechanistisch berekend wordt (Hoar 1981), dat het tekstaspect bijna altijd buiten schot blijft (Hoar 1981) en tenslotte dat geen enkele enkelvoudige taalmaat een betrouwbare index van taalvaardigheid geeft (Beheydt 1981, p. 72).

Deze interne kritiek van de toegepast taalkundigen wordt dan nog aangevuld door externe kritiek uit de hoek van de sociolinguïstiek. Sinds Hymes' theoretisch manifest over "Communicative Competence" (1971) wordt vrij algemeen aangenomen dat men taalvaardigheid niet louter op het taalgebruik kan beoordelen. Taalvaardigheid immers is niet alleen een kwestie van taalgebruik, maar evenzeer een kwestie van socio-culturele en situationele factoren. Een niet-verbaal knikje kan in bepaalde omstandigheden communicatief meer aangewezen zijn dan een volmondig ja, en een glimlach met een knipoog communicatief efficiënter dan een verbale toenaderingspoging. En als antwoord op de vraag of in België nou echt iedereen friet eet, kan een terloops-ironisch "ja!" (met de juiste intonatie!) een grotere taalvaardigheid verraden dan een rationeel-referentieel antwoord als "die dooddoener wordt langzamerhand zo goedkoop dat alleen Nederlanders hem nog gebruiken". Taalvaardigheid meten aan het taalgebruik komt in die sociolinguïstische visie neer op het ontkennen van de constitutieve rol die aan socio-culturele en situationele factoren onmiskenbaar moet worden toegekend bij de beoordeling van taalvaardigheid. Taalvaardigheid is immers ook de vaardigheid te zwijgen als dat hoort, en met sociale clichés te dialogeren waar dat gepast is.

Externe kritiek op de kwantitatieve benadering van taalvaardigheid is er ook gekomen van de aanhangers van de speech act theorie. Sinds Searle's "Taalhandelingen" is ook ten onzent de overtuiging gegroeid dat taalvaardigheid niet kan worden beschreven in termen van grammatica en vocabularium, maar dat ze moet worden beschreven in termen van "communicatieve effectiviteit van de taalhandeling". Taal wordt primair gebruikt om te overtuigen, te vragen, te bevelen, te prijzen, te smeken, te schelden, te juichen of te klagen. In de mate waarin een taalgebruiker die communicatieve intenties effectief weet te uiten kan men zijn taalvaardigheid groot of klein noemen en daar hebben taalmaten in verband met vocabulariumdiversiteit of syntactische complexiteit geen vat op.

Dit is grosso modo de kritiek die een verdediger van de kwantitatieve meettechnieken voor taalvaardigheidsonderzoek te verwerken krijgt. Wat kan daarop geantwoord worden?

2. Repliek en alternatieven

Vooreerst moet gezegd worden dat de sociolinguïstische kritiek grotendeels gefundeerd is, tenminste als men het heeft over de zogenaamde "algemene taalvaardigheid". Toch is daarmee het kwantitatieve taalgebruiksonderzoek niet afgeschreven. Het feit dat men niet de "algemene taalvaardigheid" kan meten, sluit niet uit dat men welomschreven hypothesen met betrekking tot taalvaardigheid kan meten (vgl. Beheydt 1981, p. 70). Men kan bijvoorbeeld rekening houdende met de sociale en culturele factoren tot de hypothese komen dat een kind in een bepaalde taalgemeenschap slechts schoolrijp is als het een welomschreven taalvaardigheidsniveau bereikt heeft. Op grond van een aantal uit de hypothese afgeleide taalmaten (b.v. vocabulariumtoetsen e.d.) kan men dan de graad van taalvaardigheid van elk individueel kind binnen die taalgemeenschap nagaan en beoordelen. Dat is overigens wat vrij efficiënt gebeurt in allerhande schoolrijpheidstests.

De vraag die dan natuurlijk rijst is: hoe komen we aan die initiële hypothesen. Zoals ik elders al omstandiger uiteengezet heb (Beheydt 1981, p. 71) is het uiteindelijk het gezond verstand dat op grond van empirische evidentie die hypothesen moet genereren. Als ik wil weten wat het noodzakelijke taalvaardigheidsniveau is om de basisschool aan te kunnen, dan kan ik mijn hypothese het best baseren op een corpus reëel taalgebruik van het soort waarmee de aanstaande eersteklasser zal worden geconfronteerd en dat door de gemiddelde eersteklasser wordt geproduceerd. Dat wil in concreto zeggen dat ik maar eens het taalgebruik van de onderwijzer moet gaan analyseren evenals dat van de gebruikte leerboekjes en op grond daarvan een omschrijving van taalvaardigheidsvereisten moet maken. In methodologisch jargon heet het dan dat we van *observatie* naar *inductie* gaan (De Groot 1968) maar deze jargonisering vermag niet te verhelen dat het afleiden van hypothesen een intuïtieve aangelegenheid is waarin de creatieve fantasie of – als u wilt – het natte-vingerwerk belangrijker is dan enigerlei methodologische logica. Ik ben het in dit opzicht volkomen eens met Van den Toorn waar hij stelt:

> "Er is namelijk geen heuristiek, geen methode tot het vinden van een vruchtbare hypothese; er bestaan geen regels om iets te ontdekken of uit te vinden. Zulke regels bestaan evenmin voor het vinden van een muzikaal thema of een lyrisch gedicht. Men kan dagenlang vruchteloos piekeren en onverwacht voor het inslapen of tijdens het scheren, komt een inval die de oplossing brengt. We zouden hier met een groot woord van inspiratie kunnen spreken, en in dit opzicht kan wetenschappelijke bedrijvigheid vergeleken worden met artistieke werkzaamheid" (Van den Toorn 1978, p. 102).

Betekent dit nu dat we maar meteen de hele methodologie naar de schroothoop moeten verwijzen? Helemaal niet. De methodologie kan bijvoorbeeld wel belangrijke aanwijzingen geven met betrekking tot de wijze waarop intuïtieve hypothesen tot een theorie, d.w.z. een systeem van logisch samenhangende, niet strijdige hypothesen betreffende een bepaald studieobject, kunnen worden uitgewerkt. De methodologie kan ook de voorwaarden aangeven waaraan een geldige operationalisering van de theorie moet voldoen (vgl. Beheydt 1981, p. 71). In dit opzicht is de methodologie niet alleen nuttig maar zelfs noodzakelijk, als een controle op de wetenschappelijke activiteit.

Er is dus ondanks de kritiek van de sociolinguïstiek toch wel plaats voor onderzoek met behulp van operationele taalmaten. De reactie op de kritiek van de taalhandelingstheorie gaat overigens ongeveer in dezelfde zin. Het bestaan en de eventuele primauteit van taalhandelingen doet niets af van de mogelijkheid om toetsbare hypothesen m.b.t. taalgebruik te formuleren. Een voorbeeld moge dit illustreren. Interactieve studies van de taalhandelingen op school hebben recent nogal wat aan het licht gebracht over het sociolinguïstische register van de schooltaal, met het inzicht in de discoursestructuur en met de hele achtergrond van machtsrelaties, attitudes, sociale invloeden enz. (zie: Spoelders & Van Besien 1979). Maar al deze bevindingen doen niets af van de mogelijkheid als hypothese te formuleren dat bijvoorbeeld het schoolfalen van kinderen uit de lagere sociale klassen mede kan worden toegeschreven aan de kloof die er bestaat of beter de fundamentele incompatibiliteit tussen de socialiseringstaal van de lagere klassen en de schooltaal, zoals Bernstein heeft gesteld.

Men kan inderdaad deze vage en intuïtieve hypothese herformuleren in een theorie bestaande uit een reeks scherp geëxpliciteerde deelhypothesen waarvoor dan een aantal operationele taalmaten kunnen ontworpen worden die de theorie moeten verifiëren of falsifiëren.

3. Hypothesen en taalmaten

Exemplarisch zullen wij hierna aangeven hoe door de afleiding uit de grondhypothese van een aantal taalmaten een analyserooster kan worden gecreëerd dat krachtens de grondhypothese moet bestaan uit (positief of negatief) correlerende taalmaten.

De grondhypothese bij Bernstein is dat de socialiseringstaal in de lagere klassen semantisch impliciet is en dat de socialiseringstaal in de hogere klassen semantisch expliciet is, zoals overigens ook de taal op school. Het verschil in semantische explicitering heeft natuurlijk gevolgen op het *niveau van het taalgebruik*. Er is een causale relatie tussen de lexicale en grammaticale opties en de onderliggende semantische opties. Het expliciteren van de betekenis brengt een genuanceerdere syntactische selectie en een gedifferentieerder vocabularium met zich mee. Het verbaal impliciet laten van de betekenis manifesteert zich dan in de neiging om eenvoudiger grammaticale selecties te maken, minder lexicaal te differentiëren en meer te betrouwen op situationele steun van niet-verbale communicatiemiddelen. Voor een ruimere steekproef taalgebruik mag men derhalve aannemen dat, als die semantisch expliciet is, ook de taalgebruikskenmerken van semantisch expliciet taalgebruik er in aanwezig zullen zijn, m.n. een grotere lexicale diversiteit, syntactische afwisseling, logisch-syntactisch complexere structuren, etc. In die zin kunnen die taalgebruikskenmerken dan weer als kwantitatieve indices gelden voor de kwaliteit van het taalgebruik. Met deze beperking nochtans dat één of zelfs enkele van die maten geen uitspraken toelaten over de expliciteringsgraad van het taalgebruik in het algemeen.

Immers het taalgebruikersrepertoire biedt verschillende mogelijkheden om semantisch te expliciteren. De taalmaten die men aanwendt moeten zowel grammaticaal als lexicaal zijn. Zowel grammatica als lexicon kunnen gebruikt worden om betekenissen verbaal expliciet te maken: "one cannot really separate vocabulary from grammar: the two form a single component in the linguistic system" (Halliday 1973, p. xi). Grammatica en lexicon staan m.a.w. in een complementaire verhouding en zijn tot op zekere hoogte verwisselbaar in het taalgebruik. Zinnen als:

(1) van wie is dat huis dat op de hoek staat?
(2) van wie is dat huis op de hoek?
(3) van wie is dat hoekhuis?

bereiken niettegenstaande hun verschillende lexicale en grammaticale selecties toch dezelfde graad van verbale

explicitering van intentionele betekenis. In zin (1) en (2) wordt de semantische explicitering vooral bekomen door grammaticale middelen (bijzin *resp.* nominale groep), in zin (3) daarentegen door de keuze van een semantisch complex lexicaal item.

Het spreekt dan ook vanzelf dat, als we de graad van semantische explicitering willen meten, we zowel grammatica als lexicon in de meting moeten betrekken. Aangezien ze allebei gebruikt kunnen worden om semantisch te expliciteren, is het meten van alleen één van beide misleidend. Het kan namelijk zeer wel het geval zijn dat de ene taalgebruiker meer een beroep doet op de lexicale alternatieven waar een andere meer de syntactische aanspreekt. Bovendien kan een kwantitatief verschil in het gebruik van een bepaalde variabele ook situationeel bepaald zijn of toevallig ontstaan zijn en aldus niet relevant voor het taalgebruik als geheel.

Daarmee hebben we zoals de attente lezer al zal hebben gemerkt een deel van de interne kritiek op de kwantitatieve maten beantwoord. We hebben namelijk laten verstaan dat de kritiek gelijk heeft als hij stelt dat geen enkele taalmaat een betrouwbare index van taalvaardigheid is op zichzelf, maar we hebben er meteen aan toegevoegd dat het werken met een analyserooster van taalmaten wel een oirbare praktijk is.

3.1. Syntactische taalmaten

Welke taalmaten komen nu in aanmerking voor het analyserooster? Er is onderhand in de literatuur een dusdanige diversiteit aan taalmaten voorhanden dat het voor de onderzoeker soms moeilijk kiezen wordt.

Voor de syntaxis bijvoorbeeld zijn er allerlei kwantitatieve en kwalitatieve taalmaten ontworpen die het moeten mogelijk maken taalgebruik op syntactische verschillen te evalueren.

Bij de kwantitatieve maten onderscheiden we vooreerst het type van de syntactische complexiteitsscores. Dit zijn taalmaten die op grond van bepaalde complexiteitscriteria een puntenwaarde toekennen aan zinnen: hoe hoger het aantal punten, hoe hoger de complexiteit. De complexiteitscriteria verschillen volgens de linguïstische theorieën waarop de scores geënt zijn. Een nauwkeuriger inspectie van alle ons bekende syntactische complexiteitsscores leidt ons tot de betreurenswaardige conclusie dat zulke scores onbetrouwbaar zijn of in elk geval prematuur (Beheydt, in voorber.). Zolang de linguïstiek geen bevredigende hiërarchie weet aan te brengen in de complexiteit van zinnen, is het op zijn minst misleidend numerieke waarden toe te kennen aan structuren en vormen en op basis daarvan vergelijkingen te maken en conclusies te trekken.

Als alternatief voor die complexiteitsscores zien wij twee mogelijkheden. De eerste bestaat hierin dat men syntactische maten gebruikt die niet afhankelijk zijn van een door de scorer toegekend puntensysteem, maar die vanuit het materiaal zelf een score opleveren. Wij denken hierbij vooral aan kwantitatieve maten als gemiddelde zinslengte, standaarddeviatie van de gemiddelde zinslengte, enz. De tweede mogelijkheid die wij in het huidige stadium zien is het gebruik van een beschrijvende syntactische analyse. Over die beschrijvende syntactische analyse zullen wij hier niet verder uitweiden. Wij volstaan met een verwijzing naar zo'n beschrijvende syntactische analyse in Beheydt 1979. Wat de kwantitatieve taalmaten als gemiddelde zinslengte betreft: wij zijn er ons van bewust dat die op veel tegenstand stoten. Terecht is in verband met bijvoorbeeld gemiddelde zinslengte opgemerkt dat iemand die in korte zinnen spreekt toch erg taalvaardig kan zijn en het is zeker zo dat syntactische complexiteit niet alleen aan zinslengte kan worden gemeten.

Men moet nochtans toegeven dat gemiddelde zinslengte (GZL) een aanvaardbare syntactische maat kan zijn. Inderdaad, nagenoeg elke syntactische complexering (coördinatie, subordinatie, uitbreiding van de nominale groep, toevoeging van bepalingen en van modaliteit, passivering) leidt tot vergroting van de zinslengte.

Bovendien is het ook zo dat het aantal mogelijke constructiecombinaties toeneemt met de toenemende zinslengte. De spreker "who limits all his sentences to five words in length finds he has relatively few different sentences at his command. The writer who uses fifty words in every sentence has a tremendous range of different patterns available" (Miller 1951, p. 137).

In dit opzicht blijkt dus GZL wel een bruikbare maat. Die betrouwbaarheid wordt o.i. nog geadstrueerd door bestaand experimenteel onderzoek. Zo vonden Shriner & Sherman (1967) dat van zes maten die courant worden gebruikt om het taalontwikkelingsniveau van kinderen te meten, de GZL de beste enkelvoudige maat is en bovendien is uit onderzoek van De Villiers & De Villiers (1973) nog gebleken dat GZL sterk positief correleert met morfologische complexiteit. Voorts is uit een onderzoek van Leonard e.a. (1976) duidelijk geworden dat GZL sterk verband houdt met semantische complexiteit. Er is dus voor de GZL voldoende grond om die als maat van syntactische complexiteit te gebruiken. Een andere vraag is natuurlijk hoe die zinslengte moet worden gemeten. Zelf opteer ik op allerlei gronden voor een GZL in woorden per uiting (vgl. Beheydt, in voorbereiding).

Is de gemiddelde zinslengte een vrij betrouwbare maat ze is anderzijds toch niet erg informatief. Twee steekproeven kunnen een nagenoeg identieke GZL hebben en toch vrij sterk verschillen wat betreft hun afwisseling in zinslengte. Als één van de steekproeven een vrijwel homogene zinslengte heeft en de andere daarentegen vertoont een grote afwisseling rond het gemiddelde, dan zal de laatste steekproef meestal syntactisch complexer zijn. Om nu dit verschil in spreiding rond het gemiddelde te meten beschikken we over een vrij eenvoudige statistische parameter, namelijk de standaarddeviatie. Toegepast op zinslengte geeft de standaarddeviatie ons een inzicht in de spreiding van de zinslengte rond het gemiddelde. Hoe groter de standaarddeviatie van de gemiddelde zinslengte (SD-GZL) hoe groter de syntactische complexiteit.

Een nadeel van de GZL en de SD-GZL is wel dat de invloed van een numeriek groot overwicht aan korte zinnen, de invloed van een beperkt aantal langere en complexere zinnen dusdanig indijkt dat deze laatste invloed nog nauwelijks in de cijfers voor GZL en SD-GZL tot uiting komt.

Daarom leek het ons nuttig te beschikken over een maat die een relatief grotere waarde toekent aan de langere zinnen en een relatief lagere aan de kortere zinnen. Een dergelijke maat menen wij te hebben gevonden in wat wij de *Distributiefactor* (D) genoemd hebben. De distributiefactor van de zinslengte is een empirisch gevonden maat die gebaseerd is op de frequentiedistributie van de zinslengten en die voldoet aan de hiervoor geformuleerde voorwaarde, zoals moge blijken uit de formule:

$$D = \frac{\Sigma fx.\ X^2}{N}$$

d.i. de som van de produkten van de frequentie van elke voorkomende zinslengte (fx) met het kwadraat van die zinslengte (X^2), gedeeld door het totale aantal zinnen. Aangezien de zinslengte gekwadrateerd wordt zal een hoge zinslengte een relatief grotere invloed hebben op de waarde van D dan een lage of m.a.w. een complexere zin zal meer invloed hebben dan een eenvoudige: door dit kwadraat bovendien te vermenigvuldigen met de absolute frequentie van de zinslengte zal de teller ook groter zijn naarmate er meer langere zinnen voorkomen. Natuurlijk kan de invloed van de absolute frequentie slechts geëvalueerd worden tegen de achtergrond van het totale aantal zinnen. Vandaar dat we N in de noemer plaatsen. Deze uitleg volstaat om in te zien dat naarmate het taalgebruik syntactisch complexer zal zijn in termen van de distributie van de zinslengten, ook de distributiefactor groter zal zijn.

De hiervoor gegeven bespreking moge volstaan om aan te geven op welke manier men operationele taalmaten plausibel uit hypothesen kan afleiden. Ze heeft hopelijk ook duidelijker gemaakt dat het een redelijke aanname is te veronderstellen dat de maten die het zelfde aspect van het taalgebruik meten, i.c. de syntactische complexiteit, resultaten zullen opleveren die sterk correleren. Maar zoals gezegd volstaat het niet, alleen maar de syntactische complexiteit te meten. Taalvaardigheid uit zich bijvoorbeeld evenzeer in afwisseling in het gebruik van verschillende structuren als in de complexiteit waarmee ze gebouwd zijn. We moeten dus behalve de syntactische complexiteit ook de syntactische diversiteit meten.

Voor het meten van die syntactische diversiteit kan men bijvoorbeeld gebruik maken van de relatieve frequentie van de gebruikte syntactische structuren. De relatieve frequentie van de syntactische structuren kan vrij gemakkelijk worden berekend door het totale aantal structuren te delen door het totale aantal verschillende structuren. Dit getal levert ons de gemiddelde frequentie, d.i. het gemiddeld gebruik van elke structuur. Die relatieve frequentie is omgekeerd evenredig met de syntactische diversiteit, d.w.z. dat ze daalt naarmate het taalgebruik complexer wordt.

Geeft de relatieve frequentie van de syntactische structuren aan of er *gemiddeld* evenveel verschillende structuren worden gebruikt, ze geeft anderzijds toch niets aan van het *relatieve gebruik* dat van die verschillende structuren wordt gemaakt. Immers, de relatieve frequentie van de syntactische structuren wordt berekend door het totale aantal structuren te delen door het aantal verschillende en als zodanig wordt dit cijfer niet beïnvloed door de verdeling van de gebruiksfrequenties over de verschillende structuren. Derhalve is het bijvoorbeeld onmogelijk om met de relatieve frequentie de bewering van Bernstein te toetsen dat er in het taalaanbod van de verschillende sociale klassen niet zozeer een verschil in het aantal verschillende structuren hoeft te worden verwacht, als wel een verschil in het relatieve gebruik dat ervan gemaakt wordt (Bernstein 1971, p. 183).

Als we dit verschil in het relatieve gebruik willen nagaan dan kunnen we gebruik maken van de standaarddeviatie van de frequentiedistributie van de syntactische structuren (SD-f synt.str.). Die SD is namelijk een maat van de spreiding rond het gemiddelde en dus hoe kleiner de *SD-f synt.str.* is, hoe gelijker de frequenties gespreid liggen rond de gemiddelde frequentie. Is de standaarddeviatie hoog dan betekent dat dat enkele types structuren heel frequent worden gebruikt en nogal wat structuren zelden worden gebruikt. Met andere woorden een hoge standaarddeviatie verwijst in het algemeen naar een stereotiep gebruik van syntactische structuren en een lage standaarddeviatie naar een gevarieerde syntaxis.

Een nadeel van de relatieve frequentie en van de standaarddeviatie is wel dat ze afhankelijk zijn van steekproefgrootte (Beheydt 1979, p. 206-211). Om een betrouwbaarder beeld te krijgen moeten we een syntactische diversiteitsmaat zien te vinden die onafhankelijk is van steekproefgrootte. Uit de kwantitatieve linguïstiek kennen we een maat voor de repetitiegraad van het vocabularium die onafhankelijk is van de steekproeflengte, nl. de *characteristic* van Yule. Die characteristic, de zgn. *K-factor*, is een numerieke waarde die de herhalingsgraad van het vocabularium weergeeft: een hoge K-factor wijst op een hoge herhalingsgraad. (Het omgekeerde, een lage K-factor wijst op een lage herhalingsgraad gaat – hoewel door sommigen beweerd – niet altijd op.) De formule voor de berekening van die K-factor is volgens Yule (1944):

$$K = 10000\ \frac{S_2 - S_1}{(S_1)^2}$$

Waarin S_1, het eerste moment van de distributie, de som is van alle types in elke frequentiegroep vermenigvuldigd met de frequentie van die groep (= totaal aantal tokens) en S_2, het tweede moment van de distributie, de som van al de types in elke frequentiegroep vermenigvuldigd met het kwadraat van de frequentie. Als we nu die K-factor transponeren op syntactische structuren, dan krijgen we een van steekproefgrootte onafhankelijke maat van het verschil in relatief gebruik van syntactische structuren, nl. *de repetitiegraad van de syntactische structuren.* Tegen die transpositie kan vanuit mathematisch oogpunt geen bezwaar worden gemaakt als men van de syntactische structuren een frequentiedistributie kan opstellen die volledig gelijksoortig is aan de frequentiedistributie voor het vocabularium, immers de K-factor is in principe niets anders dan een statistische parameter van een frequentiedistributie.

3.2. Lexicale taalmaten

Naast de syntactische diversiteit en de syntactische complexiteit bepaalt vooral de specificiteit van het vocabularium mee de graad van semantische specificiteit van het taalgebruik.

In de kwantitatieve linguïstiek zijn een aantal maten ontwikkeld waarmee de vocabulariumdiversiteit kan worden nagegaan. De meest gebruikte is wellicht de Type Token Ratio (TTR), dit is de verhouding tussen het aantal verschillende woorden (V,types) en het totale aantal woorden (N,tokens). Het is duidelijk dat naarmate er relatief meer verschillende woorden worden gebruikt de TTR groter is en meteen ook de lexicale diversiteit.

Nochtans is er aan deze in wezen eenvoudige maat een nadeel verbonden, ze is vrij sterk afhankelijk van steekproefgrootte. (cf. Beheydt 1979, p. 232-233). Op empirische gronden hebben wij voor die formule een correctiefactor gevonden die voor verschillende steekproefgroottes toch zowat een constante opleverde. Die correctiefactor bestond uit een logaritmische transformatie van de TTR (de logTTR) nl. de verhouding van de logaritme van het aantal types (logV) tot de logaritme van het aantal tokens (logN). Voor gegevens in de orde van grootte van N gaande van zowat 2000 tot N zowat 10000 is dit een erg stabiele maat gebleken.

Het zou onvoorzichtig zijn voor een vrij complex fenomeen als vocabulariumdiversiteit maar één maatstaf aan te leggen. Voor de betrouwbaarheid en de geldigheid verdient het, zoals gezegd, aanbeveling zo veel mogelijk en dan nog liefst ongelijksoortige maten te gebruiken die telkens andere aspecten van de diversiteit pakken.

Een maat die pretendeert beter dan de logaritmische TTR de verhouding tussen vocabularium en totaal aantal gebruikte woorden weer te geven is de "indices de Richesse" van Guiraud (1959, p. 88). Over deze indice zegt Guiraud namelijk zelf: ' J'ai en son temps proposé après des tests innombrables, de prendre comme indice de richesse le rapport $V/\sqrt{N}$, le nombre des mots différents divisé par la racine carrée du nombre total de mots, et bien que cet indice ne soit pas absolument constant je continue à croire qu'il est pratiquement le meilleur et le plus facilement utilisable dans certaines limites". Het in de laatste zin geuite optimisme kunnen wij niet volledig delen, want de *richesse* R, blijkt in elk geval voor grotere steekproeven socialiseringstaal sterk afhankelijk van steekproefgrootte. Het door die maat geïmpliceerde lineaire verband, namelijk dat het vocabularium toeneemt als de vierkantswortel uit het totale aantal gebruikte woorden, blijkt trouwens niet zo goed aan te sluiten bij de geobserveerde waarden als het logaritmisch verband.

Een totaal anders gerichte maat die de individuele specificiteit van het vocabularium meet en die bovendien het niet te miskennen voordeel heeft onafhankelijk te zijn van steekproefgrootte is de reeds in verband met de syntaxis gebruikte "Characteristic" van Yule. Deze "Characteristic" die we K-factor genoemd hebben is zoals gezegd gebaseerd op het eerste en het tweede moment van de frequentiedistributie. De formule is:

$$K = 10000\,\frac{S_2 - S_1}{(S_1)^2}$$

waarin S_1 het totale aantal gebruikte woorden is, en S_2 de som van de vocabulariumwoorden in elke frequentieklasse vermenigvuldigd met het kwadraat van de frequentie. Voor een stapsgewijze handleiding bij de berekening verwijzen wij naar het artikel van P.E. Bennett (1969).

Wat meet nu die K-factor eigenlijk? Oorspronkelijk voorgesteld als een diversiteitsmaat, is het eigenlijk toch meer een maat van de herhalingsgraad van het vocabularium (Williams 1970, p. 98-99). Het is meer een maat van uniformiteit dan van diversiteit: hij wordt groter naarmate de diversiteit kleiner wordt en is dus omgekeerd evenredig met de diversiteit. Hoe dan ook de K-factor is een zeer goede maat van lexicale stereotypie, omdat vooral de frequentie van de meest voorkomende woorden de waarden van K beïnvloedt.

Een afgeleide maat van de K-factor die een paar van zijn gebreken niet heeft, nl. de arbitraire vermenigvuldiging met een factor 10000 en de in elk geval voor het door mij verzamelde taalaanbodsmateriaal experimenteel gefalsifieerde voorwaarde van een onderliggende Poissonverdeling, is de parameter υ_m van Herdan (1964, p. 70). Die υ_m wordt berekend volgens de formule:

$$\upsilon_m = \frac{\sigma x/N}{Mx}$$

waarin σ x de standaarddeviatie is van de gemiddelde frequentie, N het totale aantal woorden en Mx de gemiddelde frequentie van elk vocabulariumwoord (= $\Sigma fx/N$). Die υ_m is een index van de relatieve fluctuatie van het gemiddelde gebruik, of nog een coëfficient van de gemiddelde variatie van de gemiddelde frequentie van de vocabulariumwoorden. Een hoge relatieve fluctuatie van de gemiddelde frequentie wijst op een lage diversiteit van de woordenschat. Dus ook deze maat is omgekeerd evenredig met de diversiteit en ook hier weer geldt dat de coëfficient eerder beinvloed wordt door de klasse van woorden met hoge frequentie: υ_m "traduit surtout la densité des mots grammaticaux" (Muller 1964, p. 92), dat wil zeggen dat υ_m vooral iets zegt over de herhalingsgraad van voorzetsels, lidwoorden en voornaamwoorden.

Een andere maat van diversiteit die in tegenstelling tot de vorige maten niet op de totale distributie steunt maar op een bepaald aspect ervan, namelijk op die klasse woorden die maar één keer voorkomen, de zgn. hapaxen, is de verhouding van het aantal hapaxen (V_1) tot het totale aantal verschillende woorden (V). Deze hapaxen zijn interessant omdat ze ons een idee geven van de voorraad woorden waaruit een spreker put. Immers, aanvankelijk wanneer iemand begint te spreken is ieder woord een hapax, maar naarmate men meer spreekt, worden steeds meer woorden herhaald. Nu is het aantal hapaxen sterk afhankelijk van steekproeflengte, zoals gemakkelijk kan worden ingezien, maar het gedraagt zich ten opzichte van de steekproefgrootte ongeveer zoals het aantal nieuwe woorden, zodat we kunnen stellen dat V_1 relatief gelijk toeneemt met V. Als V_1 echter in verhouding vlugger toeneemt, dan betekent dit dat de diversiteit van keuze groter

is en dat het lexicon waaruit geput wordt waarschijnlijk ook groter is. In die zin is V_1/V dus een maat van de semantische specificiteit van het vocabularium. Het is bovendien een maat die recht evenredig is met de diversiteit, d.w.z. hoe hoger het procent hoe groter de diversiteit.

Al de tot hiertoe besproken maten richtten zich op bepaalde aspecten van de diversiteit. Het leek ons wenselijk ook een maat te hebben die de verschillende aspecten in de meting betrok en zowat als een globale diversiteitsmaat kon fungeren. Een dergelijke maat was echter niet zo direct beschikbaar zodat wij die zelf wel moesten ontwerpen. Uiteindelijk zijn we via logische deductie en experimenteel uitproberen tot de volgende maat van de afwisseling in het vocabularium gekomen:

$$A = \frac{V_1 \times V}{N}$$

Deze afwisselingsmaat (A) heeft het voordeel dat ze een aantal deelaspecten van de diversiteit tezelfdertijd verdisconteert: nl. zowel steekproefgrootte (N), vocabulariumgrootte (V) als aantal hapaxen (V_1). Bovendien blijkt ze vrij onafhankelijk van steekproeflengte en heeft ze haar eenvoud en directe inzichtelijkheid voor. We hebben V_1 en V in de teller geplaatst omdat het logisch is dat de diversiteit groter zal zijn naarmate er méér woorden één keer worden gebruikt en naarmate er meer verschillende woorden voorkomen. De vermenigvuldiging van die twee factoren zorgt dan voor een relatief sterkere invloed van de diversiteitsaspecten. Natuurlijk moeten V_1 en V dan wel beoordeeld worden in relatie tot het totale aantal gebruikte woorden. Zowel het aantal één keer gebruikte woorden als het aantal verschillende woorden zijn immers afhankelijk van steekproefgrootte. Experimenteel is A een zeer goede globale diversiteitsmaat gebleken (Beheydt 1979, p. 243).

3.3. Tekstcomplexiteitsmaten

Na de verschillende maten voor de vocabulariumdiversiteit komen we nog op een semantisch complexiteitsaspect dat – zoals de kritiek terecht heeft aangemerkt – te veel buiten schot gebleven is in kwantitatieve benaderingen van taalgebruik en dat is het *tekstaspect.* Nochtans kan ook daarvoor gedacht worden aan kwantitatieve maten en bepaalde formules uit het leesbaarheidsonderzoek komen beslist voor omwerking tot tekstcomplexiteitsindexen van gesproken taal in aanmerking. Zo blijkt bijvoorbeeld Tuldava (1975) experimenteel een zeer betrouwbare complexiteitsindex gevonden te hebben die beantwoordt aan de volgende eenvoudige formule:

$$R = \bar{i} \lg GZL$$

R is de tekstcomplexiteitsindex, $\bar{i}$ de gemiddelde woordlengte in lettergrepen, en lg GZL, de logaritme van de gemiddelde zinslengte in woorden. Dit is maar één voorbeeld van een mogelijke benadering van de tekstcomplexiteit. Voor meer inspiratie in dit verband verwijs ik naar Van Hauwermeiren (1975) en (1977).

4. Besluit

Met dit al hebben wij aangetoond hoe men operationeel een kwantitatief analyserooster kan opbouwen waarmee men een aantal basishypothesen kan toetsen. Zo lijkt het ons zeer wel mogelijk om met behulp van het hiervoor uitgewerkte analyserooster uitspraken te doen over het verschil in socialiseringstaal tegenover kinderen uit verschillende sociale klassen. Met een dergelijke toetsbatterij kan men ook nagaan of de school op taalgebruiksniveau eisen stelt die niet in overeenstemming zijn met het soort taalgebruik dat bepaalde kinderen in hun eigen milieu gewend zijn. Daartoe volstaat het dit analyserooster over de taal van de school te leggen en de resultaten te vergelijken met de resultaten die ditzelfde rooster oplevert wanneer men het op de socialiseringstaal van die kinderen toepast. Bovendien zal dit rooster ook kunnen aanduiden waar de verschillen zitten en kan het dus als een diagnostisch instrument worden gehanteerd. Er is dus kennelijk nog wel voldoende reden om het taalgebruik kwantitatief te benaderen.

Het zou natuurlijk struisvogelpolitiek zijn te doen alsof er aan de kwantitatieve benadering geen gebreken zaten. Wij zijn er ons van bewust dat de hier gepresenteerde vocabulariummaten bijvoorbeeld geen rekening houden met verschijnselen als homonymie (*bank* om op te zitten t.o. *bank* als geldinstituut), dat de syntactische complexiteit factoren als pronominalisering, afwijkende woordvolgorde en dubbele negatie niet verdisconteert en dat de tekstcomplexiteitsindex woorden als *keukentafel* als complexer beschouwt dan *logaritme.* (vgl. Hoar 1981). Maar dergelijke gebreken beletten niet dat de voorgestelde taalmaten *gezamenlijk* én *over een representatieve steekproef,* betrouwbare resultaten opleveren.

Bibliografie:

Appel, R.: *Het meten van syntactische complexiteit van de taal van het kind.* Paper ATW Amsterdam, 1972.

Appel, R. e.a.: *Sociolinguistiek.* Utrecht, Antwerpen, 1976.

Beheydt, L.: *Variatie in taalaanbod. Een sociolinguïstisch onderzoek van de primaire socialisatie in West-Vlaanderen.* Winksele, 1979.

Beheydt, L.: "Taalvaardigheid in het geding", *Tijdschrift voor Taalbeheersing* 3 (1981), p. 67-75.

Beheydt, L.: *Kindertaalonderzoek* (in voorber.), Malle.

Bennett, P.: "The statistical measurement of a stylistic trait in *Julius Caesar* and *As you like it*". In: L. Doležel & R.W. Bailey (eds.): *Statistics and Style.* New York, 1969, p. 29-56.

Bernstein, B.: *Class, Codes and Control. Volume I. Theoretical Studies towards a Sociology of Language.* London, 1971.

Groot, A.D. de: *Methodologie. Grondslagen van onderzoek en denken in de gedragswetenschappen.* 4e druk, 's Gravenhage, 1968.

Guiraud, P.: *Problèmes et méthodes de la statistique linguistique.* Dordrecht, 1959.

Halliday, M.A.K.: "Foreword". In: B. Bernstein (ed.): *Class, Codes and Control.* Vol. II, London, 1973, ix-xvi.

Hauwermeiren, P. van: *Het leesbaarheidsonderzoek.* Groningen, 1975.

Hauwermeiren, P. van: "De ontwikkeling van Nederlandse leesbaarheidsformules" *Tijdschrift voor massacommunicatie* (1977), p. 161-181.

Herdan, G.: *Quantitative Linguistics.* London, 1964.

Hoar, N.: "Why we need linguists and not just formulas for determining the readability of documents". In: *Aila 81: Proceedings I,* 1981, p. 461-462.

Hymes, H. dell: "On Communicative Competence". In: J.B. Pride & J. Holmes (eds.): *Sociolinguistics.* Harmondsworth, 1972, p. 269-293.

Leonard, L.B. e.a.: "An examination of the semantic relations reflected in the language usage of disordered children". *Journal of Speech and Hearing Research* 19 (1976), p. 371-392.

Miller, G.A.: *Language and Communication.* New York, 1951.

Muller, C.: *Initiation à la statistique linguistique.* Paris, 1964.

Shriner, T.H. & D. Sherman: "An equation for assessing language development" *J.S.H.R.* 10 (1967), p. 41-48.

Spoelders, M..& F. van Besien: "Verbale communicatie in de klas" In: J. van den Broeck (red.): *Recent sociolinguistisch onderzoek in Vlaanderen.* Abla papers 2, Diepenbeek, 1979, p. 72-85.

Toorn, M.C. van den: *Methodologie en Taalwetenschap.* Utrecht, Antwerpen, 1978.

Tuldava, J.: "Ob izmerenii trudnosti teksta" *Acta et Commentationes Tartuensis.* Vol. 345. Tartu, 1975.

Villiers, J. de & P.A. de Villiers: "A cross-sectional study of the development of grammatical morphemes in child speech" *Journal of Psycholinguistic Research* 2 (1973), p. 257-278.

Williams, C.B.: *Style and Vocabulary: Numerical Studies.* London, 1970.

Yule, G.U.: *The Statistical Study of Literary Vocabulary.* Cambridge, 1944.