

Taalstatistiek *)

door H. Brandt Corstius **)

UDC 31:40

S u m m a r y*Discussion of the possibilities and limitations of Language Statistics.*

Sampling of „normal” prose is difficult as the „universe of written Dutch” is undefined. What predictions can be made about the contents of tomorrow’s newspaper? The letter frequencies are very stable. Frequencies of bigrams and trigrams of letters bring us towards the word level about which something may be predicted. The sentence level can only be dealt with after a mechanical sentence analysis method is available. The highest level is still outside the scientific domain and belongs to the literary critic. Literary Statistics should be founded on a solid knowledge of Language Statistics. It is considered inappropriate to begin this field with difficult historical problems. The mutual distrust between linguist and statistician requires much tact on both sides. Eventually Language Statistics may become one of the bridges between the „two cultures”. The modern computer is an indispensable tool both for practical reasons (the giant mass of material) and for theoretical ones (the need to give unambiguous definitions of concepts like „sentence”, „word” and „syllable”). The use of Language Statistics in Mechanical Translation research is discussed. Review of the activities in the Netherlands. A one million word count is on its way. The speaker pleads for a National Center for Lexicology and Language Statistics.

Tuut. Tuut tuut. Tuut tuut tuut. Tuut tuut tuut tuut. Tuut tuut tuut tuut tuut.
 Tuut tuut tuut tuut tuut tuut. Tuut tuut tuut tuut tuut tuut tuut. Tuut tuut tuut
 tuut tuut tuut tuut tuut.

Tuut boe tuut fak tuut tuut.

Tuut tuut boe tuut fak tuut tuut tuut.

Tuut boe tuut tuut fak tuut tuut tuut.

Tuut tuut boe tuut tuut fak tuut tuut tuut tuut.

Als u goed heeft geluisterd, heeft u begrepen dat in de taal die ik zojuist sprak „tuut” de eenheid van tellen voorstelt, zodat „tuut tuut” „twee” betekent en „tuut tuut tuut” „drie”. Verder heeft u uit de manier waarop deze woorden in de zinnen voorkwamen kunnen opmaken dat „boe” „plus” betekent en „fak” „is gelijk aan”, zodat „tuut boe tuut fak tuut tuut” betekent: „ $1 + 1 = 2$ ”.

Als u mij voldoende tijd zou geven, en dat zou dan wel meer kosten dan de

*) Voordracht gehouden op de Statistische Dag 1964, Mathematical Centre, communication MR 70.

**) Medewerker Mathematisch Centrum, A’dam.

drie kwartier die ik nu heb, kon ik op deze manier een nieuwe taal opbouwen, waarin u mij zou begrijpen, zonder dat ik eenmaal mijn toevlucht tot het Nederlands zou hoeven nemen om u de betekenis van de woorden uit te leggen. Die taal is er al, want de woorden waar ik mee begon vormen het eerste leeslesje van een dialect van „LINCOS”, een taal door Professor FREUDENTHAL ontworpen voor communicatie met buitenaardse wezens met wie alleen via de radio contact is gelegd. Door LINCOS zullen we met deze mensen van andere hemellichamen kunnen praten, onder de aanname dat hun hersenen ongeveer zo werken als de onze. Zo zouden zij dus de zinnen waar ik mee begon ook begrepen hebben, maar zou de rest, in gewoon Nederlands uitgesproken, abracadabra zijn.

Wat is het verband van LINCOS met ons onderwerp? Met het onderwerp van deze dag: communicatie, is het verband duidelijk, ook al denkt men daarbij waarschijnlijk niet in de eerste plaats aan communicatie met personen waarvan het bestaan nog niet vaststaat. Het verband met mijn onderwerp: de taalstatistiek is minder duidelijk. Het is niet, dat de ontwerper van LINCOS ook in statistische kringen niet onbekend is. Maar het is wel om te demonstreren wat de taalstatistiek met mijn eerste zinnen zou doen. We nemen dan een taalstatisticus die niet verder kijkt dan zijn taalstatistische neus lang is. Hij zal na nauwkeurig tellen zeggen: „Tuut kwam 60 keer voor, boe en fak ieder vier keer, de gemiddelde zinslengte is 5,8 woord.” Het is duidelijk dat zo’n uitspraak ons weinig helpt bij het begrijpen van boodschappen van andere planeten in het een of ander Lincosdialect. Toch is het tellen van woorden, letters en zinnen dagelijks werk van de taalstatisticus. Wat voor zin heeft dat tellen dan wel? Wat zijn de mogelijkheden, en wat de begrenzingen van de taalstatisticus? Welke positie bekleedt de taalstatistiek tussen de taalkunde en de statistiek? Op deze vragen wil ik trachten u een antwoord te geven.

Als u mij wilt toestaan om voor de n -e keer een definitie van statistiek te geven, dan wil ik statistiek definiëren als de kunst van het nuttig verliezen van informatie. Wanneer onze hersenen zo waren ingericht dat we duizenden getallen gemakkelijk konden overzien, dan zouden we statistiek niet zo nodig hebben (behalve als we dan juist weer miljoenen getallen tegelijk wilden overzien). Statistiek levert ons de transformatie van bomen tot bos. Het verschaft ons zinvolle uitspraken over onoverzichtelijke informatie. Onoverzichtelijk omdat die zo groot is, of omdat het informatie is die nog in de toekomst ligt: via de transformatie van bomen tot bos, kunnen we ook over het achterste gedeelte van het bos, waar we nog nooit geweest zijn, ons een idee vormen. Een van de grootste en onoverzichtelijkste dingen om ons heen is onze eigen taal. Naast vele andere bezigheden, als het maken en lezen van woordenboeken, kan de taalstatistiek ons een inzicht in dit complexe geheel verschaffen.

Taalstatistiek is geen bijzondere statistiek. Met andere woorden: u hoeft van statistiek niet meer te weten dan u nu weet, en u hoeft ook niets van wat u nu weet te veranderen. Omdat mijn gehoor hier uit statistici bestaat, zal ik hier van de taal-statistiek dus meer het eerste gedeelte, de taal benadrukken, bij een gezelschap linguïsten zou ik op het laatste gedeelte, de statistiek, dieper moeten ingaan. De toepassing van de statistiek op taalverschijnselen is heel goed te vergelijken met die op de biologie: de biometrie. We zijn alleen nog lang niet zover dat we van linguametrie kunnen spreken, daar het werk dat tot nog toe is gedaan óf nogal erg elementair, óf niet erg verantwoord is, zoals uit het vervolg zal blijken.

Net zoals in de biometrie speelt in de taalstatistiek een grote rol het begrip van de verhouding tussen type en teken. Ieder van u behoort tot het type *mens* waarvan ieder van u een afzonderlijk teken is. Als men in de biologie over „het gemiddelde aantal chromosomen” spreekt, dan heeft dat alleen zin als men erbij zegt of dat gemiddelde is genomen over de types, dus over de afzonderlijke diersoorten, of dat het gemiddelde is genomen over alle exemplaren, dus alle dieren afzonderlijk. Precies zo heeft het alleen zin te spreken over „de gemiddelde woordlengte” als men erbij vertelt of dat gemiddelde is berekend over de verschillende woorden, zoals die in het woordenboek staan, of over alle woorden uit een tekst normaal Nederlands proza. In het laatste geval wordt de gemiddelde woordlengte aanzienlijk korter, omdat juist de veel voorkomende woordjes kort zijn. Het lijkt allemaal erg voor de hand liggend, maar er wordt tegen dit onderscheid tussen type en teken helaas nog wel eens gezondigd. Zo ziet men b.v. leenwoordenonderzoeken, waarbij men dus nagaat welke woorden wij van andere talen geleend hebben, waarin over de frekwentie van de leenwoorden niet wordt gesproken.

Omdat niemand weet wanneer een woordenboek compleet is, en omdat deze vraag in het Nederlands door het optreden van onvoorspelbare samenstellingen zelfs niet is te beantwoorden, is het meestal het beste de tekens te tellen, dus de taalelementen uit een stuk lopend Nederlands proza. Wat is normaal Nederlands proza? Hoe lang moet men zo'n tekst nemen om iets verstandigs over de frekwentie van de getelde taalelementen te kunnen zeggen? Dit is het soort vragen dat u bekend in de oren zal klinken. De steekproeftheorie komt op de proppen. Maar hierbij doet zich direct een gevaarlijk punt voor: men kan wel een bepaald stuk test als steekproef nemen: maar wat is de populatie waarvan het een steekproef moet zijn? Het is praktisch onmogelijk een goede definitie van het universum van de geschreven Nederlandse taal te geven. Gesproken taal laat ik geheel buiten beschouwing, daar die op dit ogenblik nog niet goed meetbaar is. en hetzelfde woord door twee sprekers uitgesproken door onze apparaten als twee verschillende woorden wordt gehoord. Men kan natuurlijk wel door een.

typiste een geluidsbandje laten uittikken, maar dan is de gesproken taal toch weer getransformeerd tot geschreven taal en heeft de typiste al vele willekeurige beslissingen moeten nemen. Dit vooropstellen van de geschreven taal strijkt de meeste taalkundigen tegen de haren, die meestal immers de spreektaal voorop zetten. Maar het primaat van de gesproken taal wordt ook helemaal niet aangetast door de constatering dat we, in het licht van wat de wetenschap nú kan, ons beter tot de geschreven taal kunnen beperken.

Hoe moeten we nu het universum van het geschreven Nederlands definiëren? Een concreet voorbeeld: op het Mathematisch Centrum hebben we, als onderdeel van een veel ambitieuser plan, een telling verricht over tien kranten van eenzelfde dag. Is het nu wel eerlijk om een krant met een oplage van 400.000 exemplaren gelijk te stellen met een krant, die maar in 25.000 exemplaren wordt afgedrukt? Moeten we dus de resultaten van onze tellingen niet wegen met de oplagecijfers? Maar worden al die woorden uit die veelverkochte krant wel echt gelezen?

We zouden toch eigenlijk die woorden willen tellen die in het communicatieproces volledig zijn gebruikt. Of moeten we afzonderlijke statistieken maken: een van de woorden zoals die van de schrijver afkomen, en een statistiek van de woorden zoals de lezer die opneemt? Eén telling dus van de bedoeling en één van het effect. In ons krantenmateriaal kwamen b.v. wel drukfouten voor. Wij hebben die onveranderd geregistreerd, zodat men het als een telling van de lezer kan beschouwen (een lezer dus die tien kranten op een avond las).

Als we, om ons intuïtief idee van het Nederlandse taal-universum te verduidelijken, de populatie van het geschreven Nederlands zouden definiëren als alles wat in een bepaalde week aan Nederlandse tekst wordt gelezen, dan zal het leeuwen-aandeel daarin komen van de tweemiljoen televisietoestellen, waar de buitenlandse films met Nederlandse ondertitels op verschijnen. Zo'n telling zou dus een vertekend beeld geven van het Nederlandse woordenmateriaal, waaruit men wellicht een grote invloed van de misdaad op onze taal kon afleiden.

Gedeeltelijk ligt de populatie waarvoor we ons interesseren ook in de toekomst, en ik zou om onze gedachten te kunnen bepalen, willen vragen:

Is het mogelijk om op grond van tellingen bij de kranten in het afgelopen jaar iets te zeggen over de krant die u vanavond thuis vindt? De gedeeltelijk positieve beantwoording van deze vraag brengt ons tot de verschillende lagen die de taalstatisticus moet onderscheiden.

Als we uit de tellingen van het afgelopen jaar weten met welke frekwenties de letters van het alfabet voorkomen, dan kunnen we er zeker van zijn dat diezelfde frekwenties in de krant van vanavond zullen optreden. Zelfs als in die krant toevallig het verslag wordt gegeven van een voetbalwedstrijd tussen Xerxes en Ajax zal daardoor de letter X niet noemenswaard van zijn nederige

positie opstijgen. Probeer u maar eens een samenhangend betoog te schrijven waarin b.v. de letters E of N niet voorkomen. Wat we van de letters zeiden, geldt ook nog wel voor de bigrammen, dus de opvolgingen van twee letters, en misschien ook nog wel voor de trigrammen. Maar nu komen we bij de opeenvolgingen van vier of vijf letters. Een van de belangrijkste letters uit ons alfabet is de spatie, die de scheiding tussen twee woorden markeert, het is een letter, die in frekwentie zo hoog staat als de letter E. In de tetragrammen en pentagrammen komt hij dus nog al eens voor, en we zien hier dus dat we op een hoger niveau komen, dat van: welk wordeinde kan voorafgaan aan welk woordbegin? De hoge frekwentie van het trigram „N D” komt niet alleen door de hoge frekwenties van zijn samenstellende letters, maar ook omdat de combinatie VAN DE zo vaak voorkomt (Deze twee woordjes DE en VAN zorgen, hoe ongelooflijk het ook klinkt, voor meer dan tien procent van de tekst) en omdat de N in het algemeen als slotletter (meervouden) geliefd is.

We kunnen ook direct op het hele woord overgaan en vragen: in hoeverre is de woordenschat van de krant van vanavond te voorspellen op grond van die over het afgelopen jaar?

Daar valt wel iets over te zeggen, speciaal wat betreft de zeer hoogfrequentie woorden, waarvan de eerste tien (DE, VAN, HET, EEN, EN, IN, DAT, TE, IS, VOOR) al een kwart van de krant zullen vullen. Maar verdere uitspraken zijn heel moeilijk te doen. Het is dan ook niet verstandig de taalstatistiek te laten beginnen bij de hapax legomena, de woorden die maar eenmaal voorkomen, uit Homerus. Want die eenlingen zeggen statistisch niets, ze hadden ook wel de frekwentie twee of nul kunnen hebben. Er zullen zelfs zeker in de krant van vanavond woorden voorkomen, die daar nog nooit eerder in werden afgedrukt, en die zelfs in geen enkel woordenboek zijn te vinden: de eigennamen. Onze krantentellingen dateren van 19 juni 1956 en in de lijsten staan de woorden „makarios” en „cyprioten” hoog genoteerd. Acht jaar hebben deze woorden gesluimerd op onze ponsbanden, en nu zouden ze weer met dezelfde frekwentie in de krant die u vanavond leest voorkomen. Moeten we de eigennamen, die door hoofdletters te herkennen zijn dan maar niet meetellen? De ervaring heeft ons geleerd dat iedere eigenmachtige beslissing: laten we dat maar niet meetellen, later tot spijt aanleiding geeft.

Van letters, bigrammen en trigrammen via woorden tot zinnen. Is er iets zinnigs te zeggen over de zinnen in de krant van vanavond? Voor een krant toevallig wel, omdat voor de aandachtige lezer elke krant voor een groot gedeelte een herkauwing is van de editie van de vorige dag. Maar precies gelijke zinnen zijn, behoudens weerberichten en politieke credo's, in het Nederlandse krantenproza toch zeldzaam. Onze tellingen moeten hier dus gedaan worden op zinsstructuren. Helaas ontbreekt een mechanische methode van zinsontleding, die dit

op grote schaal mogelijk zou maken, maar tellingen van eenvoudige zinsstructuren zijn toch al gedaan.

Boven het zinsniveau staat dat van de artikelen, verhalen, samenhangende betogen. Hierover kan ik kort zijn, omdat er nog geen exact werk over bekend is. Het is nog geheel het gebied van de literaire criticus, die zich jammer genoeg meestal niet met krantenproza bezig houdt. Toch wil ik er op wijzen dat de criticus H. Gomperts in zijn „Schok der herkenning” informatietheoretische begrippen in zijn literaire kritiek aanwendt. Hij doet dit nog zuiver bij wijze van beeldspraak, of zoals hijzelf in zijn inleiding schrijft:

„De introductie van cybernetische begrippen lijkt mij niet per se bevorderlijk voor de waardering van literatuur, maar op zijn best voor het begrijpen van wat de kritiek doet. Zij hebben het voordeel dat zij koel en zakelijk zijn – een voordeel bij een onderwerp dat zo gemakkelijk vervaagt achter een gordijn van emotioneel gaas – en het nadeel dat zij juist daardoor een valse suggestie van exactheid en objectiviteit geven. Men moet nu eenmaal nieuwe misverstanden creëren om de oude te bestrijden.”

Dit hoogste niveau behoort nog niet tot het gebied van de taalstatistiek, alle reclame over programma's voor rekenmachines die op statistische gronden machinale uittreksels zouden vervaardigen ten spijt. Maar we moeten niet uit het oog verliezen, dat dit het niveau is waar het ons uiteindelijk om gaat.

De krant van vanavond, en die van gisteren, behoort tot „het” Nederlands, of misschien (dat kunnen we pas weten wanneer we de andere tellingen die op het programma staan hebben uitgevoerd) tot het kranten-nederlands. In een krant wordt door meer dan een persoon geschreven, ook al is de ANP-telexdienst wel een van de productiefste schrijvers. Een heel nieuw gebied van taalstatistiek komen we binnen als we ons beperken tot het taalgebruik van één mens. Daar neemt men dan meestal geen Scheveningse visser voor, maar een schrijver. Dit is een meer spectaculair gebied, en – dus – een meer riskant gebied.

Voorop wil ik stellen dat het alleen zin heeft om dit particuliere taalgebruik statistisch te onderzoeken als men over de statistische eigenschappen van de taal waarin de te onderzoeken persoon schrijft, beschikt. Wie b.v. meent dat bij Paul van Ostayen bijzonder veel O's en A's voorkomen kan die bewering alleen maar waar maken als hij beschikt over de frekwentie van die twee klinkers in „normaal” Nederlands.

Het is nonsens om te beweren dat men op grond van statistiek de auteur van een bepaald stuk proza kan aanwijzen. Bij de bekende gevallen zijn er altijd belangrijke mede-overwegingen.

In een beroemd geworden geval, het auteurschap van bepaalde bijbelgedeelten, is enige skepsis op zijn plaats. Het lijkt me niet verstandig om een pas beginnende wetenschap als de taalstatistiek te gaan toepassen op geschriften van

19 eeuwen oud, door verschillende schrijvers geschreven en met allerlei onzekerheden omtrent de herkomst van verschillende gedeelten.

Laten we, om onze methode te ijken, eens met een eenvoudig geval beginnen. Het geschenk van de boekenweek zou een goed testgeval zijn, en ik hoop dat de oude traditie van de geheim gehouden schrijver hersteld zal worden. Als men probeert na te gaan hoe een lezer beslist door wie iets dat hij las geschreven werd dan gebeurt dat op grond van het onderwerp en de toon van het verhaal. Maar daarom is het nog niet uitgesloten dat bepaalde meetbare stijleigenschappen, die we ons niet bewust maken, door statistische analyse eruit gehaald kunnen worden. Tenslotte herkennen wij elkaar wel niet aan onze vingerafdrukken, maar kan de politie dat wel. Dit zijn dan hopelijk tevens eigenschappen die de schrijver zich niet bewust was. Want elke schrijver heeft bij elk nieuw boek het voornemen „nu ga ik het eens heel anders doen”. Deze evolutie bij een schrijver kan op zich natuurlijk ook weer statistisch onderzocht worden. Op deze manier hoopt men b.v. erachter te komen in welke volgorde de werken van Plato geschreven werden. Ook hier geldt weer dat men de grote historische problemen beter kan laten wachten totdat we enige zekerheid hebben over de soliditeit van onze methodes. De discussies die hierover gevoerd worden staan nog al te vaak op het beminlijke, onwetenschappelijke peil waarop in de zeventiende eeuw de waarschijnlijkheidsrekening werd bedreven. Het onderzoek naar het taaleigen van een bepaalde schrijver wordt bovendien nog gecompliceerd door het feit dat elke schrijver tracht zijn figuren te laten spreken zoals hij dat karakteristiek voor ze acht. Hoe in dit tweedehands taalgebruik nog meetbare stijlkenmerken gevonden moeten worden, weet ik niet. Ik kan me voorstellen dat het onderzoek naar particulier taalgebruik u interesseert, maar uit het oogpunt van een harmonische ontwikkeling zou een grondig onderzoek naar „het” Nederlands de prioriteit moeten hebben.

Zoals het particuliere taalgebruik onder het algemene taalgebruik staat, zo kan men het Nederlands ook zien ten opzichte van andere talen. Vragen van vermenging van talen, afstand tussen talen. Er zijn daar ook wel statistische publicaties over maar het is een terrein waar de taalkunde zelf zo onzeker over is, dat de statistiek er beter af kan blijven. Ten slotte is de statistiek, als meestal, hulpwetenschap, en dat de taalkundigen er – nu en hier – te weinig gebruik van maken geeft de statisticus niet het recht zich als taalkundige voor te doen.

Taalstatistiek is het gebied waar de taalkundige de statisticus verwijt dat hij met zijn plompe methodes het wezenlijke van zijn onderzoeksobject voorbijschiet, en waar de statisticus de taalkundige verwijt dat hij met zijn ad hoc-methodes, ongeorganiseerde tellingen en niet bewust gemaakte hypothesen de gezonde beginselen van de statistiek met voeten treedt. Beiden hebben helaas nog al eens gelijk. De conclusies waartoe de enthousiaste promotor van de taal-

statistiek, Herdan, komt moeten voor veel taalkundigen lachwekkend of treurig zijn, al naar hun aard. De conclusies waartoe sommige taalkundigen komen, en ik noem dan maar liever geen namen, zijn voor het statistisch oog vaak ongegrond. Er zijn grote gebieden van de taalkunde waar de statistiek ons niet kan helpen (herinner u de tuuttuuttaal). Maar er zijn andere delen waar een meer kwantitatieve zienswijze alleen maar heilzaam kan zijn. Dat in de grote woordenboeken over de frekwentie van de daarin opgenomen woorden niets wordt gezegd (behalve misschien „zeldzaam”) is wel jammer, omdat de frekwentie van een woord een onderdeel van zijn betekenis is, maar te vergeven omdat de samenstellers die frekwentie niet wisten; maar dat men nergens kan vinden hoeveel woorden er nu eigenlijk in staan, is misschien niet zo belangrijk maar toch tekenend voor een tijd die voorbij zou moeten zijn. Het onderscheid tussen alfa en bèta, door velen al bestreden, maar door ons onderwijssysteem hardnekkig voortbestaand, maakt deze situatie van wederzijdse achterdocht in ons land moeilijk. Misschien dat de merkwaardige toestand dat een alfa-gymnasiast wel, een bèta-gymnasiast geen statistiekonderwijs krijgt, hier verbetering in brengen kan.

Als statisticus bent u er misschien aan gewend te werken met mensen die door traditie en wellicht aanleg het kwantitatieve denken vreemd is. In de taalkunde is dat nog zo. Maar dat verandert. Niet alleen in de vorm van statistiek, maar ook als pure wiskunde doen de exacte wetenschappen hun intree in de taalkunde. Wederzijdse tact en wijsheid zijn hierbij nodig. Het is niet goed wanneer Herdan het voorstelt of de taal een gas is waar de woorden als moleculen in rondnansen (en een gedicht dus een soort Brownse beweging wordt), en het is ook niet goed, wanneer een beroemd Nederlands taalkundige schrijft over „ingenieurslogica”. (Voor de ingenieurs onder u: dit was misprijzend bedoeld.)

De scheiding tussen alfa en bèta is zinloos. Hoe eerder ze weg is hoe beter.

Taalstatistiek is zo oud als de taalkunde, en ouder dan de statistiek. De hapaxen uit Homerus, de Romeinse geheimschriften bewijzen het. Maar een werkelijk wetenschappelijk begin werd pas in recente tijden gemaakt. En wel taalstatistiek in de elementairste vorm: *tellen*. Tellen is meestal iets waar de statisticus op neerkijkt, omdat hij conclusies wil, maar tellen is in dit vak al moeilijk genoeg. Een artikel uit de oertijd van de taalstatistiek, dat van Yule in *Biometrika* van 1938, die maar eens naïef zinslengten in Engelse proza gaat tellen is tekenend. Wat is een woord? Wat is een zin? Moet ik aangehaalde zinnen ook meetellen? (Yule doet het niet tenzij de aanhaling kort is.)

Voor dit tellen is de moderne rekenautomaat, de computer, een onmisbare hulp. En wel om twee redenen, een praktische en een principiële.

De praktische reden: De omvang van het materiaal is huiveringwekkend. Toen De la Court in 1935 begon met het tellen van zijn miljoen Nederlandse

woorden werd hij al gauw moe van het turven van het woordje DE, dat hij toen maar op oneindig zette. Een computer kent zo'n vermoeidheid niet: de Electrologica X1 kan het verlangde miljoen tekstwoorden in twintig nachten makkelijk tellen (taalstatistici zijn overal ter wereld voor computertijd op de nachten aangewezen), en bij dat tellen dan ook nog met groot gemak alfabetiseren, en alfabetiseren van achteren naar voren zodat rijmwoorden bij elkaar komen te staan, woordlengtes, lettergreepaantallen, letterfrequenties bepalen en wat u maar wilt.

De principiële reden: Een computer heeft geen idee van wat hij moet tellen. Het moet hem dus expliciet gezegd worden. Wie dat probeert te doen, merkt dat hij het zelf ook niet zo precies wist. Definities als „een woord is wat tussen twee spaties staat”, „een zin is wat tussen twee punten staat” blijken zeer onvolledig. Eenmaal op dit pad houden we bij woord en zin niet op en gaan we verder met het exact definiëren van intuïtief bekende taalelementen. Wat is een lettergreep? We hebben allemaal wel een idee, maar de bestaande definities zijn voor een computer niet bruikbaar. Toch is het mogelijk om een computer te leren woorden in hun lettergrepen te splitsen. Hoever kan men zo doorgaan? Zou men beeldspraak, ironie, monologue intérieure, zinsritme ook niet door een computer kunnen laten herkennen? Niet omdat het zo belangrijk is dat een computer het kan, maar omdat we dan pas kunnen zeggen dat wij het kunnen. Het excuus „Maar wat weet ik nou van rekenmachines?” gaat niet meer op, sinds er talen zijn bedacht die niet meer van een bepaalde machine afhangen, maar die in het algemeen geschikt zijn voor het beschrijven van gecompliceerde reeksen van bevelen. Het is hier niet de plaats om voor een van die talen propaganda te maken, en u mag in plaats van het hier te lande populaire ALGOL ook best COMIT, FORTRAN, IPL V, of COBOL gebruiken, maar als u probeert uw gedachten in voor de machine begrijpelijke precisie uiteen te zetten, dan heeft u er zelf minstens zoveel aan als de machine. Het woord „reken”machine is alleen nog historisch, de toepassingen op gebieden buiten de wiskunde beginnen pas.

En daarmee zijn we op het gebied van de vertaalmachine. Niemand zit hier te springen om een vertaalmachine die uit of in het Nederlands kan vertalen. Maar het onderzoek naar zo'n machine (dat betekent dus, naar de manier om een bestaande rekenmachine als vertaalmachine te laten optreden) is het waard om met meer kracht te worden gevoerd dan tot nu toe het geval is, omdat het een onderzoek naar onze eigen taal is. Welke diensten kan de taalstatistiek ons hierbij verlenen?

Het moeilijkste gedeelte van het vertaalproces is de fase van het „lezen” van de brontaal, waaruit vertaald wordt, m.a.w. het, nog zonder aan vertalen te denken, begrijpen van een tekst, dus kunnen zeggen waar dat HIJ op terug

slaat, weten wat het onderwerp van de zin is, weten in welke betekenis het woord „spreiding” hier wordt gebruikt. Dit lees-probleem is (de wetenschap is nu eenmaal de kunst van het in stukken hakken van problemen) te verdelen in een lexicologisch probleem en een syntactisch probleem.

Het lexicologisch probleem is kort gezegd: welk van de vertalingen in het woordenboek achter het nederlandse woord moet ik kiezen? Hier kunnen statistische middelen ons helpen. Niet, zoals sommige tegenstanders van mechanische vertaling schijnen te denken door dat woord te kiezen dat de hoogste frekwentie heeft. Dan hoeft men de andere vertalingen helemaal niet op te nemen. Maar wel op een andere manier. Het is duidelijk dat de betekenis van een woord met meer dan een betekenis (en welk woord heeft dat niet) afhangt van de woorden er om heen. Hoeveel woorden erom heen? En hoe wegen we de invloed van dichtbijgevoerde woorden t.o.v. verre woorden? Op het Mathematisch Centrum is met het onderzoek naar deze vragen een bescheiden begin gemaakt. Als u ons een zin toezendt (en wij zullen u daar dankbaar voor zijn, want het verzamelen van voldoende materiaal kostte ons meer moeite dan het maken van de programma's) waar het woordje GLAS in voorkomt, dan kan onze XI in 80 procent van de gevallen juist beslissen of het hier gaat om een drinkglas of om het materiaal glas. U mag niet zelf een zin met GLAS verzinnen want de machine is gemakkelijk om de tuin te leiden. Voor een statistisch doorkneed gehoor hoef ik niet te zeggen dat 80 procent laag is: 50 procent zou de machine al scoren door volstrekt willekeurig te raden. 80 Procent is voor vertaling niet voldoende want een zin met b.v. vijf van die dubbelzinnige woorden zou dan al een kans van $1 - (4/5)^5 = 2/3$ maken om onjuist vertaald te worden. Hier ligt m.i. een groot werkterrein voor de statistiek, want helemaal hopeloos is de situatie niet. Nog moeilijker dan het lexicologische probleem is het syntactische probleem bij de mechanische vertaling. Stel we weten van alle woorden de juiste vertaling, in welke volgorde moeten we ze dan zetten? Moeten er geen woorden in de nieuwe taal bij? Wie b.v. uit het Russisch wil vertalen zal zelf alle lidwoorden moeten invoegen, daarbij dus steeds een keus doend tussen DE en EEN.

Anders dan men wel meent geloof ik niet dat de statistiek ons bij syntactische problemen kan helpen. Wel kan, wanneer men het probleem enigszins meent opgelost te hebben, de statistiek uitspraak doen bij de beoordeling van de resultaten. Kwaliteitscontrole is bij de mechanische vertaling een belangrijk en vaak vergeten punt. Er zijn tegenstanders van mechanische vertaling die nu al jaar in, jaar uit komen met zinnen als:

„De musicus kocht een piano en een strijkstok, die hij onder de arm mee naar huis nam.” Voor vertaling in bepaalde talen moeten we weten of DIE op piano en strijkstok, of alleen op strijkstok slaat. Wij weten direct dat het laatste het

geval moet zijn omdat een piano te zwaar is om onder de arm mee naar huis genomen te worden. Maar dit zou betekenen dat de machine in zijn woordenlijst bij ieder voorwerp een index bezit die aangeeft of dat voorwerp onder-de-arm-naar-huis-draag-baar is. Dit is absurd en dus bestaat de vertaalmachine niet. Als we dit specifieke argument erkennen dan blijft toch de vraag: hoe vaak komen zulke zinnen voor? Hierop moet de statistiek op grond van een steekproef (en na deugdelijke definitie van de populatie!) een antwoord geven, maar dat kan pas nadat een eerste grove poging tot machinaal vertalen is gedaan.

Het is overigens een merkwaardige zaak met de vertaalmachine. Met de regelmaat van de nieuwe kunstmanen verschijnen er berichten dat hij er nu is, we lezen zelfs korte vertaalde stukjes al dan niet begrijpelijk Engels, en dan horen we verder niets tot het volgende bericht met soortgelijke strekking. Recent is het publiceren van het succes met de vertaalmachine uit het Chinees, die bij nader kijken niets meer blijkt te zijn dan een codeersysteem van de chinese karakters. Het chinese „alfabet” is dus nu te ponsen, zoals onze 27 letters dat allang waren, maar of men dit nu al vertalen mag noemen? Ook als men de reclame van fabrikanten en overzeese onderwijsinstellingen eraf trekt, blijft het een fascinerend researchobject waar de statistiek zeker bij behulpzaam kan zijn, maar tot nu toe nog weinig heeft bijgedragen.

Er is nog een belangrijke, te weinig gebruikte samenwerking van computer en taalstatistiek. Als men op grond van tellingen een voorlopig model van een taal heeft opgesteld, kan men door de machine stukken proza in die modeltaal laten maken. Deze vorm van „basic Dutch” kan dan met de echte taal vergeleken worden. Er bestaat hier gevaar voor een misverstand: de voorstanders van „gemechaniseerde” taalkunde propageren geen enkele vorm van „basic Dutch”. Iedereen die, al was het maar als drukproefnakijker, met taal te maken krijgt, zal wel eens verzuchten: moet dat nu zo? waarom zijn er geen twee soorten punten, een voor afkortingen en een voor het eind van de zin? waarom is de spelling zo en niet zoo? waarom is het meervoud van kind niet gewoon kinden? wat is het nut van het onderscheid tussen DE en HET?

Maar niemand denkt eraan om deze problemen op te lossen door de taal te veranderen. Taalkunde is een empirische wetenschap en net zo min als een fysicus die met modellen van atomen werkt, denkt dat hij daarmee electronen kan dwingen zich te bewegen zoals het hem goed uitkomt, net zo min zal een taalstatisticus om zijn werk te vergemakkelijken, afwijken van het taalgebruik zoals dat nu eenmaal historisch gegroeid is. Dat hij daarnaast als staatsburger, bepaalde ideeën over b.v. de spelling mag hebben, kan niemand hem ontfangen.

Ik wil nu een paar toepassingen van de taalstatistiek behandelen. In de eerste jaren van de taalstatistiek was men verzot op wetten. Het bekendste daarvan is geworden de wet van Zipf die zegt dat als men de woorden uit een

tekst op een rij zet met het meest gebruikte woord bovenaan met rangnummer 1, en men de rangnummers van de verschillende woorden vermenigvuldigt met de frekwentie waarmee ze voorkomen, dit product constant zal zijn. Bij logaritmisch uitzetten wordt de grafiek dus een rechte lijn. De wet gaat natuurlijk niet volkomen op. Bovendien zijn de grootheden waartussen het verband bestaat niet bepaald onafhankelijk. Het ergste is, dat mij nooit een toepassing van deze wet onder de ogen is gekomen.

Er zijn nog vele zulke wetten, b.v. het onloochenbare feit dat de woordlengte van hoogfrequentie woorden korter is dan van weinig gebruikte, uit een efficiëntie-oogpunt wel verklaarbaar (maar de taal is niet altijd even efficiënt). De tellingen die tot dit soort „wetten” leiden vind ik eigenlijk van meer belang dan de wetten zelf.

Zo is het, om ook eens een détailpunt te noemen, opvallend hoe het histogram van de woordlengtes, dat van woordlengte 1 (er zijn maar 26 woorden van 1 letter, namelijk de namen van 26 letters) opklimt naar drie (let wel: het gaat hier om alle woorden, dus EEN en HET zijn in hun volle zwaarte aanwezig) en dan regelmatig, langzaam naar nul terugloopt, een merkwaardige deuk vertoont bij de woordlengte 5. Deze deuk is gauw verklaard uit het feit dat woorden nu eenmaal uit een geheel aantal lettergrepen bestaan. Maar pas tellen kan aantonen of deze hypothese juist is. Welnu: na telling bleek ons dat inderdaad de woorden van 1 en die van 2 lettergrepen, die beiden een keurige verdeling met een top en glad verlopen rechterzijde hebben beide bij de lengte 5 laag zijn. Bij de lengte 4 doen de eenlettergrepige woorden nog zwaar mee, en bij de lengte 6 de tweelettergrepige, maar daartussen liggen beide grafieken te laag om als som nog veel op te leveren.

Het klassieke gebied van de taalstatistiek is de geheimschriftkunde. Geheimschriften zijn anders gecodeerde bekende talen. Daarnaast zijn er ook de onbekende talen. Op geheimschriften is de taalstatistiek zeer goed toepasbaar, op de ontcijfering van onbekende talen veel minder. De ontdekking van het lineair B was niet van statistische aard, het belangrijkste was dat men ineens zag dat het een vorm van Grieks was.

Voor de informatietheorie zijn de resultaten van de taalstatistiek van groot belang. Er is in elke taal veel redundantie. De PTT is daar b.v. in geïnteresseerd omdat zij handelt in het vervoer van berichten. Als men daar de redundantie bij de afzender zou kunnen verwijderen, en er bij de ontvanger weer zou kunnen bijdoen, zou zij minder kosten hebben. De PTT-klant past dit principe al toe bij het versturen van telegrammen. Aan de andere kant is het telegram „KOM MORGEN” niet erg duidelijk, en zal er bij verminking van een bericht zonder redundantie weinig meer van te begrijpen zijn. Men kan de mate van redundantie in verschillende talen als volgt illustreren: in de meeste kruiswoordpuzzels komen zwarte hokjes voor. Maar men kan kruiswoordpuzzels ook

maken zonder zwarte hokjes. In het Nederlands is een driedimensionale kruiswoordpuzzel zonder zwarte hokjes, waar men dus behalve horizontaal en verticaal ook nog heeft „van voren naar achteren”, niet goed denkbaar. In het Hebreeuws is dit wel mogelijk. Dat men bij de toepassing van de informatietheorie op natuurlijke talen voorzichtig moet zijn blijkt uit het volgende: er zijn mensen die zonder veel moeite een gedicht van b.v. tweehonderd regels onthouden. Maar niemand kan waarschijnlijk het rijmschema van die lengte onthouden: a a b b a c c a etc., tenzij hij om het te onthouden eerst dat gedicht (dat toch meer informatie bevat) uit het hoofd leert.

Ook talrijke specifieke taalkundige problemen kunnen statistisch worden opgelost. Ik doe maar een wilde suggestie: de verdeling van de DE- en de HETwoorden lijkt volkomen willekeurig, alleen kan men zeggen dat een woord dat eindigt op JE wel een HETwoord zal zijn. Zou het geen goed statistisch onderzoek zijn om de DEwoorden en HETwoorden eens als groepen te vergelijken, dus een factoranalyse toe te passen op allerlei kenmerken als begin, uitgang, lengte? Het is niet te verwachten dat er veel uit zal komen, maar des te verrassender als er wat uit komt.

Een merkwaardige toepassing van taalstatistiek ligt in het steeds nijpender probleem om nieuwe merknamen te verzinnen. In de Verenigde Staten zijn er nu al meer dan een miljoen geregistreerd, dat zijn er dus meer dan een taal aan verschillende woorden heeft. Fabrikanten zoeken voor hun artikelen namen die in veel talen zijn uit te spreken, niet lijken op bestaande en goed klinken. Zo leverde het Mathematisch Centrum jaren geleden 30.000 namen aan Philips-Roxane. Ze hebben nog niet bijbesteld.

Naast deze, en vele onbesproken, toepassingen, kent de taalstatistiek ook haar begrenzungen. De belangrijkste is de structuurverwoestende werking (denk aan de tuuttuuttaal). En de beperking op literair terrein is duidelijk: er is niets tegen om statistiek op literaire teksten uit te oefenen, het is aanmatiging om op grond van die statistiek alleen uitspraken over literatuur te willen doen.

Hoe internationaal de wetenschap van de statistiek ook is, de taalstatistiek is door haar object noodzakelijk nationaal. Wij interesseren ons wel voor andere talen, maar wie zou het Nederlands bestuderen als wij het niet deden?

Daarom wil ik, voor zover die mij bekend zijn, de activiteiten op dit gebied in ons land noemen.

De bekendste taalstatisticus uit Nederland valt daar dan direct buiten, want Professor Guiraud is Fransman en telt Frans. Zo is in Utrecht werkzaam Alinei die een studie heeft gemaakt van uitgangen in het Italiaans.

Tellen is moeizaam werk. In 1935 begon De la Court zijn telling van een miljoen woorden. Later is dit werk in Utrecht weer opgevat door Professor Linschoten.

Zoals u misschien weet is Professor Linschoten vandaag juist een week ge-

leden onverwacht overleden. Het is te hopen dat de zeer ontijdige dood van deze geleerde zijn belangrijke werk niet zal doen afbreken, en dat anderen het zullen voortzetten. Deze psycholoog, die voor zijn studie behoefte had aan gegevens over de taal, is ze, toen bleek dat de taalkundigen hierin tekort schoten, zelf gaan zoeken. Toen hij nog geen twee weken geleden op het Mathematisch Centrum was, waar we toch wel aan grote getallen op het gebied van informatie-verwerking gewend zijn, schokte hij ons door zijn mededeling dat hij niet minder dan een half miljoen pentagrammen op ponskaarten had staan, en door de boekhoudmachine van de universiteit in Utrecht wilde laten bewerken.

De resultaten van De la Court en Linschoten zullen op hun beurt weer het materiaal vormen voor een grootscheepse grammaticale inventarisatie die het Nederlands Instituut in Amsterdam zich voorstelt te gaan maken; het sorteren naar woordsoorten. Hiervoor kan men 200 studenten inschakelen. Bij de vele klachten die men de laatste tijd kan horen over het toenemend aantal studenten, dus ook weer een voordeel van de grotere kwantiteit. Als dit werk af is, zijn dertig jaar voorbij sinds De la Court begon te tellen, en zijn de woorden KOELIE en DIENSTMEID verwisseld van frekwentie met WASMACHINE en BABY-SIT. Net als bij de woordenboekmakers is de wedloop met de tijd bijna niet te winnen. Dit is alleen mogelijk met de modernste middelen, in casu computers. De voeding van deze apparaten moet tot nu toe geheel „met de hand” gebeuren. Een leesmethode, die dus letters kan herkennen, zou uitkomst brengen. Een dergelijk apparaat zal ook wel op de markt verschijnen. Er is nog een andere oplossing: moderne druktechnieken gaan er steeds meer toe over om zetmachines via ponsbanden te bevoorraden. Deze ponsbanden zouden na gebruik heel goed als basis voor taalstatistiek kunnen dienen.

Wat onze bezigheid op het Mathematisch Centrum betreft: in opdracht van een werkgroep van personen uit het gehele land, en met subsidie o.a. van het Delfts Hogeschool-fonds zullen wij, of liever onze computer, een miljoen Nederlandse woorden tellen, uit twintig categorieën van schrijftaal: kranten, romans, zakencorrespondentie, kinderboeken, handelingen der Staten Generaal, wetenschappelijke artikelen enz. De eerste 50.000 woorden zijn reeds binnen, uit tien verschillende kranten. Het bleken 11.000 verschillende woorden te zijn, waarvan er 4000 meer dan eenmaal voorkwamen.

Deze 4000 zijn gepubliceerd, en binnenkort zal een boekje verschijnen met het gehele getelde materiaal, plus de bigrammen, trigrammen, lettergrepen en een statistische beschouwing over de verschillen tussen de kranten onderling. Wij onderscheidden slechts 4 woordsoorten: zelfstandige woorden, bijvoegelijke woorden, werkwoorden, en de rest. Deze, toch summiere, onderscheiding, bleek het moeilijkste gedeelte van het project.

Toen ik zoëven zei: Taalkunde is een empirische wetenschap als de natuur-

kunde, heeft u misschien gedacht: er is toch verschil, want de natuurkunde onderzoekt buitenmenselijke grootheden, terwijl de taal toch door ons mensen zelf wordt gemaakt.

Laat ik dan de vergelijking met de economie maken. Die gaat beter op. Ons economisch leven is, net als onze taal, door onze voorouders en ons, gemaakt. De economen hebben allang begrepen, dat de statistiek hun kan helpen een inzicht in hun object te krijgen. De taalkundigen beginnen het te begrijpen. Economische statistiek en taalstatistiek geven ons inlichtingen over die bouw-sels die we wel zelf gemaakt hebben, maar die toch uiterst taai zijn. En de taal is nog taaier dan de economie. Want een revolutie kon in Rusland wel de economie ingrijpend veranderen, maar niet de frekwentie van het woord „meneer” dat tegen alle propagering van „kameraad” in, zo hoog bleef, dat het gebruik ervan nu weer officieel wordt toegestaan.

Zoals het Centraal Bureau voor de Statistiek onze portemonnaie in de gaten houdt, zo zou er eigenlijk ook een centraal bureau voor de taalstatistiek moeten zijn, dat onze mond in de gaten houdt. Ik sluit me dan ook van harte aan bij het pleidooi dat de redacteur van het Woordenboek van de Nederlandse taal Dr. F. de Tollenaere heeft gehouden voor een Centrum voor Lexicologie in Nederland, zoals b.v. Frankrijk en Italië dat hebben in Besançon en Gallarate. Men zou dan wel „lexicologisch” in zeer ruime zin moeten opvatten, want naast de hoofdarbeid van zo’n centrum: het up to date houden van ons nationale woordenboek, zou de kwantitatieve verwerking van taal in het algemeen het werkterrein moeten zijn.

Laten we een blik in de toekomst werpen, en een dag op het Centraal Bureau voor de Taalstatistiek rondkijken. Iemand is daar bezig met het maandelijks bulletin over de jongste aanwinsten van het Nederlands, zo mogelijk met de bron van de eerste gebruiker. Ergens anders worden de ponsbanden van een boek, die de uitgever heeft afgestaan, toegevoerd aan de Nederlandse woordenschat met bijgehouden frekwenties.

Aan de telefoon is een promovendus die wil weten welke namen van kleuren Vondel gebruikt, en hoe vaak. Hij zal ze morgen thuisgestuurd krijgen. Een filmmaatschappij belt op met de nieuwste ondertitels voor een film: ze moeten herschreven worden met zo kort mogelijke woorden om niet te veel van het filmdoek te bedekken. De computer zoekt de korste synoniemen op. Nu is er een zeepfabriek die een merknaam wil hebben op SIL eindigend, maar voor Turken goed in het gehoor liggend.

Aan zo’n instituut zullen ook statistici verbonden zijn. Ik hoop dat er onder u zijn die voor dit werk belangstelling hebben, ik hoop dat zo’n centrum er komt, ik hoop dat we dan meer zullen gaan weten over dat verschijnsel, dat mij in staat stelt tot u te zeggen: *ik heb gezegd*.

9314 NZ