STICHTING
MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM

REKENAFDELING

Rapport R 642

Onderzoek Woordfrequentie

door

J.A.Th. van Berckel

October 1961

J. A. Th. van Berckel

O N D E R Z O E K W O O R D F R E Q U E N T I E.

Inleiding.

Om een beeld te krijgen van de woordfrequentie van de Nederlandse taal, gaf een werkgroep o.l.v. prof. dr. ir. J.L. van Soest, de Rekenafdeling van het Mathematisch Centrum, de opdracht, een X-1 programma te maken, dat in staat zou zijn ca. een millioen woorden te tellen en deze dan te sorteren op frequentie.
De teksten zijn genomen uit een tiental categorieen: kranten, Kamerverslagen, toneelstukken, gedichten, romans, schoolboeken, enz. Hierin zijn de woorden, d.m.v. een gekleurde onderstreping, verdeeld in 4 groepen:

Werkwoorden	groene onderstreping	3
Bijvoegelijke woorden	blauwe onderstreping	2
Zelfstandige naamwoorden	rode onderstreping	1
Restgroep	geen onderstreping	0

De teksten moeten met behulp van de Flexowriter op een 7-gats band (heptades) geponst worden, waarbij men in plaats van de gekleurde onderstreping een cijfer achter elk woord zet.

Programma.

Het verwerken van Flexowriter-banden.
De machine leest zo'n band heptade voor heptade, vertaalt deze in interne symbolen en onderzoekt ze. Detecteert hij nu het begin van een woord, dan worden de symbolen geborgen, totdat een cijfer het woord afsluit. Uit deze symbolenrij maakt het programma een intern woord en zoekt hiermee in het geheugen, of het betreffende woord, met hetzelfde cijfer, reeds voorkomt. Is dit het geval, dan verhoogt de X-1 de telling van dat woord met 1 en anders bergt hij het woord op, waarna hij het volgende woord van de band kan lezen.
Op den duur zal het geheugen vol raken (ca. 2500 verschillende woorden), waarna nog niet eerder voorgekomen woorden door de machine, in interne representatie, worden uitgeponst (afval-band). Zijn op deze manier alle Flexowriter-banden verwerkt, dan gaat de machine de woorden, die in het geheugen staan, sorteren op afdalende frequentie en bij gelijke frequentie op alphabet, waarna hij de geheugen-woorden uitponst. (g.w.-band.)

Het verwerken van afval-banden geschiedt als volgt:
Woorden, in interne representatie, van de band lezen; daarmee in het geheugen zoeken; ophogen of bergen; bij gebrek aan ruimte afval-band maken; sorteren en g.w.-band uitponsen. Dit proces wordt zolang herhaald, totdat er geen afval-banden meer geproduceerd worden.

Uiteindelijk houden we een aantal g.w.-banden over, die elk voor zich gesorteerd zijn. Deze banden moet de machine nu in elkaar vlechten tot een gesorteerde band, die dan door de flexowriter uitgetypt kan worden.

Om te voorkomen dat de machine eigennamen verwerkt, moeten die tijdens het ponsen van de tekst, tussen haken geplaatst worden.
bijv.: <Den Bosch>.
Van een woord worden niet meer dan de eerste 22 symbolen verwerkt. Zijn er meer, dan worden de overige weggelaten; het soortcijfer wordt echter weer wel verwerkt.
Om scheidbare samenstellingen als "opvallen" aan te kunnen, eist het programma, dat een zin als: "Dit valt in de tekst nauwelijks op", als volgt geponst wordt: "DitO valt$_{10}$ inO deO tekst1 nauwelijks2 $_{10}$op3". Het deel dat met zo'n $_{10}$ is afgesloten, wordt zolang bewaard. De volgende woorden worden normaal verwerkt, totdat het bijbehorende tweede deel geopend wordt met een $_{10}$, waarna met de, dan te vormen, samenstelling in het geheugen gezocht wordt. De beperking tot 22 symbolen is hier van toepassing op de gehele samenstelling.

De machine kan zich, tijdens het verwerken van Flexowriter-banden, in diverse toestanden bevinden, o.a. rusttoestand (0), toestand van woordverwerking (1), enz. Om in de toestand van woordverwerking te komen, (detectie begin nieuw woord) moet:

1e de machine in rusttoestand zijn en
2e op de band, of een letter, of een ', of een $_{10}$ voorkomen, voorafgegaan door SPACE, TAB, CARRIAGE RETURN, UPPER CASE, l o w e r c a s e , aanhalingsteken of openingshaakje (.

De machine is nu in toestand 1 en leest de volgende heptades, totdat een cijfer of $_{10}$ hem weer in de rusttoestand terugzet, waarna hij het woord verder gaat verwerken.
In een woord mogen alleen de volgende symbolen voorkomen:
letters, spatie, punt, koppelteken en apostrophe.
Komt het programma, in de toestand van woordverwerking, een koppelteken gevolgd door een CARRIAGE RETURN tegen, dan worden beide symbolen uit het woord verwijderd.

De ruimte, die gereserveerd is om bij elk woord de frequentie bij te houden, is afhankelijk van de woordlengte. Is het aantal symbolen in een woord 6 of minder, dan kan de telling ten hoogste 65535 bedragen, bij meer dan 6 symbolen ten hoogste 2047. Wordt dit aantal overschreden, dan wordt er voor dat woord nieuwe ruimte gereserveerd, waar dan de verdere telling bijgehouden wordt.
Komt men in de resultaten-lijst, een woord met een van de bovengenoemde frequenties tegen, dan weet men wel haast zeker, dat dit woord nogmaals voor zal komen. De betrokken tellingen moeten dan bij elkaar geteld worden om de frequentie te verkrijgen.

De tekst is verdeeld in secties van ten hoogste 99 regels. Elke sectie wordt voorafgegaan door een label, bestaande uit een letter en twee cijfers. De letter geeft het hoofdstuk aan (bijvoorbeeld De Tijd, in de categorie kranten), het eerste cijfer de bladzijde (het artikel), terwijl het tweede een volgnummer per bladzijde (artikel) is. Deze label wordt uitgetypt, zodra hij gelezen is. Dit is gedaan om het fouten zoeken te vereenvoudigen. In het programma zijn n.l. diverse tests ingebouwd ten behoeve van het opsporen van fouten. De machine geeft een signaal om de aandacht te vestigen op een regel, waarin een combinatie voorkomt, die misschien of zeker fout is. De machine typt n.l. een code-getal, dat ons iets zegt over de soort fout en de regel waarin deze gededecteerd is.
Een tabel hiervan volgt hieronder.

regelnummer : Letter ₁₀ ' niet voorafgegaan door SPACE, TAB, enz.
regelnummer + 600: Spatie in een woord. Dit kan duiden op een vergeten van het cijfer achter een woord.
regelnummer + 700: Woord langer dan 22 symbolen.
regelnummer + 800: Niet toegestaan symbool in een woord.
regelnummer + 900: Er is waarschijnlijk een --------₁₀ vergeten.
regelnummer + 1000: Er is waarschijnlijk een ₁₀-------- vergeten. Dit laatste wordt gedetecteerd bij het einde van de sectie. Om nu gemakkelijker te kunnen nagaan waar de fout gemaakt is, worden altijd getypt:
r.n. + 400, indien --------₁₀, en
r.n. + 500, indien ₁₀-------- voorkomt.
regelnummer + 1100: Hetzelfde woord is reeds met een ander cijfer voorgekomen. Dit k a n duiden op een onjuiste onderstreping. Bij het verwerken van de afval-band wordt deze test niet meer uitgevoerd. De sectie aanduiding en de regeltelling zijn dan verloren gegaan.

Ponsconventies.

1. Elke band begint met een stuk blank van tenminste 25 cm. (tape feed).
2. Het eerste geponste symbool is CARRIAGE RETURN, gevolgd door de case-definition l o w e r c a s e , daarna de datum van ponsen.
3. De sectie aanduiding moet tussen v i e r k a n t e h a k e n voor de betreffende tekst getypt worden. De diverse secties scheidt men van elkaar door een roffel blank.
4. Na elke roffel blank moet e e r s t UPPER CASE of l o w e r c a s e gegeven worden, alvorens men een teken ponst, dat voor case-definition vatbaar is. (Men mag dus wel eerst CARRIAGE RETURN, TAB of SPACE geven.)
5. Afhankelijk van de kleur van de onderstreping, wordt er achter elk woord of samengesteld woord een cijfer getypt.
   3 groene onderstreping
   2 blauwe onderstreping
   1 rode onderstreping
   0 geen onderstreping
   Dit cijfer moet o n m i d d e l l i j k achter de laatste letter getypt worden. Hierop zijn maar twee uitzonderingen:
   a. Eigennamen worden zonder cijfer tussen < > getypt.
   b. Bij werkwoorden die gescheiden zijn, wordt het eerste deel afgesloten met een ₁₀ , d i r e c t a c h t e r de laatste letter; het tweede deel wordt weer met een ₁₀ geopend d i r e c t v o o r het eerste symbool; het woord wordt afgesloten met het soortcijfer 3.
6. De volgende symbolen mogen niet gebruikt worden:
   |, ∧, ∨ en punch off; terwijl [,], <, > en ₁₀ voor bovengenoemde toepassingen gereserveerd zijn en dus nergens anders voor gebruikt mogen worden. Onderstrepen is niet toegestaan (_ is geen toegestaan symbool in een woord). Wenst men op een woord bijzonder de aandacht te vestigen, dan wordt het woord toch gewoon getypt en de regel voltooid, waarna op de volgende regel, onder het betrokken woord ===== getypt wordt. Dit heeft op de wijze van woordverwerking geen enkel effect.
7. De band wordt afgesloten met STOPCODE gevolgd door een roffel blank van ± 15 cm.