bleek te zijn en niet een louter individuele affaire. Het laatste zou je verwachten. Toch voltrekt het zelf worden zich in menselijke relatie.
Mijn ervaring is nog maar heel pril.
De reden dat ik er nu al in dit uiterst prille stadium mee voor de dag kom, is dat iedereen er naar hartelust tegenaan kan schoppen.
Ik van mijn kant zal heel erg graag terugschoppen. Niet alleen op de basis van wat ik hier verteld heb, maar op de basis van de gehele ervaring met de kleine Chomskyaantjes.

Heemstede, 25 september 1971

FREQUENTIEONDERZOEK VAN HET NEDERLANDS

A. J. Vervoorn

Bij het onderwijs in het Nederlands is de laatste paar jaar de vraag wat er nu eigenlijk onderwezen moet worden sterk op de voorgrond gekomen. De bundeling van leraren op alle niveaus in de VON heeft daartoe in sterke mate bijgedragen: het is een van de weinige maatschappelijke gebeurtenissen geweest waarbij confrontatie niet in een conflict maar in constructief handelen resulteert. Het is overigens voor mij nog een vraag of het vak "taal" de speciale positie van het moedertaalonderwijs toch niet beter weergeeft dan het vak "Nederlands". De VON zou een VOM moeten worden: vereniging voor het onderwijs in de moedertaal. Maar dit terzijde.
Eén van de hoofddoelen van het moedertaalonderwijs zal uitbreiding van de taalschat zijn. Het is intussen ruim voldoende aangetoond, welke ernorme maatschappelijke consequenties een beperkte taalbeheersing, en als belangrijk onderdeel daarvan, een beperkte woordenschat voor een leerling heeft. Beschikking over een ruime voorraad woorden en zinsconstructies is dus een ideaal einddoel voor elke vorm van moedertaalonderwijs. De vraag is dan: hoe kan dat ideaal zo systematisch en dus zo efficiënt mogelijk bereikt worden?
Inzicht in wat gewoon en wat ongewoon is in de taal, in de frequenties van woorden en constructies is daarvoor een eerste vereiste. "Het frequentieonderzoek is van grote betekenis voor de didactiek van het taalonderwijs, daar het een wetenschappelijke basis verschaft voor de keuze van de woorden, die men de leerlingen het eerst moet bijbrengen" merkt Staal in zijn interessante, maar weinig gebruikte boek op. (Lit. 10, pg. 16)

Oude frequentieonderzoekingen

De behoefte aan een goed inzicht in de frequentie van het Nederlands (en dit geldt min of meer ook voor andere talen) ontstond niet vanuit het moedertaalonderwijs. Het eerste frequentieonderzoek van het Nederlands werd "in opdracht van het Departement van Onderwijs en Eeredienst" (de tijden zijn wél veranderd!) in het toenmalige Nederlands-Indië ondernomen. Het verslag van dat

onderzoek publiceerde J. F. H. A. de la Court in 1937 te Batavia onder de titel: De meest voorkomende woorden en woordcombinaties in het Nederlandsch. "Daar deze woorden moesten dienen bij het onderwijs in het Nederlands in Indië, werden eenvoudige teksten (samen 1.000 000 woorden omvattend) onderzocht, namelijk kinderlectuur en eenvoudige lectuur voor volwassenen. Op grond hiervan werden twee lijsten samengesteld: 1. een alfabetische lijst van 3296 woorden die in de teksten het frequentst waren en 2. een lijst van dezelfde woorden, gerangschikt naar frequentie en verdeeld in zeven radii. Bij elke radius worden de woorden weer in alfabetische volgorde gegeven." (Lit. 10 pg. 12) Vooral in België heeft deze lijst op indirecte wijze veel diensten bewezen.

Voor het onderwijs in het Nederlands als tweede taal heeft G. Vannes er namelijk zijn veel gebruikte Vocabulaire du néerlandais de base (Antwerpen, 1949) op gebaseerd.

Nog steeds is er voor het Nederlands geen materiaal beschikbaar dat in omvang of in methode het werk van De la Court overtreft. Noch Formal properties of newspaper Dutch, door Van Berckel, Brandt Corstius e. a. (Amsterdam, 1965) noch W. Martins werk uit 1968 (Lit. 7) geven daarvoor voldoende en voldoende gespreid materiaal. Toch zou het om verschillende redenen plezierig zijn als voor de lijsten van De la Court een beter materiaal in de plaats gesteld kan worden. "Beter"

omdat niet alleen woorden geteld moeten worden, maar ook syntactische verschijnselen;

omdat niet alleen maar kinderboeken en eenvoudige lectuur bestreken moeten worden, maar ook andere taalvelden;

beter omdat de taal van Nederlandsch-Indië anno 1937 niet meer gelijk is aan het Nederlands van nu.

In 1965 merkte prof. dr. Engels op het Vlaams filologencongres over een te ondernemen frequentieonderzoek van het Nederlands op: "Dit wetenschappelijk onderzoek wil eerst en vooral een didactisch resultaat opleveren, zo noodzakelijk in België: een deugdelijke lijst van frequente woorden en vooral van frequente constructies voor het Nederlands als tweede taal. De frequentie van de constructies in de levende talen werd tot nu toe nooit onderzocht; nochtans hebben didactische experimenten ons geleerd dat de loutere frequentie van notionele woorden geen oplossing biedt, en alleen een degelijke graduering van frequente constructies doeltreffend kan werken in het vreemde-talenonderwijs. De bestaande frequentielijsten zijn trouwens van weinig nut meer, vooral ten gevolge van het verouderde taalmateriaal dat werd onderzocht. Ook wordt in die lijsten het notionele woord alleen geteld, het structurele woord wordt buiten beschouwing gelaten." (Lit. 5, pg. 230)

Waarop moet frequentieonderzoek antwoord geven?

Hoe zou men zich nu een frequentieonderzoek van het Nederlands wensen? Welke vragen mogen er aan de uitkomsten gesteld worden en wat kan men

met die uitkomsten doen? Welke methodische problemen doen er zich zo voor bij de opzet van een frequentieonderzoek naar taalverschijnselen? Dit zijn een paar vragen, waar ik iets nader op wil ingaan.
Een eerste voorwaarde die aan een frequentieonderzoek van het Nederlands gesteld moet worden, is dat het materiaal inderdaad het Nederlands omvat. Dat wil dus zeggen niet alleen maar kinderboeken plus eenvoudige lectuur, niet alleen maar krantentaal en romans. Een volledig onderzoek zal zowel de geschreven (in de praktijk alleen de gedrukte) als de gesproken taal moeten omvatten. Elk van die twee hoofdgroepen zal weer zo goed mogelijk in taalvelden verdeeld moeten zijn, om "de ruimte van het volledige leven te bevatten". Deze taalverkaveling zal bovendien voor de gedrukte en de gesproken taal niet op dezelfde manier kunnen gebeuren. Voor de laatste categorie spelen b.v. locale en leeftijdsverschillen een veel grotere rol, terwijl bij gedrukte taal het gebruiksdoel de belangrijkste verschillen geeft.
Zowel de forumdiscussie als het buurpraatje, zowel de poëzie als het handboek voor duivenliefhebbers vormen een deel van het Nederlands.
Nu is het onmogelijk om de volledige taalproductie van een bepaalde tijd, b.v. 1971, te tellen: daarvoor zijn de hoeveelheden te groot. (Van het werk van Kaeding, die in 1898 voor het Duits een frequentiewoordenboek uitgaf op basis van ongeveer elf miljoen woorden, wordt gezegd dat hij de beschikking had over een krijgsgevangenenkamp voor het telwerk. Kaeding wilde overigens een nieuw stenografiesysteem opzetten.) Men zal dus moeten proberen uit de gekozen taalvelden een representatieve steekproef te nemen: "De sample dient zo gekozen te zijn dat hij de grotere verzameling representeert; de statistische methode is juist, als de werkelijkheid lijkt op een veelvoudig vergrote projectie van de sample", zoals J. J. M. Bakker het in zijn dissertatie formuleert (Lit. 2, pg. 40). Een moeilijkheid hierbij is, dat men niet het totaal kent en dus ook niet kan zeggen hoe representatief bijvoorbeeld een miljoen woorden eigenlijk zijn voor het Nederlands. Als men dan, zoals nu bij het in gang zijnde project "Frequentieonderzoek van het Nederlands", eenmaal zo'n totaal aantal gefixeerd heeft, moet daarbinnen weer een verdere verdeling plaatsvinden. P. C. uit den Boogaart, die met een subsidie van ZWO aan de Technische Hogeschool te Eindhoven bezig is met de uitvoering van het project, heeft over de manier om tot een verantwoorde verkaveling te komen reeds een gedetailleerd artikel geschreven (Lit. 3). Reeds eerder had trouwens prof. dr. Engels een globale schets gegeven. "We kiezen ons materiaal zorgvuldig uit volgens de principes van een goede "sampling": een voldoende aantal woorden per taalsoort, of taalveld. Daarna moeten al die taalvelden met elkaar kunnen vergeleken worden, zodat de constanten of de overlappende gedeelten voor alle of sommige taalvelden gemeenschappelijk, of voor een reeks velden afzonderlijk, aan het licht komen.
Ook de volledig eigen woordenschat of typisch eigen structuren van de verschillende taalvelden komen dan te voorschijn. Om de taalvelden met elkaar te vergelijken moeten we een even groot aantal woorden per taalveld en een

even groot aantal woorden per auteur onderzoeken. Deze groepen woorden worden blind getrokken uit het werk van een auteur. Er wordt natuurlijk rekening gehouden met semantische verschillen, met samenstellingen en collocaties. Door het feit dat we het materiaal trekken uit verschillende taalvelden en de verhoudingen willen onderzoeken van die taalvelden onderling en afzonderlijk, gaan we verder dan de vroegere onderzoekers van de frequentie, die zich blijkbaar hebben tevreden gesteld met de misleidend hoge frequentie (ongeveer 95%) van het 3000-tal frequente woordeenheden, die hun tellingen hadden opgebracht. Over de overblijvende 5%, die nochtans 297.000 woorden bevat, als de taal wordt verondersteld nagenoeg 300.000 woorden te bezitten, is er na de freqentietellingen tussen 1930 en 1940 nooit meer gesproken." (Lit. 4, pg. 86)

Gebruik van onderzoeksresultaten

Met de laatste opmerking van prof. Engels komt de vraag naar voren: wat kun je nu met uitkomsten van frequentieonderzoek doen? Wat mag ervan verwacht worden? Dat hangt natuurlijk af van de manier waarop iemand met taal bezig is. Inderdaad zal voor het onderwijs van het Nederlands-als-tweedetaal een lijst van de 1000, 2000 of 3000 meest frequente woorden een uitermate nuttig hulpmiddel zijn.
Toch zijn bij de uitkomsten van een frequentieonderzoek niet de lijsten met de meest frequente woorden het interessantst. Het onderwijs in het Nederlands als vreemde taal zal altijd wel van betrekkelijk bescheiden omvang blijven (ook al is de belangstelling ervoor groeiende). Maar er is nog iets anders. Wanneer de uitkomsten van een aantal frequentieonderzoeken voor diverse talen vergeleken worden, blijken juist de meest frequente woorden voor een groot deel dezelfde te zijn. M.a.w. de vertaalde lijst van de 1000 meest frequente Engelse woorden, vormt ook een zeer goede basis voor het onderwijs in het Nederlands. Voor de eerste 6000 woorden uit het Engels, Frans, Duits en Spaans is een dergelijke vergelijking al gemaakt door Helen S. Eaton (Lit. 11). Voor het moedertaalonderwijs lijkt me dit minder interessant (ik zeg niet: oninteressant), en wel om twee redenen.
Ten eerste omdat een groot deel van de zeer frequente woorden behoort tot de lidwoorden, voornaamwoorden, voorzetsels, voegwoorden. Wanneer het gaat om taalarmoede denkt men toch niet in eerste instantie aan "de, het, van, op, die, ik" etc. Staal (Lit. 10, pg. 18) wijst er ook al op dat juist in de kindertaal substantiva, adjectiva en werkwoorden de belangrijkste woordgroepen zijn. Ik zie geen reden om aan te nemen, dat het Nederlands af zal wijken van Frans, Duits of Engels, voor welke talen cijfers bekend zijn. Staal haalt daarbij cijfers aan uit een artikel van Margaret M. Nice. Zij "vond van ongeveer drie tot ongeveer zeven jaar een vrij vaste verhouding van de percentages der verschillende woordsoorten en wel voor zelfstandige en bijvoeglijke naamwoorden samen 50 à 60%, werkwoorden 20 à 24%".
Men kan dus (en dit ten tweede) zeggen dat voor het moedertaalonderwijs de

uitgebreidere "tweede laag" van de woordvoorraad meer perspectief biedt, dan de toplaag. Wanneer er inderdaad een goed overzicht van deze tweede laag beschikbaar is - en dan verdeeld over de verschillende taalvelden omdat het totaal onhandelbaar groot zal blijken -, kan er met meer systeem aan uitbreiding van de taalschat gewerkt worden.
Men moet dan bij het onderwijs, bij het "verwerken" van die woorden met een paar factoren rekening houden. Want "de frequentie kan of mag ... niet het enige criterium zijn om de belangrijkste woorden van een taal te bepalen. Daarnaast spelen o. m. ook de range (= het aantal verschillende teksten waar men een woord vindt), de disponibiliteit (= de bruikbaarheid van een woord binnen een semantisch veld), de valentie (= de mogelijkheid om andere woorden te vervangen; E. Coverage), en de bruikbaarheid (= voorkomen van een woord in samenstellingen en afleidingen) een rol." (Lit. 7, pg. 15) Deze opmerking maakte W. Martin met een zekere vooruitziende blik in zijn boek over "de inhoud van krant en roman". Want toen hij een jaar later (Lit. 9) de besprekingen van zijn boek op een rij zette, kon hij er direct naar verwijzen. Vanuit didactisch perspectief merkt hij trouwens op: "het lijkt ons van belang te weten welke frequentie die voegwoorden en die pronomina hebben om ze, tevens rekening houdend met hun fonetische moeilijkheidsgraad, te gepaster tijd te kunnen introduceren." (Lit. 8, pg. 28)

Andere frequentieonderzoeksterreinen

Met die "fonetische moeilijkheidsgraad" wordt terloops een enigszins andere frequentie binnengehaald. Natuurlijk is het ook, vooral voor het Nederlands voor anderstaligen, zeer plezierig een overzicht te hebben van de frequentie der fonemen in het Nederlands. Welke klanken of klankcombinaties zijn erg gewoon en welke zeldzaam? Is "Scheveningen" een fonetisch curiosum of zijn de scherpe -ch- en de -ng- belangrijk genoeg om als eerste op het lesprogram te verschijnen? Daarvoor hoeven we geen miljoen woorden te tellen: de woordvoorraad van b.v. "het groene boekje", de Woordenlijst der Nederlandse taal, is daarvoor voldoende. Voor een overzicht van de Nederlandse klanken doet het er niet toe of één bepaald woord heel veel gebruikt wordt, het gaat er om in hoeveel verschillende woorden een klank voorkomt. Een goed inzicht daarin geven de publicaties van J. J. M. Bakker. (Lit. 1 en 2)
De uitkomsten van frequentieonderzoek kunnen ook nog voor een heel ander facet van het moedertaalonderwijs interessante gegevens opleveren, nl. voor de behandeling van teksten, voor het literatuuronderwijs.
Wanneer het woord frequentie gebruikt wordt, gaan de gedachten min of meer automatisch naar grote aantallen uit. We hebben al gezien dat echter juist de iets minder grote frequenties interessant zijn. Voor het inzicht in en de waardering van een schrijver zijn zelfs de woorden die maar één keer gebruikt worden interessant. Martin zegt over deze unica uit de woordvoorraad: "In de vorige paragraaf noteerden wij dat de hapax legomena een speciale plaats in het vocabularium van een auteur innemen. Meer nog dan de andere woorden lichten zij

ons in over het lexicon waaruit de auteur put. Inderdaad, hoe groter de concentratie van de hapax in een deel van de tekst, hoe ruimer het lexicon is dat de auteur in dat gedeelte ter beschikking heeft." (Lit. 9, pg. 77). Juist wanneer we spreken over taalschat (en de uitbreiding ervan), over de rijkdom van een auteur, gaat het erom hoeveel verschillende woorden iemand gebruikt en niet hoe vaak hij sommige woorden herhaalt.
Nog veel boeiender zal inzicht in frequentieverschijnselen zijn die meer dan het woord omvatten: zinslengte, syntactische structuren, stilistische kenmerken. "Een statistische interpretatie zal uiteraard beperkter zijn in toepassingsmogelijkheden dan een literaire analyse, vermits lang niet alle kwalitatieve gegevens kwantitatief meetbaar zijn; maar de dingen die zij kan aantonen, zullen dan ook met onweerlegbare, wetenschappelijke strengheid bewezen zijn, ten minste indien de statistische spelregels in acht genomen werden."
"Statische beschrijving en interpretatie van taalkundige en literaire variabelen staat nog in de kinderschoenen. Uit de enkele aangehaalde voorbeelden mag wel blijken dat de toepassingsmogelijkheden talrijk zijn, maar dat het gebied nog weinig ontgonnen is. De moeilijkheden zijn niet zo groot voor de beschrijving van taalkundige frequenties, omdat men daar van meet af te doen heeft met kwantitatieve gegevens. De eigenlijke moeilijkheden beginnen pas bij de interpretatie ervan, vooral dan wanneer die gegevens betrekking hebben op kwalitatieve aspecten van de taal". Aldus twee citaten van A. Keuleers (Lit. 6, pg. 45 en 47) die mij de moeite waard leken: het eerste om de beperking die aangelegd wordt, het tweede om de perspectieven die geboden worden. Want in die perspectieven zit een raakvlak tussen taalkundig onderzoek aan de ene kant en taalonderwijs aan de andere kant. Het frequentieonderzoek van het Nederlands is in volle gang, wellicht zullen binnen een jaar de eerste resultaten gepubliceerd zijn. De interpretatie van de resultaten zal voor een belangrijk deel in het onderwijs van het Nederlands moeten gebeuren. Taalkundige frequenties zullen gebruikt kunnen worden als bewijsmateriaal voor literaire interpretaties. De waardering voor een auteur kan uit de subjectieve waarderingssfeer gehaald worden en "bewezen" met geconstateerde feiten. Maar bovendien zal een veel vastere basis gegeven kunnen worden voor b.v. de beschrijving van een ontwikkelingsgang bij een auteur. Door nauwkeurige vergelijking zullen invloeden precies vastgelegd en aangetoond kunnen worden. Leerlingen zullen aan de hand van frequentieverschijnselen zelf kunnen ontdekken waarin het taalgebruik van de ene auteur afwijkt van dat van de ander, en zo ook een gefundeerde waardering kunnen uitspreken. Juist in het onderwijs zullen de uitkomsten van een, uit de aard der zaak algemeen gericht, frequentieonderzoek op hun toepasbaarheid getoetst kunnen worden aan individuele auteurs. Daar zullen ook de vragen uit voortvloeien naar de niet-behandelde aspecten, naar frequenties die over het hoofd gezien zijn. The proof of the pudding is in the eating.

Gebruikte literatuur

1 Bakker, J. J. M. - Frequency in usage and in the lexicon; Lingua 21 (1968) pg. 13-22
2 Bakker, J. J. M. - Constant en variabel; Dissertatie Amsterdam 1971
3 Boogaart, P. C. uit den - Sampling van tekstfragmenten uit Nederlandse dagbladen; ITL Review for applied linguistics 10 (1970) pg. 25-33
4 Engels, L. K. - Automering en mathematisatie uit een linguistisch oogpunt; Wetenschappelijke Tijdingen 2 (1967) pg. 82-92
5 Engels, L. K. - Electrotechnische Machines en Taalonderzoek; Handelingen XXVe Vlaams Filologencongres (1965) pg. 229-237
6 Keuleers, A. - Beschrijving en Interpretatie van Linguistische Frequenties; ITL 1 (1968) pg. 33-48
7 Martin, W. - De inhoud van krant en roman. Een frequentieonderzoek; Antwerpen 1968
8 Martin, W. - Kanttekeningen bij een frequentieonderzoek; ITL 4 (1969) pg. 25-33
9 Martin, W. - Analyse van een vocabularium met behulp van een computer; Brussel 1970
10 Staal, A. J. - De methoden van psychologisch taalonderzoek; Enschede 1946
11 Eaton, H. S. - An English - French - German - Spanish Word Frequency Dictionary; Dover Publications, New York, 1961

Lexicale en syntactische codering

Van het secretariaat van de Werkgroep Frequentieonderzoek ontvingen we nog wat gegevens over de verwerking van het materiaal: het zal "lexicaal" en gedeeltelijk ook "syntactisch" gecodeerd worden. Dit laatste gedeeltelijk wegens het arbeidsintensieve karakter ervan.
De lexicale code bestaat uit drie cijfers achter het woord. De eerste twee duiden de woordsoort aan en geven soms al enige syntactische informatie; het tweede cijfer is een nadere specificatie van het eerste. Het derde cijfer biedt morfologische categorieën. Een voorbeeld:

Alle	(1) pronomina B	(2) indefinita bijv.	(3) verbogen
vogels	substantief	gewoon subst.	meervoud
vliegen	persoonsvorm	intransitief	pres. plur.

De syntactische code is gecompliceerder. Men kan er o. m. mee aangeven de zinsdelen als onderwerp en gezegde, onderdelen van het gezegde, (onderdelen van) bijwoordelijke bepalingen enz., en structuurverschijnselen als nevenschikking, inbedding, disjunctie e. d.
Het secretariaat van de werkgroep (p/a Keizersgracht 569-571, Amsterdam)