DE VERWERKING VAN LINGUISTISCHE OPGAVEN D.M.V. EEN COMPUTER

W. MARTIN

Daar ons slechts weinig tijd ter beschikking staat zullen wij ons tot een summiere weergave van de LDP (language data processing) - activiteit aan het ITL (Instituut voor Toegepaste Linguistiek, Leuven) moeten beperken.

Deze uiteenzetting zal dan ook uiteenvallen in :
1. een *schematisch overzicht* van de computer-programma's die het ITL ter beschikking heeft;
2. een bespreking aan de hand van enkele voorbeelden van de mogelijke *betekenis* van deze *bewerkingen;*
3. tot besluit zullen wij dan proberen een antwoord te geven op de vraag *hoe* de computer tot nog toe op het ITL wordt gebruikt.

1. *SCHEMATISCH OVERZICHT VAN DE COMPUTER-PROGRAMMA'S AAN HET ITL*

Het ITL werd in 1963 opgericht aan de Leuvense Universiteit. In de aanvang werd vooral aan research op het gebied van het taalonderwijs gedacht. Sinds 1967 werd bij die research overgeschakeld eerst naar mechanografische (pons-, sorteer-, en tabelleermachines), later naar elektronische apparatuur (computers). Na een aanpassings- en leerperiode kon vanaf 1968 voorgoed aan language data processing gedacht worden.
Daarbij moet er rekening gehouden worden met het feit dat de krachtlijnen bij het onderzoek sinds de aanvang enigszins verlegd zijn van pragmatische taalonderwijsresearch naar meer theoretische kwantitatieve taalstudie, contrastieve linguïstiek en computertaalkunde (hier als linguïstische automering op te vatten cfr. infra).

Uiteraard gaat het hier om een vrij jonge computerervaring. Het software-pakket (1) dat het ITL ter beschikking kan stellen is dan ook nog zeer beperkt. Schematisch zouden wij het als volgt kunnen indelen :

1. Classificatieprogramma's
   1.1. Software programma's
   1.2. Reorganisatieprogramma's

Extrait de la Revue (R.E.L.O.)
VII, 1 à 4, 1971. C.I.P.L. - Université de Liège - Tous droits réservés.

1.2.1. Translatie
1.2.2. Selectie

2. Statistische programma's
   2.1. Mathematische statistiek
   2.2. Lexicale statistiek
   2.3. Statistische hulpprogramma's

3. Linguïstische automering
   3.1. Lexicale automering
   3.2. Syntactische automering

Het is de bedoeling de gebruikte terminologie toe te lichten en de resp. hieronder ressorterende programma's te vermelden.

A. *Classificatieprogramma's.*

Hieronder laten wij alle programma's ressorteren die niet bij de statistische programma's of de linguïstische automering thuishoren.
Om het positief uit te drukken : we spreken van classificatie wanneer het gaat om de bewerking van alfanumerieke (alfabetische of numerieke) gegevens via eenvoudig-logische procedures die het inputmateriaal geheel of gedeeltelijk op een andere manier ordenen of uitdrukken.
Zo komt het dat wij hierbij zowel software-programma's (programma's die door het uitbatingssysteem worden geleverd en dus niet zelf te programmeren zijn, -enkel bepaalde parameters moeten worden ingevuld-), als zelf geprogrammeerde reorganisatieprogramma's rangschikken.
Bij de software-programma's horen dan o.m. de *copy-, sort- en print-out utilities* thuis waardoor het mogelijk wordt machineleesbare data te copiëren, te sorteren of uit te drukken. In wezen verschillen deze programma's niet van de zg. reorganisatieprogramma's, daar het ook hier gaat om het omwerken van een bestaand materiaal A tot een nieuw materiaal B. Het enige verschil is dat het in het tweede geval gaat om eigengecreëerde programma's, niet aanwezig in het software-pakket dat bij het uitbatingssysteem hoort. De reorganisatieprogramma's verlopen volgens twee principes :
- translatie,
- selectie.

Bij translatie is er een volledige commutatie mogelijk tussen A (input) en B (output); bij selectie is dit niet meer het geval. De voornaamste *translatieprogramma's* aan het ITL zijn :

a.- het *zin-woord programma* : vertrekkend van *zinskaarten* d.w.z. ponskaarten met doorlopende tekst wordt een *woordband* gemaakt, d.i. een magnetische tape ingedeeld in logische records die als voornaamste informatie het "woord" plus een aantal andere gegevens bevat. Aldus worden classificaties die het "woord" als eenheid nemen mogelijk.

b.- het programma *samentrekken* : het comprimeren van een data-set zo dat identieke tokens (samengebracht door een voorafgaande sort) per type worden opgegeven.

c.- *het indexeren* : de tokens worden niet alleen tot types gecomprimeerd, maar ook de vindplaatsen der tokens per type worden gegeven.

d.- *het concordantie*-programma : een gedeelte context der tokens per type wordt opgegeven. Een variant hiervan is de *KWIC-index* : de tokens staan per type in het midden van de bladzijde, links en rechts omgeven door eenzelfde hoeveelheid context. Niet noodzakelijk hoeven de types hier als woorden gedacht te worden, men kan ook een wat we zouden kunnen heten *KSIC-index* maken, een Key-Sentence-in-Context-index waarbij b.v. zinnen met identieke structuur één type uitmaken.

e.- *de retrogradering of invertering* : om het woord te kunnen sorteren van rechts naar links i.p.v. - zoals gewoonlijk bij alfabetische sort - van links naar rechts (nodig om een inverte lijst te bekomen) converteert de computer aan de hand van dit programma het naar links gealigneerde woord tot een naar rechts gealigneerd woord en neemt daarbij ook het spiegelbeeld van het woord op. Wanneer het inverteringsprogramma is uitgevoerd kan via een sort en een tape-print een inverte lijst bekomen worden. Het volstaat dat de computer sorteert op het spiegelwoord en in de bekomen volgorde de naar rechts gealigneerde woorden afdrukt.

f.- Tenslotte kan men ook *fusieprogramma's* en *programma's met speciale print output* als translatieprogramma's beschouwen. Hierdoor is het mogelijk resp. een data-set tussen een reeds bestaande te voegen, anderzijds data-sets volgens bepaalde formaten te laten drukken.

Een tweede reeks classificatieprogramma's hebben wij *selectie-programma's* genoemd. Hier wordt de input data-set niet meer in globo behandeld, maar één of meer elementen worden eruit geselecteerd.

Wij onderscheiden twee soorten programma's :
- *ad-hoc selecties*
- *het vergelijkingsprogramma.*

Bij ad-hoc selecties wordt één element of een verzameling elementen opgegeven die men uit een data-set wil selecteren. Aldus kan men op die wijze uit een Nederlands taalsample alle woorden verkrijgen die op *-heid* eindigen, zes grafemen tellen, met *ver-* beginnen enz. Het volstaat telkens het basisprogramma even aan te passen. Is het input-materiaal op een daartoe bestemde wijze (manueel of machinaal) gecodeerd dan is het evenzeer mogelijk b.v. woordsoorten, woordgroepen, zinsstructuren e.d. uit dit materiaal te selecteren.
Daarnaast staat een algemener programma dat wij het vergelijkingsprogramma hebben genoemd : uitgaande van twee of meer samengetrokken banden spoort dit programma alle gemeenschappelijke elementen (b.v. woorden) op ($A \cap B$: A intersectie B), of alle exclusieve elementen voor A ($A \setminus B$: A negatie B), of omgekeerd ($B \setminus A$: B negatie A).

B. *Statistische programma's.*

Statistische programma's doen niet langer een beroep op een eenvoudige translatie- of selectieprocedure, het algoritme is nu een statistische, wiskundige formule.
Wij hebben een driedeling gemaakt in :
- mathematische statistiek
- lexicale statistiek
- statistische hulpprogramma's.

Het onderscheid tussen a en b ligt hierin dat a een beroep doet op de wiskundige probabiliteitsleer, terwijl b rekening houdt met linguïstische empirie en observatie bij het opstellen van het probabiliteitsmodel of de formule. De berekeningen uit a en b kunnen rechtstreeks gebruikt worden door de taalkundige, bij c is de statistische bewerking slechts een eerste fase in de bewerking der gegevens. Vandaar de indeling in 3 subgroepen.

Laten wij op elk van deze groepen even ingaan.

Ad a : Mathematische Statistiek.

Vooraf zij gezegd dat het *software-pakket* van het uitbatingssysteem vaak ook *statistische routines* bevat. Er werd de voorkeur gegeven aan eigen geschreven programma's omdat de genoemde subroutines vaak te algemeen waren en te omslachtig in het gebruik.
Samen met de programma's die voorkomen bij de mathematische statistiek bespreken wij hier ook het zg. programma *Statistiek Voca*. In feite gaat het hierbij voornamelijk om *beschrijvende statistiek*. Van een aantal uitkomsten of waarnemingen wordt een frequentiedistributietabel met absolute en relatieve waarden afzonderlijk en cumulatief opgegeven. Ook een grafiek van de verdeling, samen met de 4 momenten (gemiddelde, standaarddeviatie, scheefheid en welving), en andere statistische grootheden als v (variatiecoëfficiënt) worden daarbij aangeduid.
Programma's die wel thuishoren in de mathematische statistiek zijn dan o.m. programma s die de *correlatie* tussen reeksen variabelen berekenen (volgens de coefficient van Pearson) ; verder de berekening van de *rangcorrelatie* (coëfficiënt van Spearman), de *regressierechte* (zo is het mogelijk de beste schatting van y voor een gegeven x af te lezen), de *Standaardfout van het verschil tussen twee proporties* (SE_{dif}), de *Chi-Kwadraat*, enz.

De Chi Kwadraat-toets is trouwens een subroutine die opgeroepen wordt bij *het testen van mathematische modellen*. Het ITL beschikt over een programma waardoor het mogelijk is frequentieverdelingen van waarnemingen te toetsen aan de normale, log-normale en Poisson-verdeling. Dit programma berekent een aantal statistische grootheden (4 momenten der verdeling, v, modus, mediaan enz.), maakt een gecombineerde grafiek van de geobserveerde en van de geexpecteerde waarden en vergelijkt die twee d.m.v. een $\chi 2$-toets.

Ad b : Lexicale Statistiek.

Hieronder kunnen we het programma dat de *entropie* berekent vermelden. Verder de berekening van de *U(= usage)-waarde* (d.i. een grootheid die de frequentie relativeert door de dispersie) (JUILLAND). Een ander programma bepaalt de *sleutelwoorden* van een tekst. Sleutelwoorden worden gedefiniëerd als woorden die een significantafwijkend frequentieverschil vertonen in het werk van de bestudeerde auteur t.o.v. een geëxpecteerde frequentie, berekend op grond van een basisvocabularium. Voor de berekeningswijze verwijzen wij naar GUIRAUD (p. 41 en vgl.).

Het Instituut beschikt tevens over een programma dat een vocabularium van N tokens kan reduceren tot een vocabularium van N' tokens. (*Reductie Voca*). Wij steunden hierbij op de binomiale verdeling of de wet van Newton. Deze berekening werd reeds eerder toegelicht o.a. door MULLER en EVRARD. Wij hebben dit programma o.m. gebruikt bij het reduceren van het vocabularium van *Het Afscheid* (42.865 tokens) tot dezelfde grootte als *Het Boek Alfa* (33.489 tokens) (beide romans van de Vlaamse auteur Ivo Michiels) en bij het berekenen van de theoretische groei van beide vocabularia. Aldus worden statistische grootheden als V (aantal types), f (gem. frequentie), V_1 (aantal hapax legomena) van verschillende vocabularia onderling vergelijkbaar, wat vóór reductie bij vocabularia met verschillende N (aantal tokens) niet het geval is. (MARTIN (a), p. 38 en vlg.). Dit programma berekende dus voor N' : V' en tevens de spreiding van V' over de f-klassen. Een belangrijke grootheid in de lexicale statistiek is de *K-factor van YULE*. Ook daar werd een subroutine voor geschreven. Aldus is het mogelijk de concentratie, de zg. repetitiegraad, van vocabularia met elkaar te vergelijken en na te gaan of deze grootheid inderdaad betrouwbaar is (bij gereduceerde vocabularia b.v. moet de K-waarde gelijk blijven aan die uit het niet gereduceerde vocabularium).

Tenslotte is er nog de *wet van ZIPF* (met de correctie die MANDELBROT aanbracht) die voor computer werd geprogrammeerd. Dit programma vergelijkt via tabellen en grafiek de geobserveerde rechte en de verwachte (op basis van de wet van Zipf (fxr =C) en de correctie van Mandelbrot).
Nog een laatste opmerking : het zal wellicht opgevallen zijn dat wij spraken van *lexicale* statistiek. Inderdaad de meeste bewerkingen die hieronder ressorteren hebben betrekking op het woord, m.m. kunnen de meeste van deze programma's echter ook op andere niveau's worden overgeschakeld.

Ad c : Statistische Hulpprogramma's.

Het voornaamste programma waarover het ITL beschikt is het zg.

Randu 1 en 2.

Randu 1 genereert randomgetallen of toevalcijfers. Het genereren zelf van deze getallen is een bewerking die hetzelfde resultaat geeft als een blinde loting van x (= het gewenste aantal) getallen uit een reeks van 1 tot en met N. (1 begin van de reeks, N = einde van de reeks).

Randu 2 neemt deze getallen als input voor het bekomen van een lijst at random gekozen linguïstische eenheden (b.v. woorden).

Aldus is een random sampling mogelijk.

C. *Linguïstische automering.*

Wij spreken van linguïstische automering wanneer het algoritme waarop het programma steunt niet langer een classificeringsprocedure of een kwantitatieve formule is, maar één of meer *grammaticale regels* omvat. Het probleem dat zich voordoet kan niet door een classificatie, noch door een kwantitatieve bewerking opgelost worden. Hiervoor is nu één of een stel grammaticale regels vereist.
Naar gelang van het niveau waar deze regels te situeren zijn hebben wij de beschikbare programma's in twee groepen onderscheiden, t.w.
- *lexicale automering*
- *en syntactische automering.*

De lexicale automering verloopt op het woord-niveau terwijl de syntactische automering te situeren is op een ruimer plan nl. op dit van de woordgroep of zin.
Het aantal programma's waarover het ITL op dit gebied beschikt is niet zo groot, o.m. ook omdat het noodzakelijk was eerst minder complexe problemen aan te vatten.
Bij de lexicale automering kunnen we allereerst een tweetal *syllaben-detectie-programma's* vermelden. Het eerste heeft betrekking op het *Nederlands,* het tweede op het *Engels.* Het principe en de werkwijze voor het Nederlandse syllabenprogramma vertoont veel gelijkenis met dit van BRANDT CORSTIUS (p. 7 en vlg.) en is er trouwens door geïnspireerd. Ook hier stelt de computer het aantal syllaben gelijk aan het aantal klinkers of klinkerclusters in het woord. Het komt er enkel op aan deze te bepalen. Dit gebeurt niet zoals bij Brandt Corstius door het extenderen van vocalen tot syllabendragers (via een matrix), maar juist andersom door het vergelijken van groepen letters beginnend met vocaal, van lengte 3, vervolgens 2 en tenslotte 1 letter met een vector die de vocalen en vocaalclusters (max. lengte 3 in het Nederlands) bevat.

Dit programma geeft als output een vector met de woordlengte in grafemen, in syllaben en met de grootheden $\bar{s}$ (gemiddelde syllabenlengte per woord in de behandelde tekst) en $\bar{g}$ (gemiddelde woordlengte van de groepen woorden met gelijke syllabenlengte in de tekst).
Deze grootheden dienen voor een onderzoek naar objectieve tekstkarakteristieken. Het programma ter detectie van het aantal syllaben in het Engels is op hetzelfde principe gebaseerd, nl. dat het aantal syllaben = aantal vocalen, of vocalenclusters. Hier moet tevens o.a. met een hele reeks "mute endings" (-ed, -ble, etc.) en met een reeks vocaliseringen van consanten (b.v. in de uitgang -ism) rekening worden gehouden.

Een ander programma is een poging tot automatische ontleding van *samengestelde Nederlandse woorden* in hun samenstellende delen. Daar dit het object is van het referaat van Dhr. VAN OVERBEKE wordt op dit programma dat o.m. een eerste aanloop biedt tot het lemmatizeren van woordvormen (tokens) tot lexemen (types) niet verder ingegaan.
Blijft tenslotte de *syntactische automering*. Hier heeft het ITL een basisprogramma uitgewerkt dat in feite in verschillende subprogramma's uiteenvalt.
De bedoeling is Engelse teksten (i.c. wetenschappelijke literatuur) te voorzien van een dubbele cijfercode : de woorden uit de tekst worden voorzien van een zg. *syntactische woordsoort* d.w.z. dat zij naar gelang van hun morfologische valentie een 1e cijfer krijgen dat varieert van 0 tot 9, naar gelang van hun syntactische functie krijgen zij een tweede cijfer eveneens variërend van 0 t.e.m.9.
Het geheel functioneert via een lexis (d.i. een gecodeerde woordenlijst van ongeveer 3000 items), een systeem van uitgangen, en een aantal woordvolgorderegels. Aldus kon tot op heden grotendeels de adjectivische, de nominale en de verbale functie, worden gedetecteerd (ENGELS (a) en (b)).

2. *NUT EN BETEKENIS DER BEWERKINGEN*

Het is de bedoeling aan de hand van enkele voorbeelden de drie reeksen programma's die hier ter sprake kwamen te illustreren en vooral het mogelijke nut van deze bewerkingen voor linguïsten te belichten.

2.1. *Classificatieprocessen.*

Op dit gebied publiceerde het Instituut tot op heden o.m. een aantal *indexen* en *inverte lijsten.*
Meestal wordt bij dergelijke werken de nadruk gelegd op het *instrumentele* karakter ervan. Ongetwijfeld terecht.
Het gaat echter niet op ze daarom ook als periferische classificaties dood te verven en ongebruikt te laten. Het komt er o.i. vooral op aan dat de linguist of literatuuronderzoeker deze werkinstrumenten gaat gebruiken en aldus juist de periferie van het classificeren en het tellen doorbreekt om tot een diepere "taalkennis" te komen.
Als men b.v. het aantal indexen en concordanties overziet die er voor het Nederlands bestaan (cfr. INL) dan lijkt het verwonderlijk hoe weinig daarvan wordt gebruik gemaakt door "outsiders" dit zijn de mensen die zelf niet bij computer-research betrokken zijn. Gaat het hier om de beruchte huiver, de negatieve sensibiliteit van de alfa-wetenschappers tegenover de techniek, of gaat het eerder om onwetendheid, een gebrek aan inzicht in de mogelijkheden van het instrument in kwestie ? Wij laten deze vraag in het midden. Het zij ons echter toegestaan op enkele mogelijkheden o.m. in verband met inverte lijsten te wijzen.
Zopas verscheen de *Inverte Frequentielijst van het Nederlands* (MARTIN, (b)). Deze lijst is gebaseerd op de frequentietelling die J.F.H.A. De la Court in 1935 maakte, en die door J. Linschoten in 1963 werd herzien.
Het materiaal bestaat uit één miljoen woordtekens en 23.463 woordtypen.

Naast de klassieke bedoelingen eigen aan elke inverte of retrograde lijst (t.w. hulp bij onderzoek naar woordvorming, bij reconstructie van teksten, en bij het samenstellen van rijmwoordenboeken) heeft deze lijst een meer specifiek doel : de voornaamste opzet is nl. de linguïst en literaire onderzoeker een instrument te bezorgen waardoor het mogelijk wordt *een probabiliteitsmodel voor suffixen en andere eindmorfemen in het taalgebruik op te bouwen.* Om de verwachte frequentie van een eindmorfeem in een tekst vast te stellen volstaat het de frequentie van de woorden in de lijst die op dat morfeem eindigen, samen te tellen en het bekomen getal tot de orde van grootte van de onderzochte tekst te herleiden.
Zo gezien, is de inverte frequentielijst relevanter dan de gewone alfabetische, daar de nadruk nu op meer onbewuste taalelementen zoals suffixen ligt, in plaats van op woorden. Op het eerste niveau zal de probabiliteit ongetwijfeld een grotere rol spelen dan op het tweede, waar slechts weinig elementen een stabiele relatieve frequentie bezitten.

2.2. *Statistische programma's.*

Als exemplum kozen wij hier het toetsen van linguïstische data aan mathematisch-statistische hypothesen of modellen. Laten wij dit met een recent onderzoek illustreren. Onlangs onderzochten wij de distributie van nieuwe woorden in *Mei* van H. Gorter (MARTIN, (c)). Wij gingen daarbij niet van de tekst uit maar wel van de volgende vaststelling : de nieuwe woorden die wij in *Mei* detecteerden kwamen op een paar uitzonderingen na alle met f = 1 voor. Welnu, de kans op voorkomen in een tekst van woorden met f = 1 is in tegenstelling met de woorden met f > 1, overal gelijk. Verdelen wij een tekst in 10 segmenten dan zal daarentegen de probabiliteit van woorden met f = 10 om voor het eerst op te treden kleiner zijn voor deel X (laatste segment), dan voor deel X-1 b.v.
Gegeven het stijlcriterium nieuwe woorden ging het dus om volgende alternatieven :
- ofwel is hun aantal homogeen verdeeld over de hele tekst;
- ofwel is hun aantal niet homogeen verdeeld. De vraag is dan hoe de verdeling is en waarom.

Vertrekkend van deze (literair niet relevante) nulhypothese ontdekten wij dat :
1. Balders lied significant afweek van de rest van de tekst (per tekstsegment (3000 woorden) verwachtten wij 45 nieuwe woorden, in Balders lied alleen (789 woorden) vonden wij er al 42!).
2. de distributie voor de rest de klassieke bouw van het gedicht volgde : stijgend van Zang 1 naar Zang 2 die het hoogtepunt vormde en dalend van Zang 2 naar Zang 3.

M.a.w. dit onderzoek bracht ons tot een nieuw inzicht in de werkwijze van de auteur. Daartoe kwamen we via een statistische hypothese die pas in tweede instantie verklaard werd door een (literaire) teksthypothese, m.a.w. het omgekeerde van wat bij stijlonderzoek gewoonlijk het geval is.
Om de zaken scherp te stellen moeten wij er nog het volgende aan toe voegen :

1. Het feit dat men b.v. "nieuwe woorden" onderzoekt veronderstelt natuurlijk reeds een stilistische hypothese,nl. deze dat een nieuw woord, hoe dat ook gedefinieerd mag worden, (cfr. MARTIN, (c)), een relevant stijlfenomeen is.
2. Het blijft echter waar dat het onderzoek niet noodzakelijk hoeft voort te spruiten uit een uit de tekst gewonnen inzicht. M.a.w. het experiment op zich is helemaal niet zo gek als dit op het eerste gezicht lijken mag. Columbus "wist" dat de aarde rond was en daarom trok hij er op uit om Indië te bereiken en dit weten voor de andere mensen inzichtelijk te maken. Was hij zonder meer uitgevaren op ontdekkingsreis dan zou hij wellicht evenzeer Amerika hebben ontdekt.
3. Het experimenteel onderzoek is thans echter een veel minder riskante en "uitzichtsloze" onderneming omdat wij over veel betere hulpmiddelen beschikken dan vroeger.
   Dit geldt ook voor literair onderzoek : dank zij de computer kan een experiment snel en efficiënt uitgevoerd worden.
4. Het blijft waar dat het interpretatieprobleem daarmee niet is weggewerkt. M.a.w. de afstand tussen observatie en evaluatie blijft bestaan, hoewel er o.i. toch een fundamenteel onderscheid bestaat tussen de traditionele stijlstudies en stijlstudies per computer : in het eerste geval begint men heel vaak met een conclusie die men bewijzen moet, computerstudies daarentegen genereren gewoonlijk deze bewijzen experimenteel zonder van een verdoorgedreven teksthypothese gebruik te moeten maken. Deze blijkt o.i. veel makkelijker achteraf, post factum, uit de computerevidentie gehaald te kunnen worden.

2.3 *Linguïstische automering.*

Een programma als dit van prof. Engels dat Engelse zinnen automatisch van een grammaticale code voorziet biedt de linguïst o.i. volgende voordelen :

a. Allereerst laat dit programma een beschrijving toe op een niveau dat hoger ligt dan dit van het woord : woordgroepen en zinnen kunnen nu worden aangepakt. Dit wil niet zeggen dat het woordniveau volledig wordt verlaten, integendeel nuttige en nodige informatie ter onderscheiding van homografen en homoniemen b.v. wordt nu verkregen.
b. De codering verloopt volledig konsekwent d.w.z. uniform, en zeer economisch d.w.z. in een minimum van tijd.
c. De computer kan automatisch een woordenboek aanleggen van de in een bepaald taalgenre gebruikte substantieven en adjectieven, en aldus ingeschakeld worden in content-analysis en information-retrieval (ENGELS, (c)).
d. Wellicht het belangrijkste aspect hierbij is de mogelijkheid tot snelle en efficiënte evaluatie van de gebruikte regels ; de computer levert een eerste interpretatie (surface-structure) vertrekkend van de "naakte" tekst.

3. *BESLUIT*

Wij hebben bij het overzicht van de programma's aan het ITL beschikbaar, gesproken van
- classificatieprogramma's,
- statistische programma's,
- linguïstische automering.

Deze indeling berust op de *aard* van het algoritme dat de kern vormt van het programma.
Als wij ons nu niet langer op het standpunt van de *programmeur* maar op dat van de *linguïst* zouden stellen die vraagt *hoe* aan het ITL de computer wordt gebruikt, dan menen wij het antwoord als volgt te mogen samen vatten :

1) als *administratie- en rekenmachine* : hierdoor kunnen allerlei nuttige basisinstrumenten gecreeerd worden en waarnemingen worden verricht, die tot dieper "taalinzicht" kunnen leiden op voorwaarde dat zij inderdaad gebruikt worden.
2) als *controle-machine* : ter verificatie van bepaalde uitkomsten (onderzoek naar significantie) of ter evaluatie van bepaalde hypothesen. Hierbij wordt de mogelijkheid tot een soort experimentele linguïstiek of stilistiek geschapen, die niet-linguïstische hypothesen (b.v. statistische) of niet volledig uitgewerkte linguïstische of literaire hypothesen, snel en efficiënt kan evalueren.
3) als *simulatiemachine* : bij sommige programma's voor linguïstische automering (b.v. syllabendetectie Nederlands) poogt men te komen tot een volledig uitgewerkt model dat werkt als de werkelijkheid en dat als theorie voor de werkelijkheid kan worden beschouwd (VAN DIERENDONCK).

Het is de bedoeling in de toekomst de research van het ITL vooral op punt 2 en 3 toe te spitsen, zonder daarom de eerste gebruiksmogelijkheid van de hand te wijzen.

(1) Zoals men verder zal merken wordt *software* in dit artikel in twee betekenissen gebruikt t.w.
1. in ruime zin : voor het totale programmapakket;
2. in engere zin : voor een onderdeel van 1, nl. voor deze programma's die de computerfabrikant c.q. het rekencentrum ter beschikking stelt.

H. BRANDT CORSTIUS, *Exercises in computational linguistics,* Amsterdam, 1970.

L.K. ENGELS, a) *Automatic Analysis of English,* in ITL, 2 (1968), 19-36;
b) *Automatische Analyse van het Engels,* Linguistica Antverpiensia, II (1968), 177-188;
c) *Automatische herkenning van het genre van een Engelse tekst,* Handelingen v. h. XXVIIe VI. Filologencongres, 250-257.

E. EVRARD, *Deux programmes d'ordinateur pour l'étude quantitative du vocabulaire,* in Revue, 1967, nr. 3, 81-95.

P. GUIRAUD, *Problèmes et méthodes de la statistique linguistique,* Dordrecht, 1959.

INL, *Informatie Nederlandse Lexikologie,* nr. 1, juni 1970.

A. JUILLAND en E. CHANG RODRIGUEZ, *Frequency Dictionary of Spanish Words,* Den Haag, 1964.

B. MANDELBROT, *Structure formelle des textes et communication,* in Word, 10 (1954), 1-27.

W. MARTIN, a) *Analyse van een vocabularium met behulp van een computer,* Brussel, 1970;
b) *Inverte frequentielijst van het Nederlands,* Leuven, 1971;
c) *De distributie van nieuwe woorden in Mei van Herman Gorter* (ter perse, verschijnt in De nieuwe Taalgids, Groningen).

Ch. MULLER, *Calcul des probabilités et calcul d'un vocabulaire,* in TLL, II (1964), 235-244.

A. VAN DIERENDONCK, *Methodologische problemen in verband met computersimulatie van cognitieve processen,* referaat gehouden op het Colloquium "Het gebruik van computers en het onderzoek in de menswetenschappen" (Brussel, 25-27 februari 1971).

M. VAN OVERBEKE, *Travail du Séminaire de Langue Néerlandaise,* referaat gehouden op boven genoemd colloquium.

G.U. YULE, *The statistical study of literary vocabulary,* Cambridge, 1944.

G.K. ZIPF, *The Psycho-Biology of Language,* Cambridge (Mass.), 1965².

SUMMARY

This report gives a survey of the LDP (Language Data Processing) at the Institute of Applied Linguistics, Louvain (Dutch sigil = ITL). In a first section the software packet available (up to 31-12-1970) at the Institute is given.

Schematically we can divide it into :

1. Classification programs
   1.1. Software programs
   1.2. Reorganisation programs
      1.2.1. Translation
      1.2.2. Selection

2. Statistical programs
   2.1. Mathematical statistics
   2.2. Lexical statistics
   2.3. Statistical Aids

3. Linguistic automation
   3.1. Lexical automation
   3.2. Syntactical automation

Terms are explained and corresponding programs given. In a second point the possible usefulness of some of these operations is dealt with, taking a few concrete investigations as an example.

Finally, as a conclusion, the question how the computer is used at ITL is answered. (As an administration and calculating machine, a control machine, and a simulation machine).