Over enkele betrekkingen tussen linguïstiek en statistiek

door

W. MARTIN

0. INLEIDING

Wij leven in een tijd waarin interdisciplinaire research, ook in de humane wetenschappen, veld wint.

De linguïstiek b.v. heeft in samenwerking met andere disciplines als de psychologie en de sociologie nieuwe gebieden voor onderzoek opengesteld, denken we maar aan de thans zo bloeiende psycho- en sociolinguïstiek.

Daarnaast is er, nu ongeveer een 25 jaar geleden, o.m. ook een toenadering ontstaan tussen linguïsten en statistici. Deze wederzijdse toenadering groeide stilaan uit tot de zg. kwantitatieve of statistische taalkunde.

Al bij al bleef dit echter werk van individuelen, van enkelingen.

Dat ik vandaag over de relatie linguïstiek-statistiek wens te spreken heeft dan ook te maken met de opvatting dat een verdere, bredere toenadering tussen linguïsten en statistici zeker voor de eerstgenoemden een vruchtbare inbreng in hun wetenschappelijk onderzoek zou betekenen. Verder ben ik zo vrij te veronderstellen dat de statistiek niet louter als dienstmeid voor de taalkunde hoeft te fungeren, maar dat er zich geleidelijk aan, na wederzijds contact en studie, een soort taalstatistiek kan ontwikkelen, uitgerust niet alleen met de klassieke maar ook met specifieke, eigen, aangepaste, methodes die uiteindelijk zowel op de praktijk als op de theorievorming van de statistiek in het algemeen, hun weerslag zouden kunnen hebben.

Dit laatste aspect is voor het ogenblik echter nog te weinig realis te veel potentialis. Ik zal mij dan ook beperken tot het schetsen, vanuit de taalkundige hoek, van enkele algemene, grote raaklijnen die, m.i., beide disciplines nu reeds met elkaar verbinden.

1. STATISTISCHE REGELS IN DE TAAL

Definities voor om het even welke wetenschapstak zijn legio. Ze variëren vaak van auteur tot auteur, naar gelang hij bepaalde aspecten wil beklemtonen. Bij J. Hemelrijk b.v. lezen we dat

„statistiek toegepaste waarschijnlijkheidsrekening is" [1]. Aanvaarden wij binnen deze context Hemelrijks definitie dan betekent dit o.m. dat enkel de fenomenen of gebeurtenissen die zich aan een puur deterministisch systeem onttrekken, door een statistische analyse gebaat kunnen worden. Statistiek immers houdt zich, om het paradoxaal uit te drukken, met de *wetmatigheid* van het *toeval* bezig.

Causaal bepaalde gebeurtenissen daarentegen zijn statistisch irrelevante stof. M.a.w. waar oorzaak en gevolg vastliggen, waar het zg. toeval geen rol speelt, is alles zeker, niets waarschijnlijk.

Zijn er echter in de taal wel statistische regels, d.w.z. regels die niet volledig determinerend werken ? [2] Is het niet veeleer zo dat het concept 'taalregel' reeds impliciet de notie van determinisme insluit, m.a.w. dat, wanneer voldaan wordt aan een aantal linguïstische voorwaarden, een bepaald taalfeit noodzakelijk moet volgen ? [3]

Dergelijke deterministische regels bestaan zeer zeker in de taal. Denken we b.v. aan het stemloos worden van stemhebbende occlusieven op het einde van een woord voor een pauze in het Nederlands. Dit kunnen we aldus formaliseren :

$$\begin{bmatrix} + \text{ occlusief} \\ + \text{ stem} \end{bmatrix} \longrightarrow [- \text{ stem}] \; / \; - \; \#$$

wat we lezen als : een foneem dat de kenmerken stemhebbend en occlusief heeft wordt stemloos wanneer het in de context komt aangegeven na de schuine streep. — duidt dan de plaats aan waar het foneem wordt aangetroffen, i.c. vóór # dit is op het einde van een woord vóór een pauze. Aldus kunnen alternanties als paard [t] – paarden [d] en heb [p] – hebben [b] worden verklaard [4].

1. J. Hemelrijk, *Statistiek en Practijk,* Amsterdam, Mathematisch Centrum, 1954, p. 1.

2. Over het statistische regelconcept zie ook R. M. Frumkina in *Exact Methods in Linguistic Research,* Berkeley, University of California Press, 1963, p. 80, v.v.

3. Daarme wil ik niet het bestaan van optionele regels in de taal negeren. Bij statistische regels gaat het echter om de *juistheid van het resultaat dat onzeker is,* (bij toepassing van de regels volgt niet noodzakelijk een correct resultaat), bij optionele regels is men onzeker *omtrent het resultaat, niet omtrent de juistheid ervan,* (het is niet zeker dat men b.v. op een bepaalde dieptestructuur een (oppervlakte-)transformatie gaat toepassen : naast *Ik sla Jan,* is ook b.v. *Jan wordt door mij geslagen* mogelijk).

4. De t/d resp. p/b alternantie wordt dus afgeleid van één enkel foneem. Of dit laatste nu t of d (resp. p of b) (of eventueel een foneem dat van beide verschilt) is, wordt bepaald door de graad van eenvoud die deze keuze voor het grammaticaal systeem impliceert. In ons geval wordt meestal voor de stemhebbende consonant als onderliggende basisvorm geopteerd.

M.a.w. als een bepaald foneem voldoet aan de verzameling voorwaarden V
nl. – een aantal karakteristieken : k (stemhebbende occlusief)
– en een bepaalde context c (i.c. vóór #)
volgt automatisch R, d.i. een voorafgekend resultaat of uitkomst (i.c. verlies van stemhebbendheid).

Zulke eenvoudige, volledig gedetermineerde, taalregels zijn echter zeldzaam. Meestal functioneren de taalobjecten onder invloed van heel wat meer factoren en is het praktisch uitgesloten met alle rekening te houden om R, het resultaat van hun interactie, vooraf te bepalen.

Een typisch voorbeeld hiervan is het gebruik van het lidwoord vóór een naamwoord in het Engels. Hoewel de keuze beperkt is, – formeel gezien kunnen alleen *the, a* en *ø* (het lidwoord „nul" = ontbreken van een lidwoord) optreden –, toch blijkt het alvast heel moeilijk te zijn om het complex van voorwaarden die de keuze van een passend lidwoord bij een naamwoord in het Engels bepalen, op een niet-ambiguë, algemeen-geldende wijze te beschrijven.

Wanneer wij een grammatica van het Engels naslaan [5] dan vinden wij niet langer een bondige strikte regel, wel ettelijke bladzijden met voorschriften, voorbeelden en uitzonderingen. Zelfs als we eraan denken onze analyse zo ver mogelijk door te drijven en geen complicatie uit de weg te gaan, dan nog zijn we niet zeker dat we ieder geval correct zullen oplossen. Een paar voorbeelden mogen dit verduidelijken.

Eén van de opgegeven regels luidt als volgt : „Plaatsnamen die enkel uit één eigennaam bestaan hebben *gewoonlijk* geen artikel, behalve wanneer het namen van rivieren, zeeën, bergketens en eilanden betreft". Aldus vindt men in het Engels : *Japan, Roumania* en *Belgium,* maar men zegt en schrijft *The Sudan* and *The Crimea.* Zo ook vindt men plaatsnamen als *Snowdon* en *Radnor* zonder artikel maar *The Wrekin* en *The Cheviot* met het lidwoord *the.*

In dezelfde grammatica lezen we verder dat : „Voor plaatsnamen van het type eigennaam + soornaam het lidwoord gebruikt wordt wanneer de soortnaam nog als dusdanig *gevoeld* wordt, *vooral* wanneer het om een meervoudsvorm gaat". Voorbeelden : *The Sharpness Bridge,* maar *London Bridge.* „Streams of traffic were reported on *the Hastings Road, the Folkestone Road,* and on *the Oxford Road*" ; maar ook *Grosvenor Road. The Black Mountain* staat tegenover *Fish Street Hill, the Yorkshire Moors* tegenover *Trinity Square Gardens,* enz., enz.

5. De geciteerde voorbeelden en regels zijn ontleend aan : G. Scheurweghs, *Present-Day English Syntax,* London, Longmans, 1959, p. 100-101.

Uit de reeks voorbeelden, uit de annotaties *gewoonlijk* en *vooral* en de rol toegekend aan het *gevoel,* kan men reeds opmaken dat het weinig zin zou hebben te streven naar exhaustief-determinerende regels : er bestaat immers geen exhaustieve lijst van regels die de lidwoordkeuze vóór een naamwoord definitief bepalen in het Engels. In dit geval kunnen wij niet anders dan verder gaan met het maken van 'fouten', d.w.z. dat gegeven de set van voorwaarden die R d.i. de keuze tussen *a, the* of *ø* bepalen, wij de mogelijkheid open laten dat R verkeerd is, of :

$$\left\{\begin{array}{c} V_1 \\ | \\ | \\ | \\ | \\ V_n \end{array}\right\} \rightarrow R = \left\{\begin{array}{c} a \\ the \\ \emptyset \end{array}\right\} + N(oun)\ ^{6}$$

De taal is dus geen volledig deterministisch systeem, wel iets systeemachtigs, een systemoïde. In het geval van de keuze van een lidwoord bij een naamwoord in het Engels zouden wij het zo kunnen stellen dat er een aantal algemene regels zijn (V), daarnaast echter een aantal buiten-de-regels-vallende, specifieke, individuele gevallen. Wanneer wij dus een volledig deterministisch complex van regels zouden willen opstellen, zouden wij over een exhaustieve lijst van die zg. on-regelmatige gevallen moeten beschikken. Precies omwille van het gebrek aan systeem, aan regelmaat is regel-vorming hier uitgesloten. Verder ook al omdat de woordgroep *A(rticle)* + *N(oun)* in se infiniet is. De klasse N (= noun) is nl. een zg. open woordklasse, d.w.z. dat het aantal naamwoorden in een taal als het Engels niet te begrenzen is, daar de mogelijkheid tot samenstelling en afleiding a.h.w. als een perpetuum mobile werkt : samenstellingen kunnen op hun beurt input zijn voor nieuwe samenstellingen, deze op hun beurt voor andere, enz., enz.

Wanneer wij nu constateren dat sommige taalfenomenen niet via strikte, eenvoudige, deterministische regels te vatten zijn betekent dit nog niet dat zij helemaal aan linguïstische descriptie ontsnappen.

Als wij een bepaalde set V (voorwaarden tot het bekomen van een linguïstisch object) hebben opgesteld zullen wij merken dat er een connectie bestaat tussen V en R. V zal niet determinerend werken in de zin dat de individuele uitkomst R automatisch is gekend, wanneer we echter onze set van regels toepassen op een

6. De aangehaalde voorbeelden slaan op een subklasse van de N(ouns) nl. de Proper Names.

grote hoeveelheid tekst zullen we merken dat het aantal gevallen dat een juiste R geeft de meerderheid vormt als onze set V goed geformuleerd is. Meer zelfs, als we V toepassen op verschillende materialen van ongeveer dezelfde grootte dan zullen we steeds een ongeveer gelijk aantal juiste en foute R's aantreffen. Er is dus wel degelijk een verband tussen V en R : niet zo dat V automatisch de juiste R oplevert, maar wel dat V in b.v. 90 % der gevallen de juiste R geeft. Er is in dit geval dan 90 % kans om, gegeven zijnde V, een juiste R te krijgen. We kunnen in zo'n geval beweren dat het voorkomen van een foute beslissing een *random of toevalsgebeurtenis* is : door het toeval alleen reeds staat het vast dat b.v. ongeveer 10 % van de R's bij een gegeven V verkeerd zullen zijn. Dit betekent dus dat er in de taal statistische regels zijn d.w.z. regels die enkel een probabiliteitswaarde hebben en dus een waarschijnlijkheid, geen zekerheid garanderen.

M.a.w. bepaalde taalfenomenen kunnen wij niet volledig voorspellen, anderzijds echter vertoont deze onvoorspelbaarheid toch een zekere regelmaat : in het geval van de juiste keuze van een lidwoord bij een naamwoord weten wij b.v. hoe groot de kans is dat wij ons bij een gegeven V aan een foute R mogen verwachten.

De studie van random gebeurtenissen, feiten die enkel door het toeval verklaarbaar zijn, die m.a.w. niet onder invloed van een bepaalde determinerende factor vallen vormen precies het onderwerp van de waarschijnlijkheidsrekening en de statistiek.

Het is dan ook evident dat, in de mate dat bepaalde fenomenen in de taal als random (of) toevalsgebeurtenissen kunnen worden beschouwd, de probabiliteitstheorie en de statistiek de taalstudie van groot nut kunnen zijn.

2. TAALGEBRUIK EN PROBABILITEITSSTRUCTUUR VAN DE TAAL

Tot nog toe hebben wij gesproken over taalregels en taal (deze laatste term gebruikten we als synoniem voor taalsysteem). In feite zijn dat abstracties. Wat wij direct observeren zijn de concrete realisaties van die systematiek, van die regels, wat wij zouden kunnen noemen : het taalgebruik of de teksten, hetzij het nu om geschreven, gedrukte of gesproken teksten gaat. Wij zouden het dan ook zo kunnen stellen dat het concrete taalgebruik, de taalrealisaties, tegelijk input en output [7] voor het taalsysteem zijn. Input : zij zijn het uitgangspunt, de beginwaarneming, nodig om te komen tot een regel, een wet of een model ; output : eenmaal

7. Althans gedeeltelijk. Het taalsysteem zoekt niet alleen een verklaring te bieden voor de geobserveerde teksten, maar tevens voor de *oordelen* van de taalgebruikers omtrent identiteit, welgevormdheid, grammaticaliteit e.d. van uitingen.

een regel, wet of model is opgesteld kunnen de concrete realisaties, de taalmanifestaties, daaruit afgeleid en voorspeld worden. In dit opzicht lag de toenadering tussen de taalkunde en de wiskunde, met name vooral de verzamelingenleer, voòr de hand. Wanneer wij vanuit het taalgebruik abstracties maken naar het taalsysteem toe dan merken wij dat dit systeem, of beter uit het taalgebruik geabstraheerde elementen van dit systeem, niet zo maar een toevallige verzameling vormen maar een gestructureerd geheel zijn waarin de elementen in bepaalde verhoudingen tot elkaar staan zodanig zelfs dat deze relaties de innerlijke structuur van het geheel mee helpen bepalen.

Een voorbeeld : in het Nederlands is er een differentiatie tussen de klinkers *ie* zoals in *viel* en *oe* zoals in *voel*. Het verschil is zelfs zo dat wanneer ik *ie* door *oe* vervang in de context v-l er dan een ander woord ontstaat. Tevens bestaat er een onderscheid tussen *ie* in *viel* en *ie* in *vier* (men kan o.m. vaststellen dat de tweede *ie* langer is dan de eerste), hier echter ontstaat er geen betekenisverschil wanneer ik de 2 *ie's* verwissel.

Wij zouden dus kunnen zeggen dat :

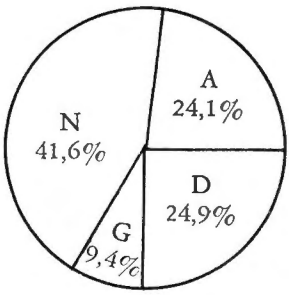
$$\begin{array}{llll} ie_1 & \varepsilon & IE & \\ ie_2 & \varepsilon & IE & \\ oe & \varepsilon & OE & \\ IE & \cap & OE & = \ \varnothing \end{array}$$

Wat wij dan kunnen interpreteren als 3 verschillende elementen ie_1, ie_2 en *oe* die echter tot slechts 2 verschillende verzamelingen IE en OE behoren. Samenvattend zouden wij dus kunnen stellen dat de taalsystematiek (althans gedeeltelijk) haar uitgangspunt vindt in het taalgebruik, het taalmateriaal, en vice versa dat vanuit deze systematiek, het taalmateriaal kan worden afgeleid. De tweede helft van deze bewering blijkt echter onvolledig te zijn en brengt ons meteen op ons thema : de relatie linguïstiek-statistiek terug. Het taalsysteem, de taal, is immers niet het enige wat uit het taalmateriaal kan worden afgeleid, en omgekeerd, uit het taalsysteem kan niet alles over het taalgebruik worden geconcludeerd. Als we b.v. het Duitse casus-stelsel onderzoeken, zien we dat er 4 naamvallen zijn in het Duits d.w.z. dat deze naamvallen een gelijke positie in het taalsysteem innemen : nominatief, genitief, datief en accusatief kunnen in het Duits ten opzichte van elkaar worden onderscheiden en voorkomen bij substantieven en hun vervangers. Wij kunnen echter uit het taalmateriaal meer opmaken dan dat.

De frequentiële belasting van deze 4 casus b.v. is geenszins gelijk : de nominatief wordt veruit het meest gebruikt, de genitief

het minst, de datief en accusatief hebben ongeveer een gelijke frequentie, maar komen voor bijna de helft minder voor dan de nominatief, enz.

De frequentieverdeling voor de 4 naamvallen in het Duits ziet er volgens H. Meier [8] immers als volgt uit :



Dit betekent dus dat bepaalde karakteristieken van het taalgebruik niet uit het taalsysteem kunnen worden afgeleid, het Duitse taalsysteem b.v. leert ons wel welke naamvallen in het Duits kunnen onderscheiden worden en in welke context ze voorkomen, niet echter welke de frequentiële belasting is van deze taalobjecten. Precies deze abstrahering van het taalmateriaal die niet door het taalsysteem wordt ondervangen zouden we de probabiliteitsstructuur van de taal willen noemen.

Daarmee zijn we terug in een typisch statistische situatie terechtgekomen. Het taalgebruik is per definitie niet of heel moeilijk exhaustief te capteren, ook als we het beperken tot het taalgebruik van één gemeenschap, één periode, één genre, één individu. Er wordt nu geconstateerd (cf. het voorbeeld voor het Duits) dat het taalgebruik een frequentiële structuur bezit. Dit is o.m. een gevolg van het feit dat het regelsysteem in de taal eindig is en dat er zich derhalve bepaalde elementen uit dit systeem herhalen : het complex van de relatieve frequenties van deze onderscheiden elementen noemen we de probabiliteitsstructuur van de taal (of subtaal die we onderzoeken). De sprong van frequentie naar probabiliteit, van taalgebruik naar taal is analoog aan de overgang in de statistiek van sample naar populatie.

Immers niet hét Duits is ons bekend, wel samples uit het Duits. Niet hét 17-eeuwse Nederlands, maar sommige teksten uit de 17e eeuw in het Nederlands geschreven. En zelfs wanneer we bij het œuvre van één auteur blijven, bestrijken we nog maar een deel van zijn populatie, i.c. van zijn taalgebruik : slechts een klein gedeelte nl. datgene wat van hem gedrukt en uitgegeven werd is ons in het gunstigste geval bekend.

8. H. Meier, *Deutsche Sprachstatistik,* Hildesheim, G. Olms, 1964, p. 260.

Dit sluit weliswaar niet in dat alle taalobjecten een stabiele relatieve frequentie zouden bezitten, het is juist de taak van de kwantitatieve taalkunde uit te maken welke taalobjecten wel en welke niet daarvoor in aanmerking komen. G. Herdan [9] heeft dan ook op enkele van deze linguïstische random variabelen gewezen o.m. op de frequentiedistributies van grafemen, fonemen en woordordestrings. Bij deze taalobjecten convergeert de relatieve frequentie naar een ideale waarde die wij de probabiliteit van voorkomen van dit fenomeen noemen. Relatieve frequentie en probabiliteit vallen nooit samen maar het verschil tussen beide wordt kleiner als we het aantal waarnemingen verhogen, wat door de onderstaande formule wordt uitgedrukt [10] :

$$\underset{\text{als}}{[\,|\, f_n(A) - p(A) \,|\, > \varepsilon\,]} \underset{n \longrightarrow \infty}{\longrightarrow} 0$$

waarbij $f_n(A)$ de relatieve frequentie van het fenomeen A in een reeks uitkomsten van grootte n aangeeft, p(A) diens (abstracte) probabiliteitswaarde, en ε een arbitrair kleiner getal is. Dit tweede punt zouden wij aldus kunnen besluiten : daar er in het taalgebruik random variabelen bestaan, d.w.z. variabelen die voor hun mogelijke waarden een vaste probabiliteit van voorkomen hebben, kunnen wij via samples (teksten, taalmateriaal) tot conclusies komen omtrent de populatie (taal). Welnu, daar „statistical methods may be described as methods for drawing conclusions about populations by means of samples" [11] ligt het dus voor de hand dat de linguïstiek bij de extrapolatie van frequentie in taalgebruik naar probabiliteit in taal gebruik gaat maken van de statistiek.

3. STIJL ALS STATISTISCH CONCEPT

De laatste jaren heeft er zich een sterke toenadering tussen taal- en literatuurstudie voorgedaan. Met name de stilistiek, de stijlstudie, is een gebied geworden waar taalkundigen en literatuurwetenschappers elkaar ontmoeten. Wat moeten wij echter onder „stijl" verstaan ?

Het zou ons te ver voeren om alle mogelijke definities van dit begrip op te sommen en te bespreken. Er is inmiddels echter een literaire statistiek gegroeid – ook wel computationele stilistiek

9. G. Herdan, *Quantitative Linguistics,* London, Butterworths, 1964, p. 5 v.v.

10. Zie ook mijn artikel *The rules of the art,* in *ITL* (Review for Applied Linguistics), 12, (1971), p. 69.

11. P. G. Hoel, *Elementary Statistics,* New York, Wiley, 1971[3], p. 2.

12. De idee van de statistische definiëring van stijl heb ik o.m. ook naar voor

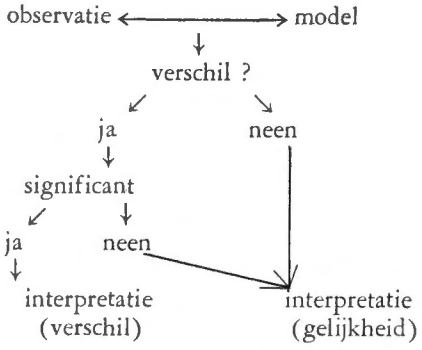
genoemd [12] – die uitgaat van de opvatting die wij in punt 2 hebben geschetst, nl. dat (althans sommige) taalelementen een vaste waarschijnlijkheid hebben, eigen aan de bestudeerde taal. Er wordt nochtans geconstateerd dat van deze waarschijnlijkheid in een concreet „sample" afgeweken kan worden. M.a.w., naast het (probabiliteits-) model van de taal postuleert deze methode de vrijheid van het individuele taalgebruik. De individuele afwijking kunnen wij dan als een stijlelement interpreteren. De norm, het model waaraan de deviaties kunnen worden getoetst, kan worden geapproximeerd d.m.v. het gemiddelde taalgebruik.

Laten wij dit even verduidelijken. Een natuurlijke taal is in feite een ambigu systeem. Dit wil o.m. zeggen dat er in de taal verschillende manieren mogelijk zijn om hetzelfde te zeggen : flexie, woordorde, synonymie, het gebruik van functiewoorden zijn vaak zo vele keuzemogelijkheden in onze taal. Er verandert niets aan de informatie als ik spreek over „de auto van Jan" of over „Jans auto", over „het gebruik van woorden als substantieven" of „het substantivische gebruik van woorden"... Door deze keuze is fluctuatie van de norm, de background waartegenover de gemiddelde lezer het taalprodukt in kwestie afweegt, zeer goed mogelijk. Theoretisch gezien is deze norm het gemiddelde taalgebruik van alle sprekers van die taal. Deze norm kan op haar beurt onderverdeeld en verder genuanceerd worden in sub-normen die rekening houden met tijd, genre, situatie, enz. Dit betekent dus dat wij de auteur of de taalgebruiker zien als enerzijds onderworpen aan bepaalde normen of conventies, anderzijds ervan bevrijd, omdat zovele regels optioneel zijn. Juist door deze keuzemogelijkheid is er afwijking van de norm mogelijk en kan de auteur een bepaald effect nastreven. Is deze fluctuatie (statistisch)-significant dan kunnen wij spreken van een stijlverschil, tenminste indien wij van een (stilistisch) representatieve norm kunnen uitgaan. Meestal ligt precies hier de grote moeilijkheid : genuanceerde frequentietellingen waarbij o.m. rekening wordt gehouden met onderwerp, situatie (dialoog of monoloog b.v.), structuur (proza, poëzie, genres), gesproken of geschreven taal, milieu of streek van de auteur (b.v. Zuid- tgo. Noord-Nederland) enz. en die de probabiliteit van diverse linguïstische objecten als fonemen, morfemen, woorden, woordgroepen en zinnen weergeven, bestaan immers vrijwel nog niet. Daarbij laten wij in het midden of deze probabiliteit of relatieve frequentie overal een even grote waarde toegekend kan worden. Hoe dan ook zolang die tellingen er niet zijn zal het

gebracht in mijn *Analyse van een vocabularium met behulp van een computer,* Brussel, AIMAV, 1970, p. 14 v.v.

moeilijk blijven de literatuur- en stijlstudie exacter en objectiever te maken.

Samengevat betekent dit alles dat wij *stijl* zien als de som van de individuele opties die de auteur maakt van de lexicale, morfologische en syntactische keuzemogelijkheden in de taal. Een typerend stijlkenmerk voor een auteur sluit dan in dat een of andere optie significant afwijkt van de verwachting die wij ons op basis van de norm daaromtrent hebben gemaakt. Als wij b.v. de woordlengte willen evalueren in (literaire) teksten zullen wij uitgaan van een zogenaamde gemiddelde taalgebruiker en op basis daarvan een mathematisch-statistisch model bouwen dat de reflectie is van de (abstracte) karakteristiek (i.c. woordlengte) die deze taal bezit. Afwijkingen in concrete teksten kunnen dan geëvalueerd worden d.m.v. significantietoetsen. Schematisch kunnen wij dit als volgt voorstellen :



De problemen die bij kwantitatieve tekstanalyse worden voorgelegd kunnen in grote lijnen meestal op die wijze worden geschematiseerd. Uit dit schema wordt het duidelijk dat een zg. exacte methode de interpretatie-fase niet uitsluit of overbodig maakt. Het is zelfs zo dat bij deze laatste fase, de rol van de statistiek bijkomstig is. Het voordeel is echter dat nu de weg geopend wordt tot wat voor interpretatie in aanmerking komt. M.a.w. niet-significante verschillen hebben geen wetenschappelijke waarde als verschil, en worden als zodanig buiten beschouwing gelaten.

In een dergelijke optiek die stijl beschouwt als afwijking t.o.v. een norm is het duidelijk dat de statistiek als „la science des écarts" een uitgelezen middel is om bij stilistisch onderzoek gebruikt te worden.

P. Guiraud heeft ooit geschreven : „La linguistique est la science statistique type, les statisticiens le savent bien, la plupart des

linguistes l'ignorent encore" [13], wij zouden gerust hetzelfde durven beweren voor de stilistiek.

4. BESLUIT

Met deze enkele beschouwingen hebben wij proberen duidelijk te maken dat de taal en het taalgebruik een statistische structuur bezitten en dat de linguïstiek derhalve een nuttig gebruik zou kunnen maken van de statistiek. Er mag echter niet uit het oog worden verloren dat het hier slechts om enkele dimensies, enkele facetten, van de taalstudie gaat. M.a.w. om het oneerbiedig uit te drukken, wij zouden niet van een linguïst een pure getallenwichelaar willen maken. In de eerste plaats gaan zoals J. F. Schouten [14] al eerder heeft beklemtoond, „aan de quantitatieve meting (...) in beginsel enige uitermate belangrijke stadia van wetenschappelijk denken vooraf". Een conditio sine qua non voor een statistisch onderzoek is immers dat de samenstellende eenheden van de populatie volkomen gedefinieerd zijn. We kunnen nooit tot bruikbare resultaten komen wanneer niet precies geweten is, waaraan datgene dat we aan een statistisch onderzoek onderwerpen beantwoordt. M.a.w. de begrippen en grootheden die wij hanteren moeten van andere onderscheiden en met elkaar vergeleken kunnen worden. Pas na deze voorbereidende stadia, kan de kwantitatieve meting en daarbij aansluitend de statistische analyse plaatsvinden. Wij menen echter dat de taalstudie in een dusdanig stadium is gekomen dat de twee : 'kwalitatieve' en 'kwantitatieve' analyse niet na maar samen met en door elkaar tot het uiteindelijke doel van alle taalstudie kunnen komen : een dieper inzicht krijgen in het proces van spreken, schrijven en verstaan, d.i. in hun uiteindelijke object : de taal [15].

K.U. Leuven
Departement Linguïstiek
Instituut voor Toegepaste Linguïstiek
Vesaliusstraat 2, Leuven.

13. P. Guiraud, *Problèmes et Méthodes de la Statistique Linguistique,* Dordrecht, Reidel, 1959, p. 15.

14. J. F. Schouten, *De methode in de verschillende wetenschappen,* in : *De gang der gedachte ; Negen voordrachten over de methodes in de natuur- en geneeskundige wetenschappen,* Den Haag, Nijhoff, 1960, p. 118.

15. Deze tekst is een licht gewijzigde versie van een reeds eerder verschenen artikel : *Enkele betrekkingen tussen linguïstiek en statistiek,* in *NIKO* (Tijdschrift van het Belgisch Centrum voor Methodiek van de Wiskunde), Brussel, 13 (1973), 69-82. Ik sta erop Prof. Dr. L. K. Engels, Dr. R. Eeckhout, en Drs. W. Smedts hartelijk te bedanken voor de opmerkingen die zij maakten bij de twee versies.