TILBURG UNIVERSITY

Onderzoek naar woordenrijkdom

Renkema, Jan

Published in:
Tijdschrift voor Taalbeheersing

Publication date:
1983

Link to publication

Citation for published version (APA):
Renkema, J. (1983). Onderzoek naar woordenrijkdom: Taalstatistische analyse via een microcomputer. Tijdschrift voor Taalbeheersing, 5, 275-289.

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy
If you believe that this document breaches copyright, please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 20. aug. 2017

Onderzoek naar woordenrijkdom*

TAALSTATISTISCHE ANALYSE VIA EEN MICROCOMPUTER

J. RENKEMA

Samenvatting

Statistische analyses en het gebruik van de computer worden steeds belangrijker in onderzoek naar taalgebruik. Naar aanleiding van enkele problemen in onderzoek naar woordenrijkdom wordt nader ingegaan op de mogelijkheden en het nut van taalstatistische analyse via een microcomputer. In onderzoek naar woordenrijkdom wordt vaak gebruik gemaakt van de zogenaamde type-token ratio (TTR). Maar de uitkomsten van TTR-onderzoek kunnen worden beïnvloed door factoren als: verbogen vorm versus standaardvorm, en de lengte van de gebruikte teksten. In dit artikel wordt beschreven hoe deze factoren kunnen worden geneutraliseerd via lemmatiseerprocedures en de zogenaamde omrekenmethode.

1 Inleiding

Onderzoek op het gebied van woordenrijkdom kan een belangrijk hulpmiddel zijn bij de analyse van literatuur en taalgebruik. Het gaat hierbij niet alleen om vragen als: Gebruikte Vestdijk in zijn latere romans meer verschillende woorden dan in zijn vroegere romans? of: Worden in triviaallectuur minder verschillende woorden gebruikt dan in 'echte' literatuur? Met behulp van onderzoek naar woordenrijkdom kan ook antwoord gegeven worden op vragen als: Hoe ontwikkelt zich de woordenschat van tweede-taalverwervers? Heeft woordenrijkdom invloed op leesbaarheid? Brengt een cliënt zijn problemen genuanceerder onder woorden naarmate de therapie meer succes heeft? Bij het onderzoek naar woordenrijkdom doen zich een aantal problemen voor die met behulp van statistische analyses en het gebruik van een computer voor een groot deel kunnen worden opgelost.

In taal- en literatuurwetenschap wordt de laatste jaren steeds meer gebruik gemaakt van statistische analyses. In de zes dissertaties uit 1981 en 1982 aan Nederlandse universiteiten, die in De Vries (1982) onder Studies over Taalgebruik[1] worden gerangschikt, zijn steeds statistische toetsen en procedures gehanteerd om significante verschillen aan te tonen tussen globaal gezegd typen spreek- en/of schrijftaal. In ongeveer een kwart van de artikelen die in de laatste jaargangen van dit tijdschrift zijn verschenen, wordt gebruik gemaakt van statistische technieken.

Veelal gaat het niet alleen om een eenvoudige chikwadraattest of 2-x-2-tabel, maar ook om bijvoorbeeld regressie-analyse en factor-analyse. Een linguïst die taalgebruik bestudeert en een statistische analyse nodig heeft, gaat dan meestal te rade bij de statistische technieken die in de sociale wetenschappen worden gehanteerd. De 'Statistical Package for Social Sciences'[2] vormt het belangrijkste handvat voor de linguïst. Maar bij taalgebruiksonderzoek kan ook nog op een andere manier gebruik worden gemaakt van statistische analyse. In dit onderzoek naar woordenrijkdom wordt hiervan een voorbeeld gegeven.

Ook de computer wordt steeds meer als hulpmiddel gehanteerd bij onderzoek in Letterenfaculteiten. Onderzoek naar woordfrequenties en variaties in taalgebruik

Tijdschrift voor Taalbeheersing 5 (1983), 4, 275-289

is nu veel eenvoudiger geworden. Onderzoek naar zinspatronen in grote hoeveelheden tekst behoort sinds enige jaren tot de mogelijkheden (zie bijvoorbeeld het Query-programma, ontwikkeld op de computerafdeling van de Letterenfaculteit van de Universiteit van Amsterdam). Meestal maakt de linguïst of letterkundige gebruik van een grote computer waarvoor standaard-programmatuur is ontwikkeld, die ook nuttig is voor taalgebruiksonderzoek. Maar sinds de opkomst van de microcomputer kan ook op een andere, en vaak eenvoudiger manier computer-linguïstisch onderzoek worden verricht. Ook hiervan wordt in dit artikel een voorbeeld gegeven. In dit artikel zal nader worden ingegaan op de mogelijkheden en het nut van taalstatistische analyse via een microcomputer. Dit zal gebeuren naar aanleiding van enkele problemen in onderzoek naar woordenrijkdom.

2 Woordenrijkdom

Het berekenen van woordenrijkdom kan van belang zijn voor onderzoek naar bijvoorbeeld kindertaalontwikkeling, tweede-taalverwerving, leesbaarheid of stijlverschillen. Veelal wordt dan als maat genomen: het aantal verschillende woorden (types) in een tekst gedeeld door het aantal woorden (tokens) in die tekst, de zogenaamde type-token ratio (TTR). De TTR is dus gelijk aan 1 wanneer in een tekst elk woord maar één keer voorkomt. Het zal duidelijk zijn dat de TTR altijd onder de 1 blijft, omdat in een tekst veel woorden meer dan een keer voorkomen. De TTR wordt als maat gebruikt in zeer verschillende disciplines. In de stilistiek is de TTR gehanteerd om verschillen aan te geven tussen individuele literaire stijlen (Calbert 1974). De TTR is ook toegepast in onderzoek naar redevoeringen van Nixon (Hart 1976). In de psychologie en de psychiatrie is de TTR veelvuldig toegepast, bijvoorbeeld als maat voor cognitieve complexiteit of als stress indicator (Meisels 1967, Höweler 1972, Ricci 1974). Silverman (1973, 1977) ziet de TTR als betrouwbare maat in schizofrenie-onderzoek: 'schizofrenen' hebben een lagere TTR dan 'niet-schizofrenen'. Dit wordt door anderen betwijfeld. Meara (1978) bijvoorbeeld toonde aan dat het taalgebruik van schizofrenen niet verschilt van dat van tweede-taalgebruikers. Een laatste voorbeeld betreft het onderscheid gesproken en geschreven taal. Gruner e.a. (1967) concludeerden dat de TTR van gesproken taal lager is dan die van geschreven taal, maar Moe (1974) kon geen verschillen op dit punt vaststellen.
Een van de conclusies die zich na literatuurstudie opdringt is de volgende. Veelal wordt de TTR gebruikt om op een betrekkelijk eenvoudige manier precieze gegevens te produceren. De betrouwbaarheid van deze gegevens wordt door anderen dan weer betwist, waarbij onduidelijk blijft of de TTR als meetinstrument juist is gebruikt.
Voor een beter begrip van de problemen in TTR-onderzoek kan de volgende illustratie dienstig zijn. Welke van de volgende twee fragmenten[3] vertoont de grootste variatie in woordgebruik?

I Kindertaal

> Daar komt vader thuis. Met de fiets. Jip loopt hem tegemoet. Hier zegt vader, jij mag de fiets wegzetten Jip. Dat is fijn. Jip mag de fiets achter het huis brengen. In het schuurtje. Voorzichtig hoor, zegt vader. En vader gaat naar binnen.
> Daar staat Jip met de fiets op straat. En Janneke komt eraan.
> Ik ga fietsen! roept Jip. Kijk maar, ik kan al fietsen. Janneke kijkt met open ogen. En Jip steekt zijn been onder de stang van de fiets. Hij houdt het stuur goed vast. En hij trapt.

Het gaat goed. O, wat gaat het goed. Jip kan heus fietsen. Het zwabbert wel een beetje. En het gaat wel van zigzag. Maar Jip komt vooruit.
Mooi! roept Janneke. Ze klapt in haar handjes. Mooi! Mag ik ook eens? Maar Jip wil nog veel meer laten zien. Hij gaat nu sturen naar links. Hij neemt een bocht. En dan gebeurt het! Dan valt de fiets om. Met een slag. Hoe! gilt Janneke. Jip ligt helemaal onder de fiets. Helemaal bedolven onder de fiets. Wat een ongeluk, wat een ongeluk.
Janneke helpt. Ze trekt de fiets weg. En dan krabbelt Jip op. Hij heeft een grote schram op zijn wang. En nog een op zijn knie.
Daar komt vader de deur uit. Wat is dat nou, zegt hij. Je zou de fiets alleen maar wegbrengen, Jip. Ja, zegt Jip. Maar ik kan toch lekker fietsen. En hij lacht door zijn tranen heen. En dan gaat Janneke mee naar binnen. En ze gaan Jip z'n gezicht wassen. Die kinderen toch, zegt vader. Dat fietst al.

II Krantentaal

Middelbare scholen met overcapaciteit kunnen het onderwijsniveau gaan verbeteren, meer individu-gericht bijvoorbeeld, of de Moedermavo's gaan aanvullen met verschillende vormen van tweede-kans onderwijs. Beide zullen nog lang nodig blijven.
De universiteiten zijn een hoofdstuk apart. Bij het publiek staan ze al enige tijd in een kwade reuk. De wetenschappen worden – ten onrechte – niet meer als bron van vooruitgang gezien, maar eerder als gevaar. Indianenverhalen over "nieuwe vrijgestelden" hebben de indruk verspreid, dat op de universiteiten naar hartelust geluierd wordt. Jaloerse bureaucraten hebben daarvan gebruik gemaakt door te pogen een prikklok op de universiteit in te voeren, wat alleen nog meer afbreuk doet aan de reputatie van de universiteit en aan de productiviteit van vele wetenschapsbeoefenaars.

Deze fragmenten zijn volstrekt willekeurig gekozen en zijn slechts bedoeld als illustratie bij het behandelen van enkele problemen.

3 Vier problemen

Het berekenen van de TTR voor de twee fragmenten gaf het volgende resultaat.

	tokens	types	TTR
I fragment kindertaal	265	123	0,46
II fragment krantentaal	114	86	0,75

Fragment II vertoont dus een veel hogere TTR. Toch kan deze uitkomst niet zonder meer vergeleken worden met resultaten van ander TTR-onderzoek. De TTR wordt namelijk beïnvloed door de aard van de gebruikte teksten, door de manier waarop type en token worden gedefinieerd en door de lengte van de teksten. Wanneer resultaten worden vermeld van TTR-onderzoek, moet worden aangegeven hoe de volgende vier problemen zijn opgelost.

1 Aantal onderwerpen

De TTR wordt beïnvloed door het aantal onderwerpen dat in de te onderzoeken teksten is behandeld. Wanneer we een aardrijkskundeboek van 10.000 woorden vergelijken met een steekproef van dezelfde omvang die is samengesteld uit 500 fragmenten uit kranten, dan is de kans groot dat de steekproef uit kranten een hogere TTR zal hebben omdat die steekproef meer verschillende onderwerpen (en dus kans op meer verschillende woorden) bevat. Ook in de voorbeeld-fragmenten

zien we verschil in aantal onderwerpen. Wanneer het fragment zo gekozen was dat alleen de universiteit ter sprake kwam, dan zou door het ontbreken van de woorden die met 'middelbare school' te maken hebben, de TTR veel lager zijn.

2 Homografen, homoniemen en polysemen
Voor een betrouwbare TTR is het van belang dat 'mijn' als bezittelijk voornaamwoord en 'mijn' als zelfstandig naamwoord' niet als tokens van één type worden beschouwd. Hetzelfde geldt voor polysemen als 'bank' (financiële instelling en zitmeubel). Met behulp van het criterium 'woordsoortverschil' zijn een aantal homografen en homoniemen redelijk gemakkelijk te onderscheiden; zie in de voorbeeldfragmenten 'een' als lidwoord en als telwoord. Maar vooral bij homoniemen en polysemen wordt het vaststellen van types aanzienlijk bemoeilijkt. In TTR-onderzoek is dit probleem nog nauwelijks aangestipt.

3 Verbogen vorm – basisvorm
Vanuit taalkundig oogpunt is het zeer onbevredigend iedere verbogen of vervoegde vorm als apart type naast een basisvorm te beschouwen. (Bij telling in de fragmenten zijn bijvoorbeeld 'ga' en 'gaat' als tokens van één type beschouwd.) De TTR wordt zuiverder wanneer de verbogen of vervoegde vormen herleid (gelemmatiseerd) worden tot de basisvorm. Voor zover bekend is nooit onderzocht of – en zo ja in hoeverre – de uitkomsten van vergelijkend TTR-onderzoek veranderen wanneer men in plaats van een ongelemmatiseerde een gelemmatiseerde TTR als uitgangspunt neemt.

4 Tekstlengte
Het zal duidelijk zijn dat de TTR in belangrijke mate wordt beïnvloed door de lengte van de tekst. Hoe langer de tekst, des te minder kans op nieuwe woorden. Dus een lange tekst heeft automatisch een lagere TTR dan een korte tekst. Wanneer we het kindertaal-fragment afkappen bij 114 woorden (de lengte van het andere fragment) krijgen we 65 types. De TTR stijgt dan van 0,46 naar 0,57; de TTR blijft lager dan die in het krantentaal-fragment, maar het verschil is veel kleiner. Voor het oplossen van problemen betreffende TTR en tekstlengte zijn in de literatuur wel oplossingen voorgesteld (zie paragraaf 6).

In de volgende paragraaf zal aan de hand van een korte beschrijving van de tekstbestanden waarvoor in dit onderzoek de TTR is berekend, worden aangegeven in hoeverre de hier geschetste problemen een rol spelen.

4 De tekstbestanden

In Nederland zijn een aantal tekstbestanden beschikbaar die al via een computer zijn bewerkt voor frequentie-tellingen. Het betreft hier de volgende typen taalgebruik:

1.	dagbladen	(DA)
2.	opiniebladen	(OP)
3.	populair wetenschappelijk proza	(PW)
4.	romans en novellen	(RN)
5.	gezinsbladen	(GB)
6.	overheidstaal	(OV)

De eerste vijf bestanden (elk ± 120.000 woorden) zijn verzameld op initiatief van de werkgroep Frequentie-onderzoek van het Nederlands (zie Uit den Boogaart 1975 voor verdere gegevens). Voor het bestand overheidstaal (ongeveer 50.000 woorden) zij verwezen naar Renkema 1981.

1. Voor het probleem 'aantal onderwerpen' levert een nadere beschouwing van de bestanden interessante gegevens. De vraag of de bestanden verschillen in aantal onderwerpen moet gezien de opbouw ervan naar alle waarschijnlijkheid met 'ja' worden beantwoord. De bestanden zijn samengesteld uit tekstfragmenten met per bestand een andere norm-lengte. (Zie hiervoor het Verslag van de Werkgroep Frequentie-onderzoek (1974) en Renkema (1981) met een motivatie voor normlengte 100.) In de volgende tabel staan per bestand de norm voor fragmentlengte en de uitkomsten van TTR-berekening op eenzelfde hoeveelheid tekst (zie voor dit laatste Renkema (1983)).

Type taalgebruik	Norm voor fragmentlengte	Types	Tokens	TTR
DA	75	11884	48242	0,25
OP	115	11161	48242	0,23
PW	250	10557	48242	0,22
RN	300	9801	48242	0,20
GZ	125	10935	48242	0,23
OV	100	7930	48242	0,16

Elk fragment bevat volledige zinnen, en wel zodanig dat de laatste zin zo dicht mogelijk eindigt bij de norm voor fragmentlengte. Wanneer we ervan uitgaan dat elk fragment één onderwerp bevat – en steeksproefsgewijze controle bevestigt dit – en dat de factor 'aantal onderwerpen' de TTR in belangrijke mate beïnvloedt, dan zijn de verschillen in TTR niet verwonderlijk. Wanneer we OV buiten beschouwing laten, zien we dat de verschillen zijn te verklaren uit de verschillen in normlengte van de fragmenten. Hoe kleiner de gemiddelde fragmentlengte, dus hoe hoger het aantal fragmenten per bestand, des te hoger de TTR. Voor de verschillen in intervalgrootte kan ik geen verklaring geven. PW bijvoorbeeld bevat slechts half zoveel fragmenten (onderwerpen) als OP maar toch verschilt de TTR slechts weinig. Opvallend is wel de plaats die OV inneemt. Gezien het aantal fragmenten in het bestand, zou de TTR 0,24 (i.p.v. 0,16) moeten zijn.

2 Voor het probleem van homografen, homoniemen en polysemen biedt het gebruik van deze bestanden slechts een gedeeltelijke oplossing. De bestanden zijn met de hand gecodeerd volgens een codeersysteem dat grammaticale en lexicale informatie bevat. Elk woord in de tekst kreeg een driecijfercode die-420 de-370 woordklasse-000 specificeerde-255 en-700 relevante-103 vormkarakteristieken-001 in-600 kaart-000 bracht-255.[4]
Door een type te herdefiniëren als woord plus code kan een groot aantal homografen en homoniemen worden onderscheiden. Het codeersysteem maakt bijvoorbeeld onderscheid tussen:

- de bal is rond-100 (bijvoeglijk naamwoord)
- rond-500 duizend gulden (bijwoord)
- hij hing maar wat rond-620 (niet-werkwoordelijk deel van een samengesteld werkwoord)

Het codeersysteem onderscheidt echter alleen verschillende betekenissen van een woordvorm voor zover die gerelateerd kunnen worden aan verschillende woordsoorten. Met nadruk zij hier vermeld dat het probleem blijft bestaan voor polysemen en voor homografen of homoniemen die tot dezelfde woordsoort behoren.

3 Het probleem 'verbogen vorm – basisvorm' kan wel worden opgelost omdat het codeersysteem in het laatste cijfer mogelijkheid geeft voor informatie over de verbogen of vervoegde vorm: In Woordfrequenties (Uit den Boogaart, 1975) wordt ook een gelemmatiseerde frequentielijst gegeven, maar in deze lijst zijn geen woorden opgenomen met een frequentie lager dan 5. Deze lijst is dus minder geschikt voor TTR-onderzoek. In paragraaf 5 wordt aan de hand van dit derde probleem het nut van een microcomputer in taalonderzoek geïllustreerd.

4 Het probleem 'tekstlengte' kan in eerste instantie worden omzeild door te werken met bestanden van exact dezelfde lengte (zie het schema op pag. 279). Maar voor verfijnder TTR-onderzoek stuit men toch weer op het probleem tekstlengte. Wanneer men onderzoek wil doen naar de woordenrijkdom binnen een woordsoort (bijvoorbeeld de TTR van bijvoeglijke naamwoorden in verschillende typen taalgebruik) dan krijgt men ondanks gelijke tekstbestanden weer te maken met verschillende aantallen tokens per woordsoort. Uiteraard kan men ook hier het probleem omzeilen door steeds eenzelfde aantal bijvoeglijke naamwoorden te nemen, maar er zijn statistische procedures die een vergelijking mogelijk maken tussen TTR's van woordsoorten die in aantallen tokens verschillen.

Na deze beschrijving van de tekstbestanden met het oog op problemen in TTR-onderzoek kan samenvattend het volgende worden opgemerkt.

1 Aantal onderwerpen
Alleen de tekstbestanden Opiniebladen en Overheidstaal kunnen met elkaar worden vergeleken, omdat hier de verschillen in fragment-lengte (en dus in aantal onderwerpen) minimaal zijn.

2 Homografen, homoniemen en polysemen
TTR-onderzoek op de hier vermelde bestanden is alleen betrouwbaar voor zover het criterium 'woordsoortverschil' homografen en homoniemen onderscheidt. Met name het polysemie-probleem blijft onopgelost. In dit onderzoek wordt ervan uitgegaan dat dit laatste verschijnsel zich in dezelfde mate voordoet in de verschillende bestanden. Bij vergelijkend TTR-onderzoek behoeft dit dus geen probleem te zijn.

3 Verbogen vorm basisvorm
Door de wijze waarop de tekstbestanden zijn gecodeerd is het in principe mogelijk verbogen vormen te herleiden tot de basisvormen. Zie hiervoor de lemmatiseerprocedure voor bijvoeglijke naamwoorden in paragraaf 5.

4 Tekstlengte
Wanneer men TTR-onderzoek wil verrichten op woordsoorten met een verschillend aantal tokens moet nagegaan worden in hoeverre verschil in tekstlengte van invloed is, en in hoeverre die invloed geneutraliseerd kan worden. Zie paragraaf 6.

5 Werken op een micro

Voor dit TTR-onderzoek zijn de bestanden die in grote computers waren opgeslagen (TH-Eindhoven en Mathematisch Centrum Amsterdam) getransporteerd naar een micro-computer. Ideaal voor computer-linguïstisch onderzoek is uiteraard een ter-

minal op eigen bureau, verbonden met een mainframe waarbij via pasklare of gemakkelijk te wijzigen programma's interactief grote bestanden kunnen worden onderzocht. Maar computer-tijd op een mainframe is nogal kostbaar. Voor dit onderzoek bleek de micro een goed alternatief. Met een redelijk gangbare configuratie (RAM 52 Kbyte, diskdrive voor 8" diskettes met een opslagmogelijkheid van 1,2 Mbyte per diskette) konden de problemen betreffende lemmatiseren en statistische bewerking worden opgelost.
Vaak wordt beweerd dat een micro niet geschikt is voor onderzoek op grote bestanden. Maar met een goede organisatie is alles wat op een mainframe kan gebeuren in principe ook mogelijk op een micro. Een aantal bewerkingen, bijvoorbeeld het alfabetiseren van grote aantallen woorden, neemt meer tijd in beslag, en voor dit onderzoek moesten bestanden worden gesplitst omdat de geheugencapaciteit onvoldoende was. Maar voordelen, zoals beschikbaarheid van de gegevens op elk gewenst moment, wegen hier ruimschoots tegenop. Bovendien kan de capaciteit van micro's gemakkelijk worden uitgebreid.[5]

5.1 de TTR van bijvoeglijke naamwoorden

Voor het probleem verbogen vorm – basisvorm zijn, in aansluiting op vorig onderzoek (Renkema 1981) de bijvoeglijke naamwoorden nader onderzocht.
Met behulp van de microcomputer kan antwoord worden gegeven op de vraag: verschillen de bestanden opiniebladen (OP) en overheidstaal (OV) in woordenrijkdom bij het gebruik van bijvoeglijke naamwoorden? De uitkomst is als volgt:

bijvoeglijke naamwoorden			(ongelemmatiseerd)
	types	tokens	ttr
OP	1383	2875	0,48
OV	876	2570	0,34

Op basis van deze tabel zou geconcludeerd kunnen worden dat OP een grotere variatie bijvoeglijke naamwoorden vertoont dan OV. Maar in hoeverre wordt deze uitkomst beïnvloed door het aantal verborgen vormen? De computer telt immers 'mooier', 'mooist' enz. als aparte types naast de basisvorm 'mooi'.
In het codeersysteem waarmee de bestanden zijn gecodeerd zijn de volgende onderscheidingen aangebracht voor bijvoeglijke naamwoorden:

100	basisvorm	duidelijk
102	genitief	duidelijks
103	overige verbogen vormen	duidelijke
104	comparatief (onverbogen)	duidelijker
105	comparatief (genitief)	duidelijkers
106	comparatief (andere verbogen vormen)	duidelijkere
107	superlatief (onverbogen)	duidelijkst
109	superlatief (andere verbogen vormen)	duidelijkste

De codes 102 en 103 gelden ook voor archaïsche vormen, bijvoorbeeld: 'zaliger-102 nagedachtenis', 'in goeden-103 doen', 'te goeder-103 trouw'.

Voor het herleiden tot basisvormen – de lemmatiseerprocedure – werden de volgende regels opgesteld:

1. $\{s, er\} \longrightarrow \emptyset$ / ——— # en (..2)

2. $\{e\{n, r\}\} \longrightarrow \emptyset$ / ——— # en (..3)

3. $\{er\{s, e\}\} \longrightarrow \emptyset$ / ——— # en (..4) of (..5) of (..6)

4. st(e) $\longrightarrow \emptyset$ / ——— # en (..7) of (..9)

De regels moeten als volgt worden gelezen: 1 als het derde cijfer een 2 is, en als de uitgang -s is, schrap dan -s; als de uitgang -er is, schrap dan -er. Hetzelfde geldt mutatis mutandis voor de regels 2, 3 en 4.
Deze vier regels produceren de juiste basisvormen voor bijvoeglijke naamwoorden waar de spelling niet verandert in de verbogen vorm. Voor de bijvoeglijke naamwoorden waar de spelling wel verandert (dik-dikker, ambitieus-ambitieuzer, groot-groter, gaaf-gaver enz.) moesten vier andere regels worden geformuleerd voor dubbele consonanten, voor 'z' en 'v' en voor vocaalverdubbeling.

5. C1 C2 → C1 / ——— # en C1 = C2 [*dikk→ dik]

De procedure is beëindigd wanneer regel 5 kan worden toegepast. Is dit niet het geval dan moet regel 6 òf 7 worden toegepast en daarna regel 8.

6. z→ s / ——— # [*ambitieuz→ ambitieus]
7. v→ f / ——— # [*gav→ *gaf]

8. $V_1 \rightarrow V_1V_1$ / (C ——— C # en $V_1 \neq$ i) [*gaf→ gaaf]
of
(i ——— C # en $V_1 \neq$ e) [*idiot→ idioot]

Regel 8 zorgt voor verdubbeling van vocaalsymbolen, uitgezonderd de 'i'. De conditie dat het vocaalsymbool tussen twee consonantsymbolen moet staan, voorkomt dat het programma types produceert als 'goeed' (goed), 'kouud' (koud), 'lauuw' (lauw) en 'leuuk' (leuk). Maar 'idiote' bijvoorbeeld heeft als basisvorm 'idioot' en niet 'idiot', vandaar de toevoeging na 'of'. De conditie dat het vocaalsymbool geen 'e' mag zijn zorgt ervoor dat het programma geen types produceert als 'viees' (vies).
Met deze set regels kon ongeveer 97 procent van de verbogen vormen worden gelemmatiseerd. De overige vormen, die interactief zijn gecorrigeerd, betroffen de volgende drie categorieën:

a. foutieve toepassing van het codeersysteem, bijvoorbeeld 'gouden-103' in plaats van 100;
b. onregelmatige vormen, goed-beter-best;
c. basisvormen met een schwa in de laatste lettergreep. 'sombere-103' werd via regel 2 correct gelemmatiseerd tot 'somber', maar toepassing van regel 8 gaf hier de incorrecte vorm 'sombeer'.

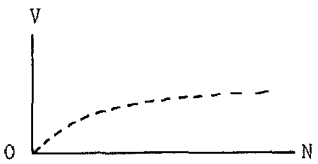
Op basis van deze lemmatiseerprocedure zijn de frequenties van de verbogen vormen opgeteld bij die van de bijbehorende basisvormen. Verbogen vormen zonder corresponderende basisvorm werden uiteraard als apart type geteld. De TTR-berekening voor de aldus gelemmatiseerde lijsten bijvoeglijke naamwoorden voor opiniebladen en overheidstaal staat in de volgende tabel:

bijvoeglijke naamwoorden			(gelemmatiseerd)
	types	tokens	ttr
OP	1073	2875	0,37
OV	676	2570	0,26

Wanneer we naar de gelemmatiseerde en de ongelemmatiseerde TTR's kijken, dan mogen we concluderen dat – althans bij dit vergelijkend TTR-onderzoek naar bijvoeglijke naamwoorden – lemmatiseren geen ander beeld oplevert. De verhouding blijft immers gelijk.[7] Uiteraard is de gelemmatiseerde TTR lager omdat verbogen vormen niet meer als aparte typen worden beschouwd, maar voor de vergelijking maakt dit niets uit.

6 Statistische analyse

Tot slot het nog niet behandelde probleem tekstlengte. Hoe meer tokens des te minder kans op nieuwe types, des te lager de TTR. Dus in het voorbeeld uit de vorige paragraaf waar OP meer tokens heeft dan OV, zou men bij vergelijking eigenlijk uit moeten gaan van hetzelfde aantal tokens, en dan zal OP waarschijnlijk een lagere TTR hebben dan hier is aangegeven. Het probleem is nu dat niet precies bekend is hoe de TTR daalt bij toename van het aantal tokens.
Een grafiek die bij elke tekstlengte (N) het aantal verschillende vormen (V) geeft ziet er over het algemeen als volgt uit:



Wanneer men in TTR-onderzoek geen rekening houdt met de factor tekstlengte, veronderstelt men eigenlijk dat V gedeeld door N een constante (c) is

(c) is ($\frac{V}{N}$ = c of V = cN). In dit geval vertoont de grafiek een rechte lijn.

In de literatuur over TTR zijn verschillende voorstellen gedaan om de factor tekstlengte te neutraliseren of om het verloop van de curve te beschrijven.[8]

1 Manschreck e.a. (1981) werkten met een TTR gebaseerd op gemiddelden van TTR's voor steekproeven van honderd woorden.
2 Herdan (1960) kwam na berekeningen voor de werken van Shakespeare tot de conclusie dat log V/log N constant is, m.a.w. dat een logaritmische TTR verschillen in tekstlengte neutraliseert.[9]
3 In andere literatuur is uitgegaan van de veronderstelling dat verschil in tekstlengte geen invloed uitoefent wanneer men het aantal verschillende woorden deelt door de wortel van tweemaal het aantal woorden. Met andere woorden dat $V\sqrt{2N}$ constant is. Zie bijvoorbeeld Van der Geest e.a. (1973) en Andolina (1980) waarin verwezen wordt naar Carroll.

Het eerste voorstel is een uitstekend middel om de factor tekstlengte te neutraliseren omdat men steeds uitgaat van honderd woorden. Het bezwaar is echter dat er dan informatie wordt weggegooid, omdat niet bekend is hoe de curve na honderd woorden verder loopt.

De overige voorstellen zijn van geheel andere aard omdat ze voorspellingen doen over het verloop van de hele curve. Deze voorspellingen kunnen getoetst worden wanneer het verloop van de curve precies bekend is, wanneer dus bij elk aantal woorden (N) het aantal verschillende woorden (V) bekend is. Met behulp van een microcomputer kan deze curve gemakkelijk worden berekend. In de tabellen hieronder staan voor een aantal waarden van N de gemeten waarden van V, dus hoeveel verschillende bijvoeglijke naamwoorden er voorkomen bij een bepaald aantal bijvoeglijke naamwoorden.[10] In de kolommmen 1 en 2 staan de voorspellingen volgens de formules van Herdan en Carroll.[11] In kolom 3 staan om typografische redenen ook de voorspellingen volgens de zogenaamde omrekenmethode. (Hierover volgt apart uitleg in paragraaf 6.1.)

I Bijvoeglijke naamwoorden: Overheidstaal. N = 3133; V = 872

Tokens (N)	Types (V)	Voorspellingen		
		1. Herdan	2. Carroll	3. Omrekenmethode
600	324	217	382	341
1200	503	389	540	524
1800	635	547	661	657
2400	738	697	763	764
3000	856	841	854	854
Gemiddelde afwijking		− 73	+ 29	+ 17

II Bijvoeglijke naamwoorden: Opiniebladen. N = 3583; V = 1436

Tokens (N)	Types (V)	Voorspellingen		
		1. Herdan	2. Carroll	3. Omrekenmethode
600	383	294	588	415
1200	674	544	831	694
1800	885	779	1018	919
2400	1089	1006	1175	1113
3000	1268	1227	1314	1290
Gemiddelde afwijking		− 90	+125	+26

We zien dat de voorspellingen volgens de formules van Herdan en Carroll voor de verschillende waarden van N nogal afwijken van de gevonden waarden voor V. De voorspellingen van Herdan zijn te laag en die van Carroll te hoog. Bij de hier niet opgenomen waarde N = 200 gaf Carroll zelfs de theoretisch onmogelijke voorspelling V = 220. Wel kan geconstateerd worden dat de voorspellingen de werkelijkheid dichter benaderen naarmate N groter wordt. Bij N = 3000 zijn de verschillen tussen gemeten en voorspelde waarden van V heel klein.
Naar aanleiding van deze uitkomsten is op het Mathematisch Centrum te Amsterdam een voorstel onderzocht [12] om een steekproef van een bepaalde omvang om te rekenen naar een steekproef van een kleinere omvang.

6.1 De omrekenmethode

Ter illustratie eerst een theoretisch voorbeeld. Gesteld dat wij willen controleren of een buitenlander die Nederlands leert na twee jaar meer verschillende voorzetsels kent dan na één jaar. We hebben als steekproeven tekst A met 20 voorzetsels (waarin 5 verschillende) en tekst B met 40 voorzetsels (waarin 7 verschillende) waarbij de frequenties van de voorzetsels bekend zijn. We kunnen pas betrouwbare conclusies trekken wanneer steekproef B is teruggebracht tot 20 voorzetsels. Hiertoe kan de omrekenmethode worden gebruikt. We gaan uit van 40 voorzetsels, en schatten dan het aantal verschillende voorzetsels bij 20 voorzetsels. Voor dit doel nummeren wij op een of andere manier (bijvoorbeeld alfabetisch, of in volgorde van aantreffen) de in de tekst voorkomende types van voorzetsels. Bijvoorbeeld:

nummering		frequentie
1	aan	1
2	door	4
3	in	8
4	met	1
5	tot	1
6	van	23
7	voor	2

De omrekenmethode (O) voorspelt nu hoeveel verschillende voorzetsels steekproef

B bevat wanneer deze wordt teruggebracht tot de lengte van steekproef A. De formule luidt als volgt:

$O(X) = V - (1 - X)^{F1} - (1 - X)^{F2} \ldots - (1 - X)^{Fv}$

In deze formule zijn de volgende symbolen en afkortingen gebruikt:

- O: omrekenen
- X: het deel van de steekproef waarvoor moet worden omgerekend. In ons geval 20/40, dus een 1/2. X mag liggen tussen 0 en 2. We mogen dus extrapoleren tot maximaal twee keer de steekproef
- V: aantal verschillende woorden
- $F_1 - F_V$: frequentie van de woorden

Wanneer we de formule toepassen op het voorbeeld krijgen we:

$O(½) = 7 - (½)^1 - (½)^4 - (½)^8 - (½)^1 - (½)^1 - (½)^{23} - (½)^2 = 5{,}18$

Aan de exponenten in de formule O valt af te lezen hoezeer de frequentie van de woorden van invloed is bij het omrekenen. De kans dat een type verdwijnt als men de tekst gaat verkleinen is bij een hoog frequent woord (van 23) veel kleiner. Dus in de steekproef B – teruggebracht tot dezelfde omvang als die van tekst A – worden 5,18 verschillende voorzetsels voorspeld.

De omrekenmethode kan worden toegepast wanneer N en V en de frequenties van de verschillende woorden bekend zijn. Resultaten voor de bijvoeglijke naamwoorden staan in kolom 3 van de tabellen op pag. 284-285. We zien hier dat de omrekenmethode een nauwkeuriger voorspelling doet over de TTR dan de formules van Herdan en Carroll.

7 Slotopmerkingen

Het doel van dit artikel was om – aan de hand van problemen bij het meten van woordenrijkdom – te laten zien hoe taalstatistisch onderzoek via een microcomputer een bijdrage kan leveren aan taalgebruiksonderzoek.

In dit onderzoek ging het alleen om bijvoeglijke naamwoorden, maar de programmatuur die hiervoor is ontwikkeld kan uiteraard ook gebruikt worden voor onderzoek (in micro-configuratie) naar andere woordsoorten. Wanneer er meer gegevens beschikbaar komen over TTR kan ook meer gezegd worden over de waarde van de gevonden verschillen. Met de gegevens uit dit artikel kunnen we vaststellen dat de TTR op een steekproef van 3000 bijvoeglijke naamwoorden in Overheidstaal 0,29 is en in Opiniebladen 0,42. Maar deze getallen zeggen weinig wanneer niet bekend is in welke mate de TTR kan variëren.[13] Pas wanneer hierover meer duidelijkheid bestaat kan bijvoorbeeld de relatie tussen TTR en oordelen over woordenrijkdom worden onderzocht.[14]

Uit dit onderzoek is gebleken dat met behulp van de microcomputer voorstellen om in TTR-onderzoek de factor tekstlengte te neutraliseren redelijk gemakkelijk op betrouwbaarheid kunnen worden getest. Nu de mogelijkheid bestaat voorspelde waarden te vergelijken met werkelijk gevonden waarden kunnen de formules van Herdan (zie noot 10) en Carroll zinvol worden bediscussieerd. Verder kan, dankzij een lemmatiseerprogramma, de vraag over verbogen vormen – basisvorm worden beantwoord.

Maar er is meer. Dit artikel toont ook aan dat vruchtbare samenwerking met statistici niet beperkt kan blijven tot het gebruik van standaard-toetsen uit het SPSS-

pakket. De woordenrijkdom-problematiek vertoont verrassend veel overeenkomst met problemen die statistici krijgen voorgelegd uit de biologie en zelfs uit de numismatiek.[15] De omrekenmethode is niet alleen voor het berekenen van woordenrijkdom ontworpen. Ook bij het gebruik van de computer is er meer dan alleen het ontwerpen van lemmatiseerprogramma's. Tot nu toe was het nauwelijks mogelijk gedetailleerd TTR-onderzoek te verrichten op woordsoorten. In dit onderzoek is nog gewerkt met tekstbestanden die met de hand zijn gecodeerd op woordsoorten. Maar de mogelijkheid van (half-)automatisch coderen komt steeds dichterbij.[16] Dan wordt het veel gemakkelijker om nieuwe teksten te onderzoeken.

Noten

* Met dank aan dr. R.D. Gill en R. in 't Veld (Mathematisch Centrum Amsterdam) voor hulp bij statistische analyse en programmeerwerk, en aan J. Portier voor ontwerpen en implementeren van algoritmen en kritisch commentaar. Verder hebben kritische kanttekeningen van dr. F.H. van Eemeren, dr. R. Grootendorst, dr. W. van Peer en drs. E.J. van der Spek veel onduidelijkheden verhelderd. Met dank ook aan het INL te Leiden voor het belangeloos beschikbaar stellen van een micro-computer in de eindfase van dit onderzoek.

1 Zie J.W. de Vries, in Neerlandica extra Muros, nr. 39, pp. 50-57.

2 De meeste gebruikers van SPSS gaan ervan uit dat deze algemeen gebruikte programma-set betrouwbaar is. Hierop valt echter wel wat af te dingen, men leze 'De Softwarecrisis/2' van G.P. van der Vorst in Intermediair van 24 december 1982.

3 Fragment I komt uit de bundel Jip en Janneke van Annie M.G. Schmidt; fragment II uit een krante-artikel over de bezuinigingspolitiek (G. van Benthem van den Berg in NRC-Handelsblad van 13-1-1983).

4 De coderingen bevatten de volgende informatie: 420 – zelfstandig gebruikt betrekkelijk voornaamwoord; 370 – bijvoeglijk gebruikt aanwijzend voornaamwoord; 000 – 'gewoon' zelfstandig naamwoord in de basisvorm (001 – idem voor het meervoud); 255 – persoonsvorm van overgankelijk werkwoord, verleden tijd enkelvoud; 600 – nevenschikkend voegwoord; 103 – 'gewoon' bijvoeglijk naamwoord in verbogen vorm.

5 Met nadruk zij hier nog vermeld dat er veel min of meer triviale bewerkingen nodig zijn voor computerlinguïstisch onderzoek van start kan gaan. Hier slechts één voorbeeld. De computer telt 'Hij' en 'hij' als twee verschillende types. Voor dit TTR-onderzoek moest dus eerst een programmaatje worden geschreven dat hoofdletters door kleine letters vervangt.

6 Het is opvallend op hoeveel verschillende plaatsen in Nederland pogingen zijn ondernomen voor het ontwikkelen van lemmatiseerprocedures. Enige coördinatie op dit punt zou veel geld en mankracht kunnen besparen. Zie hiervoor de activiteiten van de Werkgroep Corpuslinguistiek en ook het verslag van een vooronderzoek *Hal*fautomatische *T*eksta*n*alyse.

7 Dit geldt ook voor de andere bestanden: dagbladen, romans, enz.

8 Ook de discussie over de wet van Zipf-Mandelbrot is hier van belang. Zie bijvoorbeeld Mandelbrot (1961) en Brainerd (1982). In dit artikel beperk ik mij echter tot voorstellen die al zijn toegepast in TTR-onderzoek op het gebied van letteren.

9 Volgens Weitzman (1971) is deze conclusie onjuist. Maar Ratkowsky (e.a.) 1980 zijn positiever over Herdan's voorstel en ontwikkelden in zijn voetspoor een iets nauwkeuriger maat: $N = V \log_2 (V/2)$. Deze maat leverde in dit onderzoek echter voorspellingen op die weinig verschillen van die van Herdan; daarom is ze hier achterwege gelaten.

10 In deze tabel staan de getallen voor ongelemmatiseerde bijvoeglijke naamwoorden. Gelet op de uitkomsten van paragraaf 5.1 mogen we dezelfde resultaten verwachten voor de gelemmatiseerde bijvoeglijke naamwoorden. Verder zijn ook de bijwoordelijk gebruikte bijvoeglijke naamwoorden meegeteld. De getallen voor V en N zijn dus hoger dan het voorbeeld uit paragraaf 5.1. Voor het toetsen van de voorspellingen volgens de formules maakt dit uiteraard geen verschil.

11 De voorspellingen volgens Herdan en Carroll zijn als volgt berekend (c staat voor constante):

1. Herdan $\frac{\log V}{\log N} = c \rightarrow \log V = c \log N \rightarrow V = N^c$

   I $c = \frac{\log 872}{\log 3133} = 0{,}8411$ Dus $V = N^{0{,}8411}$

   II $c = \frac{\log 1436}{\log 3583} = 0{,}8883$ Dus $V = N^{0{,}8883}$

2. Carroll $\frac{V}{\sqrt{2N}} = c \rightarrow V = c\sqrt{2N}$

   I $c = \frac{872}{\sqrt{2 \times 3133}} = 11{,}02$ Dus $V = 11{,}02 \times \sqrt{2N}$

   II $c = \frac{1436}{\sqrt{2N \times 3583}} = 16{,}96$ Dus $V = 16{,}96 \times \sqrt{2N}$

12 Voor nadere informatie zij verwezen naar het artikel van Good en Toulmin uit 1956.

13 Voor dit probleem is inmiddels een oplossing gevonden. Op het Mathematisch Centrum Amsterdam zijn formules ontwikkeld, waarvoor op dit moment programmatuur wordt ontwikkeld. Zie Gill, Renkema, In 't Veld (te verschijnen).

14 In dit artikel gaat het dus alleen om TTR-onderzoek als methode om eventuele kenmerken van typen taalgebruik op te sporen. In onderzoek waarvan dit artikel een uitloper is (De Taal van Den Haag, Renkema 1981), is de suggestie gewekt dat zo ook oordelen over taalgebruik worden getoetst doordat oordelen over ambtelijke stijl als richtsnoer zijn gebruikt bij het opsporen van kenmerken van die stijl. De correlatie tussen TTR-uitkomsten en bijvoorbeeld oordelen over moeilijkheidsgraad zou een goed onderwerp zijn voor een vervolgonderzoek.

15 Het gaat hier om o.a. de volgende vragen: Hoe kan men bepalen of de bermflora langs snelwegen verarmt? Hoe groot is de kans dat men nog een nieuwe munt vindt uit een bepaalde periode, gegeven het aantal verschillende gevonden exemplaren?

16 In het kader van de samenwerking tussen de TH-Eindhoven (Toegepaste Taalkunde) en de KH-Tilburg (Tekstwetenschap) zal binnenkort het project Halfautomatische Tekstanalyse (HATA) van start gaan. Dit project heeft als doel een bijdrage te leveren aan een systeem dat op basis van invoer van ongecodeerde teksten – na automatische bewerking – een uitvoer oplevert van eenduidig gecodeerde teksten.

Bibliografie

Andolina, Ch., 'Syntactic Maturity and Vocabulary Richness of Learning Disabled Children at Four Age Levels'. In: *Journal of Learning Disabilities* (13), pp. 373-377, 1980.

Brainerd, B., 'On the Relation between the Type-Token and Species-Area Problems'. In: *Journal of Applied Probability* (19), pp. 785-793, 1982.

Calbert, J.P., *Dimensions of Style and Meaning in the Language of Trakl and Rilke.* Contributions to a Semantics of Style. Tübingen: Niemeyer, 1974.

Geest, T. van der, R. Gerstel, R. Appel, B.Th. Tervoort, *The Child's Communicative Competence: Language Capacity in Three Groups of Children from Different Social Classes.* The Hague/Paris: Mouton, 1973.

Gill, R.D., J. Renkema, R. in 't Veld (te verschijnen) Type Token Ratio and Statistics.

Good, F.J. en G.H. Toulmin, 'The Number of New Species, and the Increase in Population Coverage, when a Sample is Increaesed'. In: *Biometrika* (43), pp. 45-54, 1956.

Gruner, C.R. e.a., 'A Quantitative Analysis of Selected Characteristics of Oral and Written Vocabularies'. In: *Journal of Communication* (17), nr. 2, 1967

Halfautomatische Tekstanalyse, een vooronderzoek, Eindhoven-Tilburg (ongepubliceerd), 1983.

Hart, R., 'Absolutism and Situation: Prolegomena to a Rhetorical Biography of Richard M. Nixon'. In: *Communication Monographs* (43), pp. 204-228, 1976.

Herdan, G., *Type-Token Mathematics.* A Textbook of Mathematical Linguistics. 's-Gravenhage: Mouton, 1960.

Höweler, M., 'Diversity of Work Usage as a Stress Indicator in an Interview Situation'. In: *Journal of Psycholinguistic Research* (1), pp. 243-248, 1972.

Mandelbrot, B.B., 'On the Theory of Word Frequencies and on Related Markovian Models of Discourse'. In: *Proceedings of Symposia in Applied Mathematics*, Vol. XII, pp. 190-219, 1961.

Manschreck, T.C., B.A. Maher, D.N. Ader, 'Formal Thought Disorder, the Type-Token Ratio and Disturbed Voluntary Motor Movement in Schizophrenia'. In: *British Journal of Psychiatry* (139), pp. 7-15, 1981.

Meara, P., 'Schizophrenic Symptoms in Foreign Language Learners'. In: *UEA Papers in Linguistics* (7), pp. 22-49, 1978.

Meisels, M., Text Anxiety, Stress, and Verbal Behavior'.
In: *Journal of Consulting Psychology*
31, pp. 577-582, 1967.

Moe, A.J., 'A Comparative Study of Vocabulary Diversity: the Speaking Vocabularies of First-Grade Children, the Vocabularies of Selected First-Grade Primers, and the Vocabularies of Selected First-Grade Trade Books'. Paper presented at the Annual Meeting of the American Education Research Association (Chicago, April 15-19), 1974.

Ratkowsky, D.A., H.H. Halstead, L. Hantrais, 'Measuring Vocabulary Richness in Literary Works: a New Proposal and a Re-assessment of some Earlier Measures'. In: R. Grotjahn (ed.), *Glottometrika* (2), pp. 125-147. Bochum: Brockmeyer, 1980.

Renkema, J., *De taal van 'Den Haag':* Een kwantitatief-statistisch onderzoek naar aanleiding van oordelen over taalgebruik. 's-Gravenhage: Staatsuitgeverij, 1981.

Renkema, J., 'On Functional and Computational LSP-Analysis. Officialese as an Example'. In: Pugh, A.K. and Ulijn, J.M. (eds.), *Reading for Professional Purposes*: Studies in Native and Foreign Languages, pp. 109-119. London: Heinemann Educational, 1983.

Ricci, A.M., 'Content Analysis of Interviewee Verbal Communication: Type-Token Ratio as a Function of Repression-Sensitization and Self-Disclosure'. In: Dissertation Abstracts International (34), pp. 5642-5643, 1974.

Silverman, G., 'Redundancy, Repetition and Pausing in Schizophrenic Speech'. In: *British Journal of Psychiatry (122)*, pp. 407-413, 1973.

Silverman, G., 'The Effect of Topic and Repetitition on Type-Token Ratios of Spoken Monologues'. In: *Language and Speech* (20), pp. 232-239, 1977.

Uit den Boogaart, P.C. (ed.), *Woordfrequenties in geschreven en gesproken Nederlands*. Utrecht: Bohn, Scheltema en Holkema, 1975.

Verslag Werkgroep Frequentie-Onderzoek van het Nederlands. Een onderzoek naar kwantitatieve, lexicale en grammaticale verschijnselen in het Nederlands. Eindhoven (niet gepubliceerd), 1974.

Weitzman, M., 'How useful is the logarithmic type/token ratio?' In: *Journal of Linguistics* (7), pp. 237-243, 1971.