Statistiek en tekstanalyse

door

PROF. DR. W. MARTIN

1. *Uitgangspunt*

In één van zijn talrijke uiteenzettingen over mathematische taalkunde heeft H. Brandt Corstius ooit eens beweerd dat „iedere taalkundige een elementair begrip van de statistiek zou moeten hebben". Tijdens de bloeitijd van het transformationalisme mag deze uitspraak dan wel als een boude bewering hebben geklonken, vandaag de dag hebben hoe langer hoe minder linguïsten er moeite mee. Misschien liggen de zaken (nog steeds) anders als we bovenstaande uitspraak zouden transponeren naar het domein van de tekststudie, meer in het bijzonder naar de studie van literaire teksten. Vandaar dat wij, binnen het bestek van deze korte bijdrage, zouden willen aantonen dat „iedereen die zich bezig houdt met tekststudie een elementair begrip van de statistiek zou moeten hebben". Anders geformuleerd : wij zouden willen duidelijk maken dat een statistische tekstanalyse één van de sleutels is om teksten te openen. Tenslotte, meer bescheiden uitgedrukt : aan de hand van een paar voorbeelden zouden wij graag een inzicht willen geven in het soort inzichten dat een statistische tekstanalyse te bieden heeft.

2. *Definities, Basishypotheses, Strategieën*

Ursula Pieper parafraserend, kunnen wij een tekst definiëren als „das Produkt eines Prozesses der Sprachverwendung wodurch unterschiedliche Informationen vermittelt werden. Das bedeutet, dass dem Text eine kommunikative Funktion zukommt wobei der jeweilige Zweck der kommunikativen Funktion den Prozess der Texterstellung steuert" [1].

Centraal in deze definitie is het feit dat een tekst een verzameling is die bestaat uit één of meer linguïstische data (Produkt eines Prozesses der Sprachverwendung) en dat deze verzameling geen geïsoleerd, op zichzelf staand, fenomeen is maar ingeschakeld is in een communicatieproces, m.a.w. gerelateerd is aan een zender (spreker, schrijver) en een ontvanger (hoorder, lezer). Binnen het kader van een statistische tekstontleding krijgen teksten daaren-

1. U. Pieper, Differenzierung von Texten nach numerischen Kriterien, *Folia Linguistica,* VII, 1975, 62-63.

boven nog één belangrijk kenmerk meer toegewezen. Ook hier zijn teksten verzamelingen van één of meer linguïstische objecten met een communicatieve functie, daarenboven wordt er van uitgegaan dat deze objecten recurrent zijn, m.a.w. dat zij ook in andere teksten voorkomen of kunnen voorkomen en dat deze recurrentie niet toevallig is maar gestructureerd. M.a.w. een statistische tekstontleding gaat er van uit dat een tekst bestaat uit één of meer talige objecten die functioneren binnen een communicatieproces en (waarvan er althans sommige) daarenboven een voorkomensprobabiliteit bezitten. In die zin is een statistische tekstanalyse te definiëren als het soort tekstonleding dat zich beroept op de basisprincipes van de kwantitatieve of statistische taalkunde.

Het is hier niet de plaats om op de gegrondheid van deze principes in te gaan, deze belangrijke kwestie hebben wij in andere publikaties uitvoerig behandeld [2], wel lijkt het ons nuttig om vanuit de kwantitatieve taalkunde de notie kwantitatieve of statistische tekstanalyse nader te belichten.

In een eerste benadering zou men kunnen stellen dat een kwantitatief taalkundige zich interesseert voor de frequentie van linguïstische objecten (b.v. fonemen, morfemen, woorden enz.) of van karakteristieken van deze objecten (b.v. duur, lengte, categorie enz.), m.a.w. niet alleen *wat* voorkomt, maar ook *hoe vaak* iets voorkomt vindt een kwantitatief linguïst belangrijk. Een eerste conclusie m.b.t. een statistische tekstanalyse zou dan kunnen zijn dat een dergelijke analyse in hoofdzaak neerkomt op het vaststellen van de *frequentie* van bepaalde linguïstische fenomenen in een tekst. Deze conclusie is misleidend en foutief : het gaat bij kwantitatieve taalkunde en tekststudie niet in de eerste plaats om *observatie van linguïstische frequenties,* wel om de *waarschijnlijkheid* waarmee deze linguïstische fenomenen kunnen voorkomen. Concreet : het gaat niet (enkel) om het vaststellen dat een woord W n keer voorkomt in tekst T, wel om het vaststellen van de frequentie waarmee W in analoge, niet onderzochte teksten T', T'' enz. zou kunnen voorkomen.

2. Zie o.m. W. Martin, Over enkele betrekkingen tussen linguïstiek en statistiek, *Handelingen der Kon. Zuidned. Mij. voor Taal- en Letterkunde en Geschiedenis,* 27, 1973, 231-241.

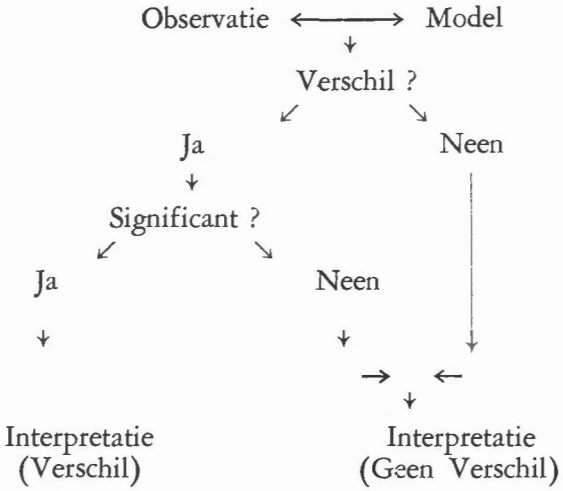
Random-Variablen, in : H. Stammerjohann (Hrsg.), *Handbuch der Linguistik,* München, 1975, 351-352.

Taal en Tal, *Handelingen van het 34ste Nederlands Filologencongres,* 1976, 179-197.

Preliminaire Opmerkingen over Kwantitatieve Lexicologie, in : P. van Sterkenburg (Uitg.), *Lexicologie, opstellen voor F. de Tollenaere,* Groningen, 1977, 189-196.

Möglichkeiten und Grenzen der quantitativen Linguistik beim Studium der wissenschaftlichen Fachsprachen, in : T. Bungarten (Hrsg.), *Wissenschaftssprache,* München, ter perse.

Het belangrijkste onderscheid tussen kwantitatieve taalkunde enerzijds en kwantitatieve of statistische tekstanalyse anderzijds bestaat er dan in dat eerstgenoemde tot probabiliteitsmodellen voor linguïstische fenomenen probeert te komen, terwijl de tweede zich interesseert voor de „reactie" van concrete teksten t.o.v. dergelijke modellen. M.a.w. de strategie gevolgd bij een statistische tekstanalyse kan men als volgt schematiseren [3] :



Bij een statistische tekstanalyse wordt de frequentie van een bepaald taalfenomeen in een concrete tekst vastgesteld en met de verwachting die men daaromtrent heeft geconfronteerd. Het eventuele verschil tussen geobserveerde en geëxpecteerde waarden wordt daarna op zijn significantie getoetst, d.w.z. men gaat na of dit verschil al dan niet nog door het toeval verklaarbaar is. N.g.v. de uitkomst zal het bestudeerde fenomeen ook verschillend geduid worden. Samengevat : een statistische tekstanalyse werkt niet noodzakelijk met numerieke fenomenen wel met probabiliteitswaarden (van kwalitatieve of kwantitatieve fenomenen). Aan de hand van deze cijfers probeert men dan tot uitspraken te komen die de tekst in kwestie moeten typeren. Hierbij kan men o.i. twee grote mogelijkheden onderscheiden.

3. *De twee wegen bij een statistische tekstanalyse*

Bij een statistische tekstanalyse kan men in feite twee kanten uit :

3. Zie onze lezing vijf jaar geleden gehouden voor de Zuidnederlandse Mij., afd. Taalkunde (*Handelingen*, 27, 1973, 240).

- *ofwel* gaat aan het zoëven voorgestelde schema een *bepaalde, voor de tekst in kwestie z.g. relevante intuïtie of hypothese vooraf.* De bekomen cijfers dienen dan precies om deze intuïtie te toetsen (te verwerpen of te aanvaarden) ;
- *ofwel* gaat men *niet van een bepaalde teksthypothese uit,* maar veeleer van een pure kwantitatieve vraagstelling (b.v. bevat fragment x evenveel woorden als de fragmenten y en z ?), of van een statistische hypothese. De cijfers dienen dan niet zozeer om teksthypotheses of -intuïties te toetsen, maar wel om deze te genereren.

De eerste mogelijkheid is bij tekstanalyse en, m.n. bij analyse van literaire teksten, vrij goed bekend. Gewoonlijk echter wordt ze als iets secundairs aangezien, als een confirmatie van wat men van te voren reeds wist, iets in de aard van „dat wisten we reeds, maar het is niet kwaad dat ook nog eens met cijfers gestaafd te zien". Een typische uitspraak in dit verband is b.v. deze van René Wellek (in een commentaar op onderzoek van Seymour Chatman over de vergelijking van de satiren van John Donne en imitaties ervan door Alexander Pope). Schrijft Wellek : „Mr. Chatman's paper makes a careful comparison between the satires of Donne and their revision by Pope. He seems to me in general correct, although his method may be excessively cumbersome (Chatman maakt gebruik van kwantitatieve methodes) : we could easily predict that Pope would regularize the meter of Donne's lines, would introduce endstops, avoid harsh consonant clusters, and so on, just from our knowledge of the poetics of Pope's time, of the whole change from the Baroque to Neoclassicism. But it is good to see it demonstrated in such detail" [4].

De tweede mogelijkheid waarbij men geen teksthypothesen toetst, maar deze poogt te genereren lijkt van haar kant dan weer uitzichtsloos, langdradig, of riskant. Dat dit geenszins zo hoeft te zijn zouden wij in wat volgt willen aantonen.

4. *Mei 1*

Hierboven werd betoogd dat men bij een statistische tekstanalyse *twee kanten* uitkan :

- ofwel gebruikt men statistiek om *a-priori hypothesen te toetsen* ;
- ofwel gebruikt men statistiek om *hypothesen a posteriori te genereren.*

4. R. Wellek, Closing statement from the viewpoint of literary criticism, in : T. A. Sebeok (Ed.), *Style in Language,* Cambridge, Mass., 1960, 413.

Zoals gezegd gaat men vooral aan de tweede mogelijkheid voorbij.

In wat volgt zullen wij aan de hand van dezelfde tekst, en van hetzelfde object uit die tekst (t.w. rijmwoorden) beide mogelijkheden summier demonstreren en illustreren [5].

In beide gevallen gaat het om *Mei* van Gorter, in beide gevallen gaat het tevens om functionele aspecten van de rijmwoorden in deze tekst.

Bij het eerste onderzoek wordt van een vooraf bepaalde tekst-hypothese uitgegaan, nl. van de veronderstelling dat de woordrijkdom in *Mei* in feite zeer sterk door het rijmschema (binair rijmende verzen) werd beïnvloed.

Zoals men weet publiceerde Herman Gorter in 1889 zijn lyrische epos *Mei* waarin hij het leven en de dood van Mei beschrijft : Mei verrijst uit de zee terwijl haar zuster April op sterven ligt, zij dwaalt door Nederland en is gelukkig om alle zinnelijke schoonheid die zij ervaart. Na een tijdje echter probeert zij het puur esthetisch-sensitieve te overstijgen en iets of iemand (in het gedicht door Balder voorgesteld) te bereiken die duurzaam en blijvend is. Weldra echter ondervindt zij dat dit doel niet in haar bereik ligt. Ontgoocheld en droef sterft zij, opgevolgd door haar zuster Juni.

Dit gedicht telt 32.235 woordtekens waarvan 5.665 verschillende (z.g. woordtypes), het is vrijwel volledig in binair rijmende verzen geschreven. Als men weet dat een „normale" prozatekst van ongeveer dezelfde lengte zowat 4.000 woordtypes telt dan ligt het voor de hand *Mei* qua vocabularium als rijk en abundant te typeren. Een hypothese die hierbij ter verklaring kan gebruikt worden is dat deze woordrijkdom (het excessieve aantal verschillende woorden voor een tekst van dergelijke lengte) grotendeels te wijten is aan het feit dat het om een binair rijmend gedicht gaat. Uitgaand van deze hypothese kan een statistische tekstanalyse deze stelling als volgt argumenteren :

a. Stel dat de tekst verminderd wordt met 4.380 woordtekens (= woorden) ; dit is namelijk het totale aantal binaire rijmwoorden en dus ook binaire verzen in de tekst. In dergelijk geval verwachten wij een vermindering van het vocabularium met 520 woordtypes (= verschillende woorden). Dit aantal bekomen wij door een beroep te doen op het binomiale model. Dit model simuleert een reductie van de tekst d.m.v. een toevals-

5. Voor een uitvoeriger behandeling van de eerste mogelijkheid zie W. Martin, Some quantitative vocabulary aspects of a Dutch poem, in : A. J. Aitken, R. W. Bailey & N. Hamilton-Smith (Eds.), *The computer and literary studies,* Edinburgh, 1973, 61-68.

steekproef. M.a.w. indien wij een at random getrokken steekproef van 4.380 woordtekens uit onze tekst van 32.235 woorden zouden weglaten, (en dus een tekst van 27.855 woordtekens zouden hebben) verwachten wij een reductie van 520 woordtypes t.o.v. het oorspronkelijke vocabularium van 5.665 (wij verwachten dus 5.145 verschillende woorden in onze gereduceerde tekst). Een dergelijke toevalssteekproef verkrijgen wij b.v. wanneer wij per vers elk tweede woord weglaten, dit laatste resulteert in een verlies van 547 woordtypes.

b. Stel dat wij onze tekst vervolgens verminderen met de rijmwoorden zelf : opnieuw laten wij nu 4.380 woordtekens weg (= het aantal rijmwoorden). Dit resulteert nu echter in een reductie van 1.349 woordtypes. T.o.v. de verwachting (gebaseerd op de stelling dat er geen verband bestaat tussen de rijmwoorden en het aantal verschillende woorden in de tekst) is dit een surplus van 829 woordtypes ! Waar we in een prozatekst van dezelfde lengte gemiddeld ongeveer 4.000 woordtypes verwachten kunnen we dus stellen dat zowat de helft van het woordensurplus dat we in *Mei* vinden, verklaard kan worden door het feit dat het hier om een binair rijmend gedicht gaat.

5. *Mei 2*

Tot nog toe hebben wij het over de eerste mogelijkheid gehad : het toetsen van a-priori hypothesen d.m.v. een statistische tekstanalyse. Daartegenover staat de tweede mogelijkheid, nl. de generatieve kracht van een statistische tekstanalyse : hypothesen worden nu niet langer getoetst, maar voortgebracht, gegenereerd.

Ter illustratie van deze mogelijkheid kiezen wij opnieuw het fenomeen rijmwoorden in *Mei* [6]. Zoals in par. 4 werd gezegd bevat dit gedicht 4.380 binaire verzen of 2.190 binaire rijmparen. Een vraag die daarbij rijst is : is de dichter in staat altijd „goede" („volmaakte", „perfecte") rijmwoorden te vinden, of vervalt hij af en toe in z.g. „defectieve" rijmen ? En indien dit laatste het geval is, hoe is dit dan te verklaren ?

Als „goede" rijmen werden beschouwd : rijmen waarin de laatste betoonde klinker of tweeklank en al wat daar eventueel op volgde gelijk was. B.v. kinderen – hinderen ; geluid – uit.

Als „defectief" werden beschouwd : al die rijmen waarvan de laatste betoonde klinker of tweeklank niet gelijk was en/of wan-

6. Dit fenomeen werd o.m. tijdens het college Kwantitatieve Taalkunde (K.U.L., Germaanse Filologie, Academiejaar 1977-78) onderzocht. Ik dank alle studenten die het college volgden en m.n. Guido van Damme die een paper – waarvan ik passim gebruik heb gemaakt – aan dit onderwerp wijdde.

neer hetgeen op die klinker of tweeklank volgde niet gelijk was. Voorbeelden : rinkelden – kiezel en ; wateren – de hen ; mooi – bloei [7].

Ook hier zou een bepaalde teksthypothese als uitgangspunt voor het onderzoek kunnen dienst doen : de veronderstelling b.v. dat de dichter eerder op het einde dan bij het begin in defectiviteit zou vervallen (wat dan in elk geval ook nog de vraag naar de chronologie der delen zou doen rijzen).

In plaats echter van uit te gaan van een bepaalde teksthypothese werd het gedicht in 10 delen van elk 438 verzen (219 binaire rijmparen) verdeeld. Vervolgens werd er een *louter statistische hypothese* aangenomen, de z.g. nulhypothese : volgens deze hypothese verdelen zich de defectieve rijmparen gelijkmatig over de 10 delen (m.a.w. er is geen verschil tussen de 10 delen onderling inzake aantal defectieve rijmparen, vandaar de naam nulhypothese). In dergelijk geval kunnen wij het aantal defectieve rijmparen per tekststuk van 438 verzen met een zekerheidsgraad van 95 % bepalen als liggende tussen 7 en 21 [8].

De volgende, onderstaande tabel geeft aan :

in kol. 1 : het tekstgedeelte ;
in kol. 2 : het aantal geobserveerde defectieve rijmparen ;
in kol. 3 : de verschillen tussen observatie en verwachting ; 0 duidt aan dat het om een niet-significant verschil gaat, + duidt op een significant surplus (t.o.v. de verwachting), – op een significant tekort.

Tekstgedeelte	Aantal Def. Rijmparen	Verschil
1- 438	11	0
439- 876	7	0
877-1.314	8	0
1.315-1.752	14	0
1.753-2.190	18	0
2.191-2.628	20	0

7. Zoals men kan merken is hier enkel met „echte" defectieven (die een klankafwijking vertonen) rekening gehouden ; rijmen waarbij de normale woordvorm onder rijmdwang werd aangepast (cf. schaûw-nauw ; Oberoon-kroon ; wei-mei- (waarbij „doorn" dan op de volgende regel kwam) of verzen waarbij dezelfde rijmwoorden werden herhaald (b.v. lief-lief) zijn hier buiten beschouwing gelaten.

8. Het totale aantal defectieve rijmparen bedroeg 138, d.w.z. gemiddeld 13,8 defectieve rijmparen per tekstgedeelte van 438 (binaire) verzen. Nemen wij afwijkingen t.o.v. dit gemiddelde aan ten gevolge van steekproeffluctuaties en stellen wij de onbetrouwbaarheidsdrempel (α) op 5 % dan liggen de kritieke waarden bij : resp. 13,8 — (1,96 × sd) [sd = $\sqrt{pqN}$ = $\sqrt{0.10 \times 0.90 \times 138}$ = 3.52] = 13,8 — (1.96 × 3,52) = 6,9 afgerond 7 en 13,8 + (1.96 × 3,52) = 20,7 afgerond *21*.

2.629-3.066	14	0
3.067-3.504	31	+
3.505-3.942	10	0
3.943-4.380	5	−

Tot welke conclusies kan deze tabel nu leiden ? Allereerst spreekt de verdeling der defectieve rijmparen de statistische hypothese niet tegen behalve dan op twee plaatsen (8ste en 10de fragment), waarbij zich significante afwijkingen t.o.v. de verwachting voordoen.

De afwijking is vooral erg duidelijk in het 8ste tekstgedeelte. Welnu dit is het fragment waarin *Mei* aan het einde van haar zoektocht haar geliefde Balder vindt, maar haast onmiddellijk door hem weer in de steek wordt gelaten. Deze afwijking valt dus samen met een situatie waarin het hoofdpersonnage het toppunt van geluk en vreugde, maar ook haar diepste ontgoocheling ondervindt. Deze gegevens samen vormen dan ook de aanleiding om de hypothese aan te nemen dat de dichter in momenten van grote emotie zich minder aan rijmdwang gelegen laat, of nog : hoe groter de emotie, hoe vrijer de dichter staat tegenover het rijm, ergo hoe meer defectieve rijmparen.

6. *Generaliserende Conclusies*

Mede op grond van het voorgaande menen wij tot de volgende algemene conclusies te kunnen komen :

a. Wat hier aan de hand van een literaire tekst werd gedemonstreerd geldt m.m. evenzeer voor niet-literaire, b.v. „wetenschappelijke" teksten.

b. De kern van dit betoog houdt in dat een statistische tekstanalyse niet alleen een soort proef op de som hoeft te zijn voor a-priori hypothesen, maar ook een heuristische methode kan zijn leidend tot het opstellen van nieuwe hypothesen.

c. Deze nieuwe hypothesen worden op hun beurt niet volkomen toevallig of willekeurig gegenereerd. In het voorbeeld in par. 5 gaat het om defectieve rijmparen, een fenomeen dat door de onderzoeker als tekst-relevant ervaren wordt.

d. In dit opzicht is het risico van deze heuristische methode ook beperkt : ten eerste is de strategie niet volledig blind, vervolgens zijn er thans ook technische hulpmiddelen (zoals computers) die ons kunnen helpen gelijkaardige experimenten (waarbij eventueel veel data moeten verwerkt en/of complexe berekeningen worden gemaakt) op een snelle en efficiënte manier uit te voeren.