DE MOGELIJKHEDEN VAN HET STATISTISCH ONDERZOEK IN DE LINGUISTIEK

door L.K. ENGELS,

Katholieke Universiteit te Leuven

Elke wetenschap is in essentie onafhankelijk van gelijkwelke andere wetenschap. Ze is een volledige wezenheid op zichzelf. Twee verschillende wetenschappen hebben gewoonlijk verschillende wetenschappelijke theorieën. Men kan dus de methodes en de theorieën van bepaalde wetenschappen niet zonder meer toepassen of aanwenden in een andere wetenschappelijke discipline. Tegen deze regel werd de laatste jaren maar al te dikwijls gezondigd : als we spreken van mathematisatie in de linguïstiek, dan bedoelen we slechts, dat de mathematica kan *helpen* bij het descriptief, of statistisch bewerken van linguïstisch materiaal, maar we beweren niet dat we aan zogenaamde *mathematische taalkunde* zouden doen, omdat zulke taalkunde door haar notionele vermenging van twee totaal verschillende disciplines op een mislukking moet uitlopen. Het is onmogelijk de linguïstiek, die zoals de sociologie en de economie, een DESCRIPTIEVE wetenschap is, te beoefenen vanuit het standpunt van de mathematica, die een PREDICTIEVE wetenschap is, zoals de scheikunde, de fysica en de sterrekunde.

Het is ook onmogelijk de linguïstiek te beoefenen vanuit het standpunt van de psychologie of de logica, die voor een groot deel NOTIONELE wetenschappen zijn, omdat ze werken met *formele begrippen.*

De eerste voorstanders van de automatische taalkunde en automatische vertaling b.v., waren meestal mathematici en logistiekers (1). We kunnen over hun pogingen, die reeds vijftien jaar aan gang zijn, of waren, in verscheidene instituten voor research, over de gehele wereld verspreid, slechts zeggen, dat ze alle mislukt zijn : de formele aanpak van taal door middel van wiskundige predictieve modellen in plaats van door observatie van de taalfeiten, kon niet bewerken, dat de problemen van de mechanisatie en automatie van analyse en vertaling werden opgelost. Als men die problemen met predictieve modellen wil oplossen dan neemt men tegenover het te bereiken resultaat zo ongeveer dezelfde houding aan als de taalgeleerde die zou zoeken naar de taal die bepaalde

(1) Bar-Hillel, Carnap, Yngve, Hays, Booth, Oettinger e.a.

wezens op de planeet Mars zouden spreken. De linguïst kan zich niet veroorloven om voor zijn onderzoek van de bestaande taalfenomenen eerst een taal te creëren van buitenaards karakter (2).

Zo zullen de statistici ook niet beweren dat hun wetenschap, als empirische discipline, een onderdeel is van de mathematica. De statistiek berust wel op mathematische methodes, vooral dan die van de probabiliteitstheorie, maar ze richt zich vooral tot descriptie en verklaring van geobserveerde patronen, terwijl de mathematica veeleer een totaal abstraherende discipline is, die naar een THEORIE, naar een MODEL zoekt van de werkelijkheid. Soms stelt men de mathematica voor, als zou ze de statistiek omvatten, wat ook weer een valse opvatting is. Al die verwarringen leiden dan in de linguïstiek soms tot de naam *mathematische linguïstiek,* wanneer in feite slechts bedoeld wordt, dat de linguïstiek gebruikt maakt van bepaalde statistische methodes bij haar taalkundig onderzoek, maar het zal dan weer, zoals verder zal blijken, geen totale transpositie zijn van statistische methodes op de taalkunde, we zullen moeten zoeken naar aangepaste *statistisch-linguïstische methodes.* De evolutie in de wetenschap is oorzaak van de verwarring van de begrippen. We hebben immers nood aan elkaar : fysici interesseren zich voor de fonetiek van de taal, de taalkundige begint het nut in te zien van het gebruik van een computer om zijn steekkaarten te vervangen, de statisticus wou graag weten hoe het met sommige taalproblemen gesteld is. In het begin ziet men niet meer klaar : noties en methodes worden vermengd en verward, totdat men tot bezinning komt en de zaken weer duidelijker begint te onderscheiden (3). We zijn aan zo een inzichtelijk keerpunt geraakt, zodat we met een gerust gemoed kunnen handelen over de mogelijkheden van statistisch onderzoek in de linguïstiek.

Taalstatistiek is nog een zeer jonge wetenschap : het eerste statistische onderzoek van grote omvang, op de Duitse taal toegepast, niet met een zuiver linguïstisch doel voor ogen, maar ten dienste van de stenografie, werd gedaan door *Kaeding* in 1898 : het was een frequentielijst van de woorden der Duitse taal op bronnen die zo maar 10.910.777 lopende woorden omvatten (4). Tussen 1920 en 1940 werd er over de gehele wereld en voor verscheidene talen (vooral het Engels) naar woordfrequentie gezocht ; het beste resultaat werd bereikt in 1936 met een lijst voor het Engels : *Interim Report on Vocabulary Selection* (5). De principes, waarmee deze lijst werd uitgewerkt, vonden dan hun toepassing op vele talen.

---

(2) Hans Freudenthal, *Lincos, Design of Language for Cosmic Inetrcouise.* Amsterdam, 1960.

(3) Robert M.W. Dixon, *Linguistic Science and Logic,* Mouton, The Hague, 1963.

(4) Kaeding, *Häufigkeitswörterbuch der deutschen Sprache,* Steglitz, 1897, 1898.

(5) *Interim Report on Vocabulary Selection,* IRET, Tokyo, 1936. Leiding : H.E. Palmer.

Zo verscheen in 1939 een frequentielijst voor het Nederlands door De la Court en Vannes. Na de wereldoorlog geraakte het echt statistische onderzoek van de taal een beetje in de schaduw door de *formele* aanpak van de linguïstiek, die zou leiden naar de overmoedige proeven op automatische analyse, automatische vertaling, automatisch opvangen en ordenen van informatie. Wij weten niet of de mislukking van de pogingen tot mechaniseren van een wezen, zo wispelturig als de taal, de rechtstreekse aanleiding is geweest voor de kentering die duidelijk merkbaar is vanaf 1960 in de linguïstiek : de taalstatistiek, waarvoor de vorsers op het gebied van de automatische vertaling niets anders dan misprijzen kenden, krijgt weer meer waardering. In 1961 begint Karlgren te Stockholm met een tijdschrift voor linguïstische statistiek (6). De theoretische studies over taalstatistiek van Herdan verschijnen in 1960 en 1962 (7). In 1963 maakte Juilland zijn zevenjarenplan bekend om een aantal Romaanse talen statistisch te onderzoeken (8). Bewonderenswaardig is het resultaat van veertig jaar inspannende arbeid (zonder de hulp van een computer om de berekeningen te maken), geleverd door Helmut Meier in zijn *Deutsche Sprachstatistik* (9).

Als reactie tegen de voorstanders van de formele taalanalyse, die een uiting is van romantisch streven naar de niet te bereiken « blauwe Blume », betekent de taalstatistiek een realistische aanpak, moeizaam, langdurig, maar met zekere resultaten, omdat het descriptieve karakter van de taalkunde in wezen wordt erkend, en omdat de taalkundige zo nederig geworden is om te erkennen dat het domein van zijn wetenschap, de levende taal, slechts nauwkeuriger zal kunnen benaderd worden, als vele elementen erin MEETBAAR zijn geworden.

Een primordiale vraag vooraleer we van taalstatistiek kunnen spreken houdt dus verband met het quantitatieve aspect van de taal : WAT is meetbaar in de taal ? Als we Istvan Fodor moeten geloven in zijn werk *The Rate of Linguistic Change,* dan blijft er schijnbaar niet veel meer over om te meten in de taal. Fodor beweert, dat de invloed van de geschiedenis, de cultuur, de geografie, de nabijheid en contact met andere talen, het nationaal karakter van een volk, hoewel EXTERNE invloeden, onmogelijk kunnen gemeten worden of quantitatief voorgesteld. Hij ziet de zaken vooral in een historisch perspectief, d.w.z. het zijn externe invloeden die men noodzakelijk meetbaar zou moe-

---

(6) *Statistical Methods in Linguistics,* Skriptor, Stockholm. 1 : 1961, 2 : 1963, 3 : 1964.

(7) Herdan : *Typen Token Mathematics,* Mouton, 1960 - *The Calculus of Linguistic Observation,* Mouton, 1962.

(8) A. Juilland, *Dictionnaire inverse de la langue française,* Mouton. *Frequency Dictonary of Spanish Words,* Mouton.

(9) Helmut Meier, *Deutsche Sprachstatistik,* G. Olm, Hildesheim, 1964.

ten maken om de taalveranderingen in de loop van de geschiedenis te kunnen verklaren. Maar waarom ons reeds druk maken over de verklaring van de ontwikkeling van de taal, als onze huidige geschreven of gesproken uitingen hun geheim nog niet hebben prijsgegeven ? Er zijn heel wat terreinen — ik beperk me van nu af tot onze eigen Nederlandse taal — die wachten op een quantitatieve aanpak (om te beginnen met louter *descriptieve* statistische methodes) : distributie van lettertekens, syllaben, fonemen ; woordfrequentie, veranderingen in woordgebruik, neologismen, ontlening aan vreemde talen ; frequenties van taalconstructies ; onderzoek van morfemen (buigingsresten, derivatie, woordvorming en dergelijke).

Vermits er voor het Nederlandse taalgebied buiten de frequentielijst van Vannes (1939) nog vrijwel niet aan taalstatistiek werd gedaan zal het interessant zijn te vernemen, wat er op dit ogenblik gebeurt en welke plannen er gesmeed worden voor de toekomst. Aan het rekencentrum van de Amsterdamse universiteit maakte men bij wijze van proef een lijst met afnemende frequentie uit een reeks van 50.000 lopende woorden (uit kranten van 1 bepaalde dag). Het was een voorbeeld van descriptieve taalstatistiek, uitgevoerd door middel van een electronische rekenmachine. De lijst bood weinig linguïstische relevantie, omdat het aantal woorden te klein was om gelijkwat te besluiten uit de resultaten ; de hele operatie was dus waarschijnlijk bedoeld als een proef met een computerprogramma.

In het Instituut voor Toegepaste Linguïstiek te Leuven wordt er in samenwerking met het rekencentrum van de universiteit (onder de leiding van Professor Florin) geëxperimenteerd met computerprogramma's, die in staat moeten blijken om de taal te analyseren en te reconstrueren ; zodra we erin geslaagd zijn door vernuftige maar eenvoudige coderingen aan de machine de opdracht te geven om voor ons woorden, woordverbindingen, constructies, zinsstukken, zinsdelen en dergelijke op bevel uit te zoeken, dan is al wat de machine ons op haar tabelleerlijsten bezorgt meetbaar, telbaar geworden en vatbaar voor descriptieve statistieken, die we naar gelang het werk vordert, zonder ophouden bekomen.

Onze eerste publicatie wordt een werk dat bepaalde statistische onderzoeken kan voorbereiden : het is een retrograad woordenboek zoals het reeds voor het Duits en het Frans is verschenen. De woorden staan er gerangschikt volgens hun uitgangen, zodat het een mooi hulpmiddel wordt voor linguïsten die belangstellen in de uitgangen van de Nederlandse woorden. De vele morfologische kenmerken van het Nederlands, voor zover ze op het einde van de woorden voorkomen kunnen dan gemakkelijker onderzocht worden.

Onze tweede opgave bestaat uit de vervaardiging van een nieuwe frequentielijst van het Nederlands. We zijn met dit groot werk begonnen, omdat

we overtuigd zijn (met Herdan) dat linguïstische elementen onderworpen zijn aan de wetten van de kansverdeling. In de woordenschat is het distributiepatroon gelijk, als een voldoende hoeveelheid woorden werd onderzocht, hoewel in iedere zin de woordenschat afhangt van inhoud, auteur, stijl. Er bestaat een grote uniformiteit van de massa, ondanks de diversiteit van de onderdelen. Fonemen en letters vertonen een grote stabiliteit van distributie, zelfs zo groot dat individuele verschillen totaal worden uitgewist. Er zijn zo'n belangrijke combinatorische wetmatigheden dat het mogelijk is informatietheoretische vergelijkingen te maken, gesteund op een binaire benadering van de onderzochte quanta. Het is begrijpelijk dat een werk zoals dit wordt aangevat, omdat de vooruitgang van de wetenschap ons daartoe aanzet. De linguïsten die de theoretische statistiek van de woordenschat hebben uitgewerkt konden bepaalde beschrijvende formules in wiskundige vorm vastleggen (Herdan, Josselson, Zipf) : dat werd gedaan voor de frequentiedistributie van de tekst, voor het rekenkundig gemiddelde, voor de standaarddeviatie, voor het variatiecoëfficiënt. We hebben de probabiliteitsberekeningen laten uitvoeren voor de resultaten van de frequentielijst van De la Court-Vannes, onze voorganger van 1939 (10), en we stelden vast, dat er gebreken waren aan die frequentielijst : de zeven radii van frequenties samen genomen telden 3.296 woorden ; de zevende radius heeft nog een frequentie van 25 met een fout van 40 %. Gesteld echter dat we een fout van 30 % willen aanvaarden, dan komt op een totaal van 1.000.000 lopende woorden het aantal woorden waarvan de frequentie nog relevant is op slechts 2.200. De zevende radius van de vroegere frequentielijst zou dus moeten vervallen, omdat hij niet voldoende zekerheid biedt. Er zijn nog andere bezwaren tegen deze lijst : Het materiaal is van zeer verschillende aard en het werd zeer ongelijk gemengd. Men had afzonderlijke lijsten moeten berekenen voor elke taalsoort of taalveld. De keuze in de tijd is niet verantwoord : naast *De Kleine Johannes,* geschreven in 1885 (met 21.000 woorden vertegenwoordigd), vinden we krantenartikels van 1930. Er is geen eenheid in de spreiding over 50 jaar. Er zijn vertalingen vertegenwoordigd uit het Engels, het Duits en het Scandinaafs ; er valt geen maatstaf vast te stellen om het materiaal te kiezen : romanliteratuur varieert van 800 tot 21.000 woorden voor I auteur.

Het huidig frequentie-onderzoek gebruikt totaal andere maatstaven. Er worden twee verschillende werkwijzen gebruikt ; maar steeds door middel van electronische rekenmachines. De eerste werkwijze schijnt aantrekkelijk op het eerste gezicht, omdat er zovele bewerkingen automatisch kunnen geschieden : men ponst het materiaal op band of kaarten, nummert de woorden per regel

(10) *Vocabulaire du néerlandais de base*. De Sikkel, 1939.

of per zin, samen met een code voor auteur, soort tekst en dergelijke. Men kan dan een frequentielijst maken van *vormen :* alles staat op woordkaarten die door de machine volledig automatisch kunnen gealfabetiseerd worden, geteld en hun frequenties berekend. De bewerking gaat zeer vlug, maar er is niet het minste onderscheid tussen de homoniemen. Men kan natuurlijk de woorden met hun context opnemen ; de meeste homoniemen kunnen dan achterhaald worden, maar niet automatisch. De tangvormingen van het Nederlands (de gescheiden opstellingen) ontsnappen totaal aan de telling. Er is een enorme geheugencapaciteit nodig in de machine.

In het Instituut voor Toegepaste Linguïstiek te Leuven gaat men uit van een ander principe : men is daar niet zozeer bekommerd om het automatiseren van het proces, als dat moet uitgeboet worden door gemis aan degelijkheid en precisie. Er wordt gewerkt op de tabelleerlijsten die zowel het woord als de context duidelijk weergeven. Daarop wordt dan de zogenaamde pre-editie aangebracht, die door velen als zeer tijdrovend wordt beschouwd, maar die uiteindelijk betere resultaten oplevert, omdat de linguïstische beschrijving correcter en vollediger kan gebeuren en omdat er machines met kleinere capaciteit kunnen gebruikt worden. Ons linguïstisch standpunt verschilt ook sterk van dat van onze voorgangers : wij zijn van oordeel, dat het onmogelijk is bij de analyse van een tekst lexicologie en syntaxis volledig te scheiden. Zelfs bij zogenaamd zuiver lexicologisch werk (het samenstellen van een woordenboek) wordt telkens de syntaxis betrokken. Bij ons onderzoek van de lexicale constructies hebben wij dan ook syntactische elementen betrokken, en bij de behandeling van de zinsconstructie hebben we rekening gehouden met lexicale tekens. De grote moeilijkheid is steeds het vinden van een coderingsformule die soepel genoeg is om allerlei nuances te registreren, om onvoorziene gevallen op te vangen, maar tevens eenvoudig genoeg om geen al te grote inspanning te eisen van het menselijk geheugen en aangepast aan de mogelijkheden van de mechanische behandeling door een electronische rekenmachine.

We zijn er in geslaagd om de machine alle woorden, woordgroepen, zinsdelen, bijzinnen en hoofdzinnen van de tekst te laten *reconstrueren* op bevel. Het zou ons te ver voeren om dit alles hier te tonen. Het volstaat om u te zeggen dat de hele codificering slechts vier kolommen van de ponskaart in beslag neemt : I nummercode van zin tot zin voor elk woord van de zin, die door de machine automatisch wordt aangebracht ; twee kolommen voor een *lexico-syntactische woordsoort* die, in cijfers uitgedrukt, 100 verschillende woordsoorten kan differentiëren ; een kolom voor een *dependentiecode* van woorden

en zinsdelen onderling in de volzin. Door het feit dat de machine kan *reconstrueren,* kan ze ook al het gevondene tellen en de descriptieve frequenties ervan berekenen en tabelleren. Het gebruikte materiaal bestaat op dit ogenblik uit geschreven taal tussen 1960-65. Op dit ogenblik zijn we met proefnemingen bezig met 5.000 zinnen van 20 verschillende Z.N. romanschrijvers, 250 zinnen per auteur (gerekend van punt tot punt). Later breiden we het materiaal uit naar andere *taalvelden* (essays, toneel, luisterspel, T.V.-interview, krantenartikels, wetenschappelijke proza van diverse aard, gevulgariseerde wetenschap, zelfs de *kindertaal* wordt niet uitgesloten) (11). In iedere rubriek wordt natuurlijk nog onderscheiden per auteur of per bron. Het voordeel van deze handelwijze ligt in de *spreiding :* voor ieder woord kunnen we die spreiding berekenen : gemiddelde en standaardafwijking σ en variatiequotient v. Het variatiequotient moet ons de maatstaf geven om een bepaalde spreiding te aanvaarden. Wat het resultaat zal zijn van deze bewerkingen kan nu nog niet voorspeld worden. Het is zeer waarschijnlijk dat niet alles in één frequentielijst zal kunnen opgenomen worden. Er moet ongetwijfeld een gemeenschappelijke woordenschat bestaan voor alle taalvelden — een gemeenschappelijke overlapping. Maar waar zal de grens liggen ?

We hebben ons tot nog toe beperkt tot geschreven taal. Allicht zullen we later trachten ook de *gesproken* taaluitingen in dit onderzoek te betrekken, als we de technische middelen hebben gevonden om werkelijk gesproken uitingen « in de vlucht » te capteren.

Het spreekt vanzelf dat we ook een onderscheid zullen maken tussen de bekomen woordenschat-frequentielijst, die alle woorden met hun frequenties en spreiding bevat, en een zogenaamde basiswoordenschat, waarin slechts een beperkt aantal woorden zal opgenomen worden, nl. de *nuttigste,* naar gelang het doel, waartoe de lijst zal moeten dienen.

De taalstatistiek krijgt na 1960 stilaan burgerrecht in een domein van de taalwetenschap, dat zelf met moeite zijn plaats als wetenschappelijke discipline aan onze universiteiten kon veroveren : de *toegepaste linguïstiek.* Zij zal meer en meer erkenning genieten, omdat ook de beschouwende filologen hebben ingezien, dat ze best met hun redeneringen vertrekken vanop vaste grond, d.w.z. vanuit precies gemeten en quantitatief berekende gegevens uit de taalwerkelijkheid. Misschien lukt het de *automatische vertaling* nog ooit door middel van de taalstatistiek tot reële in plaats van ingebeelde algoritmen te komen, die nodig zijn om transformatieregels op te stellen. Op een groot

(11) Op dit ogenblik is een onderzoek aan gang in het I.T.L. over de geschreven taalbeheersing moedertaal (opstel) van alle lln. 1e middelbaar (technisch, 4e graad, M.O.) in het vrij ondw., Vlaams België. De statistische staaltrekking werd gemaakt door de diensten van Dr. R. Van den Bosch : *Dienst voor Statistiek en Planning,* N.S.K.O., Brussel.

aantal wetenschappelijke gebieden zal de taalstatistiek in de toekomst haar woordje kunnen plaatsen : de *dialectologie* zou er zeker goed aan doen de statistische methodes in haar onderzoek te betrekken, voor de nieuwe richting in de algemene taalwetenschap, nl. de *taaltypologie,* is statistisch onderzoek onontbeerlijk. In de psychologie, vooral dan in haar toegepaste gebieden zoals de *audiometrie* en de *logopedie* beklaagt men zich maar al te dikwijls over het gebrek aan taalstatistische gegevens om gehoorstesten, verstaanbaarheidstesten of taalvaardigheidstesten te kunnen opstellen. De linguïst moest verlegen zijn als hij psychologen moet wandelen sturen met lege handen : het onderzoek is pas begonnen, hij kan nog geen geldige resultaten meedelen. Ook de *litteraire kritiek* zou gerust haar intuïties eens mogen toetsen aan de objectieve gegevens over stijl, die door middel van quantitatieve methodes kunnen bekomen worden.

Het hoeft geen betoog dat de taalstatistiek vooral het vreemde-talenonderwijs kan ten goede komen. De opzet van de tellingen van het Nederlands zoals ze aan de universiteit te Leuven gebeuren hebben eerst en vooral tot doel aan het onderwijs van het Nederlands als tweede taal meer objectieve gegevens te verschaffen over woord en vooral structuurfrequentie om de doeltreffendheid van dat onderwijs te kunnen opvoeren. De taalstatistiek kan ook diensten bewijzen aan het onderwijs van het Nederlands als moedertaal : bij het opstellen van de eerste leesboekjes voor de eerste studiejaren zou men rekening moeten houden met de frequenties van lettertekens, met fonetische schrijfwijzen van de woorden, met gelijkvormigheden en ongelijkvormigheden van de spellingsconventie. Als de taalstatistiek verder gevorderd was, dan zou het veel gemakkelijker zijn een *spellingshervorming* voor te stellen voor talen met internationale spreiding maar hopeloos verouderde schrijfwijze zoals het Engels en het Frans. Laat ons hopen dat bij een volgende spellingswijziging de taalstatistiek eindelijk een helpend handje zal mogen toesteken. De taalstatistiek is ontstaan uit de stenografie — althans in Duitsland — en heeft daar haar bekroning gevonden in een 40-jaar lang onveranderlijk bewaarde eenheidsstenografie. Het spreekt vanzelf dat de huidige Duitse stenografen klagen over het feit, dat de Kaedinglijst niet meer beantwoordt aan de evolutie van de Duitse taal, vooral dan na de tweede wereldoorlog. Maar dat valt in het voordeel uit van de taalstatistiek : men wil een nieuwe Duitse frequentielijst ! Tenslotte kan de taalstatistiek grote diensten bewijzen aan de verbetering van de schrijfmachines: men komt over het algemeen tot het inzicht dat niet een internationaal systeem voor alle talen nuttig is, maar dat een klavier de beste diensten bewijst als de opstelling van de toetsen beantwoordt aan de statistische spreiding van de lettertekens in een bepaalde taal. Zo bewijst de taalstatistiek diensten aan de *psychotechniek,* die zich o.a. bezig houdt met de rationalisatie van de schrijfmachine en haar bediening.

De taalstatistiek heeft tot nu toe niet vele beoefenaars. De taalstudie werd tot nu toe meestal gerekend tot de zogenaamde geesteswetenschappen : ze wordt ingedeeld aan onze universiteiten bij de *wijsbegeerte en letteren*. Het is niet gemakkelijk een *notionele* denkwijze in te ruilen tegen objectieve zienswijzen door middel van statistische methodes ; het is voor de linguïst een hele stap om zijn aloude steekkaart te laten vervangen door ponskaarten, electronische banden of tabelleerlijsten ; het valt hem nog moeilijker zijn regels en inzichten te abstraheren en te beschrijven door middel van mathematische formules. Maar de tijden veranderen en de jeugd die nu aan onze universiteiten met haar studies begint, denkt anders dan vroeger : het zal mogelijk zijn om bij die jongeren enthousiasme op te wekken voor een hulpwetenschap, die in de nabije toekomst heel wat diensten zal kunnen bewijzen aan de maatschappij.