HET WOORDGEBRUIK MEESTER

Een vergelijking van enkele kwantitatieve aspecten van het woordgebruik in juridische en algemeen Nederlandse teksten

LEGAL WORD USE

A comparison of some quantitative aspects of the word use in legal and general Dutch texts

PROEFSCHRIFT
TER VERKRIJGING VAN DE GRAAD VAN DOCTOR
aan de Erasmus Universiteit Rotterdam
op gezag van de Rector Magnificus
Prof. dr. P.W.C. Akkermans M.A.
en volgens besluit van het College voor Promoties

De openbare verdediging zal plaatsvinden op
vrijdag 17 november 1995 om 16.00 uur

door

CORNELIS VAN NOORTWIJK
geboren te Schiedam

Koninklijke Vermande BV
Lelystad

PROMOTIE-COMMISSIE

PROMOTOR:	Prof. mr. R.V. De Mulder
OVERIGE LEDEN:	Prof. dr. W. Martin
	Dr. T. van Haaften
	Prof. dr. D.J. Hessing

ISBN 90 5458 261 8
NUGI 699/901
1995, Koninklijke Vermande BV - Lelystad

Omslag: Willem Mieras

Alle rechten voorbehouden. Niets uit deze uitgave mag worden verveelvoudigd, opgeslagen in een geautomatiseerd gegevensbestand of openbaar gemaakt, in enige vorm of op enige wijze, hetzij elektronisch, mechanisch, door fotokopieën, opnamen, of enige andere manier, zonder voorafgaande schriftelijke toestemming van de auteur.

Voor zover het maken van kopieën uit deze uitgave is toegestaan op grond van artikel 16B Auteurswet 1912 j○ het Besluit van 20 juni 1974, Stb. 351, zoals gewijzigd bij het Besluit van 23 augustus 1985, Stb. 471 en artikel 17 Auteurswet 1912, dient men de daarvoor wettelijk verschuldigde vergoedingen te voldoen aan de Stichting Reprorecht (Postbus 882, 1180 AW Amstelveen). Voor het overnemen van gedeelte(n) uit deze uitgave in bloemlezingen, readers en andere compilatiewerken (artikel 16 Auteurswet 1912) dient men zich tot de auteur of de uitgever te wenden.

INHOUDSOPGAVE

Voorwoord vii

Hoofdstuk 1 **Juridisch taalkundig onderzoek** 1
1.1 Inleiding en probleemstelling 1
1.2 Definities 8
1.3 De stand van zaken met betrekking tot 'recht en taal' 10

Hoofdstuk 2 **Kwantitatieve taalkunde** 16
2.1 Inleiding 16
2.2 Historisch overzicht 19
2.3 Statistische wetmatigheden en maatgetallen betreffende het woordgebruik 24
2.4 Frequentie-onderzoek 32
2.5 Subtaal en vaktaal 35

Hoofdstuk 3 **De drie corpora** 40
3.1 Inleiding 40
3.2 Toekomstig gebruik van gegevens uit de corpora 40
3.3 Het corpus teksten van Nederlandse wetgeving 41
3.4 Het corpus jurisprudentieteksten 51
3.5 Het corpus 'algemeen Nederlands' 56
3.6 Het samenstellen van woordenlijsten 59

Hoofdstuk 4 **De belangrijkste kwantitatieve eigenschappen van het corpus wetgevingsteksten** 64
4.1 Inleiding 64
4.2 Absolute en relatieve woordfrequenties 64
4.3 Frequentieverdelingen 77
4.4 De verhouding tussen woordtokens en woordtypen 83
4.5 Woordlengtes 93

Hoofdstuk 5 **De belangrijkste kwantitatieve eigenschappen van het corpus jurisprudentieteksten** 106
5.1 Inleiding 106
5.2 Absolute en relatieve woordfrequenties 106
5.3 Frequentieverdelingen 113
5.4 De verhouding tussen woordtokens en woordtypen 116
5.5 Woordlengtes 123

Hoofdstuk 6 De belangrijkste kwantitatieve eigenschappen van het corpus algemeen Nederlands 131
6.1 Inleiding 131
6.2 Absolute en relatieve woordfrequenties 131
6.3 Frequentieverdelingen 138
6.4 De verhouding tussen woordtokens en woordtypen 142
6.5 Woordlengtes 148

Hoofdstuk 7 Een vergelijking van kwantitatieve eigenschappen van de drie corpora 158
7.1 Inleiding 158
7.2 Het compenseren van de verschillen in afmeting tussen de corpora 159
7.3 Linguïstische constanten 163
7.4 Andere overeenkomsten en verschillen tussen de corpora . 166
7.5 De kop van de frequentielijsten 205
7.6 Conclusies met betrekking tot overeenkomsten en verschillen in het woordgebruik 218

Hoofdstuk 8 De similariteit van documenten 221
8.1 Het begrip similariteit 221
8.2 Basisbegrippen met betrekking tot similariteit 228
8.3 Het berekenen van een similariteits-score 237
8.4 Het vormen van document-clusters met behulp van de similariteits-score 245
8.5 Document-clusters in het corpus wetgevingsteksten 248

Hoofdstuk 9 Samenvatting en conclusies 266

Summary 272

Literatuurlijst 277

Bijlagen 285
1. Inhoud corpus algemeen Nederlands 286
2. Lijst van woorden van meer dan 30 karakters 296
3. Frequentielijsten (eerste 300 rangnummers) 303
4. Vergelijkingslijst van typen uit de kop van de frequentielijsten 310

Register van persoonsnamen 315

Zakenregister 317

Curriculum vitae 319

VOORWOORD

Door verschillende personen is mij in de afgelopen jaren de vraag gesteld hoe ik er toch toe ben gekomen om juist onderzoek te doen naar het taalgebruik van juristen. Eigenlijk zijn voor die keuze verschillende redenen aan te wijzen. De voornaamste reden heeft te maken met de 'Rotterdamse' visie op het vakgebied 'Informatica en Recht' en vooral met de rol die de *jurimetrie* daarin speelt. In het kort komt deze visie hierop neer, dat belang wordt gehecht aan het tot stand brengen van een *empirische basis* voor juridisch onderzoek. Het verrichten van onderzoek naar bepaalde eigenschappen van het taalgebruik van juristen draagt daartoe bij, gegeven het feit dat de taal in het recht een buitengewoon belangrijke rol speelt.

Naast deze, misschien wat theoretisch aandoende reden is het zo dat kwantitatieve gegevens met betrekking tot juridisch taalgebruik zonder meer noodzakelijk zijn voor allerlei vervolg-onderzoek dat bij het Centrum voor Informatica en Recht op stapel staat (en waar voor een deel al een aanvang mee is gemaakt). Voorbeelden daarvan zijn de analyse van rechterlijke beslissingen aan de hand van jurisprudentieteksten, en het ontwikkelen van conceptuele juridische informatiesystemen. Het onderzoek staat dus niet op zichzelf, maar maakt deel uit van een groter onderzoeksplan op het gebied van de jurimetrie en de juridische informatica.

Bij het uitvoeren van het onderzoek dat in dit proefschrift wordt beschreven en bij het schrijven van het proefschrift zelf zijn velen mij tot steun geweest. Elk van hen wil ik hierbij bedanken. In het bijzonder wil ik mijn promotor, professor Richard De Mulder, dank zeggen. Hij was het die mij heeft gemotiveerd om met dit onderzoek te beginnen en zijn kritische instelling, zijn vele ideeën en raadgevingen en zijn ondersteuning bij praktisch ieder aspect van het onderzoek hebben ervoor gezorgd dat ik het ook tot een goed einde heb kunnen brengen. Verder wil ik de leden van mijn promotiecommissie, in het bijzonder professor dr. W. Martin, dr. T. van Haaften en professor dr. D.J. Hessing, bedanken voor hun waardevolle opmerkingen ten aanzien van het manuscript.

Een woord van dank is ook op zijn plaats voor alle collega's van de vakgroep Rechtshistorie, Rechtsvergelijking en Jurimetrie. Zij hebben mij in staat gesteld dit onderzoek af te ronden, onder andere door een aantal andere taken van mij over te nemen, en zelfs door hun computers beschikbaar te stellen voor het uitvoeren van bepaalde programma's. In het bijzonder wil ik ook Peter van der Landen bedanken. Zonder zijn hulp zou het similariteitsprogramma nu nog niet tot een eind zijn gekomen. Van de student-assistenten wil ik Natalie van Ham bedanken, die veel literatuur voor me heeft opgezocht.

Van groot belang is ook de hulp geweest die ik ontving van het Instituut voor Nederlandse Lexicologie, in het bijzonder van dr. van Sterkenburg en van dr. Kruyt. Zij hebben ervoor gezorgd dat ik de beschikking kreeg over een vergelijkingscorpus met algemeen Nederlandse teksten, een onmisbaar element in het onderzoek.

Tenslotte wil ik Marleen, Laura en Koen bedanken. Ze hebben me de afgelopen tijd lang en vaak moeten missen. Maar op de vraag "Papa, is je boek nu nog niet klaar?" kan ik nu dan ook antwoorden: "Ja hoor, het is nu echt af!".

Kees van Noortwijk
Dordrecht, november 1995

HOOFDSTUK 1 - JURIDISCH TAALKUNDIG ONDERZOEK

1.1 Inleiding en probleemstelling

1.1.1 Recht en taal

In het verleden is door verschillende auteurs al gewezen op de essentiële rol die de taal speelt in wetenschap en praktijk van het recht[1]. In de eerste plaats aan wat we kunnen noemen de 'input'zijde: rechtsbronnen, zoals wetten, beschikkingen en rechterlijke uitspraken zijn allemaal in taal verwoord. Een jurist die van deze bronnen kennis wil nemen kan dat alleen doen door de taal waarin zij zijn gesteld te lezen en te interpreteren. Maar ook aan wat kan heten de 'output'zijde vormt de taal een onmisbaar instrument, dat een jurist in staat stelt zijn benadering van het recht aan anderen kenbaar te maken. Het juridisch probleemoplossen is dan ook in belangrijke mate een 'talige activiteit'[2]. Aangezien het belang van de taal voor het recht dus bijzonder groot is, is het niet verwonderlijk dat ook in het verleden al onderzoek op het gebied van 'recht en taal' heeft plaatsgevonden. Paragraaf 1.3 bevat daarvan enkele voorbeelden. Om zowel dit onderzoek dat in het verleden heeft plaatsgevonden, als het onderzoek dat in dit proefschrift zal worden beschreven en het onderzoek dat nog in de toekomst zou kunnen worden verricht in een kader te plaatsen, is het zinvol eerst te bezien welke aspecten aan het onderwerp 'recht en taal' kunnen worden onderscheiden. Van den Bergh noemt hier bijvoorbeeld de volgende "thema's" (Van den Bergh en Broekman 1979, p. 11 e.v.):

- recht als factor in de taal;
- taal als factor in het recht;
- de verwantschap van recht en taal;
- de verhouding van rechtswetenschap en taalwetenschap.

Zij geven daarnaast aan dat binnen elk thema nog verschillende invalshoeken kunnen worden gekozen, zoals een historische, een rechtswetenschappelijke, een taalwetenschappelijke, een theoretische of een praktische invalshoek. Een probleem met deze indeling in thema's is dat elk thema in feite een zeer breed scala aan onderwerpen kan omvatten. Zo zou bijvoorbeeld het tweede thema, taal als factor in het recht, zowel betrekking kunnen hebben op taalkundig onderzoek naar specifieke eigenschappen van rechtstaal, op rechtsfilosofisch onderzoek naar de invloed van taal op het juridisch handelen, als op rechtshistorisch onderzoek naar de herkomst van bepaalde juridische termen. De thema's zijn daarom in feite te breed, te veelomvattend om onderzoeksactiviteiten op het gebied van recht en taal zinvol te kunnen indelen.

---

1 Zie bijvoorbeeld Van den Bergh (Van den Bergh en Broekman 1979, p. 3), Loth (1991, p. 9) en De Mulder (1984, p. 79).

2 Loth (1991, p. 9).

Loth (1991, p. 9) geeft een indeling die meer is gericht op een "wetenschappelijke werkverdeling", en onderscheidt:

- empirische aspecten, die het studieterrein vormen van de empirische taalwetenschappen, de taalkunde en de linguïstiek,
- filosofische aspecten, onderzocht door rechtsfilosofie en taalfilosofie, en
- methodologische aspecten, waaronder hij vooral onderzoek begrijpt naar een methode voor de oplossing van talige problemen, indien en voorzover van belang voor juridische problemen. Dit zou met name raken aan het werkterrein van de analytische filosofie.

Hij noemt voorbeelden van onderzoek naar elk van deze aspecten. Eén daarvan, het onderzoek van Bax (1986), zal in paragraaf 1.3 nog aan de orde komen.

In dit proefschrift zal met name aandacht worden besteed aan bepaalde *empirische aspecten* van recht en taal, de andere twee aspecten zullen niet of nauwelijks aan de orde komen. Empirische aspecten hebben te maken met de werkelijkheid van recht en taal[3], en dus met het taalgebruik in de juridische praktijk. Voorbeelden hiervan zijn: specifieke kenmerken van taalgebruik in bijvoorbeeld de rechtszaal[4], en in juridische geschriften, zoals wetteksten en teksten van rechterlijke uitspraken[5]. Een overzicht van onderzoek dat in het verleden is verricht, met name op het gebied van taalgebruik in juridische geschriften, is te vinden in paragraaf 1.3. Hier kan reeds vermeld worden dat een gemeenschappelijk kenmerk van de meeste (let wel, niet van alle) onderzoeksprojecten die daar worden vermeld is, dat *kwantitatieve aspecten* van juridisch taalgebruik daarin een ondergeschikte rol spelen. De aandacht blijkt in veel gevallen vooral uit te gaan bepaalde semantische of pragmatische aspecten[6].

Het is opvallend dat nog erg weinig systematisch onderzoek is verricht naar bijvoorbeeld het *woordgebruik* in verschillende soorten juridische teksten, onder gebruikmaking van kwantitatieve methoden, waarbij dit woordgebruik wordt gerelateerd aan dat in algemeen (in dit geval Nederlands) tekstmateriaal. In Nederland is waarschijnlijk het enige project op dat gebied dat van De Mulder en Oskamp (1979), waarin het woordgebruik in het Wetboek van Strafrecht en het Wetboek van Strafvordering werd vergeleken met algemeen Nederlands woordgebruik[7]. Met behulp van dergelijk kwantitatief onderzoek kunnen weliswaar lang niet alle aspecten van juridisch taalgebruik in bijvoorbeeld wet- en jurisprudentieteksten worden beschreven, maar toch valt te verwachten dat in de toekomst het belang van dit type onderzoek alleen maar zal toenemen. Die verwachting is vooral hierop gebaseerd, dat ook in het recht steeds meer wordt ingezien hoe belangrijk het op de juiste wijze omgaan met *informatie* is. Met name onder invloed van de steeds toenemende automatisering, die het mogelijk maakt zeer omvangrijke gegevensbestanden aan te

---

3 Onderzoek kan 'empirisch wetenschappelijk' worden genoemd wanneer het zich richt op vermeerdering van kennis over objecten in de ervaringswereld (De Mulder 1984, p. 18).

4 Loth (1991, p. 10) noemt hier bijvoorbeeld een studie van P.L. Bal: *Dwangcommunicatie in de rechtszaal* (Gouda Quint, Arnhem 1988).

5 Hofhuis (1988, p. 54 e.v.) geeft een aantal voorbeelden van taalgebruik in rechterlijke uitspraken, terwijl Witteveen (1988, p 3 e.v.) in dezelfde bundel aandacht besteedt aan taalgebruik in wetteksten.

6 Zie voor de begrippen syntaxis, semantiek en pragmatiek paragraaf 1.2.

7 Onderzoek dat hieraan enigszins verwant is, is dat van Renkema (1981). Deze analyseerde met behulp van kwantitatieve methoden zogenaamd ambtelijk taalgebruik.

leggen en te raadplegen, is het besef ontstaan dat het van belang is om te beschikken over *kennis van de eigenschappen* van de gegevens die worden opgeslagen. Zulke kennis kan niet alleen een rol spelen bij het op efficiënte en eenduidige wijze inrichten van (juridische) informatiesystemen, maar ook bij het ontwikkelen van methoden en technieken voor het raadplegen van dergelijke systemen, zodat precies die informatie uit het systeem kan worden verkregen die in een bepaald geval nodig is. In juridische informatiesystemen is in het algemeen voornamelijk juridisch *tekstmateriaal* opgeslagen. Met behulp van kwantitatief taalkundig onderzoek kan kennis over en inzicht in de eigenschappen van dat tekstmateriaal worden verkregen.

1.1.2 Doelstelling en opzet van het onderzoek

Dit laatste sluit zeer nauw aan bij de doelstelling van het onderzoek dat in dit proefschrift wordt beschreven. Het onderzoek heeft in belangrijke mate een descriptief karakter. Centraal staat het verzamelen en analyseren van kwantitatieve gegevens over het woordgebruik in verschillende juridische tekst-typen, te weten teksten van wetgeving en van jurisprudentie. Aan de hand van deze gegevens worden de juridische teksten met elkaar vergeleken, en met algemeen Nederlands tekstmateriaal. Daarnaast wordt beproefd of aan de hand van gegevens met betrekking tot het woordgebruik in bepaalde omstandigheden ook *inhoudelijke* samenhang tussen bepaalde juridische teksten kan worden aangetoond. De gegevens die dit onderzoek oplevert vormen onder andere noodzakelijk basismateriaal voor diverse vervolg-onderzoeksprojecten, waarover in paragraaf 1.1.3 meer.

Puntsgewijs kan de doelstelling van het onderzoek als volgt worden geformuleerd:

1. Het samenstellen van tekstcorpora die een goed beeld geven van het taalgebruik in teksten van wetgeving en in jurisprudentieteksten, en van een corpus algemeen Nederlands dat geschikt is voor vergelijkingsdoeleinden.

2. Het in kaart brengen van het woordgebruik en van bepaalde kwantitatieve aspecten daarvan in de afzonderlijke corpora, en het aan de hand daarvan vergelijken van de corpora.

3. Het illustreren, aan de hand van een uitgewerkt voorbeeld, hoe kennis van kwantitatieve eigenschappen van juridisch tekstmateriaal kan worden benut bij het ontwikkelen van bepaalde computerprogrammatuur. Met behulp van deze programmatuur kan samenhang tussen de afzonderlijke documenten in omvangrijke juridische tekstbestanden worden aangetoond, hetgeen zowel - voor de rechtswetenschap als voor de rechtspraktijk interessant zou kunnen zijn.

Deze doelstelling is, met name voor wat betreft het tweede punt, ruim te noemen. Voor het 'in kaart brengen van het woordgebruik' is in principe een zeer groot aantal technieken denkbaar. Aangezien het niet mogelijk is om al deze technieken toe te passen, is het nodig daaruit een gefundeerde keuze te maken door aan opzet en uitvoering van het onderzoek een aantal beperkingen te stellen. De belangrijkste van deze beperkingen zijn:

1. de tekstcorpora die zijn samengesteld en bestudeerd bevatten uitsluitend *Nederlandse schrijftaal*, teksten in de Nederlandse taal die in het verleden in drukvorm zijn verschenen. Er is geen poging gedaan tot het registreren en analyseren van gesproken taal. Redenen hiervoor zijn met name het in onvoldoende mate beschikbaar zijn van materiaal - weergaven van gesproken (juridisch) Nederlands waren niet c.q. niet in voldoende mate aanwezig - maar ook de wens om te werken met reconstrueerbare brongegevens[8].

2. het onderzoek beperkt zich tot de *vorm* van de teksten, dat wil zeggen tot aspecten als: hoe lang zijn de teksten, uit hoeveel verschillende woorden bestaan ze en welke zijn dat, etc. De vorm van teksten wordt gerekend tot de *syntaxis*. Daartoe behoren ook de regels die in een taal gelden voor het onderling combineren van woorden, maar dat aspect zal hier buiten beschouwing worden gelaten. Geabstraheerd wordt ook van alle semantische en pragmatische aspecten van de teksten (aspecten die betrekking hebben op de betekenis c.q. het effect ervan). De reden voor deze beperking tot vorm-aspecten is dat op dit niveau (in tegenstelling tot bijvoorbeeld het semantisch niveau) *automatische verwerking en analyse* van het tekstmateriaal meestal nog goed mogelijk is (zij het soms met enkele beperkingen, zie ook hoofdstuk 2, paragraaf 4).

3. het onderzoek beperkt zich tot het *woord-niveau* van de bestudeerde teksten. Dat betekent dat geen analyses zullen worden uitgevoerd met betrekking tot het gebruik van bijvoorbeeld letters of lettergrepen (lager-dan-woord niveau) en ook niet naar het voorkomen danwel de structuur van woordcombinaties, zinnen, alinea's, hoofdstukken, etc. (hoger-dan-woord niveau)[9]. De reden daarvoor is vooral dat woorden kunnen worden beschouwd als de kleinste 'complete taaluitingen' in een taal[10]. Ze vormen de kleinste eenheden waarin een tekst kan worden gesplitst die ook zelfstandig een betekenis hebben. Het is daardoor, wanneer in teksten overeenkomsten of verschillen op woord-niveau worden aangetroffen, in sommige gevallen mogelijk deze te verklaren aan de hand van bijvoorbeeld het onderwerp waarover de teksten handelen[11]. Bij het lager-dan-woord niveau is dit meestal niet mogelijk. Verder geldt bij dat niveau dat grote verschillen tussen teksten, bijvoorbeeld in het lettergebruik, niet zo snel te verwachten zijn wanneer die teksten in dezelfde taal zijn gesteld. De relatieve frequenties van de letters A-Z en de cijfers 0-9 verschillen bij de drie hier onderzochte tekstcorpora steeds minder dan 0,75% van elkaar. Voor het hoger-dan-woord niveau geldt dit bezwaar niet, maar daar is weer een probleem dat het aantal mogelijke combinaties van woorden dat zou kunnen worden vergeleken, in teksten van enige omvang zeer groot is terwijl de meeste combinaties maar erg weinig zullen voorkomen. Een zeer groot aantal elementen dus, waarvan de meeste met lage frequentie.

---

8 Op het belang hiervan wordt bijvoorbeeld gewezen door Uit den Boogaart (1975, p. 11).

9 Vgl. Brandt Corstius (1978, p. 84, 143 en 198) en Dik en Kooij (1988, p. 169 e.v.).

10 Dik en Kooij (1988, p. 170).

11 Daarbij kunnen nog wel problemen optreden wanneer een woord verschillende betekenissen kan hebben. Zie Brandt Corstius (1978, p. 84).

4. Verder is het zo dat uitsluitend wordt gewerkt met de woordvormen zoals die *voorkomen in de bestudeerde corpora*. Er wordt geen poging gedaan woorden tot bijvoorbeeld hun stam-vorm terug te brengen (lemmatisering). Daarvoor bestaan tegenwoordig weliswaar vrij geavanceerde computerprogramma's (vaak aangeduid met de term 'taggers') die zonder menselijke tussenkomst werken, maar deze begeven zich wanneer een woordvorm, afhankelijk van de betekenis, herleid kan worden tot verschillende lemmata onvermijdelijk ook op het niveau van de semantiek. Dergelijke analyses vallen buiten het bestek van dit onderzoek (zie hierboven, onder 2). Om dezelfde reden zal geen gebruik worden gemaakt van programma's om de teksten taalkundig te ontleden (vaak aangeduid met de term 'parsers'). Er zullen dan ook (op enkele uitzonderingen na, zie bijvoorbeeld paragraaf 4.5.4) geen uitspraken worden gedaan over de verschillende woordsoorten - zelfstandige naamwoorden, bijvoeglijke naamwoorden, etc. - die in de teksten voorkomen.

Naast deze keuzes die zijn gemaakt ten aanzien van de samenstelling van de corpora en de toe te passen analyses, hebben ook de volgende punten nog een rol gespeeld bij de wijze waarop het onderzoek is uitgevoerd:

1. Het onderzoek heeft voor een belangrijk deel een descriptief karakter. De gegevens die in het kader ervan worden verzameld en die in dit proefschrift zijn opgetekend vormen een noodzakelijke basis voor vervolgonderzoek dat binnen het Centrum voor informatica en recht en elders wordt verricht, onder andere voor projecten die tot doel hebben de ontwikkeling van nieuwe technieken voor het zoeken in juridische tekstbestanden (zie ook paragraaf 1.1.3, hierna). Tijdens het onderzoek zal er daarom als regel zorg voor worden gedragen dat woordenlijsten, frequentiegegevens en andere onderzoeksresultaten op zodanige manier worden ingericht c.q. opgeslagen dat daarvan ook in vervolgonderzoek optimaal gebruik kan worden gemaakt, en ook de verslaglegging in dit proefschrift zal daarop zijn gericht.

2. Voor de verwerking van gegevens is uitsluitend gebruik gemaakt van personal computers (PC's). Dit type computer heeft de afgelopen tien jaar een stormachtige ontwikkeling doorgemaakt. De verwerkingssnelheid en opslagcapaciteit zijn sterk toegenomen, terwijl de prijzen zijn gedaald. De hoeveelheid achtergrondgeheugen die nodig is om te kunnen werken met omvangrijke corpora, zoals bij dit onderzoeksproject het geval is (de corpora beslaan gezamenlijk ongeveer 400 miljoen lettertekens, ofwel 400 megabytes), was bij de aanvang van het project nog enigszins een probleem (met veel moeite kon een PC worden uitgerust met twee maal 300 megabytes aan diskruimte), maar is op dit moment (medio 1995) al niets bijzonders meer. Een voordeel van het werken met PC's, naast het feit dat ze goedkoop en in ruime mate voorhanden zijn, is dat tevens wordt aangetoond dat taalkundig onderzoek op omvangrijke corpora tegenwoordig binnen het bereik van een brede groep gebruikers c.q. onderzoekers ligt.

1.1.3 Plaats van het onderzoek

Na wat hiervoor is gezegd over de opzet van dit onderzoek naar woordgebruik in juridische teksten, dient nu nog enige aandacht te worden besteed aan de plaats die onder

zoek als dit inneemt binnen zowel de rechtswetenschap als de taalwetenschap. Daarvoor moet echter eerst een enkele opmerking worden gemaakt over empirisch, rechtswetenschappelijk onderzoek in het algemeen.

In het verleden is al meer dan eens betoogd dat de rechtswetenschap slechts in beperkte mate een *empirische wetenschap* kan worden genoemd[12]. Met 'empirische wetenschap' wordt in dat verband bedoeld: een wetenschap die is gericht op kennisvermeerdering van objecten in de ervaringswereld. De Mulder (1984, p. 19) noemt de rechtswetenschap "traditioneel een 'normatieve wetenschap'", de beoefenaren richten zich meer op het hanteren van voorschriften dan op het doen van systematische waarnemingen van objecten in de ervaringswereld. Doordat rechtswetenschappelijk onderzoek zich steeds zo weinig heeft gericht op het verkrijgen van kennis over de werkelijkheid, is dit soort kennis in het recht ook slechts in geringe mate aanwezig.

De laatste decennia lijkt hierin echter verandering te komen. In het proefschrift van Kerkmeester (1989) wordt uitvoerig ingegaan op de vraag of het mogelijk is om een empirische rechtswetenschap te bedrijven, welke vraag (onder zekere voorwaarden) bevestigend wordt beantwoord. Kerkmeester gebruikt, in navolging van De Mulder (1984, p. 240) voor 'empirische rechtswetenschap' de term *jurimetrie*[13]. Wanneer wij deze term aanvaarden, kan onderzoek naar het woordgebruik in juridische teksten worden betiteld als jurimetrisch onderzoek, aangezien dit onderzoek tot doel heeft de vermeerdering van kennis over bepaalde objecten uit de ervaringswereld, te weten bepaalde soorten Nederlandse juridische teksten.

Kan onderzoek naar juridisch tekstmateriaal dus *jurimetrisch* worden genoemd omdat het object van onderzoek juridisch is, en het onderzoek tot doel heeft kennisvermeerdering over bepaalde objecten uit de werkelijkheid, het onderzoek kan evenzeer (kwantitatief) *taalkundig* worden genoemd aangezien de meeste van de gebruikte methoden uit die wetenschap afkomstig zijn. Aangezien de taalkunde voor de meeste juristen onbekend terrein is, zal daaraan in het volgende hoofdstuk afzonderlijk aandacht worden besteed.

Het onderzoek dat in dit proefschrift wordt beschreven kan dus bijdragen aan de vermeerdering van kennis over juridisch tekstmateriaal. Een van de voornaamste redenen waarom dat van belang is werd al genoemd in paragraaf 1.1.1: er bestaat een verband met de manier waarop in de toekomst juridische documentatie zal worden opgeslagen en geraadpleegd. Nu al is het zo dat computertechnologie overal ter wereld in steeds toenemende mate wordt ingezet voor het opslaan en raadplegen van gegevensverzamelingen. Met het beschikbaar komen van steeds omvangrijker achtergrondgeheugens en het dalen van de prijzen daarvan is het tegenwoordig niet alleen voor grote organisaties, maar zelfs voor particulieren mogelijk geworden zeer grote databanken op te bouwen. Met het ontstaan van meer van dergelijke databanken (ook op juridisch gebied) wordt echter ook het daaruit opvragen van gegevens een steeds groter probleem. De methoden die daarvoor

---

[12] De Mulder (1984, p. 18 en 19), Kerkmeester (1989, p. 6).

[13] Deze term werd waarschijnlijk voor het eerst gebruikt door de Amerikaanse jurist Lee Loevinger in zijn artikel 'Jurimetrics, the next step forward' uit 1949 (herdrukt in 1971): Loevinger (1971, p. 31). Zie ook Kerkmeester (1989, p. 26).

tot nu toe worden gebruikt bieden voor het selecteren van documenten meestal weinig andere mogelijkheden dan het opgeven van een of meer trefwoorden, waarna de documenten worden opgezocht waarin die woorden voorkomen.

Aangezien deze traditionele techniek (ook wel aangeduid met de term 'booleaans zoeken') ook (of misschien wel: juist) bij het zoeken in juridische gegevensverzamelingen niet langer voldoet wordt momenteel onderzoek verricht naar zoektechnieken die het mogelijk maken op een andere manier documenten te selecteren. Gewerkt wordt aan zogenaamde conceptuele zoektechnieken, waarbij niet langer uitsluitend wordt geselecteerd op basis van het al dan niet voorkomen van woorden, maar waarbij documenten worden gezocht die voldoen aan een bepaald zoek-concept. Zo'n concept kan bijvoorbeeld betrekking hebben op het juridische onderwerp c.q. vraagstuk dat in het document behandeld wordt. Simpel gesteld kan conceptueel zoeken dus worden omschreven als het zoeken op behandelde onderwerpen of concepten, in plaats van het zoeken op trefwoorden waarvan de traditionele technieken gebruik maken.

Dit onderzoek staat voor een belangrijk deel in het teken van het verzamelen van gegevens en het beantwoorden van voorvragen die van belang zijn voor het onderzoek naar conceptuele juridische informatiesystemen. Zo zal onder andere worden nagegaan hoe *omvangrijk* de onderzochte juridische teksten zijn, hoeveel *verschillende woorden* ze bevatten, hoe deze woorden over de samenstellende documenten zijn *gespreid*, etc. Verder zal ondere andere de vraag worden beantwoord of juridische documenten in bepaalde opzichten (met name voor wat betreft vorm en woordgebruik) *afwijken van andere documenten*, waarna kan worden beoordeeld of dit gevolgen kan hebben voor het vervaardigen van zoeksystemen. Een andere vraag, waarop in het laatste hoofdstuk van dit proefschrift zal worden ingegaan, is of documenten die bepaalde overeenkomsten vertonen voor wat betreft het woordgebruik, ook inhoudelijk verwantschap bezitten. Uiteindelijk is het niet ondenkbaar dat in de komende jaren zal blijken dat het woordgebruik in juridische documenten een belangrijke aanwijzing kan leveren om na te gaan of een bepaald document uit een gegevensverzameling past binnen een door de gebruiker aangegeven zoekconcept[14]. Mocht dat inderdaad het geval zijn, of mocht blijken
dat andere in dit proefschrift onderzochte aspecten die te maken hebben met de vorm van documenten hiervoor van belang zijn, dan zal ook deze kennis mogelijk van invloed zijn op het ontwikkelen van conceptuele juridische informatiesystemen.

1.1.4 Indeling van dit proefschrift

Aan het einde van deze eerste paragraaf zal tenslotte nog worden aangeven hoe de rest van dit proefschrift is ingedeeld.

- Dit inleidende hoofdstuk zal worden vervolgd met het geven van enkele definities. Verder zal aandacht worden besteed aan onderzoek naar 'recht en taal' dat in het verleden heeft plaatsgevonden.

---

[14] De begrippen 'concept' en 'conceptueel juridisch informatiesysteem' worden besproken in De Mulder *et al.* (1993), en in Wildemast en De Mulder (1992).

- Hoofdstuk 2 bevat een inleiding tot de kwantitatieve taalkunde, het onderdeel van de taalkunde dat zich bezighoudt met de kwantitatieve eigenschappen van natuurlijke taal. Hierbij passeert een aantal begrippen de revue die voor dit onderzoek van belang zijn.
- Hoofdstuk 3 behandelt de samenstelling van de drie tekstcorpora die in dit onderzoek centraal staan. Ook wordt ingegaan op de manier waarop deze corpora zijn verwerkt, en waarop het tekstmateriaal in afzonderlijke woorden is verdeeld.
- De hoofdstukken 4, 5 en 6 bevatten een beschrijving van de kwantitatieve eigenschappen van respectievelijk het corpus wetgevingsteksten, het corpus jurisprudentieteksten en het corpus algemeen Nederlands. In de hoofdstukken 5 en 6 wordt daarbij af en toe al even teruggekeken naar de eerder behandelde corpora, en wordt al een aantal gegevens vergeleken (voorzover op dat moment al mogelijk).
- In hoofdstuk 7 wordt het woordgebruik in de drie corpora in detail vergeleken. Een punt van belang is daarbij dat veel eigenschappen van de drie corpora niet zonder meer kunnen worden vergeleken, doordat bij de bepaling van sommige van die eigenschappen de *grootte van de corpora* een rol speelt. Deze grootte loopt bij de drie corpora nogal uiteen, van 4 miljoen woordtokens (zie voor dit begrip de volgende paragraaf) in het corpus algemeen Nederlands, tot 40 miljoen in het corpus jurisprudentieteksten. Om de invloed van de verschillen in grootte te neutraliseren, worden in hoofdstuk 7 uit de twee grootste corpora steekproeven genomen, ter grootte van het kleinste corpus. De kwantitatieve gegevens worden vervolgens met elkaar vergeleken. Tevens vindt vergelijking plaats met de gegevens die betrekking hebben op de volledige corpora.
- In hoofdstuk 8 wordt het onderzoek afgesloten met een uitgewerkt voorbeeld van hoe woordfrequentiegegevens in de praktijk kunnen worden gebruikt in een juridisch documentatiesysteem. Gedemonstreerd wordt hoe aan de hand van het woordgebruik in de documenten waaruit een corpus bestaat, met behulp van een computermodel clusters kunnen worden gevormd van documenten die veel met elkaar gemeen hebben. Dit voorbeeld biedt mogelijk aangrijpingspunten voor vervolgonderzoek.
- Hoofdstuk 9 bevat tenslotte een samenvatting van het onderzoek, met de conclusies die eruit getrokken kunnen worden.

1.2 Definities

Het is aan het begin van dit onderzoeksverslag nodig om enkele begrippen te definiëren.

1.2.1 Woordtype, woordtoken en woordlemma

Wanneer we het hebben over 'woorden', bedoelen we daarmee niet altijd hetzelfde. Brandt Corstius geeft om dat duidelijk te maken de volgende voorbeeldzinnen[15]:

1. De bijbel bevat een miljoen woorden
2. De gemiddelde Nederlander gebruikt vijfduizend woorden
3. Van Dale bevat 220 000 woorden

[15] Brandt Corstius (1978, p. 46).

In elk van deze zinnen heeft 'woorden' een andere betekenis. Voor de betekenis van 'woorden' in de eerste zin wordt hier de term *woordtoken* gebruikt. Andere termen voor hetzelfde begrip zijn *woordteken* en *woordvoorkomen*.

> Onder een woordtoken wordt verstaan een rijtje lettertekens dat in een tekst tussen twee woordscheiders staat, waarbij niet van belang is of het betreffende rijtje ergens anders ook staat.

Voor de betekenis van 'woorden' in de tweede zin wordt hier de term *woordtype* gebruikt:

> Onder een woordtype wordt verstaan een verzameling van gelijke woordtokens, waarbij het begrip 'gelijk' wordt opgevat als: bestaande uit eenzelfde rijtje letters.

Voor de betekenis van 'woorden' in de derde zin is in gebruik de term *woordlemma*:

> Onder een woordlemma verstaan we een klasse van woordtypen, die in taalkundig opzicht als gelijk beschouwd worden (de werkwoordsvormen van één stam, meervoud en enkelvoud van één zelfstandig naamwoord).

De term woordlemma wordt hier verder niet meer gebruikt, het onderscheid tussen woordtokens en woordtypen is echter op een aantal plaatsen in dit onderzoek van groot belang. Een frequentielijst, zoals die wordt opgesteld voor elk van de drie onderzochte corpora, is een lijst van woordtypen. Een woordtype kan op de frequentielijst worden geplaatst door er een van de woordtokens van dat type in af te drukken. 'Alle woordtypen' uit een corpus wil zeggen: alle verschillende woorden, 'alle woordtokens' betekent: alle rijtjes lettertekens (gescheiden door woordscheiders) waaruit de tekst bestaat. Het corpus wetgevingsteksten bevat 12807659 woordtokens en 143156 woordtypen, het corpus jurisprudentieteksten 40190212 woordtokens en 265841 woordtypen, en het corpus algemeen Nederlands 4616455 woordtokens en 142879 woordtypen.

1.2.2 Syntaxis, semantiek en pragmatiek

Een onderscheid dat kan worden gemaakt voor ieder tekensysteem, dus ook voor natuurlijke taal, is dat tussen de *syntaxis*, de *semantiek* en de *pragmatiek*[16]. Wanneer we ons beperken tot geschreven taaluitingen, kan onder deze begrippen het volgende worden verstaan:

- de *syntaxis* betreft de vormelijke opbouw van taaluitingen, uit welke elementen deze bestaan en hoe die elementen zijn gecombineerd. De opbouw van een zin uit woorden en van een tekst uit zinnen betreft allemaal de syntaxis van een taaluiting. Daarnaast behoren tot de syntaxis de *regels* die beschrijven hoe een taaluiting moet zijn opge-

[16] Dik en Kooij (1988, p. 27) geven aan dat dit onderscheid oorspronkelijk werd gemaakt door filosofen die zich met de eigenschappen van tekensystemen in het algemeen bezighielden. Zij noemen in dit verband een werk van Charles W. Morris: *Foundations of the theory of signs* (International Encyclopedia of Unified Science. University of Chicago Press, Chicago 1938).

bouwd, wil deze bijvoorbeeld kunnen gelden als een Nederlandse volzin. De opbouw van de *woorden* waaruit een taaluiting bestaat uit elementen als letters en dergelijke wordt door taalkundigen in het algemeen gerekend tot de *morfologie*. Aan dit aspect zal hier slechts in beperkte mate aandacht worden besteed, met name in de vorm van het analyseren en vergelijken van woord*lengtes*.

- de *semantiek* betreft de betekenis die aan een bepaalde taaluiting kan worden toegekend.

- de *pragmatiek* betreft de situatie waarin een taaluiting wordt gedaan, dus bijvoorbeeld wie de taaluiting doet, in combinatie met welke andere taaluitingen dit gebeurt, hoe de taaluiting tot stand komt en welk effect ervan uitgaat.

In dit geval zal het onderzoek zoveel mogelijk worden beperkt tot een bepaald gedeelte van de syntaxis van op schrift gestelde taaluitingen, namelijk tot de woorden waaruit deze taaluitingen zijn opgebouwd. Van deze woorden wordt een aantal kwantitatieve gegevens verzameld, welke enerzijds betrekking hebben op de woorden zelf (bijvoorbeeld: hoe lang zijn ze), en anderzijds op de manier waarop worden gebruikt (bijvoorbeeld: hoe vaak komen ze voor, en hoe is hun spreiding). Zaken die algemeen tot de syntaxis worden gerekend, maar die *geen* onderdeel uitmaken van dit onderzoek zijn bijvoorbeeld de wijze waarop woorden worden gecombineerd in een zin, de opbouw en lengte van zinnen, en de grammaticale constructies die in de taaluitingen voorkomen.

1.3 De stand van zaken met betrekking tot 'recht en taal'

1.3.1 Een kort historisch overzicht

In paragraaf 1 werd al opgemerkt dat taal in het recht een grote rol speelt. Het is dan ook niet verwonderlijk dat in het verleden in Nederland en daarbuiten al op vrij uitgebreide schaal is nagedacht en geschreven over specifieke eigenschappen en problemen van juridisch taalgebruik. Van den Bergh geeft daarvan een kort historisch overzicht[17]. Daaruit blijkt onder andere dat de aandacht zich aanvankelijk - we spreken dan over de tweede helft van de 19e eeuw - vooral richtte op specifieke, ongewenst geachte elementen in het juridisch taalgebruik. Men was van mening dat juist in het recht onzuiver taalgebruik, te lange zinnen, archaïsch taalgebruik, juristenlatijn, vreemde woorden etc. niet te pas kwamen. Later[18] ontstond ook wat ruimere belangstelling voor een meer fundamentele rechtstaal-filosofie, waarin de invloed van de taal op het menselijk denken en handelen een centrale rol speelde. Op dat onderwerp kan ik hier helaas echter niet dieper ingaan. In de periode die volgde verschenen nauwelijks meer serieuze beschouwingen over

---

17 Van den Bergh en Broekman (1979, p. 3-6).

18 Met name onder invloed van het proefschrift van Jacob Israel de Haan (1916) over rechtskundige significa.

juridisch taalgebruik[19], tot ongeveer 20 jaar geleden de belangstelling voor het onderwerp weer groeide. Begin jaren zeventig verschenen enige publikaties over het belang van de zogenaamde Sapir-Whorf-hypothese voor de praktijk van het recht, waarover in de volgende subparagraaf meer. In 1976 verscheen in het Nederlands Juristenblad een artikel van Reinsma en Reinsma[20], dat weliswaar niet vernieuwend te noemen was - het vervolgde in feite de discussie over onjuist juridisch taalgebruik die aan het begin van de eeuw werd gevoerd - maar mogelijk toch een stimulans is geweest voor ander onderzoek op dit gebied. In 1979 verscheen het al aangehaalde werk van Van den Bergh en Broekman, waarschijnlijk het breedste overzicht van de raakvlakken tussen recht en taal tot dat moment. Dit rapport was geschreven als een preadvies (of eigenlijk als twee afzonderlijke preadviezen) voor de Koninklijke Notariële Broederschap. De auteurs gaan naast kenmerken en problemen van juridisch taalgebruik bijvoorbeeld ook in op de betekenis van de taal en van stromingen binnen de taalkunde voor het *denken over* het recht, op overeenkomsten tussen het systeem van het recht en het systeem van de taal en op juridische consequenties van het gebruik van taalelementen. Op dit laatste onderwerp zal in de volgende subparagraaf nog worden terugkomen. Verder kwam, eveneens in 1979, het onderzoek van De Mulder en Oskamp tot stand. Hierin werd, voorzover kan worden nagegaan voor het eerst, ook aandacht besteed aan kwantitatieve aspecten van juridisch taalgebruik.

In de jaren tachtig is de belangstelling voor 'recht en taal' verder toegenomen, tenminste als we mogen afgaan op het aantal verschenen publikaties. Enkele daarvan kunnen hier worden genoemd, zonder dat daarbij wordt gestreefd naar volledigheid. In de eerste plaats het boek van Bax[21], waarin deze betoogt dat de taal en wereld van het recht een relatief gesloten systeem vormen, autonoom ten opzichte van algemeen Nederlands taalgebruik en de wereld van alledag. Aan de hand van de juridische constructie van 'conversie van rechtshandelingen' toont hij aan dat de interpretatie en kwalificatie van taaldaden - noodzakelijk om de juiste *betekenis* van die taaldaden vast te stellen - in het recht anders gebeuren dan daarbuiten, aangezien zij gebonden zijn aan de ruimte die wettelijke regels, jurisprudentie en dogmatiek daarvoor bieden. De studie heeft dus met name betrekking op (het vaststellen) van *betekenis*-aspecten van juridische uitspraken.

In 1988 verscheen de bundel 'Meesterlijke taal', uitgegeven ter gelegenheid van het achtste lustrum van de beroepsvereniging de Jonge Balie te Utrecht. Deze bundel bevat een aantal opstellen over de rol van taal in de rechtspraktijk. In deze bundel komen enerzijds bijdragen voor die zich vooral begeven op het terrein van de semantiek en pragmatiek van juridische taaluitingen, zoals 'De taal van de wetgever' van de hand van W.J. Witteveen[22]. Daarnaast treffen we echter ook enkele bijdragen aan waarin aandacht wordt besteed aan syntactische aspecten van (bijvoorbeeld) rechterlijke uitspraken, zoals die van H.F.M. Hofhuis over 'De taal van de rechter'[23].

---

19 Van den Bergh geeft aan dat slechts sprake was van "verspreide en anecdotische opmerkingen" (Van den Bergh en Broekman 1979, p. 6).

20 Reinsma en Reinsma (1976).

21 Bax (1986).

22 Witteveen (1988).

23 Hofhuis (1988).

Vermeld kan nog worden het begin jaren tachtig verschenen proefschrift van Renkema over 'De taal van Den Haag'[24], hoewel daarin niet zozeer juridisch, alswel 'ambtelijk' taalgebruik centraal staat. Tussen deze twee bestaat echter wel een bepaalde relatie, terwijl dit onderzoek zich bovendien onderscheidt doordat het zich ook richt op sommige kwantitatieve aspecten van taalgebruik. Daarnaast bestaan er ook overeenkomsten tussen de methode die Renkema gebruikt om stilistische kenmerken van ambtelijk taalgebruik aan het licht te brengen, namelijk het uitvoeren van vergelijkingen met een corpus algemeen Nederlandse teksten (het z.g.n. Eindhovens corpus, waarover in het volgende hoofdstuk meer) en de methode die in dit onderzoek naar juridisch woordgebruik wordt toegepast.

Een laatste studie die in dit overzicht vermeld kan worden is teven de meest recente. Het betreft het onderzoek van Maes naar het gebruik van z.g.n. nominale anaforen[25]. Op zich gaat het hier om algemeen taalkundig onderzoek. De relatie met het recht bestaat echter hierin, dat Maes in zijn onderzoek voor een belangrijk deel gebruik maakt van juridisch tekstmateriaal, namelijk rechterlijke uitspraken. Deze teksten zijn volgens hem in zoverre bijzonder, dat *praktisch alle anaforen die erin voorkomen nominaal zijn*. Denk bijvoorbeeld aan veelgebruikte formuleringen in deze teksten zoals '*Eiser* blijft in gebreke ...', '*Partijen* zijn overeengekomen ...', etc. Het gebruik van dit hoge aantal nominale anaforen is volgens Maes voornamelijk te verklaren uit twee factoren. Ten eerste uit bepaalde functionele aspecten van het recht: juridische concepten hebben vaak een kunstmatige, van de dagelijkse werkelijkheid afwijkende betekenis. Op het moment dat naar zo'n concept wordt verwezen wordt kennelijk liever geen persoonlijk of aanwijzend voornaamwoord gebruikt, om aan te geven dat het hier om het *juridische begrip* gaat. De tweede verklaring kan volgens hem zijn dat aan juridisch taalgebruik specifieke voorwaarden kunnen worden gesteld, zoals bijvoorbeeld eenduidigheid, die het gebruik van nominale anaforen in de hand werken. Interessant in verband met het onderzoek dat in dit proefschrift centraal staat is dat uit het onderzoek van Maes blijkt dat in juridische teksten bepaalde woordtypen voorkomen volgens een patroon dat duidelijk afwijkt van dat in andere teksten.

1.3.2 Iets over de invloed van de taal op het recht

Hoewel het eigenlijk buiten het kader van dit onderzoek naar juridisch woordgebruik valt, wordt in deze subparagraaf enige aandacht besteed aan een onderwerp dat te maken heeft met de invloed van taal in de praktijk van het recht. De reden waarom dit onderwerp hier behandeld wordt is dat aspecten van het woordgebruik, zoals de lengte van woordtypen, er een belangrijke rol in spelen. De belangstelling voor dit specifieke onderwerp dateert van het begin van de zeventiger jaren. Hoewel de hele discussie die in die tijd werd gevoerd nu nauwelijks meer serieus wordt genomen, was er toen sprake van een 'hot issue', met name onder beoefenaren van de strafrechtwetenschap en de criminologie. Het

---

[24] Renkema (1981).

[25] Een nominale anafoor is een terugverwijzing in een tekst, niet bestaande uit bijvoorbeeld een aanwijzend of een persoonlijk voornaamwoord, maar uit een zelfstandig naamwoord. Een voorbeeld van een nominale anafoor in een zin: Koningin Beatrix verscheen onlangs op de televisie. *De vorstin* maakte een ontspannen indruk.

ging allemaal stigmatisering (het verschijnsel dat de samenleving afwijkende individuen op een bepaalde manier zichtbaar en herkenbaar maakt, in gunstige of - en daar gaat het in dit verband vooral om - in ongunstige zin), en de invloed die bepaalde vormen van taalgebruik daarop zouden kunnen hebben. In 1971 verscheen een studie van Prof. dr. H. Bianchi, genaamd *Stigmatisering*[26]. Als criminoloog was Bianchi met name geïnteresseerd in de negatieve aspecten van dit verschijnsel - zoals bijvoorbeeld het stigma dat veel criminelen krijgen toegediend wanneer zij, na het uitzitten van hun straf, weer terugkeren in de maatschappij. Aangezien deze vorm van stigmatisering dikwijls gepaard gaat met - of misschien is het beter om te zeggen: bestaat uit - het gebruik van bepaalde woorden, zoals 'dief', 'tbr-klant' etc., zag Bianchi bij de bestrijding van dit ongewenste verschijnsel een rol weggelegd voor de taalkunde. Kort weergegeven kwam zijn betoog erop neer dat wanneer de maatschappij erin zou slagen woorden die een stigmatiserend effect kunnen hebben in het geheel niet meer te gebruiken, het hele verschijnsel stigmatisering op den duur zou verdwijnen (of in ieder geval sterk in omvang zou afnemen), bij gebrek aan mogelijkheden om stigma's in verbale en schriftelijke communicatie tot uitdrukking te brengen.

Deze opmerkelijke stelling baseerde Bianchi mede op de zogenaamde Sapir-Whorf hypothese. Door middel van deze hypothese trachtten de taalkundigen Sapir en Whorf een verband te leggen tussen de taal die door bepaalde volkeren wordt gebruikt en de cultuur van die volkeren. Het blijkt bijvoorbeeld zo te zijn dat mensen de neiging hebben de taal aan te passen aan hun specifieke behoeften. Zo hebben Eskimo's naar verluidt vele woorden voor sneeuw[27], en gebruiken bij ons boeren een aantal verschillende woorden om koeien van verschillende leeftijden aan te duiden. Sapir en Whorf gingen verder door te stellen dat omgekeerd, wanneer mensen voor een bepaald begrip geen woord kennen, zij ook de grootste moeite hebben dat begrip van andere begrippen te onderscheiden. Zo zouden bijvoorbeeld bepaalde Indianenstammen geen aparte woorden hebben voor de kleuren geel en oranje, en als gevolg daarvan ook geen onderscheid tussen deze kleuren kunnen maken. Leerden deze Indianen echter Engels, dan konden zij het onderscheid wel maken. Dat zou dus betekenen dat een taal de perceptie van haar gebruikers mede kan bepalen c.q. vastleggen.

Volgens Bianchi wees dit alles erop dat er in een taal verschillende typen woorden kunnen bestaan, te weten woorden met een hoge 'codabiliteit', die in staat zijn in een keer een grote hoeveelheid ongedifferentieerde informatie over te brengen, en woorden met een lage codabiliteit die minder, maar gerichte informatie overbrengen. Woorden met een hoge codabiliteit zijn daarbij meestal kort ('hoge communiceerbaarheid'). Juist deze woorden lenen zich volgens Bianchi voor gebruik bij stigmatisering. Verder geeft de Sapir-Whorf hypothese volgens hem aanleiding te veronderstellen dat mensen als het ware gedwongen kunnen worden een bepaald onderscheid wel of juist niet te zien, door in een taal bepaalde woorden te introduceren of door er bepaalde woorden uit te verwijderen. Wanneer we er dus naar zouden streven bepaalde woorden met een hoge codabiliteit, waarvan de praktijk uitwijst dat ze stigmatiserend werken, te vervangen door neutrale

---

[26] Prof. Dr. H. Bianchi, *Stigmatisering*. Kluwer, Deventer 1971.

[27] In feite is sprake van twee afzonderlijke woordstammen ('sneeuw op de grond' en 'sneeuw in de lucht') die met andere elementen (morfemen) gecombineerd kunnen worden.

woorden met een lage codabiliteit dan zou dit er volgens Bianchi toe moeten leiden dat het verschijnsel stigmatisering wordt ingeperkt.

Tegen de mening van Bianchi is veel ingebracht. Een uitgebreide kritiek staat op naam van Van Reenen en Siertsema (1972). Zij wezen erop dat de Sapir-Whorf hypothese door taalkundigen in brede kring was verworpen, en gaven daarnaast ook een aantal voorbeelden waaruit bleek dat het principe van de taal-wereldbeeld-correlatie niet juist was. Met betrekking tot het ontbreken van een specifieke term om een bepaald begrip aan te duiden merkten zij bijvoorbeeld op dat het ontbreken van aparte woorden voor geel en oranje in bepaalde Indianentalen beslist niet hoeft te betekenen dat de gebruikers van die taal *in het geheel geen verschil zien* tussen die kleuren. Meerduidigheid van woorden is een zeer veel voorkomend verschijnsel; zinsverband, context en situatie waarin een woord wordt gebruikt geven het pas zijn eenduidige betekenis. Voorts merkten zij op dat woorden in feite *tekens* zijn die naar een bepaald begrip (of naar een bepaalde zaak) *verwijzen*. Eerst is er dus een begrip (of een zaak), dan wordt er (soms) een woord gemaakt dat het aanduidt. Dat betekent dat de taal nooit de *oorzaak* kan zijn van het ontbreken van een bepaald begrip bij een groep personen. En dat betekent weer dat de kans nihil is dat taal een sturingsinstrument kan zijn om maatschappelijk ongewenst gedrag te beïnvloeden.

Tenslotte plaatsten Van Reenen en Siertsema de nodige vraagtekens bij de door Bianchi genoemde relatie tussen codabiliteit, mate van stigmatiserend effect en woordlengte (korte woorden zouden vaak een hoog stigmatiserend effect hebben). De woorden die Bianchi zelf aanduidde als "stigmatiserende woorden met de hoogste codabiliteit" (zoals "misdadiger" en "t.b.r.-klant") waren in feite een stuk langer (4 lettergrepen) dan het éénlettergrepige woord "dief" dat hij als minder stigmatiserend beschouwde. Ook Van Reenen en Siertsema noemden nog enkele voorbeelden waaruit blijkt dat juist de langere woorden een hoger stigmatiserend effect hebben, waardoor de bewering van Bianchi in feite iedere grond lijkt te missen.

De belangstelling voor de rol van de Sapir-Whorf hypothese in het recht ligt inmiddels ver achter ons. Niet onvermeld kan echter blijven dat er recenter en veel belangrijker werk is waarin de invloed van taal wordt behandeld, niet op verschijnselen in de juridische praktijk, maar op de *juridische methodologie*. Hier moet het boek van Loth[28] genoemd worden, dat voor het eerst verscheen in 1984. Daarin wordt beschreven hoe taalanalytische methoden een rol kunnen spelen in de juridische methodologie, bij het oplossen van juridische problemen. Dit onderwerp is nauw verbonden met de toepassingen van de *argumentatieleer* in het recht.

1.3.3 Juridische argumentatieleer

Een laatste gebied dat genoemd kan worden waar sprake is van een succesvolle integratie van recht en taalkunde is dat van de argumentatietheorie. Het besef dat het voor (aankomende) juristen belangrijk is om op de juiste wijze een betoog te kunnen opbouwen en om begrijpelijke en correcte teksten te kunnen produceren is ondertussen in brede kring

[28] Loth (1991).

doorgedrongen. Dit heeft niet alleen geleid tot een intensivering van onderzoek en tot de publikatie van een aantal leerboeken op dit gebied (zoals bijvoorbeeld Van Eemeren *et al.* 1991), maar ook tot het ontstaan van verplichte- of keuzevakken aan verschillende Nederlandse universiteiten, waarin studenten ervaring kunnen opdoen in bijvoorbeeld het schrijven en houden van betogen en het vervaardigen van beleidsteksten.

HOOFDSTUK 2 - KWANTITATIEVE TAALKUNDE

2.1 Inleiding

2.1.1 Taalkunde

Zoals in het voorgaande hoofdstuk al werd aangegeven speelt taal een belangrijke rol in de rechtspraktijk en de rechtswetenschap. Gegeven dat feit is het juist voor juristen interessant om kennis te nemen van de wetenschap die zich in het algemeen met taaluitingen bezighoudt, de taalkunde of linguïstiek, ook wel genoemd de Algemene Taalwetenschap[1].

Binnen een bepaald segment van de Algemene Taalwetenschap speelt de wiskunde een belangrijke rol. Dit segment wordt in het algemeen aangeduid als "wiskundige taalkunde"[2]. Tot de wiskundige taalkunde worden gerekend:[3]

1. de *algebraïsche taalkunde*, die de formele middelen onderzoekt en ontwikkelt die nodig zijn voor het beschrijven van een taal. In een dergelijke beschrijving moet bijvoorbeeld de systematiek worden blootgelegd die in een taal ten grondslag ligt aan het samenstellen van zinnen uit losse woorden. Deze systematiek kan worden beschouwd als een systeem van regels die, om bij dit voorbeeld te blijven, alle manieren beschrijven om uit elk denkbaar rijtje woorden iedere grammaticaal correcte zin te vormen. De taalwetenschap kan daarbij gebruik maken van de formele systemen die in de wiskunde en de mathematische logica zijn ontwikkeld[4].

2. de *kwantitatieve taalkunde* of statistische linguïstiek, die zich bezighoudt met de kwantitatieve eigenschappen van natuurlijke talen, zoals bijvoorbeeld de frequentieverdeling van woorden en de omvang van de 'woordenschat' (waaronder in dit verband wordt verstaan het aantal verschillende woorden dat in een bepaald tekstgedeelte wordt gebruikt). Martin (1981) wijst erop dat onderzoek naar (relatieve) frequenties van taalkundige fenomenen in de praktijk veelal wordt verricht met het doel inzicht te krijgen in de *probabiliteit*, de waarschijnlijkheid waarmee deze fenomenen voorkomen. Hij definieert daarom de kwantitatieve taalkunde als "de studie van de taalkundige waarschijnlijkheidsstructuur"[5]. Kwantitatieve taalkunde speelt een belangrijke rol in dit onderzoeksproject.

3. de *computer-taalkunde*. Hierbij gaat het om (een vorm van) taalkunde, met als extra gereedschap de computer en de bijbehorende op taalkundige problemen gerichte

---

1 De Mulder (1984, p. 79).
2 vgl. Brandt Corstius (1974, p. 6).
3 Brandt Corstius (1978, p. 16); Dik en Kooij (1988, p. 67).
4 Dik en Kooij (1988, p. 67).
5 Martin (1981, p. 276).

> methoden en technieken. Gestreefd wordt niet alleen naar modellen en verzamelingen regels die expliciet en volledig zijn (zoals in gewone taalkunde), maar bovendien naar omzetting daarvan in een werkend (geautomatiseerd) systeem[6].

Er bestaat een direct verband tussen de algebraïsche taalkunde en de computer-taalkunde. Wanneer bijvoorbeeld een grammaticasysteem is gevonden dat voor natuurlijke talen geschikt is, kan een algoritme worden ontwikkeld waarmee zinnen kunnen worden ontleed of voortgebracht met behulp van een computer. Ook de kwantitatieve taalkunde en de computer-taalkunde hebben veel met elkaar te maken: door de grote aantallen waar we over praten - een flink tekstcorpus bevat al gauw meer dan 1.000.000 woorden; de corpora die in dit onderzoek zullen worden gebruikt bevatten in totaal ruim 56 miljoen woorden - is bij het verrichten van kwantitatief taalkundig onderzoek een computer in feite onontbeerlijk.

De algebraïsche taalkunde zal in dit proefschrift verder niet meer aan bod komen. Het grootste deel van het onderzoek in de volgende hoofdstukken kan worden gerekend tot de kwantitatieve taalkunde, maar aangezien vrijwel alle analyses met behulp van een computer zullen worden uitgevoerd - alleen al gezien de omvang van het materiaal is dat beslist noodzakelijk - is feitelijk ook sprake van computer-taalkunde.

2.1.2 Corpustaalkunde

Een benaming voor bepaalde taalkundige activiteiten die tegenwoordig steeds vaker wordt gebruikt is *corpustaalkunde*. Het betreft hier een vrij nieuwe term, die pas vanaf het einde van de jaren zeventig in zwang is. Brandt Corstius is waarschijnlijk in Nederland een van de eersten die aandacht besteedt aan deze discipline[7]. Meestal wordt met de term corpustaalkunde iets bedoeld als: taalkunde die gebaseerd is op waargenomen, historische taaluitingen, samengebracht in een of meer corpora. Aarts en Meijs (1988) geven een omschrijving van het doel van de corpustaalkunde:

> Zij wil, door de bestudering van specifieke talen en met name van het gebruik van die talen, komen tot een dieper inzicht in taal en taalgebruik in het algemeen. Zij doet dit door lopende teksten, in geschreven of gesproken vorm, te bestuderen. Daarmee wordt ook gezegd dat er teksten onderzocht moeten worden die geproduceerd zijn door individuele gebruikers, in een bepaalde sociale situatie, met een bepaald doel. Met andere woorden, de corpustaalkunde houdt zich tevens bezig - moet zich bezig houden - met alle aspecten van taalvariëteit, zowel de individuele als de maatschappelijke en regionale.
>
> Aarts en Meijs (1988, p. 8)

In deze omschrijving zijn verschillende elementen terug te vinden die ook een rol spelen bij het onderzoek naar het woordgebruik in juridische teksten waarover dit proefschrift gaat. Met name het feit dat ook taalvariëteit voorwerp van onderzoek is voor de corpustaalkunde is interessant. Het is juist de bedoeling om met dit onderzoek aan te tonen dat juridische teksten wat hun vorm betreft afwijken van 'algemene' Nederlandse teksten, een bepaald soort taalvariëteit dus (zie hiervoor verder paragraaf 2.5). Dat betekent dat dit

---

6 Neijt en Bakker (1988, p. 3).
7 Brandt Corstius (1978, p. 70).

onderzoek niet alleen kan worden gerekend tot de kwantitatieve taalkunde en de computer-taalkunde, maar ook tot de corpustaalkunde. Daarmee is tevens aangetoond dat deze disciplines elkaar op een aantal punten duidelijk overlappen. Het onderzoek op het gebied van de corpustaalkunde richt zich momenteel, in ieder geval in Nederland, met name op het ontwikkelen van een verfijnder instrumentarium voor verrijking, exploratie en exploitatie van computer-corpora en op het gebruik van dergelijke corpora. Dit alles voor specifiek taalkundige doeleinden, zoals het vervaardigen van woordenboeken en concordanties en het samenstellen en uittesten van formele grammatica's die de taal in een corpus zo goed mogelijk beschrijven[8]. Een en ander ligt grotendeels buiten de reikwijdte van dit onderzoeksproject, daarom zal aan de corpustaalkunde verder slechts zijdelings aandacht worden besteed.

2.1.3 Kwantitatieve taalkunde

In dit onderzoek gaat het er in de eerste plaats om een aantal kwantitatieve gegevens te verzamelen omtrent een tweetal juridische corpora. Vervolgens worden deze gegevens met elkaar vergeleken, en met kwantitatieve gegevens ontleend aan een derde corpus met 'algemeen Nederlandse' teksten. Aangezien dus met name kwantitatieve aspecten worden bestudeerd, kan het onderzoek het beste worden gekarakteriseerd als behorende tot de kwantitatieve taalkunde.

De kwantitatieve taalkunde kent al een vrij lange traditie. Zo werd bijvoorbeeld al in de vorige eeuw onder andere door Mendenhall onderzoek gedaan naar kwantitatieve aspecten van de gebruikte stijl in bepaalde teksten[9]. Vooral Herdan propageerde in de jaren '60 het belang van wat hij noemde de 'statistische taalkunde'[10]. Hij meende echter dat voor de taalkunde een heel speciaal soort statistiek nodig is, hetgeen volgens Brandt Corstius niet het geval is: de gewone statistiek die door andere takken van wetenschappen wordt gehanteerd is ook bruikbaar voor de taalkunde[11]. Met name vanaf de tweede helft van de zestiger jaren tot het einde van de jaren zeventig is de belangstelling voor kwantitatieve taalkunde gering geweest, waarschijnlijk vooral onder invloed van het werk van Chomsky. Deze pionier van de transformationeel-generatieve taaltheorie zorgde ervoor dat de aandacht van taalkundigen verschoof van het taal*gebruik* naar het taal*vermogen*, en dus van het bestuderen van bestaande taaluitingen - onder andere met behulp van statistische technieken - naar het ontwikkelen van structuren voor het beschrijven van taal en het zoeken naar een grammatica waarmee iedere mogelijke zin in een bepaalde taal kan worden geproduceerd. Ook Brandt Corstius schrijft in zijn boek Computer-taalkunde nog dat "op het ogenblik (...) alleen wat de structuur niet aan kan, aan de statistiek wordt gegund"[12].

---

8 Aarts en Meijs (1988, p. 12 e.v. en p. 22).
9 Zie Bailey (1969, p. 217).
10 Herdan (1960, 1966).
11 Brandt Corstius (1978, p. 38).
12 Brandt Corstius (1978, t.a.p.).

De tachtiger jaren gaven echter een duidelijke opleving te zien van de belangstelling voor kwantitatief taalkundig onderzoek. Het onderscheid tussen taalgebruik en taalvermogen was ook binnen de theoretisch georiënteerde taalkunde wat aan het vervagen, en de verwachting was intussen dat tekstcorpora en statistische gegevens die daaraan zijn ontleend ook een bijdrage konden leveren aan het verkrijgen van inzicht in de structuur van een natuurlijke taal. Het is in deze periode dat de term 'corpustaalkunde' is ontstaan (zie hierboven). Opvallend is dat in deze meest recente periode de belangstelling voor kwantitatieve analyse van stijl (bijvoorbeeld van een bepaalde auteur) veel geringer is dan in de jaren '50 en '60. In plaats daarvan worden nu onder andere kwantitatieve gegevens afkomstig uit tekstcorpora gebruikt in onderzoek naar *morfologische* aspecten, aspecten van de interne grammatische structuur van woorden. Voorbeelden hiervan, waarbij onder andere gebruik wordt gemaakt van woordfrequentie-gegevens, zijn te vinden in Baayen (1989) en in Van Reenen (1990).

In dit hoofdstuk zal een kort historisch overzicht worden gegeven van kwantitatief taalkundig onderzoek (paragraaf 2.2). Daarna zal aandacht worden besteed aan onderzoek op enkele deelterreinen binnen de kwantitatieve taalkunde die voor dit onderzoek van bijzonder belang zijn, zoals onderzoek naar statistische wetmatigheden, naar woordfrequenties en naar kwantitatieve aspecten van 'subtaal' (of vaktaal) (paragrafen 2.3, 2.4 en 2.5).

2.2 Historisch overzicht

Na de opmerkingen die in de vorige paragraaf zijn gemaakt over kwantitatief taalkundig onderzoek dat in het verleden heeft plaatsgevonden, zal in deze paragraaf een wat gedetailleerder overzicht van de geschiedenis van deze discipline worden gegeven. Een volledige beschrijving van alle projecten tot nu toe die elementen van kwantitatieve taalkunde bevatten is helaas niet te geven; daarvoor is het aantal projecten gewoonweg te groot, terwijl ze onderling ook te verschillend zijn om een behandeling per categorie mogelijk te maken. Daarom zal volstaan worden met die activiteiten uit het verleden die voor dit onderzoek het meest van belang zijn, en met die personen die het vakgebied de voornaamste impulsen hebben gegeven. Verder zal de 20e eeuw centraal staan, en zullen activiteiten die daarvóór hebben plaatsgevonden voor het grootste deel buiten beschouwing worden gelaten.

2.2.1 Woordtellingen

Door de jaren heen, en ook nu nog, is de belangrijkste taalstatistische toepassing gebleven het *tellen van taalelementen*. Frequentiegegevens bleken al aan het begin van deze eeuw ook van belang in bijvoorbeeld de psychologie. Zo stelden Thumb en Marbe al in 1901 vast dat proefpersonen sneller reageren op frequent voorkomende woordtypen dan op zeldzame woordtypen[13]. Ook binnen andere disciplines, zoals bijvoorbeeld de communica-

[13] Brandt Corstius (1978, p. 41).

tiewetenschappen en de informatica, is in de loop der jaren op uitgebreide schaal gebruik gemaakt van frequentiegegevens bij het zoeken naar technieken om gegevens sneller te versturen en zo compact mogelijk op te slaan (hierbij gaat het dan meestal meer om letter- dan om woordfrequenties).

Tellingen van taalelementen, en dan met name woordtellingen, vormen dus de belangrijkste toepassing binnen de kwantitatieve taalkunde. Brandt Corstius en Aarts en Meijs noemen bijvoorbeeld de volgende[14]:

- Voor het Duits werd al in 1898 door Keading een woordfrequentielijst vervaardigd. Veel recenter, en ook betrouwbaarder, is het werk van Meier (1967).
- Voor het Engels zijn belangrijke corpora waarvan frequentielijsten zijn samengesteld het corpus Amerikaans-Engels (het 'Brown Corpus') dat begin jaren '60 door Nelson Francis en Henry Kučera werd samengesteld en de Brits-Engelse tegenhanger daarvan, het Lancaster-Oslo-Bergen (LOB) Corpus dat eind jaren '70 tot stand kwam.
- Voor het Italiaans bestaan verschillende frequentielijsten, bijvoorbeeld van Juilland and Traversa (1973) en van Sciarone (1977).
- Voor het Frans en enkele andere Romaanse talen bestaan onder andere frequentielijsten waaraan eveneens de naam van Juilland verbonden is, zoals Juilland *et al.* (1970) (Frans), Juilland and Chang-Rodriguez (1964) (Spaans) en Juilland *et al.* (1965) (Roemeens).
- Ook voor het Russisch bestaat een frequentiewoordenboek: Štejnfel'dt (1963).

Voor het Nederlands kunnen worden genoemd:

- De eerste frequentielijst van een omvangrijk corpus (ongeveer 1 miljoen woordtokens) werd in 1935 samengesteld door De la Court (1937). Hij liet schoolkinderen lijsten maken van alle woorden uit een groot aantal verschillende teksten. Het doel was vooral om gegevens te verkrijgen ten behoeve van het onderwijs in Nederlands Indië. De telling is echter onvolledig, aangezien de typen met de hoogste frequentie (de 'ontelbaren') om praktische redenen werden overgeslagen. Toch werd het werk in 1963 bewerkt en opnieuw uitgegeven door Linschoten (1963), bij gebrek aan goede alternatieven.
- In de jaren zestig werden diverse frequentielijsten van corpora met een beperkte omvang vervaardigd, onder andere door Van Berckel *et al.* (1963), Martin (1968 en 1970) en De Vriendt - De Man (1971).
- De voornaamste gepubliceerde frequentietelling voor geschreven Nederlands is op dit moment nog steeds het rapport Woordfrequenties van de Werkgroep Frequentieonderzoek van het Nederlands uit 1975, onder redactie van P.C. uit den Boogaart (1975). De teksten die aan deze telling ten grondslag liggen worden tegenwoordig meestal aangeduid als het 'Eindhovens corpus'. Het gaat daarbij om teksten met een totale lengte van ongeveer 600.000 tokens, zorgvuldig geselecteerd uit een vijftal bronnen: dagbladen, opiniebladen, gezinsbladen, literatuur en populair-wetenschappelijk proza. Daarnaast bevat het corpus een onderdeel 'spreektaal' van in totaal 120.000 tokens. Aangezien de woordenlijsten van dit corpus zijn verrijkt met een codering op woordniveau, waarin de grammaticale categorie van ieder type is

[14] Brandt Corstius (1978, p. 65 e.v.); Aarts en Meijs (1988, p. 7).

vastgelegd, is het geschikt om te worden gebruikt voor allerlei taalkundig vervolgonderzoek.

- Het meest omvangrijke Nederlandstalige corpus dat tot nu toe werd vervaardigd - en waaraan nog steeds wordt gewerkt - is dat van het Instituut voor Nederlandse Lexicologie in Leiden, dat het basismateriaal heeft geleverd voor het Woordenboek der Nederlandse taal dat door dit instituut wordt vervaardigd. Er wordt naar gestreefd dit corpus in de toekomst uit te breiden tot een omvang van 100 miljoen woordtokens[15]. Het corpus algemeen Nederlandse teksten dat in dit onderzoek als vergelijkingsmateriaal voor de juridische teksten wordt gebruikt is een speciaal voor dit onderzoek geselecteerd gedeelte van dit zeer omvangrijke INL-corpus.

2.2.2 Statistiek en stijl

Een deelgebied binnen de kwantitatieve taalkunde waar eveneens sprake is van een lange historie kan worden aangeduid als 'Statistiek en stijl': het onderzoeken en vergelijken van schrijfstijlen van auteurs, en eventueel van veranderingen daarin door de jaren heen, aan de hand van statistische gegevens betreffende een of meer teksten. Hierbij gaat het niet alleen om woordfrequentie-gegevens, maar bijvoorbeeld ook om gemiddelde woord- en zinslengte, interpunctie en het gebruik van 'vreemde' woorden (afkomstig uit een andere taal).

Veel onderzoek op het gebied van statistiek en stijl is verricht met het doel om van bepaalde teksten, waarvan de herkomst wel werd vermoed maar niet met zekerheid was na te gaan, vast te stellen wie de meest waarschijnlijke auteur was. Aanvankelijk werd hierbij met name aandacht besteed aan de woord*lengte*, zoals bijvoorbeeld in het onderzoek van T.C. Mendenhall, een Amerikaans geofysicus. Deze publiceerde in 1887 een artikel in Science[16] waarin hij de schrijfstijlen van een groot aantal auteurs probeerde te karakteriseren met behulp van een soort 'woordlengte-spectrogrammen', grafieken gebaseerd op de relatieve frequenties van woorden met een bepaalde lengte[17].

Gemiddelde woordlengtes noch woordlengte-patronen bleken echter voldoende om de stijlen van diverse auteurs met honderd procent zekerheid van elkaar te kunnen onderscheiden, zelfs niet wanneer deze gegevens werden gecombineerd met cijfers omtrent de gemiddelde zinslengte. Vandaar dat steeds meer onderzoekers aandacht gingen besteden aan het *woordgebruik* in de tekst, en nagingen welke woorden in teksten van een bepaalde auteur voorkwamen, en met welke frequenties. Aan de hand van deze gegevens is door de jaren heen door menig onderzoeker getracht de omvang van de *woordenschat* van een bepaalde auteur vast te stellen. Voorbeelden hiervan zijn te vinden in Thomson and Thompson (1915), Carroll (1938), Yule (1944), Bennett (1969), Martin (1970) en Martin (1973). In veel gevallen werd daarbij gebruik gemaakt van een bepaalde maat, waarin de omvang van de woordenschat kon worden uitgedrukt. Op enkele van deze maatgetallen zal in een volgende paragraaf worden teruggekomen.

---

15 Aarts en Meijs (1988, p. 22).
16 Mendenhall (1887).
17 Bailey (1969, p. 217).

Pogingen die in de afgelopen decennia werden ondernomen om het verschijnsel stijl met behulp van kwantitatieve, aan teksten ontleende gegevens zo nauwkeurig mogelijk te specificeren en op die manier te ontraadselen waren niet altijd even succesvol. Dit onderzoek moet worden geplaatst tegen de achtergrond van de ontdekking, in de loop der jaren, van enkele (veelal statistische) 'wetmatigheden' die ten grondslag liggen aan het woordgebruik in iedere tekst (of aan dat in bepaalde soorten teksten). De invloed van deze 'algemene' wetmatigheden moet natuurlijk zoveel mogelijk worden weggefilterd bij het onderzoek naar individuele stijlen. Met name in de beginjaren dat onderzoek op het gebied van statistiek en stijl werd verricht hebben lang niet alle onderzoekers zich daarvan rekenschap gegeven, en dit verklaart wellicht het geringe succes dat in die tijd werd geboekt. Pas in de tweede helft van deze eeuw is het besef ontstaan dat alleen met behulp van een volledig gespecificeerde theorie met betrekking tot stijl de oorzaken kunnen worden blootgelegd van de variaties in taalgebruik in literaire en andere teksten.

Hoewel een dergelijke theorie met betrekking tot stijl nog door niemand is geformuleerd, kunnen toch enkele projecten worden genoemd waarin door een zeer specifieke aanpak duidelijke successen zijn geboekt. In de eerste plaats is daar bijvoorbeeld het werk van Milic (1967), en daarop gebaseerd vervolgonderzoek zoals dat van Cherubini (1975) en Cluett (1976). Milic bracht in een aantal teksten van Swift en enkele van diens tijdgenoten met de hand codes voor de syntactische categorieën van woorden aan, en gebruikte vervolgens een computer om de percentages uit te rekenen van de diverse woordklassen en van zinnen die met elk van die woordklassen begonnen, en daarnaast om de verdeling vast te stellen van diverse patronen van groepen van drie woorden. Op die manier stelde hij stilistische kenmerken van het werk van Swift vast[18].

Een tweede voorbeeld van dergelijk onderzoek vormt het werk van Ellegård (1962a en b) en dat van Mosteller and Wallace (1964). In beide gevallen stelden de onderzoekers vast dat de relatieve frequentie van sommige woorden in de teksten van een bepaalde auteur duidelijk hoger lagen dan in die van anderen, ook wanneer dat tijdgenoten van de auteur waren. Zulke woorden werden door de onderzoekers beschouwd als de 'vingerafdruk' van de auteur in kwestie[19]. Deze techniek is daarna nog door een aantal anderen toegepast, in de meeste gevallen eveneens om na te gaan wie de meest waarschijnlijke auteur van bepaalde teksten was. Voorbeelden hiervan zijn Austin (1970) en Kenny (1977, 1978).

2.2.3 Ander kwantitatief taalkundig onderzoek

De kwantitatieve taalkunde kent nog een groot aantal toepassingen. Dikwijls gaat het daarbij om het verzamelen van cijfermateriaal en het ordenen van gegevens waarmee anderen verder kunnen werken. Brandt Corstius (1978) noemt in dit kader bijvoorbeeld het vervaardigen van indexen en concordanties, die onder andere van groot nut kunnen zijn om de vindplaats in de tekst op te kunnen zoeken van woordtypen die in een frequentielijst worden aangetroffen. Index en frequentielijst hebben veel met elkaar gemeen. Het zijn allebei lijsten van woordtypen, in het ene geval voorzien van de vindplaats(en) (bladzijden, regels) waar het type in te tekst voorkomt, en in het andere van de

[18] Bailey (1969, p. 221).
[19] Butler (1985, p. 22).

frequentie van het type. Bij het vervaardigen van een frequentielijst kan meestal heel eenvoudig ook meteen een index worden gegenereerd (die daarna alleen nog alfabetisch moet worden gesorteerd), en andersom. Het vervaardigen van dit soort lijsten met behulp van een computer was tot het begin van de jaren zestig ook zo'n beetje de enige activiteit waarmee die computer taal- of letterkundigen die grote tekstcorpora wilden bestuderen van dienst kon zijn[20].

Op dit moment is dat zeker niet meer het geval: de sterk gedaalde prijzen van computergeheugens en de hoge verwerkingssnelheid, zelfs van relatief eenvoudige microcomputers, hebben ervoor gezorgd dat het tegenwoordig mogelijk is omvangrijke corpora op te slaan en interactief te raadplegen, indien gewenst zelfs lokaal, op de eigen Personal Computer. Het ontwikkelen van methoden en technieken die hierbij gebruikt kunnen worden behoort tot het terrein van de corpus-taalkunde, die al werd beschreven in paragraaf 2.1.2.

De afgelopen decennia is ook regelmatig kwantitatief taalkundig onderzoek verricht ten dienste van andere disciplines binnen de taalkunde. Voorbeelden hiervan zijn:

- het gebruik van woordfrequentiegegevens voor het ontwerpen van een zogenaamde probabilistische grammatica, een grammatica waarbij iedere herschrijfregel voorzien is van een 'gewicht' dat aangeeft in hoeveel procent van de gevallen dat de regel *kan* worden toegepast deze ook daadwerkelijk *moet* worden toegepast[21]. Op die manier kan een grammatica worden vervaardigd die nauwkeuriger overeenkomt met het menselijk taalgedrag.
- het gebruik van tekstcorpora en de daaruit afgeleide woordfrequentieverdelingen voor het onderzoeken van de morfologische produktiviteit van bepaalde categorieën van woorden, dat wil zeggen van de mate waarin in de praktijk nieuwe woorden ('taalbouwsels') kunnen ontstaan door gebruik te maken van een bepaalde woordstructuur, bijvoorbeeld het gebruiken van achtervoegsels als -heid of -te[22].

De kwantitatieve taalkunde levert hierbij de nodige statistische technieken. De uitkomsten van dergelijke onderzoeksprojecten hebben soms ook een duidelijk kwantitatief karakter (dat is bijvoorbeeld het geval bij het hierboven aangehaalde onderzoek naar morfologische produktiviteit), maar dat hoeft niet het geval te zijn.

Als laatste punt in dit beknopte historisch overzicht van de kwantitatieve taalkunde kan nog worden genoemd de betrokkenheid van de kwantitatieve taalkunde bij pogingen om te komen tot geautomatiseerde vertaalsystemen. Dat deze pogingen - tot nu toe - nog weinig succes hebben gesorteerd mag wel als bekend worden verondersteld, en het is eigenlijk een wonder dat er nog steeds onderzoekers zijn die de moed om tot een goed werkend systeem te komen nog niet definitief hebben opgegeven. Het is echter wel interessant om te weten dat bij wat er op dit moment nog aan 'vertalen per computer' gedaan wordt, er in veel gevallen gebruik wordt gemaakt van statistische technieken. Onlangs verschenen bijvoorbeeld berichten in de krant dat enkele belangrijke computerfabrikanten, waaronder IBM, momenteel werken aan systemen die gebruik maken van

---

[20] Brandt Corstius (1978, p. 70).
[21] Zie Brandt Corstius (1978, p. 75-79).
[22] Zie Baayen (1989).

materiaal dat tot nu toe is blijven liggen, namelijk *alle bestaande vertalingen*. De omvangrijke vertaalbestanden die aan de hand van dat materiaal worden samengesteld bevatten gegevens waaruit voor iedere te vertalen zin de statistisch meest waarschijnlijke vertaling wordt bepaald. Waarschijnlijk is het niet zo dat alle problemen met het automatisch vertalen met behulp van de statistiek kunnen worden opgelost, maar het lijkt niet onwaarschijnlijk dat een dergelijk systeem dat werkt met 'meest waarschijnlijke vertalingen' in de toekomst tot heel redelijke resultaten zal komen.

2.3 Statistische wetmatigheden en maatgetallen betreffende het woordgebruik

In de loop der jaren hebben verschillende beoefenaren van de kwantitatieve taalkunde pogingen gedaan om statistische wetmatigheden te formuleren waarmee bepaalde waargenomen verschijnselen in tekstmateriaal kunnen worden verklaard. Een van de bekendste resultaten op dit gebied is zonder twijfel de 'Wet van Zipf', waarin door G.K. Zipf een verband werd gelegd tussen rangnummer en frequentie van een woord[23]. Verder besteedde Zipf aandacht aan de omgekeerd evenredige relatie tussen woordlengte en frequentie[24]. Hij was trouwens niet de eerste die over deze onderwerpen publiceerde: al in 1916 werd door J.B. Estoup aandacht besteed aan de relatie woordlengte-frequentie, en E.U. Condon beschreef in 1928 de constante verhouding tussen rangnummer en frequentie[25]. De naam van Zipf is met genoemde relaties echter het sterkst verbonden gebleven. De door hem geformuleerde wet over de verhouding rangnummer-frequentie kan als volgt worden omschreven: *In natuurlijke talen geven woordtypen uit een omvangrijk tekstcorpus, gerangschikt naar aflopende frequentie, een constante daling in frequentie te zien.* Met andere woorden, wanneer aan alle woordtypen rangnummers worden toegekend, nummer 1 aan het meest voorkomende type, nummer 2 aan het type met de op een na hoogste frequentie, enzovoort, en deze rangnummers worden in een grafiek afgezet tegen de frequentie onder gebruikmaking van een logaritmische schaalverdeling op beide assen, dan is het resultaat een dalende rechte lijn. In formulevorm komt dit neer op

$$f_i \,.\, i = K \qquad (2\text{-}1)$$

ofwel: de frequentie f_i van woorden uit een corpus vermenigvuldigd met het rangnummer i van die woorden levert een (min of meer) constante waarde op. Alleen voor de typen met de hoogste frequentie gaat dit verband meestal niet helemaal op; daarbij is volgens Zipf meestal sprake van een verhoudingsgewijs wat lagere frequentie.

Naast deze wet met betrekking tot frequentie en rangnummer beschreef Zipf zoals gezegd de omgekeerd evenredige relatie tussen de woordlengte en de frequentie van woorden in een tekst - in feite een tweede statistische 'wet'. In een bepaalde steekproef van Engelse

---

[23] Zipf (1935, p. 39-48). Enkele van de hier aangehaalde passages uit het werk van Zipf worden ook besproken door Baayen (1989, p. 125 e.v.).

[24] Zipf (1935, p. 20-39).

[25] Rapoport (1982, p. 5-6).

teksten uit Amerikaanse kranten telde hij 2976 *hapax legomena*[26], die gemiddeld uit 6,66 fonemen (taalklanken) bestonden. Woorden die 5 keer voorkwamen in het corpus bleken al minder lang te zijn: gemiddeld nog maar 5,59 fonemen. Woorden die 10 keer voorkwamen telden gemiddeld 5,38 fonemen, woorden die 20 keer voorkwamen 4,69, enzovoort, waarbij woorden die meer dan 60 keer voorkwamen gemiddeld uit slechts 2,66 fonemen bestonden[27]. Zipf veronderstelde dat hier sprake is van een causaal verband, in die zin dat woorden die vaak worden gebruikt automatisch steeds verder worden afgekort. Hij noemde dit aanvankelijk de 'wet van het afkorten'. Later classificeerde hij het verschijnsel als een van zijn 'principes van de minste inspanning' (Zipf 1949). Zipf meende dat deze wet (c.q. dit principe) in een taal het volgende effect oplevert:

> In view of the evidence of the stream of speech we may say that the length of a word tends to bear an inverse relationship to its relative frequency; and in view of the influence of high frequency on the shortenings from truncation and from durable and temporary abbreviatory substitution, it seems a plausible deduction that, as the relative frequency of a word increases, it tends to diminish in magnitude. This tendency of a decreasing magnitude to result from an increase in relative frequency, may be tentatively named the Law of Abbreviation.
>
> Zipf (1935, p. 38)

De door Zipf beschreven wetmatigheden zullen waarschijnlijk nooit met 100% nauwkeurigheid het woordpatroon in een bepaald corpus beschrijven. Hij ging er echter zelf vanuit dat met name een beschrijving van de punten waar het woordgebruik in een tekst *afwijkt* van het door hem beschreven patroon samen een soort karakteristiek van het taalgebruik in die tekst kunnen opleveren.

Door veel taalkundigen is intussen met name onderzoek gedaan naar de mogelijkheden die de door Zipf beschreven principes bieden voor het voorspellen van de frequentie*verdeling* van woorden in een tekst. Een woordfrequentieverdeling van een tekst is een lijst getallen die aangeven *hoeveel typen in de tekst voorkomen met een bepaalde frequentie*. De lijst is gewoonlijk geordend naar frequentie, de laagste frequentie (meestal 1) eerst. Er ontstaat dan een lijst van frequentie*klassen*, waarbij het rangnummer *r* van een klasse gelijk is aan de frequentie van de woorden in die klasse.

r	n_r	r	n_r	r	n_r	r	n_r	r	n_r	r	n_r
1	496	10	10	19	3	30	1	43	1	73	1
2	175	11	5	20	2	31	2	45	1	80	1
3	89	12	9	21	2	33	1	46	1	95	1
4	60	13	5	22	3	34	1	47	2	109	1
5	30	14	3	23	2	35	1	50	1	189	1
6	24	15	3	24	2	37	1	57	1	229	1
7	15	16	2	26	1	40	1	59	1	368	1
8	13	17	5	28	1	41	1	61	1		
9	9	18	1	29	1	42	1	65	1		

Tabel 2-1 Voorbeeld van een frequentieverdeling

[26] Grieks: 'eenmaal gezegd'. Woordtypen die in een bepaald tekstcorpus slechts één keer voorkomen.
[27] Zipf (1935, p. 28).

Een voorbeeld van een woordfrequentieverdeling, te weten die van het eerste document uit het in dit onderzoeksproject gebruikte corpus jurisprudentieteksten (zie het volgende hoofdstuk), is te vinden in tabel 2-1.

Zipf meende dat frequentieverdelingen kunnen worden beschreven met behulp van de volgende formule[28]:

$$n_r = \frac{K}{r^2} \qquad (2\text{-}2)$$

ofwel, het aantal woordtypen *n* behorend tot een bepaalde frequentieklasse *r* zou omgekeerd evenredig zijn met het kwadraat van de frequentie. Zipf merkte daarbij wel op dat een en ander helaas niet opgaat voor de meest voorkomende woordtypen; die hadden een lagere frequentie dan uit de wet volgt. Het ging daarbij ongeveer om de 2 tot 3% meest voorkomende woordtypen, die echter toch wel in totaal zo'n 40 tot 60% van alle woord-*tokens* uitmaakten. Onderzoekers na hem, zoals Gustav Herdan (1964, 1966) en Charles Muller (1969, 1977, 1979) hebben er echter op gewezen dat de frequentieverdeling van de meeste teksten nogal sterk afwijkt van dit model, dat daarom meestal niet meer dan een grove benadering van de echte frequentieverdeling kan worden genoemd. In de literatuur zijn diverse alternatieve modellen voorgesteld, onder andere door Mandelbrot (1962), die zich baseerde op Zipf's verdeling, en door Herdan en Muller[29] die gebruik maakten van een wiskundige reeks beschreven door Waring. Het zou te ver voeren om hierop in detail in te gaan, maar vermeld kan wel worden dat het resultaat van laatstgenoemd onderzoek, de zogenaamde Waring-Herdan-Muller verdeling, in de meeste gevallen aanzienlijk beter overeen blijkt te komen met de werkelijke frequentieverdeling van tekstcorpora dan het hierboven weergegeven model van Zipf, hoewel intussen is aangetoond[30] dat bij deze verdeling (net als bij Zipf's model en dat van Mandelbrot) de mate waarin het model de juiste waarden oplevert vrij sterk afhangt van de steekproefgrootte. Dit duidt erop dat al deze modellen in feite slechts bij één specifieke steekproefgrootte de meest optimale waarden zullen opleveren, hetgeen hun praktische toepasbaarheid beperkt.

Van Naelten (1991) heeft er, weliswaar in iets ander verband, op gewezen dat de 'wet van Zipf' in feite een fundamentele onjuistheid bevat. Volgens hem is het niet meer dan logisch dat tussen de twee grootheden woordfrequentie en rangnummer een sterke correlatie, en mogelijk in bepaalde gevallen een constant verband bestaat, aangezien deze twee grootheden allerminst onafhankelijk zijn. Het rangnummer is immers voor honderd procent geconstrueerd uit de woordfrequentie. Beide grootheden representeren in feite dezelfde frequentiegegevens, welke in het geval van de woordfrequentie op rationiveau worden gemeten, en in het geval van het rangnummer op het veel lagere ordinale niveau[31]. Het door Zipf waargenomen verband is daaruit volgens hem in ieder geval voor een deel te verklaren, zodat de conclusie gerechtvaardigd lijkt dat terughoudendheid geboden is bij het toepassen van deze 'wet'.

---

28 Zipf (1935, p. 41).
29 Herdan (1964, p. 85-88); Muller (1969: 42-56; 1977: 110-114; 1979: 399-425).
30 Zie o.a. Baayen (1989, p 128 e.v.).
31 Van Naelten (1991, p. 14).

Een verschijnsel dat vooral door onderzoekers is bestudeerd in verband met zijn consequenties voor de schrijfstijl in een tekst (zie ook de voorgaande paragraaf) is de *woordenschat*, het aantal woordtypen dat in een bepaald tekstgedeelte wordt gebruikt eventueel gecombineerd met gegevens over de herhalingsfrequentie van elk type. In een aantal gevallen heeft dit geresulteerd in een min of meer objectief 'maatgetal' voor de omvang van de woordenschat. Een van de bekendste voorbeelden daarvan is de door Yule (1944) geïntroduceerde karakteristiek *K*, die in feite een maat is voor de herhalingsfrequentie van woordtypen. Volgens Yule is K^{-1} een goede maat voor de omvang van de woordenschat. De definitie van *K* luidt als volgt: laat $S_1 = \sum r.n_r$ (de som van de produkten van alle rangnummers van frequentieklassen en het aantal typen in die klasse, ervan uitgaande dat het rangnummer van een klasse gelijk is aan de frequentie in het corpus van de typen die erin vertegenwoordigd zijn) en $S_2 = \sum r^2.n_r$ (de som van de produkten van de *kwadraten* van alle rangnummers en het aantal typen met dat rangnummer). Dan is

$$K = 10.000 \,.\, \frac{S_2 - S_1}{S_1^{\,2}} \qquad (2\text{-}3)$$

K blijkt inderdaad een redelijk stabiele karakteristiek in de zin dat de waarde niet of nauwelijks afhangt van de grootte van het corpus (mits dat een zekere minimum omvang heeft). Een nadeel is echter wel dat de waarde van *K* alleen goed bepaald kan worden door steekproeven van flinke omvang te nemen; de waarde wordt pas echt betrouwbaar, zoals in een volgend hoofdstuk nog zal blijken, wanneer de steekproef enkele miljoenen woordtokens groot is. Verder is de karakteristiek wel geschikt voor het meten van de gemiddelde herhalingsfrequentie van typen - hoe lager de herhalingsfrequentie, en dus hoe groter het aantal woorden dat weinig voorkomt in de tekst, des te lager wordt de waarde van *K* - maar veel minder goed voor het meten van de grootte van de woordenschat, aangezien zij vooral wordt bepaald door de typen met hoge frequentie; het rangnummer, gelijk aan de frequentie, wordt immers gekwadrateerd in de formule. Het is echter te verwachten dat teksten waarin sprake is van een omvangrijke woordenschat juist meer woorden met een lage frequentie zullen tellen[32]. Tenslotte wordt het door sommigen (onder andere door Herdan) als een nadeel gezien dat Yule er bij de afleiding van *K* vanuit ging dat het voorkomen van woorden in een tekst wordt beheerst door bepaalde statistische wetten (er zou sprake zijn van een Poisson-verdeling).

Herdan besteedde later in zijn "Advanced Theory of Language as Choice and Chance" (Herdan 1966) uitgebreid aandacht aan de karakteristiek *K* [33]. Hij leidde *K* op een wat andere manier af dan Yule. Aangezien het voor zijn methode niet nodig was te veronderstellen dat woorden in een tekst voorkomen volgens een Poisson verdeling, werden de gebruiksmogelijkheden van 'The Characteristic' volgens Herdan vergroot. Wanneer wordt uitgegaan van omvangrijke steekproeven, waarbij het resultaat van de breuk $1/S_1$ verwaarloosbaar klein wordt, en wordt afgezien van de vermenigvuldigingsfactor van 10.000, kan *K* volgens Herdan als volgt worden gedefinieerd:

$$K = \frac{S_2}{S_1^{\,2}} = \frac{v_r^{\,2}}{N} \qquad (2\text{-}4)$$

waarbij v_r staat voor de variatie-coëfficiënt van *r*, gelijk aan $\sigma_r / \bar{r}$, de standaarddeviatie-

---

[32] Baayen (1989, p. 75). Zie ook Herdan (1966, p 101-107).

[33] Herdan (1966, p. 101 e.v.).

gedeeld door het gewogen gemiddelde van de in het corpus voorkomende frequenties. Hij gebruikt voor deze 'nieuwe' *K* verder het symbool ν_m, en geeft aan dat de karakteristiek in feite een maat is voor de relatieve fluctuatie van het gewogen gemiddelde van alle woordfrequenties bij veranderende steekproefgrootte. In die zin is de karakteristiek geschikt om verschillen tussen schrijfstijlen tot uitdrukking te brengen. Bij een stijl waarbij het gewogen gemiddelde weinig verandering te zien geeft bij een wisselende steekproefgrootte zijn woordfrequenties die sterk afwijken (positief of negatief) van het gemiddelde relatief zeldzaam, terwijl ingeval van grotere schommelingen in het gewogen gemiddelde dergelijke afwijkende woordfrequenties juist sterker vertegenwoordigd zullen zijn. Herdan meent dat deze 'mate van fluctuatie' van het gewogen gemiddelde van alle woordfrequenties, die dus in feite aangeeft hoe sterk de woordfrequenties van elkaar afwijken, kan worden gebruikt om een schrijfstijl te karakteriseren[34]. Van de karakteristiek *K* zal gebruik worden gemaakt bij het vergelijken van de structuur van het woordgebruik in de diverse corpora.

Een tweede karakteristiek die beoogt de omvang van de woordenschat in een tekst weer te geven is de zogenaamde karakteristiek *R*, geïntroduceerd door Guiraud (1954, 1959). Deze wordt door hem als volgt gedefinieerd:

$$R = \frac{V}{\sqrt{N}} \qquad (2\text{-}5)$$

waarbij *V* (Vocabulaire) het aantal woordtypen, en *N* het aantal woordtokens in de steekproef weergeeft. Guiraud gaf wel aan dat *R* enigszins aan fluctuatie onderhevig leek te zijn, en stelde proefondervindelijk vast dat de karakteristiek het betrouwbaarst was voor teksten met een lengte van zo'n 10.000 tot 50.000 woorden[35]. Hij meende dat, gegeven die begrenzing, fluctuaties in *R* van weinig belang waren, zodat sprake was van een stabiel statistisch maatgetal. Opgemerkt moet worden dat de drie corpora uit dit onderzoek, en ook steekproeven daaruit, in het algemeen aanzienlijk omvangrijker zijn dan de door Guiraud genoemde optimale waarden. Zoals hierna zal blijken heeft dit geen merkbaar effect op de wijze waarop de waarde van *R* zich ontwikkelt bij toenemende steekproefgrootte.

Baayen (1989, p. 81-83) bespreekt de karakteristiek van Guiraud, en haalt daarbij onderzoek aan van Orlov[36], waaruit zou blijken dat de waarde van *R* bij toenemende steekproefgrootte eerst toeneemt, dan een maximum bereikt en vanaf dat punt voortdurend afneemt. Bij nadere beschouwing blijkt echter dat het door Orlov besproken gedrag van *R* alleen geldt binnen het raamwerk van diens theorie over de wet van Zipf (en de daarop door Mandelbrot (1962) gemaakte aanvullingen). Op deze theorie kan hier niet al te diep worden ingegaan, maar de essentie ervan is de stelling dat de wet van Zipf slechts geldt bij één bepaalde steekproefgrootte uit een bepaald corpus. Orlov noemt deze grootte *Z*, de Zipf-omvang. Alleen bij deze specifieke steekproefgrootte zou de frequentieverdeling van een tekst voldoen aan de door Zipf voorgestelde formule, terwijl ook de karakteristiek van Guiraud alleen met behulp van het aantal woordtypen dat een steekproef van de Zipf-omvang bevat zou kunnen worden berekend. Voor het berekenen van *R* bij een

---

34 Herdan (1966, p. 103).
35 Guiraud (1959, p. 89).
36 Orlov (1982, p. 215).

steekproef met grootte *N*, welke afwijkt van *Z*, dient volgens Orlov niet zonder meer het aantal woordtypen in die steekproef in de formule te worden ingevuld, maar dient dit aantal te worden afgeleid uit het aantal typen bij steekproefgrootte *Z*. De door hem gehanteerde formule voor de berekening van *R* luidt dan:

$$R(N) = \frac{V(Z) \cdot \ln \frac{X}{X-1}}{\sqrt{N}} \qquad (2\text{-}6)$$

waarbij *V*(*Z*) het aantal woordtypen weergeeft bij steekproefgrootte *Z*, en $X = Z / N$ (de grootte van *Z* in relatie tot die van *N*)[37]. Dit levert inderdaad een kromme op, die aanvankelijk stijgt, zijn hoogste punt bereikt bij $N = Z$ en daarna weer daalt. De 'gewone' curve van *R*, berekend volgens (2-5) bij steeds toenemende *N* levert echter in de meeste gevallen een lijn op die over het hele bereik een stijgend verloop heeft; zie ook figuur 2-1, waarin de curves van *R* voor de drie corpora uit dit onderzoek zijn afgebeeld. Deze lijn zal die van Orlov snijden, en wel precies waar deze zijn hoogste punt bereikt, dat wil zeggen daar waar $N = Z$.

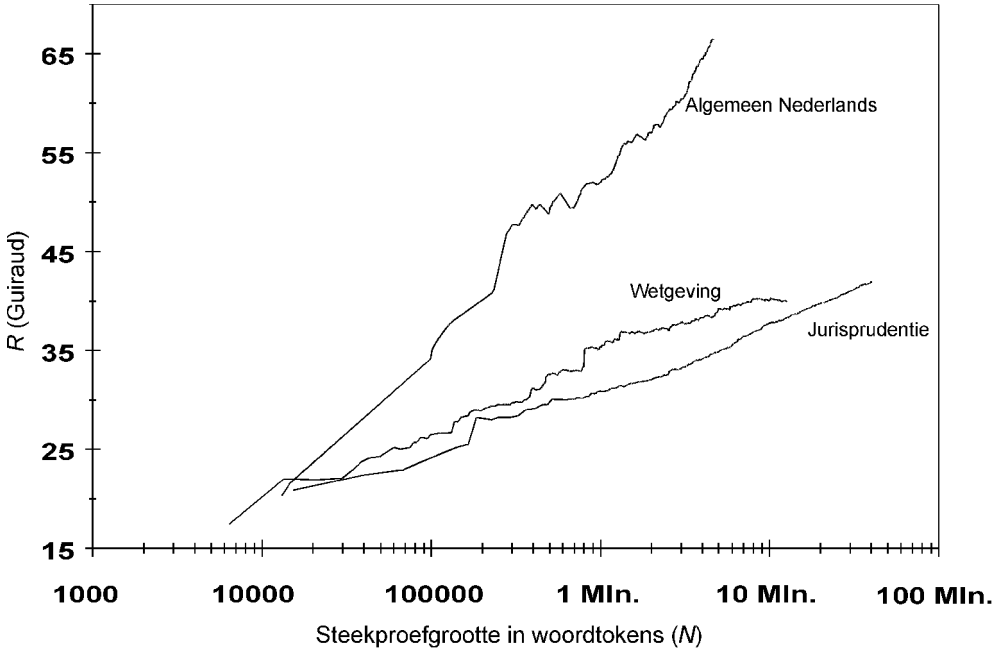
Baayen gaat er, in afwijking van het hiervoor gestelde, van uit dat ook een op de 'normale' manier berekende *R* een verloop zal kennen zoals dat door Orlov wordt geschetst, welke veronderstelling steun lijkt te vinden in de door hem bestudeerde data, maar niet in de drie corpora uit dit onderzoek. Hij geeft aan dat het feit dat *R* een maximale waarde heeft kan worden gebruikt om onderscheid te maken tussen klassen van woorden met een hoge en met een lage morfologische produktiviteit. In het door hem gepresenteerde onderzoek is er namelijk voor klassen met een hoge morfologische produktiviteit veel minder verschil tussen de maximale *R* en de (in zijn geval dus meestal lagere) *R* van de gehele tekst dan voor klassen met een lage morfologische produktiviteit[38]. Hoewel dus enige twijfel rijst met betrekking tot de basis waarop deze bevindingen zijn gegrond, kan natuurlijk niet worden ontkend dat de gevonden resultaten interessant zijn, en nadere studie door taalkundigen verdienen.

Het verloop van Guiraud's *R* voor de drie corpora uit dit onderzoek wordt weergegeven door in grafiek 2-1. Wat in de grafiek opvalt is met name dat de curve van het corpus Algemeen Nederlands een duidelijk afwijkend verloop heeft. Na aanvankelijk op ongeveer hetzelfde niveau te zijn begonnen loopt de waarde van *R* al snel op tot een hogere waarde dan de twee andere curves, zelfs bij de hoogste waarden van *N*, bereiken. Dit lijkt te wijzen op het gebruik van een groter aantal woordtypen, ofwel op een grotere woordenschat, in de documenten van dit corpus. In de volgende hoofdstukken zal worden nagegaan of dit het geval is. Mocht dat inderdaad zo zijn, dan zou dat in ieder geval een indicatie zijn dat Guiraud's *R*, hoewel misschien niet de meest stabiele karakteristiek, toch een zinvolle relatie legt tussen het aantal woordtypen en woordtokens in een tekst.

---

[37] Orlov hanteert een iets andere berekening wanneer *N* en *Z* praktisch dezelfde waarde hebben; wanneer $X = 1$ is het resultaat van de hier genoemde formule immers ongedefinieerd. Zie Orlov (1982, p. 155).

[38] Baayen (1989, p. 81-83).



Afb. 2-1 Guiraud's *R* berekend voor de drie gebruikte corpora, waarbij telkens (willekeurige) documenten worden toegevoegd.

Ook Herdan (1964,1966) besteedde aandacht aan het verband tussen het aantal woordtokens en het aantal woordtypen in een tekst. Hij gaf aan dat wanneer uit een tekst steekproeven van steeds toenemende omvang worden genomen, de type-token ratio (het aantal woordtypen in de steekproef gedeeld door het aantal woordtokens dat deze telt) weliswaar steeds toeneemt, maar dat dit gebeurt volgens een constant patroon. Wanneer namelijk het aantal typen wordt afgezet tegen het aantal tokens in een grafiek met logaritmische schaalverdeling, dan ontstaat bij toenemende steekproefgrootte een vrijwel rechte lijn door de oorsprong. Dit verband kan in formulevorm als volgt worden weergegeven[39]:

$$\log V = C \log N \qquad (2\text{-}7)$$

waarbij V wederom staat voor het aantal typen, en N voor het aantal tokens in de steekproef. In de praktijk blijkt echter dat voor vrijwel geen enkele tekst een volledig rechte lijn door de oorsprong resulteert wanneer de type-tokenratio's volgens bovenstaande methode grafisch worden afgebeeld. Meestal ontstaan in plaats daarvan curves, die steeds vlakker gaan verlopen bij afnemende steekproefgrootte[40]. Ook deze formule kent dus tekortkomingen voor wat betreft het beschrijven van de relatie tussen het aantal woordtokens en -typen in tekstgedeelten.

Wanneer we de formules van Guiraud en van Herdan nog eens wat nader bekijken, dan valt het volgende op:

---

[39] Herdan (1966, p. 76).

[40] Zie bijvoorbeeld Baayen (1989, p. 80).

- De formule van Guiraud kan als volgt worden herschreven:

$$V = R \,.\, N^{\frac{1}{2}} \qquad (2\text{-}8)$$

- Herdan's formule houdt in dat

$$V = N^{C} \qquad (2\text{-}9)$$

  waarbij te verwachten is dat $0 < C \leq 1$, aangezien het aantal typen ten hoogste gelijk kan zijn aan het aantal tokens en bekend is dat wanneer steekproeven van steeds grotere omvang worden genomen uit een corpus, het aantal nieuwe typen dat daarin wordt gevonden steeds een dalende lijn vertoont.

Er vanuit gaande dat deze twee formules de verhouding tussen het aantal woordtokens en -typen redelijk, maar niet 100% correct weergeven, ligt het voor de hand om na te gaan of een formule die beide met elkaar combineert de type-token-ratio juister kan weergeven. Deze formule zou er dan als volgt kunnen uitzien:

$$V = R \,.\, N^{C} \qquad (2\text{-}10)$$

wederom met $0 < C \leq 1$. Deze verhouding tussen het aantal types en tokens is in het verleden door verschillende onderzoekers op het gebied van de taalwetenschap genoemd, onder andere (in een vorm met enkele extra constanten) door Mandelbrot (1953). Ook Erikstad (1980) past exact dezelfde verhouding toe bij het beschrijven van eigenschappen van juridische teksten. Deze vermeldt echter geen bron voor de formule, en geeft evenmin aan dat er verband bestaat met de verhoudingen die door Herdan en Guiraud worden gehanteerd. Wel past Erikstad curve-fitting toe om na te gaan welke de best passende waarden zijn voor de twee constanten *R* en *C*, voor een corpus bestaande uit de Engelse tekst van 86 conventies van de Raad van Europa. De beste waarde voor *R* blijkt dan te zijn 5,2253, terwijl de exponent *C* uitkomt op 0,6260.[41] Aangezien uit deze waarden volgt dat bij een stijging van *N* de toename van *V* slechts langzaam minder wordt, is het volgens Erikstad duidelijk dat het aantal verschillende onderwerpen dat in de tekst van een corpus aan de orde komt - het corpus dat hij gebruikte telde kennelijk een groot aantal verschillende onderwerpen, wat gezien de aard van de teksten ook voor de hand ligt - in wezen bepalend is voor de hoeveelheid woordtypen dat in dat corpus kan worden aangetroffen. Deze conclusie zou ik zelf niet voor mijn rekening willen nemen, hoewel het voor de hand ligt dat er tussen deze twee zaken (aantal verschillende onderwerpen en 'woordenschat' in een corpus) een verband bestaat. Erikstad gebruikt zijn - wellicht dus wat vergaande - conclusie als basis voor vervolgonderzoek, waarin hij een methode ontwikkelt om de 'similariteit' van documenten (de mate waarin documenten bepaalde eigenschappen gemeen hebben) vast te stellen, welke maat dan weer kan worden gebruikt in een retrieval-systeem. In hoofdstuk 8 zal aan een maat voor similariteit aandacht besteden. Voorts zal in het vervolg van dit onderzoek worden nagegaan of formule 2-10 geschikt is voor het beschrijven van de verhouding tussen woordtokens en -typen in de drie gebruikte corpora.

---

[41] Erikstad (1980, p. 223).

2.4 Frequentie-onderzoek

Zoals al opgemerkt is binnen de kwantitatieve taalkunde het tellen van taalelementen nog steeds de belangrijkste toepassing. Daarom, en omdat frequentiegegevens zo'n belangrijke rol spelen in dit onderzoek, zal er hier nog apart aandacht aan worden besteed. Daarbij mag duidelijk zijn dat het bij frequentie-onderzoek lang niet altijd alleen maar gaat om het aanleggen van woordfrequentielijsten; ook het verzamelen van kwantitatieve gegevens op een lager niveau - frequenties van leestekens, letters of combinaties van letters - of een hoger niveau - lengte van zinnen, omvang en indeling van documenten - kan eronder vallen.

Frequentie-onderzoek binnen de taalkunde wordt meestal verricht met een duidelijk omschreven doel. Dat doel kan bijvoorbeeld zijn het aanleggen van een woordenlijst of woordenboek voor een bepaalde taal, zoals in het geval van het Eindhovens corpus (Uit den Boogaart 1975) en het corpus van het Instituut voor Nederlandse Lexicologie. Verder kan frequentie-onderzoek ten dienste staan van het onderwijs, zoals in het geval van de woordtelling van De la Court (1937), en speelt het een belangrijke rol bij (kwantitatief) onderzoek naar schrijfstijlen. Vooral op dit laatste gebied is veel gepubliceerd; in paragraaf 2.2.2 werden daarvan al enkele voorbeelden genoemd. Een uitgebreider overzicht van allerlei onderzoek naar stijlkenmerken is te vinden in Bailey (1969, p. 217-236), en in Butler (1985, p. 18-24). In praktisch elk van de aldaar beschreven onderzoeksprojecten is in mindere of meerdere mate sprake van frequentie-onderzoek.

De methode die wordt gekozen bij het verrichten van frequentie-onderzoek hangt in veel gevallen af van het doel waartoe het onderzoek wordt verricht. Dat begint al met het samenbrengen van het basismateriaal. Zo zal het bij het doen van onderzoek naar de schrijfstijl van één bepaalde auteur wellicht niet eens nodig zijn een steekproef uit zijn werk te nemen, maar kan wanneer dat werk niet te omvangrijk is en/of al geheel of gedeeltelijk in computer-leesbare vorm beschikbaar is[42] een corpus worden samengesteld dat alle teksten van zijn hand omvat. Heel anders kan dat liggen bij het samenstellen van een corpus voor lexicografische doeleinden[43]. Wanneer een dergelijk corpus de basis moet vormen voor een werkelijk compleet woordenboek van een bepaalde taal zou het idealiter alle geschreven en gesproken taaluitingen in die taal, eventueel over een bepaalde tijdspanne, moeten omvatten, of in ieder geval een representatieve steekproef daaruit. Het eerste is in ieder geval op dit moment niet te realiseren (en het is de vraag of dat ooit het geval zal zijn), het tweede levert ook nog veel problemen op[44]. Het samenstellen van corpora speciaal voor taalkundige doeleinden, zoals bijvoorbeeld het doen van onderzoek naar de frequenties van taalelementen, wordt wel gerekend tot de corpus-taalkunde, waaraan al in paragraaf 2.1.2 enige aandacht werd besteed.

---

[42] Beschikbaarheid in computer-leesbare vorm van materiaal uit boeken en tijdschriften komt steeds vaker voor, mede als gevolg van moderne zettechnieken waarvan uitgevers gebruik maken.

[43] Zie Meijs (1991, p. 323 e.v.).

[44] Martin (1988, p. 92 e.v.) wijst er bijvoorbeeld op dat het nemen van aselecte steekproeven uit een taal problemen oplevert, niet eens in de eerste plaats doordat de taal een *oneindige* populatie vormt, maar vooral doordat deze populatie zeer *heterogeen* is. Een en ander is volgens hem reden om niet te streven naar een *representatief*, maar naar een *exemplarisch* corpus. Zie ook hoofdstuk 3, paragraaf 5.

Na het samenstellen van het basismateriaal voor frequentie-onderzoek, wat dus al de nodige hoofdbrekens kan kosten, ontstaan dikwijls nieuwe problemen bij het verrichten van de eigenlijke analyses. Dat begint al met het isoleren uit het tekstmateriaal van de taalelementen die geteld c.q. onderzocht moeten worden. Brandt Corstius geeft daarvan een aantal voorbeelden[45], waarvan hier genoemd kunnen worden:

- is *'s avonds* één woord of twee?
- is *Kolen en Staalgemeenschap* één woord of drie?
- is in de zin *De was was droog* sprake van twee verschillende woorden *was*?
- moeten hoofd- en kleine letters aan elkaar gelijk worden gesteld of niet?
- moeten woorden die aan het einde van een regel zijn afgebroken gelijk worden gesteld aan dezelfde woorden die dat niet zijn?
- wat moet worden gedaan met cijfers, typografische en andere symbolen, en met kennelijke drukfouten?

Aangezien dit soort onderzoek tegenwoordig vrijwel uitsluitend met behulp van computers plaatsvindt, is het van belang om al van tevoren met al deze problemen rekening te houden bij het schrijven van het desbetreffende programma. Bij tellen met de hand kan eventueel de beslissing over ieder probleem worden uitgesteld tot het zich werkelijk in de tekst voordoet. De problemen die hierboven werden opgesomd hebben met name betrekking op het isoleren van *woorden*; ook bij het isoleren van andere onderdelen van een tekst kan echter van alles fout gaan, zoals bijvoorbeeld bij het splitsen in *zinnen*. Wanneer dat automatisch moet gebeuren, moet de computer voor iedere zin in een tekstcorpus kunnen vaststellen waar die begint en eindigt. In eerste instantie lijkt dat eenvoudig - een zin begint met een hoofdletter en eindigt met een punt - maar moeilijker wordt het wanneer zinnen voorkomen als *Mevrouw W.M. Pietersen deed iedere donderdag boodschappen*. In feite is het nodig dat een computer die de lengte van zinnen in een corpus moet bepalen in staat is om die zinnen taalkundig te ontleden, wat nog steeds een tamelijk zware opgave is. Zo kan gezegd worden dat voor het isoleren van elk soort taalelementen specifieke problemen kunnen gelden, die moeten worden opgelost voordat tot frequentie-onderzoek kan worden overgegaan.

Lang niet alle problemen die hierboven werden opgesomd kunnen worden opgelost uitsluitend door het ontwerpen van het juiste algoritme. In een aantal gevallen kan alleen automatisch worden gesplitst wanneer het corpus van tevoren wordt bewerkt. Dat is bijvoorbeeld het geval bij de homoniemen (*De was was droog*); wanneer die als verschillende woordtypen moeten worden beschouwd is het nodig om bij het woord een (grammaticale) code op te nemen in het corpus (zoals bijvoorbeeld in het Eindhovens corpus het geval is). Vroeger gebeurde dat coderen meestal met de hand, waardoor het nauwelijks mogelijk was corpora te verwerken die erg omvangrijk waren of waarvan de inhoud dikwijls veranderde. Tegenwoordig behoort ook automatisch (dat wil zeggen, met behulp van een computerprogramma) coderen tot de mogelijkheden, waarbij echter altijd rekening moet worden gehouden met mogelijke fouten (woorden die door het programma onjuist worden geïnterpreteerd, en daardoor ook onjuist worden gecodeerd).

---

[45] Brandt Corstius (1978, p. 44 e.v.).

Deze twee factoren kunnen allebei een rol spelen bij corpora waarop frequentie-onderzoek wordt verricht in het kader van het opzetten van een retrieval-systeem voor dat corpus. Het gaat dan meestal om omvangrijke corpora (meer dan 1 miljoen woordtokens), waaraan telkens documenten worden toegevoegd of waarin regelmatig documenten worden vervangen. Bij dergelijke corpora moet automatisch een nieuwe index kunnen worden gegenereerd, zonder dat eerst bewerking van documenten met de hand behoeft plaats te vinden. Dat betekent wellicht dat in zulke gevallen genoegen moet worden genomen met een frequentielijst die enkele fouten of versimpelingen bevat, omwille van de mogelijkheid om snel updates in te kunnen voeren.

Het is wellicht zinvol om hier nog enkele woorden te wijden aan resultaten die frequentieonderzoek tot nu toe heeft opgeleverd. Daarbij kan in eerste instantie weer worden gedacht aan de diverse woordfrequentielijsten die op basis van een aantal corpora zijn gemaakt. Voor een overzicht daarvan kan worden verwezen naar paragraaf 2.2.1, eerder in dit hoofdstuk. Onderzoek naar stijlkenmerken, waarvoor dikwijls frequentieonderzoek is verricht, heeft zoals in paragraaf 2.2.2 beschreven naast de nodige teleurstellingen ook positieve resultaten opgeleverd. In het kader van stijl-onderzoek zijn daarnaast enkele statistische wetmatigheden en maatgetallen ontdekt (zie paragraaf 2.3), die ook van belang zijn gebleken voor de kwantitatieve taalkunde in het algemeen. De laatste jaren is door taalkundigen ook nogal eens frequentie-onderzoek verricht dat kan worden gerekend tot de corpustaalkunde. Dit onderzoek heeft nuttige informatie opgeleverd over allerlei aspecten, variërend van nieuwe inzichten in de structuur van zelfstandige naamwoordgroepen tot mogelijkheden voor het opsporen van onderwerpen of thema's in tekstmateriaal[46]. In het algemeen kan gesteld worden dat frequentie-onderzoek door de jaren heen steeds van belang is gebleven binnen de taalkunde. Sterker nog, de belangstelling voor deze vorm van onderzoek lijkt de laatste jaren weer duidelijk toe te nemen.

Enkele onderwerpen die in deze paragraaf aan de orde zijn gesteld zijn in hoge mate relevant voor het onderzoek dat hierna zal worden beschreven. Zo zal onder andere de nodige zorg moeten worden besteed aan de samenstelling van de drie gebruikte corpora. Voor wat betreft de juridische corpora zal dit met name neerkomen op het controleren van de volledigheid, en van de aanwezigheid in het corpus van de juiste documenten. In het geval van het corpus Algemeen Nederlands ligt de zaak echter wat complexer; daarbij moet beoordeeld worden of de documenten in dit corpus inderdaad een goede doorsnede vormen van de Nederlandse schrijftaal. Ook het splitsen van teksten in afzonderlijke taalelementen (in dit geval voornamelijk woorden) zal nog wel de nodige problemen opleveren. Daarbij is onder andere een punt van overweging of woordvormen in dezelfde vorm(en) als waarin zij in de tekst voorkomen moeten worden opgeslagen, of dat het wellicht verstandig is om te abstraheren van het woord-voorkomen en woorden te herleiden tot hun basisvorm (lemma), waarna alleen deze woordlemmata worden opgeslagen. Aan deze punten zal in het volgende hoofdstuk aandacht worden besteed.

[46] Meijs (1991, p. 319).

2.5. Subtaal en vaktaal

Zoals in het vorige hoofdstuk al werd vermeld, vertonen juristen volgens sommigen een bijzonder taalgedrag. Het is opvallend dat opmerkingen die daarover worden gemaakt vaak de vorm hebben van klachten, zoals bijvoorbeeld in het al aangehaalde rapport Recht en Taal (Van den Bergh en Broekman 1979). Deze klachten komen er in het algemeen vaak op neer dat juristen ongewone *woorden* gebruiken (of gewone woorden in een ongewone betekenis), en dat juridische teksten vaak worden gekenmerkt door een ongewone *zinsbouw*. Een en ander leidt er volgens velen toe dat het taalgebruik in juridische teksten een voor leken ontoegankelijk jargon is[47]. Hofhuis nuanceert in zijn bijdrage aan de bundel "Meesterlijke taal" dit standpunt enigermate, en geeft aan dat in de teksten van rechterlijke vonnissen - teksten waarin volgens sommigen bij uitstek sprake is van het gebruik van juridische vaktaal - algemeen gangbaar Nederlands de hoofdmoot vormt. Daarnaast kennen deze teksten dan een aantal woorden en begrippen die in het recht een afwijkende betekenis (kunnen) hebben, en een aantal termen die we kunnen rekenen tot de 'echte' juridische vaktaal. Dit gebruik van vaktaal acht Hofhuis onvermijdelijk, maar hij geeft wel aan dat op een rechter, gezien het karakter van de beslissingen die hij neemt, een bijzondere plicht rust om te proberen zo duidelijk en begrijpelijk mogelijk te zijn[48].

Het lijkt er dus op dat ook juristen, net als vele andere beroepsbeoefenaars, beschikken over een eigen vaktaal en daarvan ook gebruik maken bij het opstellen van schriftelijke stukken. Dit is ook het uitgangspunt van de "Werkgroep vorm en taal van vonnissen en beschikkingen"[49]. Maar ook deze werkgroep wijst erop dat er alle reden is juist in de rechtspraktijk scherp te letten op begrijpelijkheid voor de burger. Het feit dat juristen zich van een eigen vaktaal bedienen behoeft op zich geen probleem te zijn, wanneer er maar voor wordt gewaakt dat een en ander ook voor buitenstaanders nog begrijpelijk is, en dat typisch juridische termen en begrippen, waarvoor in het algemene spraakgebruik geen synoniemen bestaan, zonodig van uitleg worden voorzien. Ook Van Eemeren *et al.* (1991) gaan uit van dit standpunt; zij wijzen erop dat het gebruik van vaktermen voor juristen noodzakelijk kan zijn om effectief met elkaar te kunnen communiceren. Daarbij worden sommige vaktermen gebruikt om nieuwe begrippen te introduceren, en andere om op een consistente manier over bepaalde zaken te kunnen spreken[50].
Met name in overheidsteksten die tot doel hebben de burgers voor te lichten moet echter de nodige aandacht aan begrijpelijkheid worden besteed, en moeten ambtelijke formuleringen zoveel mogelijk worden vermeden.

Nu dus in brede kring, ook onder juristen, de mening is aanvaard dat er iets bestaat als een 'juridische vaktaal', kan de vraag gesteld worden wat daarvan de kenmerken zijn. Gaat het alleen om het gebruik van bepaalde specifieke termen, of is er sprake van taalgebruik dat in (veel) meer opzichten afwijkt van algemeen Nederlands? En mocht dit laatste inderdaad

---

47 Van den Bergh en Broekman (1979, p. 14).

48 Hofhuis (1988, p. 62).

49 Deze werkgroep werd in 1979 opgericht op initiatief van de van de Nederlandse Vereniging voor Rechtspraak (sectie raadsheren en rechters) en van de Staatscommissie Herziening Rechterlijke Organisatie. Het eerste rapport van de sectie Civiel Recht van deze werkgroep verscheen in april 1983 in Trema", p. 151-161.

50 Van Eemeren *et al.* (1991, p. 169).

het geval zijn, waaruit bestaan dan die afwijkingen? Gaat het bijvoorbeeld om het gebruik van andere woorden, of worden dezelfde woorden anders gebruikt, bijvoorbeeld in een andere volgorde of onder gebruikmaking van afwijkende zinsconstructies? In de taalkunde worden al deze verschilpunten gerekend tot de zogenaamde *taalvariëteit*. Voor dat onderwerp bestaat de laatste 15 à 20 jaar een groeiende belangstelling, onder andere in het kader van de bestudering van (oorzaken van) taal*verandering*, waarvoor taalvariëteit als een van de meest algemene voorwaarden wordt gezien (naast allerlei sociale factoren). Als gevolg van die belangstelling voor taalvariëteit heeft ook het onderzoek naar subtalen c.q. vaktalen (op het onderscheid tussen deze twee zal hierna nog worden teruggekomen) in de jaren zeventig nieuwe impulsen gekregen. Martin en Ten Pas (1991) noemen nog enkele andere gronden voor een toegenomen interesse op dit gebied, zoals:

- toegenomen internationale contacten op het gebied van wetenschap en technologie, wat geleid heeft tot diverse studies naar 'Language for Specific Purposes', ofwel vaktaal.
- bij onderzoek op het gebied van natuurlijke-taalverwerking (dit onderzoek behoort tot het terrein van de computer-taalkunde) is gebleken dat op dit moment eigenlijk alleen resultaten kunnen worden bereikt wanneer men zich beperkt tot de taal van een bepaald (kennis)domein.

Wanneer we het over subtaal of vaktaal hebben, is een goede definitie van deze begrippen wenselijk. Het opstellen van zo'n definitie levert echter wel enige problemen op, aangezien daarin aangegeven zal moeten worden wat de relatie is tussen sub- of vaktaal en 'algemene taal'. Echter, om dat te kunnen doen moet eigenlijk eerst een definitie van 'algemene taal' worden opgesteld, wat buitengewoon lastig is. Sommige onderzoekers menen zelfs dat het principieel onmogelijk is 'algemene taal' te definiëren, onder andere in verband met het dynamisch karakter van taal. Zij zijn daarom van mening dat het beter is niet te spreken over subtaal en algemene taal, maar slechts over 'situationele varianten' binnen een taal. Anderen hanteren de begrippen subtaal en algemene taal wel; algemeen aanvaard is op dit moment echter het standpunt dat subtaal en algemene taal niet als twee onafhankelijke en op zichzelf staande grootheden kunnen worden beschouwd. Zij bevatten daarvoor, hoewel ze verschillen vertonen op lexicaal, grammaticaal en pragmatisch niveau, toch teveel overlappingen en gemeenschappelijke kenmerken[51]. Daarom kan de relatie tussen subtaal (en ook vaktaal, zoals we hierna zullen zien) en algemene taal het beste als volgt worden gekarakteriseerd: vrijwel iedere subtaal bevat kenmerken van algemene taal. De twee overlappen elkaar dus, zij het slechts gedeeltelijk. Subtaal heeft namelijk meestal ook kenmerken die niet in de algemene taal worden gevonden. De verschillen tussen beide zijn voornamelijk functioneel-pragmatisch van aard[52], dat wil zeggen, de punten die nu juist specifiek zijn voor een bepaalde subtaal hebben meestal te maken met de functie welke die subtaal voor haar gebruikers heeft, dan wel met het doel waarom zij is ontstaan c.q. wordt gebruikt.

---

[51] Martin en Ten Pas (1991, p. 362). De auteurs behandelen vervolgens nog enkele meningen over de eigenschappen van subtaal, voordat zij tot een eigen definitie komen.

[52] Martin en Ten Pas (1991, p. 363).

De begrippen subtaal en vaktaal hebben veel met elkaar gemeen. In beide gevallen gaat het om het taalgebruik op een in semantisch en pragmatisch opzicht beperkt domein. Toch zijn beide begrippen geen volledige synoniemen. Vaktaal wordt in het algemeen gerelateerd aan een bepaald vakgebied, zoals biologie, scheikunde of recht, terwijl subtaal ieder in semantisch opzicht beperkt terrein kan betreffen. Vaktaal is daardoor een begrip dat veelal 'op voorhand' exacter kan worden gespecificeerd aangezien taaluitingen die eronder vallen samenhangen met de stand van zaken (bijvoorbeeld: de stand van de techniek) binnen een vakgebied. Welk taalgebruik precies valt onder een bepaalde subtaal kan daarentegen pas worden bepaald wanneer eerst duidelijkheid bestaat over de voor die subtaal geldende grenzen en parameters. Aangezien dit onderzoeksproject zich in sterke mate richt op het taalgebruik op één bepaald vakgebied (het juridische), verdient in dit geval het begrip vaktaal toch de voorkeur. Daarom zal in het vervolg worden gesproken over juridische *vak*taal. Wellicht is het echter mogelijk om binnen deze vaktaal nog enkele subtalen (sub-vaktalen?) te onderscheiden, zoals bijvoorbeeld een subtaal voor jurisprudentie, voor wetgeving en voor juridische dogmatiek. Een van de doeleinden van dit onderzoeksproject is te trachten voor het al dan niet bestaan van deze subtalen aanwijzingen te zoeken. Het onderdeel juridische dogmatiek zal daarbij echter niet aan de orde kunnen komen, niet alleen omdat teksten die hiertoe kunnen worden gerekend vrijwel niet in computer-leesbare vorm beschikbaar zijn, maar ook omdat het, mocht in de toekomst aan deze eerste voorwaarde wel zijn voldaan, nog bijzonder moeilijk zal zijn een corpus van teksten samen te stellen dat 'de juridische dogmatiek in Nederland' omvat, gezien de grote verscheidenheid aan teksten (zoals boeken, tijdschriftartikelen, annotaties, etc.) en de lange periode welke deze bestrijken. Voorlopig is het bestaan van een juridische vaktaal nog niet meer dan een hypothese; het is één van de hypothesen die in het vervolg van dit onderzoek zal worden getoetst. Punten die bij deze toetsing aan de orde zullen komen zijn bijvoorbeeld:

- bestaan er verschillen in woordgebruik tussen algemene taal en juridische taal?
- bestaan er verschillen in woordgebruik tussen de twee juridische corpora?
- zijn deze laatste verschillen wellicht talrijker of omvangrijker dan de eerstgenoemde? In dat geval kan mogelijk worden gesproken van *twee* juridische vaktalen.

Het zal intussen duidelijk zijn dat het geven van een definitie van het begrip vaktaal - in relatie tot algemene taal - niet eenvoudig is. Martin en Ten Pas doen daartoe toch een poging door gebruik te maken van een model, waarin vijf assen worden onderscheiden waarlangs variatie in taalgebruik kan optreden, te weten een regionale, een sociale, een tijds-, een inhouds- en een communicatieve situatie-as[53]. Het is dan mogelijk om aan te geven waar vaktaal afwijkt van algemene taal, te weten met name op de communicatieve situatie-as en op de inhoudsas: vaktaal wordt gebruikt in een beperkt aantal communicatieve situaties, waarbij het meestal specifieke onderwerpen of kennis betreft. De indeling op de communicatieve situatie-as heeft met name betrekking op het *kennisniveau* van de taalgebruikers, en op het *communicatiedoel*. Met name de eerste van deze twee parameters speelt in het geval van vaktaal een rol, er is dikwijls sprake van taaluitingen tussen vakgenoten. De tweede parameter heeft te maken met een andere eigenschap van vaktaal: er is dikwijls sprake van gestandaardiseerde tekstsoorten, die een specifieke functie hebben. Van Sterkenburg (1989) wijst er in dit verband op dat woorden die tot een vaktaal behoren veelal ondubbelzinnig zijn qua betekenis, geen bijbetekenis hebben en

[53] Dit model wordt besproken door Martin (1988, p. 94 e.v.).

derhalve gevoelsneutraal zijn[54]. Kenmerken als deze kunnen nuttig zijn bij het in kaart brengen van een sub- of vaktaal. Zoals duidelijk moge zijn is het daarbij niet zo dat met een handvol criteria eenvoudig kan worden nagegaan of een bepaalde serie taaluitingen als vaktaal kan worden aangemerkt, maar is er veeleer sprake van een aantal kenmerken, die wanneer ze in sterkere of juist in geringere mate aanwezig zijn in bepaalde teksten de kans kunnen vergroten dat deze teksten vaktaal bevatten[55].

Wanneer we de kenmerken van vaktaal meer in detail willen bekijken, zal blijken dat sommige kenmerken betrekking hebben op het woordniveau; voorbeelden van zulke kenmerken zullen hierna worden besproken. Andere kenmerken hebben betrekking op het zins- of het meer-dan-zins-niveau; deze vallen echter buiten de reikwijdte van dit onderzoek, ze zullen daarom verder niet worden behandeld. Deville (1989) geeft een categorie-indeling voor alle kenmerken van vak- en subtalen, met het doel om juist de verschillen ten opzichte van algemene taal aan het licht te brengen:

- restricties: in sommige opzichten vormt vak- of subtaal slechts een subset van algemene taal;
- deviaties: vak- of subtaal kan eigenschappen hebben die algemene taal niet heeft;
- preferenties: sommige gemeenschappelijke kenmerken van vak- of subtaal en van algemene taal hebben in een van beide taalvarianten een hogere frequentie of meer gewicht dan in de andere.

Martin en Ten Pas geven, gebruik makend van deze categorie-indeling, een aantal vrij gedetailleerde kenmerken van vaktaal[56]. Daarvan hebben de volgende betrekking op het woordniveau:

- restricties:
  - het lexicon van een vaktaal is relatief gesloten, dat wil zeggen dat het meer dan in het geval van algemene taal uitputtend beschreven kan worden;
  - in tegenstelling tot algemene taal zijn in vaktaal woordcombinaties veel meer semantisch bepaald. Gegeven de concepten die de vaktaal moet kunnen beschrijven, en gegeven het feit dat synoniemen voor woorden minder vaak worden gebruikt in vaktaal liggen de manieren waarop woorden met elkaar kunnen worden gecombineerd veel meer vast dan bij algemene taal.
- deviaties:
  - vaktaal kent in veel gevallen specifieke - 'eigen' - woorden en woordvormen;
  - woorden worden in vaktaal soms op een andere manier met elkaar gecombineerd - zie ook de tweede van de hierboven genoemde restricties;
- preferenties:
  - frequenties van bepaalde woorden en woordcombinaties wijken in vaktaal af van die in algemene taal. Ook de frequentie*verdeling* is vaak anders.

---

[54] Van Sterkenburg (1989, p. 16).

[55] Een 'echte' definitie van het begrip vaktaal is dit natuurlijk niet. Vaktaal kan ook worden beschouwd als een 'concept', een abstracte notie die op zich geen onderdeel uitmaakt van de ervaringswereld. Net als het geval is bij de meeste andere concepten zijn de grenzen van het concept vaktaal enigszins vaag. Zie over het begrip concept o.a. Rescher (1969) en Hempel (1967).

[56] Martin en Ten Pas (1991, p. 366-372).

De hierboven genoemde kenmerken bieden diverse aanknopingspunten voor het onderzoek naar juridisch woordgebruik. Dat is eveneens het geval met enkele kenmerken die genoemd worden door Biber and Finegan (1986). Zij trachtten met behulp van statistische technieken (onder andere factor-analyse en cluster-analyse) een corpus bestaande uit zeer uiteenlopende Engelstalige teksten (variërend van weergaven van telefoongesprekken tot officiële overheidsdocumenten en wetenschappelijke artikelen) in te delen in automatisch herkende 'tekst-typen'. Voorzover uit hun beschrijving valt na te gaan, werden praktisch alle documenten waarvan de inhoud kan worden aangemerkt als 'juridisch taalgebruik' geclassificeerd als "Formal Exposition" (formele uiteenzetting). Dit is op zich al interessant, maar Biber en Finegan geven vervolgens ook nog een beschrijving van de kenmerken van dat specifieke tekst-type. Het gaat volgens hen om tekst met veel lange woorden, en met een relatief lage type-token ratio (d.w.z.: een relatief hoog aantal woord-typen in een tekst met een gegeven lengte)[57]. Het tweede element, het grote aantal verschillende woorden, wekt enige bevreemding, aangezien uit ander onderzoek, bijvoorbeeld dat van Davis (1986), blijkt dat met name (Engelse) wetteksten, en in iets mindere mate jurisprudentieteksten juist relatief minder verschillende woorden tellen dan teksten met een algemene inhoud[58]. Ook in een eerdere studie van De Mulder en Oskamp naar het woordgebruik in de wetboeken van Strafrecht en Strafvordering bleek dat in deze wetteksten in ieder geval de veel voorkomende woordtypen een relatief hogere frequentie hebben dan veelvoorkomende typen in algemeen Nederlands[59]. Beide studies lijken erop te wijzen dat juristen neigen tot een taalgebruik met relatief minder verschillende woorden, die dan elk vaker gebruikt worden. Het is natuurlijk mogelijk dat de omschrijving van Biber en Finegan toch een grond van waarheid bevat, bijvoorbeeld doordat in juridische teksten veel zeldzame woordtypen (typen met een frequentie van vijf of minder) voorkomen waardoor het totaal aantal woordtypen toch oploopt, en misschien uiteindelijk zelfs hoger uitkomt dan in algemeen taalgebruik. Andere door hen genoemde kenmerken, zoals een relatief hoog aantal voorzetsels en bijvoeglijke naamwoorden, en een frequent gebruik van passieve vormen en van de tegenwoordige tijd, hebben vooral betrekking op syntactische aspecten van teksten die hier buiten beschouwing worden gelaten. Van andere hier besproken kenmerken op woordniveau zal in de volgende hoofdstukken worden nagegaan of deze ook voor juridisch - Nederlands - taalgebruik gelden.

---

57 Biber and Finegan (1986, p. 40).

58 Davis (1986, p. 475).

59 De Mulder en Oskamp (1979, p. 75), De Mulder (1984, p. 150).

HOOFDSTUK 3 - DE DRIE CORPORA

3.1 Inleiding

Dit hoofdstuk bevat een beschrijving van de drie in dit onderzoek gebruikte tekstcorpora:

- een corpus met de teksten van alle wet- en regelgeving die in het jaar 1989 in Nederland in werking was,
- een corpus met de teksten van alle rechterlijke uitspraken die in de periode 1965-1989 in het tijdschrift Nederlandse Jurisprudentie zijn gepubliceerd, en
- een corpus 'algemeen Nederlands', samengesteld door het Instituut voor Nederlandse Lexicologie uit een aantal verschillende bronnen, zoals romans, studieboeken en artikelen, zowel wetenschappelijke als niet-wetenschappelijke.

In afzonderlijke paragrafen zal telkens worden aangegeven hoe een corpus is verkregen, wat de samenstelling ervan is en welke werkwijze gevolgd is voor het verkrijgen van gegevens uit het corpus. Voor het lezen van de teksten moest in alle gevallen speciale software worden geschreven. De werking daarvan zal eveneens worden beschreven. Bovendien moest bij het lezen elk van de corpora al rekening worden gehouden met het gebruik van gegevens uit dit onderzoek in vervolgprojecten. Hierop zal nu eerst kort worden ingegaan.

3.2 Toekomstig gebruik van gegevens uit de corpora

Zoals al in het eerste hoofdstuk werd aangegeven, zal de komende jaren door het Centrum voor Informatica en Recht onderzoek worden gedaan om na te gaan of bepaalde aspecten van de vorm van documenten een rol kunnen spelen bij het lokaliseren of categoriseren van die documenten in een omvangrijke (elektronische) gegevensverzameling. Ook in dit onderzoeksproject zal hieraan aandacht worden besteed, en dan met name in hoofdstuk 8, waar zal worden nagegaan of het op basis van de aan- of afwezigheid van woorden in een document mogelijk is 'clusters' van documenten samen te stellen, die bepaalde overeenkomsten voor wat betreft de woordtypen die erin voorkomen vertonen.

Om gebruik te kunnen maken van vorm-aspecten, en met name van woordfrequentiegegevens van juridische documenten is het nodig dat deze documenten aan bepaalde eisen voldoen. Naast eisen als de beschikbaarheid in elektronische vorm en dergelijke is daarbij met name van belang dat de documenten een bepaalde omvang hebben. Die omvang moet niet te groot zijn, aangezien anders het aantal woorden in het document erg groot wordt en het terugvinden van het document meestal zou betekenen dat er erg veel overbodige informatie wordt gegeven, maar zeker ook niet te klein, aangezien dan het analyseren van het document te weinig bruikbare gegevens oplevert; met name het aantal woordtypen zou dan te laag kunnen zijn om vergelijking met andere documenten mogelijk te maken.

De eisen met betrekking tot de omvang afwegend, hebben wij voorlopig besloten te streven naar een documentgrootte van 100 tot 500 woorden (ongeveer 6 regels tot 1 bladzijde tekst), waarbij moet worden aangetekend dat 100 woorden eigenlijk nog erg weinig is, en dat ook van belang is in hoeverre documenten in een gegevensverzameling in omvang verschillen: voor het optimaal vergelijken van vormaspecten zouden eigenlijk alle documenten ongeveer even groot moeten zijn. Met deze eisen is zoveel mogelijk rekening gehouden bij het ontsluiten van de diverse corpora die in het kader van dit onderzoek zijn gebruikt.

3.3 Het corpus teksten van Nederlandse wetgeving

3.3.1 Inleiding

In dit onderzoek naar kwantitatieve eigenschappen van juridische schrijftaal spelen wetgevingsteksten een belangrijke rol. Onder 'wetgeving' wordt in dit verband verstaan: wetgeving in materiële zin, dat wil zeggen alle door of vanwege het rijk vastgestelde algemeen verbindende voorschriften[1]. Het gaat dus naast wetten in formele zin om algemene maatregelen van bestuur en ministeriële verordeningen, voorzover deze laatste algemeen verbindende voorschriften bevatten. Ook internationale verdragen kunnen onder bepaalde omstandigheden tot deze vorm van wetgeving worden gerekend. Juist in wetsteksten kunnen zinsneden en woorden worden aangetroffen die (zeker op het eerste gezicht) afwijken van het normale Nederlandse spraakgebruik, terwijl ook de opbouw (bijvoorbeeld: de aanwezigheid van artikelen) karakteristiek is.

Bij het analyseren van wetteksten moet een selectie worden gemaakt. Het materiaal is anders gewoonweg te omvangrijk en te divers. Met name het verwerken van de veelheid aan internationale verdragen waarbij Nederland partij is, van E.G. regelgeving, etc. zou leiden tot een te omvangrijk corpus, dat met de ter beschikking staande apparatuur en programmatuur niet meer te verwerken zou zijn. Het bronnenmateriaal zou bijvoorbeeld aanzienlijk kunnen worden ingeperkt door alleen Nederlandse wetgeving in formele zin[2] in beschouwing te nemen. Deze benadering is voor dit onderzoek als te beperkt te beschouwen, aangezien dan een belangrijk deel van de regelgeving, zoals bijvoorbeeld de Algemene Maatregelen van Bestuur, niet is vertegenwoordigd. Het is onder andere interessant om te zien of zich tussen deze vormen van wetgeving verschillen in opbouw of woordgebruik voordoen. Verder speelde voor dit onderzoek ook de beschikbaarheid van het tekstmateriaal een rol. Om analyse met behulp van een computer mogelijk te maken moesten de teksten in machine-leesbare vorm (ofwel: in de vorm van tekstbestanden) beschikbaar zijn, aangezien faciliteiten om op grote schaal teksten over te typen of te 'scannen' ontbraken.

In de loop van 1989 werd het probleem van de beschikbaarheid van tekstmateriaal opgelost doordat uitgeverij Vermande zich bereid verklaarde de inhoud van de door haar

---

1 Van der Pot - Donner (1989, p. 524 e.v.).
2 Van der Pot - Donner (1989, p. 511 e.v.).

uitgegeven databank van Nederlandse wetgeving[3] (NLEX) integraal ter beschikking te stellen. Deze databank, die in 1980 tot stand kwam, bevatte een zo compleet mogelijke weergave van algemeen verbindende regelgeving die op het moment van uitgifte in Nederland in werking was, met uitzondering van provinciale en gemeentelijke verordeningen. Bij het samenstellen van de databank werd onder andere ingespeeld op onderzoek uit de tweede helft van de jaren zeventig naar informatiebehoeften binnen juridische beroepsgroepen. Een overzicht van deze behoeften, en van de eisen die aan geautomatiseerde informatiesystemen op juridisch gebied moeten worden gesteld, zijn te vinden in het rapport "Een nationaal geautomatiseerd systeem van wetgevings- en rechtspraakinformatie", opgesteld door een door de Minister van Justitie ingestelde werkgroep onder voorzitterschap van mr. R. Scherpenzeel[4].

Bij het samenstellen van de NLEX werden voor de wetten in formele zin exemplaren van het Nederlandse Staatsblad geraadpleegd waarin de wetteksten voor de eerste maal waren gepubliceerd, en werd de integrale tekst van de wet daaruit overgenomen. Door de jaren heen gepubliceerde wetswijzigingen werden eveneens in de databank verwerkt. Voor niet in het Staatsblad gepubliceerde wetgeving werd een vergelijkbare procedure gevolgd, zij het dat hier overname uit de Staatscourant plaatsvond, en wel van al die regelgeving die algemeen verbindende voorschriften bevatte voor alle Nederlandse ingezetenen. Nadat zo alle officieel gepubliceerde wetten (in materiële zin) waren verwerkt, bevatte de NLEX dus alle regelgeving op nationaal niveau die op dat moment in werking was. Enige tijd na het operationeel worden van de databank werden nog enkele belangrijke verdragen waarbij Nederland partij is aan de databank toegevoegd, zoals bijvoorbeeld het verdrag tot bescherming van de rechten van de mens en de fundamentele vrijheden met enige protocollen, het verdrag tot oprichting van de Europese Economische Gemeenschap en een aantal verdragen op het gebied van internationaal milieubeheer. In het vervolg zal voor al deze regelgeving kortweg de term 'wet' worden gebruikt. De databank werd daarna maandelijks bijgewerkt. Dit gebeurde voor het laatst in 1992; in de zomer van dat jaar werd helaas de uitgave van de databank gestaakt. De NLEX is sinds die tijd dus niet meer voor het publiek beschikbaar.

Aangezien dus alle - in werking zijnde - Nederlandse regelgeving, aangevuld met de voornaamste verdragsteksten, in de NLEX te vinden was, vormen de teksten in dit bestand een ideale basis voor onderzoek naar vormaspecten van de huidige Nederlandse wetgeving in het algemeen. Ook allerlei wetten en regelingen die in Nederland nog officieel van kracht zijn (want nooit ingetrokken) maar waarvan praktisch iedereen het bestaan al lang is vergeten zijn op deze manier bij het onderzoek betrokken. Zo is bijvoorbeeld een aantal wetten en regelingen met betrekking tot de mijnbouw in Nederland opgenomen (de meeste

---

[3] In het vervolg zal steeds de term 'wetgeving' worden gebruikt, in de betekenis van: alle vormen van algemeen verbindende voorschriften welke uiteindelijk tot een wet (in formele zin) zijn te herleiden. 'Wetgeving' omvat daarmee dus met name wetten in formele zin, algemene maatregelen van bestuur en ministeriële regelingen houdende algemeen verbindende voorschriften, alsmede internationale verdragen welke d.m.v. een goedkeuringswet zijn bekrachtigd.

[4] Werkgroep, ingesteld bij besluit van de Minister van Justitie d.d. 28 maart 1979. Interimadvies, 's Gravenhage, april 1980.

daterend van rond 1900 of nog eerder, en zelfs nog een aantal wetten en regelingen uit de Franse tijd (daterend van rond 1810)[5].

Het aanbod van Vermande resulteerde erin dat in december 1989 een kopie van deze databank, echter zonder de STAIRS zoekprogrammatuur die normaal gesproken voor het raadplegen wordt gebruikt, door de uitgever op computertapes werd gezet. De gegevens op deze tapes werden vervolgens met assistentie van het Erasmus Rekencentrum gekopieerd naar een optische schijf (een z.g.n. WORM - Write Once Read Many times - disk) met een capaciteit van 400 Megabytes. Alle gegevens samen bleken één enkel bestand van ruim 281 Megabytes te vormen.

3.3.2 Het lezen van de NLEX databank

Om het verdere verloop van het onderzoek te begrijpen is het nodig iets te weten over de opbouw van de NLEX databank. Deze databank is ingedeeld in *documenten*, waarvan er een of meer kunnen worden geselecteerd en bekeken op het scherm met behulp van een *zoektaal*. Deze zoektaal - in het geval van de NLEX is dat STAIRS, een andere bekende zoektaal is bijvoorbeeld STATUS - biedt de gebruiker de mogelijkheid om bepaalde zoektermen - expressies bestaande uit een of meer woorden en eventueel een of meer logische operatoren - op te geven, en zo de verzameling van gewenste documenten (bij benadering) te omschrijven. Nadat de gebruiker een zoekterm heeft geformuleerd wordt de database afgezocht en worden de documenten geselecteerd die voldoen aan de opgegeven zoekterm, dat wil zeggen dat ze de opgegeven (combinatie van) zoekwoorden bevatten. Het zoeken gebeurt daarbij volgens het 'full text' systeem, dat wil zeggen dat op ieder willekeurig woord dat in de databank voorkomt kan worden gezocht, echter met uitzondering van een aantal 'stopwoorden', woorden als 'de', 'het', 'een', 'van', etc., waarvan te verwachten is dat ze praktisch in ieder document zullen voorkomen, en die dus geen selecterende werking hebben. De NLEX databank blijkt in totaal 5875 wetten[6] te bevatten, waarbij iedere wet die een of meer artikelen bevat (bij sommige regelingen is dat nl. niet het geval) bestaat uit een aantal afzonderlijk opvraagbare 'documenten'. Ieder document bevat één wetsartikel met daarbij in afzonderlijke 'secties' informatie betreffende dat wetsartikel. Zo bevat ieder document een sectie 'volledige titel', die zoals de naam al zegt uit de volledige titel van de wet waartoe het artikel behoort bestaat, en een sectie 'datum van publikatie', waarin wordt aangegeven wanneer (de laatste wijziging van) het artikel is gepubliceerd in Staatsblad of Staatscourant. Tabel 3-1 bevat een lijst van de secties die kunnen voorkomen. Bij veel documenten ontbreken een of meer secties (zoals bijvoorbeeld de secties met toelichting, of de sectie trefwoorden). Wanneer een wet niet in afzonderlijke artikelen is ingedeeld, bevatten de NLEX-documenten een compleet onderdeel (bijvoorbeeld een hoofdstuk of een paragraaf) van een wet, of - bij afwezigheid van een hoofdstuk- of paragraafindeling - de tekst van de gehele wet.

---

[5] Een voorbeeld van de eerste categorie vormt de "Wet van den 27 April 1904, houdende nadere bepalingen betreffende de mijnontginning, met wijziging der wet van 21 April 1810", en van de tweede categorie het "Decret imperial du 22 janvier 1808, qui declare l'article 7 du titre 28 de L'Ordonnance de 1669, applicable a toutes les Rivières navigables de l'Empire".

[6] Zoals hierboven reeds aangegeven wordt voor alle in de databank aanwezige regelgeving de term 'wet' gebruikt.

Code	Inhoud van de sectie
90	volledige titel van de wet/regeling ('Wet van ... inhoudende ....')
91	officiële citeertitel ('Wet toezicht kredietwezen')
92	korte titel (alias), indien in gebruik
100	volledige plaatsaanduiding van dit artikel: hoofdstuk, titel, paragraaf, etc. (voorzover van toepassing)
110	onderwerp (trefwoorden; kantlijntekst uit wetboek)
120	tekst van het artikel
190	toelichting: datum laatste wijziging van dit artikel, opmerkingen, etc. In de meeste wetten gebruikt.
191	toelichting: subset van 190, uitsluitend met datum laatste wijziging van dit artikel. In enkele wetten.
192	toelichting: subset van 190, uitsluitend overige opmerkingen (niet onder 191 vallend). In enkele wetten.
200	nummer van het artikel in cijfers
201	aanduiding van het type regeling: Wet, AMvB, etc.
202	datum van publikatie: Staatsblad ....
210	volgnummer van de wet in de databank, in cijfers

Tabel 3-1 Sectie-indeling NLEX documenten

Voor een wetgevingsdatabank is deze indeling in documenten, en de manier waarop ieder document is opgebouwd bijzonder praktisch. Zij zorgt er namelijk voor dat ieder afzonderlijk artikel gezocht kan worden, terwijl toch het verband tussen de artikelen blijft bestaan doordat ze allemaal vergezeld gaan van de titel van de wet, en indien van toepassing zelfs van de plaats (hoofdstuk, titel, paragraaf) binnen die wet waaruit ze afkomstig zijn. Zo kunnen ook de andere artikelen uit dezelfde wet gemakkelijk worden teruggevonden. Voor taalkundig onderzoek is de opbouw van de databank minder ideaal. Het laat zich raden dat zeer veel tekstgedeelten die in de oorspronkelijke wettekst slecht één keer voorkomen, nu tientallen zo niet honderden malen zijn opgenomen; denk bijvoorbeeld aan de naam van een wet met enkele honderden artikelen. Een eerste taak is dan geweest ook om uit de databank een tekstbestand samen te stellen waaruit dubbele tekstgedeelten zijn verwijderd, en dat ook overigens zoveel mogelijk gelijkenis vertoont met de oorspronkelijke wetteksten. Een bijkomend voordeel hiervan is natuurlijk dat de omvang van het materiaal is teruggebracht, van ongeveer 280 megabytes tot iets minder dan 100 megabytes. Een vereiste is wel dat bij het nieuwe tekstbestand een index wordt aangelegd, waardoor het mogelijk wordt elk van de 5875 wetten apart op te vragen, bijvoorbeeld voor het aanleggen van woordenlijsten per wet.

Aangezien het de bedoeling is om dit wetgevingsbestand uiteindelijk te integreren in een juridisch totaal-zoeksysteem, zoals in de voorgaande paragraaf en in hoofdstuk 1 beschreven, is tenslotte nog besloten om een mogelijkheid te creëren om onderdelen van wetten apart op te vragen uit het nieuwe bestand (en niet alleen de integrale tekst van een wet). Om de documentgroottes zoveel mogelijk in overeenstemming te brengen met de richtwaarden uit paragraaf 3.2 is er daarbij voor gekozen om de wetten niet meer in artikelen in te delen, maar in het eerstvolgende niveau boven dat van de artikelen waaruit een wet is opgebouwd. Voor sommige wetten betekent dit dat de *paragrafen* de nieuwe documenten worden, voor andere dat alle artikelen binnen een *hoofdstuk* moeten worden samengenomen, terwijl voor weer andere er helemaal geen indeling mogelijk is, aangezien die wet geen hoofdstuk- of paragraafindeling kent. De nieuwe indeling komt er op neer dat telkens wanneer in de tekst een sectie nummer 100 wordt aangetroffen (nieuwe titel van hoofdstuk, paragraaf of subparagraaf) wordt begonnen aan een nieuw document. Voorts is besloten om in ieder geval de volledige titel van iedere wet (de zogenaamde intitule), de

aanhef boven de wet (de formele tekst boven een wet in formele zin die begint met "Wij Beatrix, enz. en vervolgens zeer in het kort de inhoud van de wet beschrijft) en de ondertekening onder de wet (de laatste twee onderdelen natuurlijk alleen wanneer ze aanwezig zijn) als aparte documenten te behandelen. Hiervoor zijn verschillende redenen; in de eerste plaats is het dan eenvoudig mogelijk om later aan ieder opgevraagd document de titel van de wet waaruit het afkomstig is toe te voegen, en in de tweede plaats kan dan indien gewenst alle informatie die niet direct betrekking heeft op het juridische onderwerp van de wet (nl. de intitule, de aanhef en de ondertekening) worden weggelaten of apart behandeld.

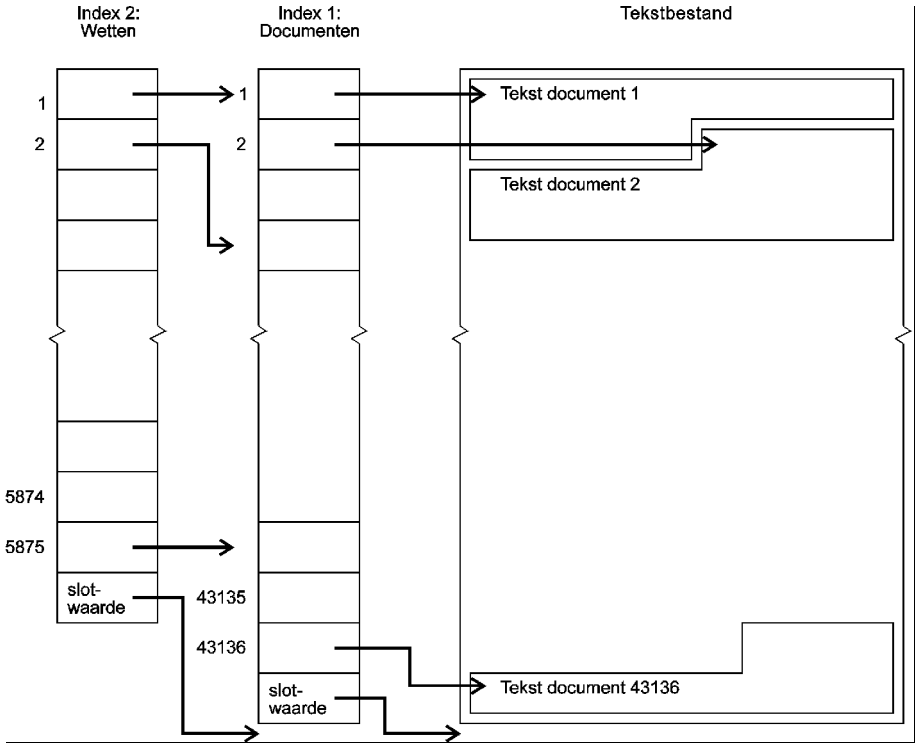
Aan de hand van bovenstaande criteria is een programma geschreven dat in staat is de oorspronkelijke databank-records te lezen, en een nieuw tekstbestand op te bouwen. Dit tekstbestand bevat dus voorzover mogelijk de complete tekst van alle wetten, zonder dat tekstgedeelten dubbel zijn opgenomen. De titel van de wet wordt in een apart document geplaatst, samen met alle andere secties (tekstgedeelten) die voor ieder document in de wet dezelfde inhoud hebben. Het gaat daarbij meestal om de secties 90, 91 en 92 (titel, citeertitel en 'alias'), 201 (aanduiding soort regeling) en 202 (datum publikatie). Besloten is sectie 210 (het volgnummer van de wet in de databank) niet meer op te nemen, aangezien een nieuwe index is samengesteld (zie hierna). Bij het opnemen van secties met nummer 100 (titels van hoofdstuk/paragraaf/subparagraaf) is ervoor gekozen om de in de databank gehanteerde methodiek te handhaven om bijvoorbeeld bij iedere subparagraaf de overige plaats-gegevens (zoals de naam van de paragraaf en het hoofdstuk waar de subparagraaf in thuishoorde) ook op te nemen[7]. Het voordeel hiervan is dat bij ieder apart op te vragen tekstgedeelte direct alle referentiegegevens beschikbaar blijven, zonder dat daarvoor een ingewikkelde index behoeft te worden aangelegd.

Tegelijk met het opbouwen van het bestand met wetteksten is een tweetal indexbestanden gecreëerd (zie afbeelding 3-1).

Het eerste indexbestand bevat voor ieder document 'nieuwe stijl' een getal, dat aangeeft op welke plaats in het tekstbestand (gerekend in lettertekens vanaf het begin van het bestand) de tekst van het document begint. Deze index bevat dus één getal voor elk van de volgende tekstgedeelten:

- de titel van een wet
- de intitule van een wet
- de aanhef van een wet
- een nieuw hoofdstuk, of een nieuwe paragraaf of subparagraaf
- de ondertekening van een wet
- tekst die volgt op de ondertekening van een wet (meestal een bijlage, soms enige referentiegegevens of noten).

[7] In de later ontwikkelde leesprogrammatuur werd een functie ingebouwd om tekstgedeelten die zodoende dubbel werden opgenomen bij het inlezen van wetteksten weer weg te filteren.



Afb. 3-1 Nieuwe indeling NLEX databank

Aan de hand van deze eerste index kan dus de tekst van ieder document 'nieuwe stijl' worden opgezocht en gelezen (er blijken er 43136 te zijn). Een document eindigt daarbij precies één positie voor het beginpunt van het volgende document.

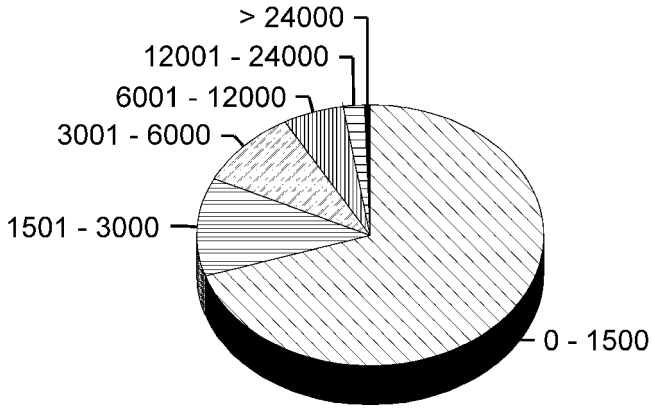
Het tweede indexbestand dat is aangelegd bestaat uit een lijst van getallen (5875 stuks) die aangeven met welk document elk van de wetten in het bestand aanvangt. Dit getal ligt dus in de range 1..43136. Op vergelijkbare wijze als bij de eerste index (een wet eindigt één document voor het document waarmee de volgende begint) kunnen met deze tweede index de nummers van alle documenten waaruit een bepaalde wet bestaat worden opgezocht. Vervolgens kunnen deze documenten met behulp van index nummer 1 uit het tekstbestand worden ingelezen. Op deze manier kan de complete tekst van elke wet afzonderlijk uit de databank worden gelezen. Om ervoor te zorgen dat kan worden uitgerekend waar de tekst van het *laatste document* in het tekstbestand eindigt, en wat het nummer van het laatste document in de *laatste wet* is, is aan elk van de twee indexen tenslotte nog een 'slotwaarde' toegevoegd, een getal gelijk aan het totaal aantal lettertekens in het tekstbestand + 1 respectievelijk het totaal aantal documenten + 1.

3.3.3 Grootte van documenten in het corpus wetgevingsteksten

Aan de hand van de documenten-index (index 1 uit de vorige subparagraaf) is elk van de afzonderlijke documenten uit de databank gelezen, waarbij onder andere de afmetingen - gemeten in aantal lettertekens en in aantal woorden - van elk document zijn beoordeeld. Besloten is daarbij om uit ieder document alleen de wettekst zelf te gebruiken voor verder onderzoek, en niet de gegevens die daaraan later, onder andere door de uitgever, zijn toegevoegd, zoals trefwoorden, noten, etc. De reden hiervoor is dat dergelijke toege-

voegde informatie per uitgave kan verschillen, en derhalve als specifiek voor deze databank-wetteneditie moet worden beschouwd, terwijl de wettekst zelf als een constant gegeven kan gelden. Voor onderzoek naar woordfrequenties zou verder nog nadelig kunnen zijn dat bepaalde woorden - met name woorden die de kern van een bepaald wetsartikel raken - 'kunstmatig' een hogere frequentie zouden krijgen wanneer ze behalve in de tekst zelf ook in een rubriek trefwoorden zouden voorkomen. Uiteindelijk is om die reden besloten alleen de secties 90 (volledige titel), 91 (citeertitel), 92 (korte titel), 100 (hoofdstuk/paragraaf titel) en 120 (tekst wetsartikelen) verder te verwerken. Wanneer in het vervolg van deze paragraaf dan ook over de tekst in een document wordt gesproken moet daaronder worden verstaan: de tekst die voorkomt in de secties 90, 91, 92, 100 en 120 van dat document.

De verwachting was dat de documenten die op deze manier tot stand zijn gekomen in grote lijnen zouden voldoen aan de eisen zoals in paragraaf 3.2 opgesomd, en met name aan de eis dat de documentgrootte (van de in aanmerking genomen tekstgedeelten) zou liggen tussen de 100 en de 500 woorden (ongeveer 600 tot 3000 lettertekens). Dit blijkt echter niet het geval; het kleinste document bevat in het geheel geen tekst - dat is kennelijk een document dat alleen bestaat uit noten, trefwoorden e.d., die zoals gezegd niet meetellen - terwijl het grootste uit bijna 170000 lettertekens (ongeveer 26000 woorden) bestaat. Afbeelding 3-2 geeft de afmetingen van de documenten grafisch weer.



Afb. 3-2 Afmetingen wetgevingsdocumenten in bytes (eerste methode, 43136 documenten)

Te kleine documenten vormen een probleem bij het implementeren van retrieval systemen die gebruik maken van gegevens als het al dan niet aanwezig zijn en de frequentie van woorden. Een documentgrootte van ongeveer 80 à 100 woorden (±500 lettertekens) kan als het absolute minimum worden beschouwd om nog zinvolle uitspraken te kunnen doen over overeenkomsten tussen het woordpatroon van twee documenten. Te grote documenten zijn evenmin ideaal, aangezien de precisie waarmee kan worden gezocht afneemt naarmate het aantal grote documenten toeneemt; een groot document bevat immers naast de gewenste vaak ook veel ongewenste informatie. Het woordpatroon van grote documenten kan echter, anders dan dat van zeer kleine, wel zinvol worden geanalyseerd. Vandaar dat ervoor is gekozen het wetgevingsbestand in te delen in documenten 'nieuwe stijl' door een flink aantal kleine documenten samen te nemen. Daarvoor zijn verschillende methoden beproefd:

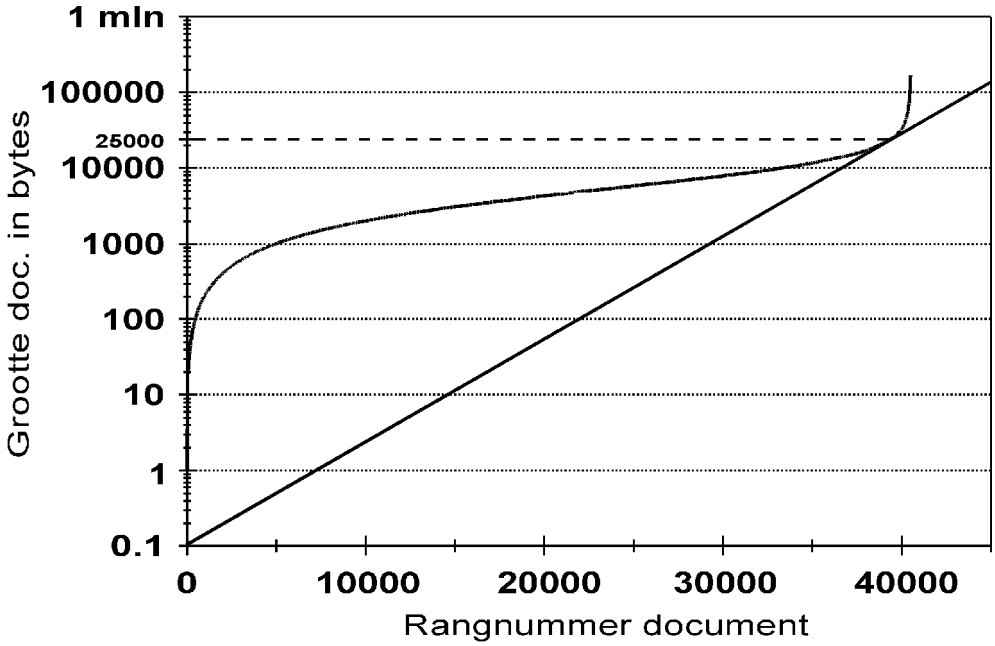
1. De titel van de wet, de intitule, de aanhef en de ondertekening (voorzover deze onderdelen in een wet aanwezig zijn) zijn in het algemeen allemaal documenten van minder dan 80 woorden. Aangezien de inhoud van de meeste van deze onderdelen, met uitzondering van de titel, weinig of geen verband houdt met het juridische onderwerp van de wet in kwestie ligt het voor de hand ze samen te voegen tot één document, dat we de titel 'formeel gedeelte van een wet' zouden kunnen geven. Nadat deze

samenvoeging is uitgevoerd blijkt echter dat er nog steeds te veel kleine documenten (van minder dan 80 woorden) over zijn. Uit bestudering van het tekstbestand blijkt dat het hierbij vaak gaat om zeer korte (sub)paragrafen (bijvoorbeeld bestaande uit 1 wetsartikel), of om hoofdstukken of paragrafen waarvan de tekst geheel of gedeeltelijk 'vervallen' is (in de wet staat dan bijvoorbeeld: "Titel V. Vervallen.").

2. Vervolgens is getracht het aantal kleine documenten te verminderen door documenten 'met de hand' samen te voegen, dat wil zeggen door de documenten uit iedere wet een voor een te bekijken, en wanneer een wet een aantal te kleine documenten bevat deze indien mogelijk samen te voegen tot documenten van het naast hogere niveau. (bijvoorbeeld: een aantal subparagrafen samenvoegen tot een paragraaf). Dit blijkt op zich goed mogelijk, maar de samengestelde documenten die hierbij worden gevormd hebben de neiging nogal groot te worden (meer dan 12000 lettertekens). Vaak is het bijvoorbeeld zo dat paragrafen naast enkele kleine subparagrafen ook een of meer grote subparagrafen blijken te bevatten. Het is in zo'n geval niet erg logisch om drie van de vier subparagrafen samen te voegen tot een nieuw document, en de vierde apart te laten staan. Het proces blijkt ook bijzonder arbeidsintensief te zijn.

3. Tenslotte is besloten af te zien van handmatig ingrijpen, en documenten samen te voegen volgens een strak algoritme. Dit algoritme luidt als volgt:
   - Alle documenten die horen bij wetten die in hun geheel uit minder dan 25000 lettertekens (ongeveer 4000 woorden) bestaan worden zonder meer samengevoegd tot één document. De ratio achter de grens van 25000 is dat wanneer alle wetten op grootte worden gesorteerd er tot dat punt sprake is van een geleidelijke toename in grootte, terwijl bij wetten van meer dan 25000 lettertekens de toename in grootte niet meer geleidelijk verloopt. Zie afbeelding 3-3; het bedoelde punt komt overeen met de plaats waar een rechte lijn door de oorsprong aan de curve raakt. Het ligt dan ook voor de hand om in ieder geval wetten van meer dan 25000 lettertekens in documenten te blijven verdelen.
   - Bij wetten van meer dan 25000 lettertekens worden in ieder geval wetstitel, intitule, aanhef en ondertekening samengevoegd tot één document.
   - Tenslotte worden binnen een wet alle documenten van minder dan 500 lettertekens samengevoegd met het document ervoor. Wanneer echter het *eerste* document van een wet (het samengestelde document met titel, intitule etc.) kleiner is dan 500 lettertekens wordt dit samengevoegd met het *daarop volgende* document.

Een groot voordeel van het gebruiken van een strak algoritme is natuurlijk dat het veel gemakkelijker is om de gegevens up to date te houden, bijvoorbeeld in het geval van wetswijzigingen. Er kan dan worden volstaan met het opnieuw draaien van het computerprogramma dat de documenten-indeling maakt, terwijl in het geval van samenvoeging 'met de hand' nauwkeurig moet worden nagegaan welke wetten zijn gewijzigd en of dat gevolgen heeft voor de opbouw van documenten.



Afb. 3-3 Grafische weergaven van de afmetingen in bytes van de 43136 documenten, gesorteerd op grootte.

'Formele' tekstgedeelten (bijvoorbeeld aanhef, intitule, ondertekening) en 'inhoudelijke' tekstgedeelten worden via deze methode in een aantal gevallen weer samengevoegd. Dit gebeurt om precies te zijn:

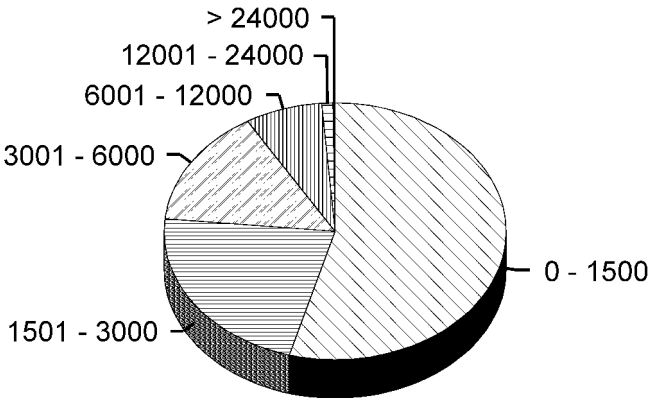
- bij alle wetten met een totale omvang van minder dan 25000 lettertekens, en
- bij alle wetten groter dan 25000 lettertekens, wanneer de formele tekstgedeelten samen minder dan 500 lettertekens tellen. In dat geval worden ze samengevoegd met het eerste inhoudelijke document uit de betreffende wet.

In ieder geval worden alle formele documenten uit een en dezelfde wet altijd tot een enkel document samengevoegd. Aangezien het formele gedeelte van een wet, ook wanneer dat niet met inhoudelijke documenten wordt samengevoegd, toch altijd nog een geringe hoeveelheid inhoudelijke tekst bevat, met name in de vorm van de titel van de wet, kan dus worden gesteld dat de formele documenten die via bovengenoemde methode ontstaan altijd enige inhoudelijke tekst bevatten, en de inhoudelijke documenten soms enige formele tekst. Dat op deze manier formele en inhoudelijke documenten in een aantal gevallen weer vermengd raken is een nadeel. Het voordeel van een geringer aantal documenten, die qua afmetingen beter geschikt zijn voor gebruik in retrieval systemen weegt hiertegen in dit geval echter op.

Een ander gevolg van deze methode dat apart vermelding verdient is dat soms documenten zullen worden samengevoegd die logisch niet goed bij elkaar passen. Zo is het mogelijk dat de eerste paragraaf van het tweede hoofdstuk van een wet, wanneer die uit minder dan 500 lettertekens bestaat, bij hoofdstuk 1 (of bij de laatste paragraaf daarvan, wanneer die een apart document vormt) wordt gevoegd. Dit kan natuurlijk een nadeel zijn, met name wanneer in het eerste hoofdstuk een totaal ander onderwerp wordt behandeld dan in het tweede. De enige manier om dit te voorkomen is om toch handmatig in te grijpen, maar dat brengt, zoals hierboven aangegeven, weer andere nadelen met zich mee. Uiteindelijk is besloten om, mede gezien de geringe hoeveelheid tekst die de

documenten in kwestie bevatten in verhouding tot de afmetingen van het gehele corpus, de nadelen voor lief te nemen, en het hierboven omschreven algoritme onverkort toe te passen.

Bij samenvoeging van documenten volgens het derde algoritme zijn uiteindelijk 18803 documenten ontstaan. De afmetingen van deze documenten worden grafisch weergegeven in afbeelding 3-4. Zoals uit deze grafiek blijkt is het aantal kleine documenten drastisch afgenomen. Er zijn nog steeds enkele documenten met minder dan 80 woorden, maar het gaat daarbij altijd om documenten die een complete wet omvatten (en dus niet met andere kunnen worden samengevoegd). Er zijn ook nog enkele zeer grote documenten; het grootste bevat zelfs bijna 26.000 woorden, dat is ruim 50 bladzijden. Het beste zou natuurlijk zijn deze documenten te splitsen. Voor dit splitsen is echter geen logisch punt in de wet aanwezig; er is gewoon sprake van een grote hoeveelheid tekst zonder hoofdstuk- of paragraafindeling. Daarom is besloten in het kader van dit onderzoek de grote documenten voorlopig te laten zoals ze zijn.



Afb. 3-4 Afmetingen wetgevingsdocumenten in bytes (tweede methode, 18803 documenten)

	Oorspronkelijk (artikelen)	Nieuw (paragrafen)	Nieuw, kleine doc. samengevoegd	Wetten
Corpus:				
Grootte in bytes	132026193	87993972	87993972	87993972
Totaal aantal tokens	18786675	12807659	12807659	12807659
Totaal aantal typen	154651	143156	143156	143156
Aantal documenten	129015	43136	18803	5875
Grootste document:				
Grootte in bytes	38528	169848	170064	838863
Aantal tokens	5874	25794	25825	114490
Aantal typen	1363	2688	2690	6152
Kleinste document:				
Grootte in bytes	88	0	32	190
Aantal tokens	9	0	4	25
Aantal typen	8	0	4	23
Gemiddeld:				
Grootte in bytes	1023,34	2039,92	4679,78	14977,70
Aantal tokens	145,62	296,91	681,15	2180,03
Aantal typen	77,83	104,83	210,58	432,94
Standaarddeviatie:				
Grootte in bytes	1034,68	4411,24	6457,37	31472,91
Aantal tokens	148,80	640,26	933,69	4602,25
Aantal typen	46,39	124,60	167,42	447,04

Tabel 3-2 Statistische gegevens corpus wetgevingsteksten

Tabel 3-2 bevat een aantal statistische gegevens betreffende de documenten zoals deze uiteindelijk zijn vastgesteld. Bij de cijfers in de kolommen 2 t/m 4 is al rekening gehouden met het weglaten van de tekstgedeelten die niet tot de oorspronkelijke wettekst kunnen worden gerekend, en die daarom in het vervolg van het onderzoek ook buiten beschouwing zullen blijven; zie daarover ook paragraaf 3.6.1.

In paragraaf 3.6 zal de werkwijze bij het vervaardigen van woordenlijsten uit dit en de andere twee corpora worden besproken, waarna in hoofdstuk 4 het woordgebruik in de wetgevingsteksten verder zal worden geanalyseerd. In dit hoofdstuk zal nu eerst de samenstelling van de andere twee corpora die in dit onderzoek worden gebruikt worden beschreven.

3.4 Het corpus jurisprudentieteksten

3.4.1 Inleiding

Voor juristen spelen, naast teksten van wetgeving, teksten van jurisprudentie een belangrijke rol. Hoewel Nederland geen precedentenstelsel kent vormen de uitspraken van de Hoge Raad en van lagere rechters een belangrijke rechtsbron, aangezien juristen daaruit kunnen opmaken hoe bepaalde wettelijke regels in de praktijk geïnterpreteerd worden. Dit is met name van belang wanneer in die regels door de wetgever - al dan niet met opzet - vage termen zijn gebruikt of zaken ongeregeld zijn gebleven. Verder vormt de jurisprudentie een belangrijke bron van nieuw recht in gevallen dat de wet door de praktijk is achterhaald. Knottenbelt en Torringa[8] schrijven over jurisprudentie:

> "Na de wet is de jurisprudentie de belangrijkste rechtsbron. (...) Wij weten intussen dat niet alle recht in de wet staat en dat ook de rechter rechtsregels vormt. Regels van geldend recht die niet in de wet staan worden door de rechtspraak geformuleerd wanneer ze in een aan de rechter voorgelegd geschil ter sprake komen."

Het ligt daarom voor de hand om in een onderzoek als dit plaats in te ruimen voor de analyse van teksten van rechterlijke uitspraken. Een probleem is daarbij echter wel dat het aantal rechterlijke uitspraken die in Nederland dagelijks worden opgesteld buitengewoon groot is, zodat daaruit noodzakelijkerwijs een keus moet worden gemaakt. Een benadering om eenvoudigweg alle jurisprudentie die in een bepaald tijdvak door rechters is 'geproduceerd' in de analyse te betrekken is alleen al vanwege de omvang van het materiaal eenvoudig niet haalbaar.

Een goed alternatief is echter gelegen in het beperken van het tekstmateriaal tot alleen die uitspraken die in een jurisprudentie-tijdschrift zijn gepubliceerd, mede met het argument dat in de praktijk alleen deze uitspraken ter kennis komen van juristen 'in den lande'. Uiteindelijk is in dit onderzoek gekozen voor het verwerken van die uitspraken die zijn gepubliceerd in het tijdschrift "Nederlandse Jurisprudentie", jaargangen 1965 t/m 1989. De voornaamste reden voor deze keuze is dat dit de meest brede en gezaghebbende

[8] Knottenbelt en Torringa (1986, p. 49).

selectie van rechterlijke uitspraken in Nederland is. Daarnaast speelde bij de keuze een rol dat deze uitspraken in elektronische vorm beschikbaar zijn, namelijk in de vorm van de Kluwer Juridische Bibliotheek. Dit is een verzameling van publikaties op juridisch gebied, die door uitgeverij Kluwer wordt aangeboden op een aantal CD-ROM schijven, in combinatie met een full-text zoeksysteem[9]. Het verwerken van de informatie uit deze 'databank'[10] leverde wel enige problemen op. Hierop zal in de volgende paragraaf kort worden ingegaan.

3.4.2 Het lezen van de NJ-databank

Bij het lezen van de teksten uit de NJ-databank is geen gebruik gemaakt van de door Kluwer geleverde zoekprogrammatuur (Status), omdat deze voor ons doel te traag zou zijn en niet beschikt over enkele functies die nodig zijn voor het onderzoek. De NJ-databank zoals deze in dit onderzoek is gebruikt is bijgewerkt tot en met december 1989 (dat wil zeggen dat de hele jaargang 1989 er nog in is opgenomen), en bestaat uit twee bestanden, die samen op een enkele CD-ROM schijf zijn opgeslagen. Het eerste bestand bevat alle uitspraken die gepubliceerd zijn in "Nederlandse Jurisprudentie" tussen 1 januari 1965 en 31 december 1975, het tweede de uitspraken die gepubliceerd zijn tussen 1 januari 1976 en 31 december 1989. Voor elk van deze twee bestanden is op de CD-ROM bovendien een index-bestand (een zogenaamde "inverted file"[11]) opgenomen. Deze index-bestanden zijn niet gebruikt in het kader van dit onderzoek.

De twee bestanden met jurisprudentieteksten blijken een "record-structuur" te bezitten, dat wil zeggen dat ze zijn opgebouwd uit een (groot) aantal elementen (records) die dezelfde afmeting hebben, en niet uit regels tekst (die natuurlijk een steeds wisselende lengte hebben). Het voordeel van een dergelijke record-structuur is dat ook tekstgedeelten die ergens midden in het bestand staan direct kunnen worden opgeroepen (wanneer tenminste bekend is in welk record ze beginnen), een nadeel is dat voor het lezen speciale programmatuur noodzakelijk is, die zorg draagt voor het opzoeken van de gewenste uitspraak, en bij het lezen de oorspronkelijke tekst-layout weer herstelt. De manier waarop de databank-records precies zijn opgebouwd wordt door Kluwer niet gepubliceerd, en moest dus eerst worden geanalyseerd. Om op een flexibele manier arresten uit beide bestanden te kunnen lezen en verwerken is vervolgens een serie procedures in de hogere programmeertaal Pascal geschreven. Met behulp hiervan is ook een nieuwe index samengesteld, aan de hand waarvan de computer elke gewenste uitspraak in het bestand kan opzoeken.

---

9 Kluwer Juridische Bibliotheek op CD-ROM, Kluwer Datalex, Deventer 1989.

10 Een databank is een gegevensverzameling, waarin in dit geval de volledige tekst van documenten is opgeslagen. De gegevens in de databank kunnen worden geraadpleegd met behulp van een geautomatiseerd zoeksysteem.

11 Met de term 'inverted file' wordt bedoeld een lijst van alle woordtypen (alle verschillende woorden) in een bepaald bestand, waarin voor ieder woordtype in het algemeen bovendien de frequentie (het aantal malen dat het type voorkomt in het bestand) en de pagina's (of de documenten, afhankelijk van hoe het bestand is ingedeeld) waarop het voorkomt worden vermeld.

Het bestand leek in eerste instantie 16539 uitspraken te bevatten. Elk van die uitspraken is, net zoals in het geval van de wetgevingsdatabank, ingedeeld in secties. Tabel 3-3 bevat een opsomming van deze secties en hun inhoud.

Naam sectie	Inhoud van de sectie
nummer	jaartal en volgnummer in Nederlandse Jurisprudentie (NJ 1980 nr 10)
rechter	aanduiding van rechtscollege dat deze uitspraak wijst
kop	zeer korte aanduiding van rechtsvraag in deze uitspraak
cursief	samenvatting van casus (soms) en uitspraak
artikel	wetsartikel(en) waar de uitspraak met name over gaat
tekst	volledige tekst van de uitspraak
mening	conclusie van Advocaat Generaal / Procureur Generaal (indien van toepassing)
noot	annotatie door rechtsgeleerd schrijver / deskundige

Tabel 3-3 Sectie-indeling NJ documenten

3.4.3 Fouten in de databank

Nadat de software voor het lezen van de NJ-bestanden was voltooid, is met behulp daarvan allereerst getracht na te gaan in hoeverre de bestanden compleet waren, dat wil zeggen of ze alle uitspraken bevatten die ook in de "papieren" versie zijn verschenen. Hiertoe werd een lijst samengesteld bestaande uit alle "nummers" (combinaties van jaartal en NJ-nummer) van de uitspraken uit de twee bestanden. Deze lijst werd vervolgens nauwkeurig geanalyseerd om na te gaan of misschien nummers ontbraken, of dat bepaalde nummers meer dan één keer voorkwamen. Inderdaad werd hierbij een aantal inconsistenties geconstateerd, met name in het bestand met uitspraken vanaf 1976. Deze inconsistenties zijn in te delen in de volgende categorieën:

- dubbele uitspraken. Een aantal uitspraken (97 stuks) bleek fysiek twee of zelfs drie maal in de databank voor te komen. Naderhand heeft Kluwer verklaard dat dit in de meeste gevallen met opzet is gebeurd, omdat deze uitspraken zouden vallen onder meer dan één van de "delen" (o.a. burgerlijk recht, strafrecht) waarin de databank is ingedeeld.

- uitspraken met één of meer fouten. Enkele uitspraken (in totaal 10 stuks) bleken voorzien van een onjuist NJ-nummer of een onjuist jaartal. Aan de hand van de tekst van de uitspraak en de gedrukte versie van de NJ is nagegaan wat het juiste nummer is. Vervolgens is ook de door de nieuwe inlees-software gebruikte index gecorrigeerd.

- ontbrekende uitspraken. Een aantal uitspraken bleek niet in de databank opgenomen (in totaal 14 stuks).

De geconstateerde fouten werden aan Kluwer gerapporteerd. Dit heeft ertoe geleid dat de meeste fouten in een latere editie van de databank (er verschijnt enkele malen per jaar een nieuwe versie) zijn hersteld. Tevens stelde Kluwer de teksten van de ontbrekende 14 uitspraken in machine-leesbare vorm ter beschikking. Deze zijn in het kader van dit onderzoek aan de databank toegevoegd, terwijl de dubbele uitspraken zijn verwijderd. Daarna is verder gewerkt met dit 'gecorrigeerde' bestand. Alle gegevens met betrekking tot het corpus jurisprudentieteksten die in het vervolg van dit hoofdstuk en in de hoofd-

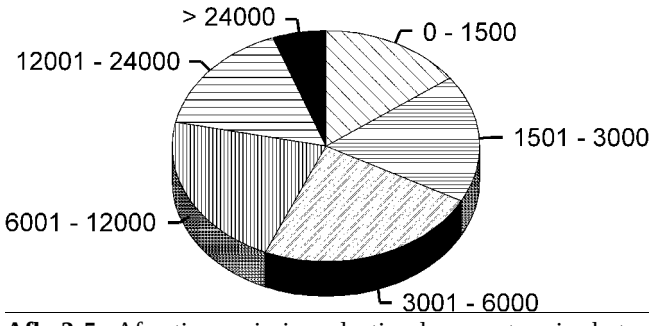
stukken hierna worden vermeld hebben betrekking op de gecorrigeerde versie, met uitzondering van de eerste kolom van tabel 3-4, waar ter vergelijking (en voor de volledigheid) enkele statistische gegevens van het oorspronkelijke, ongecorrigeerde bestand zijn opgenomen. Na correctie en hertelling bleek het corpus jurisprudentieteksten 16430 rechterlijke uitspraken te bevatten.

Bedacht moet worden dat op deze manier natuurlijk niet alle fouten die mogelijkerwijs in de databank voorkomen konden worden gevonden en gecorrigeerd. Fouten die niet blijken uit de nummering van uitspraken komen niet aan het licht. Wanneer bijvoorbeeld de nummers van twee uitspraken verwisseld zouden zijn zou dit niet zijn opgemerkt. Aangezien het nog steeds zou gaan om dezelfde teksten zou een dergelijke fout voor dit onderzoek geen gevolgen hebben; voor een jurist die de uitspraak aan de hand van het nummer zou opvragen zou dat natuurlijk anders liggen. Een fout die in theorie wel gevolgen voor dit onderzoek zou kunnen hebben zou zijn wanneer de *tekst* van een of meer uitspraken onjuistheden zou bevatten, zoals bijvoorbeeld het ontbreken of het dubbel aanwezig zijn van een bepaalde passage. Wanneer bepaalde woordtypen in zo'n passage vaak zouden voorkomen en/of wanneer de passage omvangrijk zou zijn, zou dit de totaaltellingen kunnen beïnvloeden. Gezien het grote aantal uitspraken is echter niet te verwachten dat een dergelijke beïnvloeding ooit significant zou kunnen zijn.

3.4.4 Grootte van documenten in het corpus jurisprudentieteksten

In het vervolg zullen we wanneer we het over de inhoud van de jurisprudentiedatabank hebben de algemene term 'documenten' gebruiken, waarbij ieder document de tekst van één rechterlijke uitspraak bevat; het aantal documenten in het gecorrigeerde corpus jurisprudentieteksten bedraagt dus 16430.

De afmetingen van deze documenten lopen nogal uiteen, maar niet in dezelfde mate als in eerste instantie het geval was bij het corpus wetgevingsteksten (zie paragraaf 3.3.3 bovenaan). In afbeelding 3-5 worden deze afmetingen grafisch weergegeven.



Afb. 3-5 Afmetingen jurisprudentie documenten in bytes (16430 documenten)

Duidelijk zichtbaar in de grafiek is dat de groep van de allerkleinste documenten (tot 1500 lettertekens) veel minder omvangrijk is dan in het geval van het wetgevingscorpus (of we nu vergelijken met de eerste of de tweede documenten-indeling van dat corpus). Ook is zichtbaar dat de spreiding van documenten over de grootte-klassen veel gelijkmatiger is - er komen bijvoorbeeld ongeveer evenveel documenten voor met afmetingen tussen de 3000 en de 6000 lettertekens als met afmetingen tussen de 6000 en de 12000 lettertekens, terwijl deze laatste groep in het geval van de wetgevingsteksten slechts half zo groot is als de eerstgenoemde.

Het kleinste document bevat 66 lettertekens, het grootste ruim 220000 (11 resp. ruim 34000 woorden). Wanneer deze aantallen worden getoetst aan de eisen in paragraaf 3.2, net als in het geval van het corpus wetgevingsteksten, blijkt dat de kleinste documenten zo goed als zeker te weinig woorden bevatten om te kunnen worden gebruikt in een retrieval systeem dat gebruik maakt van woordfrequenties. Het is in dit geval echter, anders dan bij het corpus wetgevingsteksten, niet goed mogelijk om deze kleine documenten met andere samen te voegen; anders dan in het geval van wetgevingsteksten gaat het in dit corpus immers om volkomen op zichzelf staande tekstgedeelten, die met de documenten die eraan vooraf gaan en die erop volgen juridisch-inhoudelijk waarschijnlijk heel weinig gemeen hebben. Hoe vervelend het daarom ook is, we zullen moeten leven met deze onvolmaaktheid; het is niet goed mogelijk om uit het corpus documenten samen te stellen die optimale afmetingen hebben voor gebruik met een op woordfrequenties gebaseerd retrieval-systeem, zoals in paragraaf 3.2 omschreven. Het gevolg zal waarschijnlijk zijn dat de informatie in deze kleine documenten niet met een dergelijk retrieval systeem kan worden ontsloten.

	Oorspronkelijk	Gecorrigeerd
Corpus:		
Grootte in bytes	302059358	261604695
Totaal aantal tokens	46212418	40190212
Totaal aantal typen	283903	265841
Aantal documenten	16539	16430
Grootste document:		
Grootte in bytes	235253	221171
Aantal tokens	35218	34117
Aantal typen	4318	3842
Kleinste document:		
Grootte in bytes	0	66
Aantal tokens	0	11
Aantal typen	0	10
Gemiddeld:		
Grootte in bytes	18263,46	15922,38
Aantal tokens	2794,15	2446,15
Aantal typen	654,34	599,11
Standaarddeviatie:		
Grootte in bytes	19239,06	17148,29
Aantal tokens	2946,22	2631,22
Aantal typen	450,40	414,75

Tabel 3-4 Statistische gegevens corpus jurisprudentieteksten

De grootste documenten leveren mogelijk ook wat problemen op (zie de opmerkingen die hierover werden gemaakt in paragraaf 3.3.3). Toch werd besloten niet tot splitsing van deze arresten over te gaan, aangezien daarvoor eigenlijk geen logisch punt in de tekst aanwezig is. In ieder geval is in de grafiek zichtbaar dat ruim een derde van alle documenten niet groter is dan een bladzijde (ongeveer 3000 lettertekens), en dat meer dan driekwart afmetingen heeft die de 12000 lettertekens (ongeveer 4 bladzijden) niet te boven gaan.

Het merendeel van de documenten voldoet dus redelijk tot goed aan de in paragraaf 3.2 genoemde afmetingen. Tabel 3-4 bevat nog enige aanvullende statistische gegevens met betrekking tot de documentgroottes. In de kolom 'gecorrigeerd' is, net als bij de wetgevingsteksten, al rekening gehouden met de tekstgedeelten die buiten het onderzoek blijven (zie paragraaf 3.6.1).

3.5 Het corpus 'algemeen Nederlands'

Een van de voornaamste doelstellingen van dit onderzoek is om na te gaan in hoeverre er verschil bestaat tussen het woordgebruik in (verschillende soorten) juridische teksten en in niet-juridische teksten. Om hierover uitspraken te kunnen doen is het noodzakelijk om de juridische corpora te kunnen vergelijken met een verzameling andere teksten van voldoende omvang. Bij het zoeken naar een dergelijk corpus met vergelijkingsmateriaal is er bovendien naar gestreefd dat alle teksten tezamen representatief zouden zijn (voorzover mogelijk) voor hedendaags Nederlandse schrijftaal. In eerder onderzoek door De Mulder en Oskamp[12] werd gebruik gemaakt van gegevens uit het rapport 'Woordfrequenties in geschreven en gesproken Nederlands'[13], het 'Eindhovens corpus'. De gegevens in dit rapport vormen weliswaar een behoorlijke dwarsdoorsnede van Nederlands taalgebruik, maar zijn intussen ook tamelijk gedateerd: ze zijn afkomstig van teksten uit de jaren 1960-1973.

Vandaar dat werd besloten gebruik te maken van nieuw vergelijkingsmateriaal. Hiertoe werd in de loop van 1991 contact opgenomen met het Instituut voor Nederlandse Lexicologie (INL) in Leiden, waar onder andere gewerkt wordt aan het opbouwen en up to date houden van diverse elektronische tekstcorpora, waaronder enkele corpora met hedendaags Nederlandse teksten. Dit resulteerde uiteindelijk in het beschikbaar stellen, begin 1993, van een door dit instituut samengesteld corpus hedendaags Nederlandse schrijftaal, bestaande uit ruim 4,6 miljoen woorden (tokens). Het corpus omvat onder andere romans, korte verhalen, studieboeken en artikelen, zowel wetenschappelijk als niet-wetenschappelijk, over uiteenlopende onderwerpen.

Het corpus bestond in eerste instantie uit in totaal 403 documenten. Een lijst met de titels daarvan is te vinden in bijlage 1. Twee documenten, te weten de documenten die in de bijlage vermeld staan onder de nummers 368 en 387, bleken bijzonder omvangrijk (761464 resp. 862817 lettertekens, ofwel 119318 resp. 125641 woordtokens). Om praktische redenen (de woordenlijst voor ieder document moest in het werkgeheugen van de gebruikte computer passen) werd besloten deze documenten te splitsen, het eerste in twee en het tweede in drie stukken. Door die correctie bevat het corpus dus nu 406 documenten.

Met name bij een corpus als het onderhavige, dat bedoeld is als een dwarsdoorsnede van Nederlandse schrijftaal, dient de nodige aandacht te worden geschonken aan de samenstelling. In eerste instantie ligt voor de hand om als maatstaf voor de kwaliteit c.q. bruik-

[12] De Mulder (1979).
[13] Uit den Boogaart (1975).

baarheid van een dergelijk corpus te hanteren de representativiteit, ofwel de mate waarin de teksten waaruit het bestaat in hun eigenschappen de te beschrijven verzameling, in dit geval dus alle geschriften in de Nederlandse taal, vertegenwoordigen. Martin (1988, p. 92) betoogt echter op overtuigende wijze dat het samenstellen van een representatieve steekproef uit de Nederlandse taal in feite niet mogelijk is, niet eens in de eerste plaats vanwege de *oneindigheid* van de populatie, die het moeilijk maakt om alle elementen een gelijke kans te geven in de steekproef vertegenwoordigd te zijn, maar vooral vanwege de *heterogeniteit* ervan. Volgens Martin kent het Nederlands op ten minste vijf verschillende wijzen variatie: regionaal, sociaal, individueel of communicatief-situationeel, inhoudelijk en temporeel[14]. De overgangen tussen varianten verlopen daarbij veelal geleidelijk, sommige varianten van het Nederlands en van bijvoorbeeld het Duits liggen zeer dicht tegen elkaar aan. Daardoor is het zo goed als onmogelijk de precieze begrenzingen van de populatie Nederlands vast te stellen. Om aselecte steekproeven uit een (taal)populatie te kunnen trekken, teneinde een representatief corpus op te bouwen, is het echter noodzakelijk dat alle elementen die tot die populatie behoren gedefinieerd kunnen worden. Martin concludeert dat om die reden geen representatief corpus van het Nederlands kan worden samengesteld. Het is volgens hem echter wel mogelijk een *exemplarisch* corpus op te bouwen, dat een afspiegeling is van bepaalde vormen van Nederlands taalgebruik. Het hoeft hierbij dus niet te gaan om een 'algemeen' corpus, het is evengoed mogelijk dat gekozen wordt voor een corpus waarin bijvoorbeeld in het geheel geen regionale variatie voorkomt. De gewenste opbouw van het corpus kan worden bepaald door op bovengenoemde vijf taalvariatie-assen (de regionale as, de sociale as, de communicatieve situatie-as, de inhoudsas en de tijdsas) *referentiepunten* te kiezen, en vervolgens teksten ten opzichte van deze punten in te delen. Uit de op die manier ingedeelde teksten kan vervolgens een keuze worden gemaakt, op basis van de positie die een tekst inneemt ten opzichte van de referentiepunten. Door de juiste referentiepunten te kiezen kan op die manier een corpus met de gewenste eigenschappen worden samengesteld, met dezelfde natuurlijke geleding als de taal zelf.

Hoewel het te ver zou gaan om alle 403 teksten uit het door het INL aangeleverde corpus volgens de door Martin aangegeven criteria in te delen langs de 5 variatie-assen en te evalueren aan de hand van referentiepunten, zal hieronder wel worden getracht om na te gaan in hoeverre teksten uit het corpus verschillende waarden op elk van deze assen vertegenwoordigen, met andere woorden, in hoeverre de diverse vormen van taalkundige variatie ook in het corpus terug te vinden zijn.

Op de *regionale as* bevindt zich aan de ene kant de standaardtaal, terwijl aan de andere kant de diverse dialecten kunnen worden geplaatst. Met betrekking tot deze as moet worden geconcludeerd dat het grootste gedeelte van het corpus bestaat uit standaardtaal. Dialect is in feite niet vertegenwoordigd. Enkele van de korte verhalen, bijvoorbeeld tekstnummers 307 en 308, bevatten een hoeveelheid 'volkstaal' (enigszins plat taalgebruik). Deze tekstgedeelten zouden waarschijnlijk ergens in het midden van de regionale as kunnen worden geplaatst.

[14] Zie ook hoofdstuk 2, paragraaf 5.

Op de *sociale as* is de samenhang tussen taal(gebruik) en maatschappelijke factoren af te lezen. De punten op deze as drukken de sociale afstand tussen taalgebruikers uit. In het geval van literaire teksten bestaat de indeling op deze as uit een schatting van het 'niveau' van de auteur. Hier zijn zowel hoge als lage waarden in het corpus vertegenwoordigd (hoewel over de kwalificatie zeker discussie mogelijk is):
naast een verhaal van Maarten 't Hart is ook een stuk van Heere Heeresma opgenomen, dat door sommigen als pornografisch zal worden ervaren. Voor wat betreft algemene en vaktaal kan worden vastgesteld dat het corpus zowel populair-wetenschappelijke stukken van wat minder bekende auteurs bevat als werken van erkende experts (veelal wetenschappers) op een vakgebied.

Op de *communicatieve situatie-as* kunnen verschillen in taalgebruik die veroorzaakt worden door de context of situatie waarin dat taalgebruik plaatsvindt in kaart worden gebracht. Het criterium voor indeling van teksten op deze as is de mate van *algemeenheid* van het taalgebruik: van zeer algemeen, tot zeer specifiek en alleen in bepaalde situaties bruikbaar. Ook hier zijn diverse waarden in het corpus vertegenwoordigd. Zeer algemeen taalgebruik is bijvoorbeeld te vinden in het werk over wetenschapsbeoefening aan de Vrije Universiteit (teksten 19 t/m 31). Wat formeler, meer aan de situatie aangepast taalgebruik bevatten de gedenkboeken voor Koningin Wilhelmina en Dries van Agt (teksten 65 t/m 95 resp. 148 t/m 161). Zeer specifiek taalgebruik, weliswaar in nauwe samenhang met het onderwerp van de tekst, is bijvoorbeeld te vinden in werken over oppervlaktebehandelingen van constructiestaal (tekst 370), en over diverse sporten, zoals de duivensport, karate en snooker (teksten 369, 391 en 402).

Op de *inhoudsas* worden teksten ingedeeld naar gelang van de inhoud, het onderwerp waar ze over gaan. Een specifiek onderwerp vereist soms een specifieke woordenschat. Het corpus bevat werken met zeer verschillende inhoud, en over onderwerpen die een zeer specifieke woordenschat vereisen. Te denken valt bijvoorbeeld aan de werken over sporten, maar ook aan die over sociale geografie, taalkunde en autokennis.

Met behulp van de *tijdsas* kunnen verschillen in taalgebruik op verschillende momenten in de tijd worden zichtbaar gemaakt. Voor het corpus geldt dat de periode die wordt omvat door de verschillende teksten eruit vrij beperkt is, namelijk van 1972 t/m 1988.

Uit het bovenstaande kan de conclusie worden getrokken dat de teksten waaruit het corpus algemeen Nederlands bestaat langs drie van de vijf door Martin genoemde variatie-assen een duidelijke spreiding vertonen. Deze variatie langs de sociale as, de communicatieve situatie-as en de inhoudsas is gezien het beoogde 'algemene' karakter van het corpus te billijken. Regionale variatie (o.a. gebruik van dialect) en temporele variatie zijn slechts in geringe mate aanwezig. Wat het eerste betreft, ook van de juridische corpora mag verwacht worden dat deze weinig regionale variatie zullen vertonen, zodat het corpus algemeen Nederlands in dit opzicht ongeveer dezelfde geleding heeft. De geringe mate van temporele variatie vormt misschien een nadeel, aangezien de periode die door de twee juridische corpora wordt beslagen aanzienlijk langer is. Anderzijds zal nu bij vergelijking tussen de corpora zogenaamd archaïsch taalgebruik, waarvan zich naar verwachting vooral in de juridische corpora voorbeelden bevinden, wellicht duidelijker kunnen worden aangetoond. Alles bijeen genomen kan worden geconcludeerd dat het corpus algemeen Nederlands weliswaar misschien geen representatief, maar dan toch in ieder geval een

bruikbaar exemplarisch corpus van algemeen Nederlands taalgebruik is, dat geschikt is voor de vergelijking op woordniveau met de juridische corpora zoals die in het vervolg van dit onderzoek zal worden uitgevoerd.

Gezien het bijzondere karakter van het corpus, en met name gezien het feit dat het de bedoeling was om het corpus te behandelen als zijnde exemplarisch voor 'algemeen Nederlandse' schrijftaal, werd besloten om geen gebruik te maken van de afzonderlijke documenten, maar bij het vervaardigen van woordenlijsten en het uitvoeren van vergelijkingen uitsluitend uit te gaan van het gehele corpus. Volledigheidshalve worden hieronder nog wel de belangrijkste statistische gegevens van het corpus vermeld, ook van de afzonderlijk documenten.

Corpus:	
Grootte in bytes	30889416
Totaal aantal tokens	4616455
Totaal aantal typen	142879
Aantal documenten	406
Grootste document:	
Grootte in bytes	656836
Aantal tokens	99194
Aantal typen	10215
Kleinste document:	
Grootte in bytes	3179
Aantal tokens	502
Aantal typen	215
Gemiddeld:	
Grootte in bytes	76082,31
Aantal tokens	11370,58
Aantal typen	2020,66
Standaarddeviatie:	
Grootte in bytes	115955,98
Aantal tokens	17631,14
Aantal typen	1869,37

Tabel 3-5 Statistische gegevens corpus algemeen Nederlands

3.6 Het samenstellen van woordenlijsten

3.6.1 Het vervaardigen van een alfabetische woordenlijst

Nadat de drie corpora waren ontsloten en de inhoud tot op zekere hoogte was gecontroleerd en waar nodig gecorrigeerd (met name in het geval van het corpus jurisprudentieteksten), werd vervolgens een computerprogramma geschreven waarmee alfabetische woordenlijsten van alle afzonderlijke documenten in de twee juridische corpora vervaardigd konden worden, en vervolgens een alfabetische totaal-woordenlijst van elk van de drie corpora. In deze lijst werd bij ieder woordtype tevens vermeld in hoeveel documenten het voorkwam, en wat de totale frequentie van het type binnen het corpus was. Bij de

twee juridische corpora werden, zoals in de paragrafen 3.3 en 3.4 beschreven, alleen die gedeelten van ieder document in aanmerking genomen die kunnen worden gerekend tot de originele tekst. Zo werden bij de wetgevingsteksten alleen de volgende secties opgenomen:

- 90 volledige titel van wet/regeling
- 91 (officiële) citeertitel
- 92 alias (korte naam of afkorting, bijvoorbeeld "WIR", "BW")
- 100 hoofdstuk/paragraaf/subparagraaf titel
- 120 tekst van het wetsartikel c.q. de regeling

De belangrijkste secties die werden overgeslagen zijn de secties met trefwoorden en noten (nrs. 110 en 190 t/m 192). Bij de jurisprudentieteksten werden opgenomen:

- nummer jaartal en volgnummer in Nederlandse Jurisprudentie
- rechter aanduiding van het rechtscollege dat de uitspraak wijst
- tekst volledige tekst van de uitspraak
- mening conclusie van Advocaat Generaal / Procureur Generaal (indien aanwezig)

Hier zijn met name de z.g.n. kop (in de gedrukte uitgaven van de Nederlandse Jurisprudentie vet weergegeven), het cursieve gedeelte bovenaan iedere uitspraak en een eventuele noot eronder de opvallendste onderdelen die werden overgeslagen. Deze tekstgedeelten, die een samenvatting van de belangrijkste punten uit de uitspraak resp. commentaar op de beslissing van de rechter bevatten, zijn door de redactie van de Nederlandse Jurisprudentie toegevoegd.

Het samenstellen van de woordenlijst vond voor ieder corpus plaats in twee fasen:

1. eerst werd het tekstbestand doorgelezen, en werd voor ieder document afzonderlijk een lijst van de in dat document aangetroffen woordtypen aangelegd. Bij ieder type werd ook de woordfrequentie in het document in kwestie opgenomen. Deze afzonderlijke lijsten werden (nog ongesorteerd) onder elkaar in één bestand geplaatst.
2. vervolgens werd het totale bestand van ieder corpus dat in de eerste fase was aangelegd alfabetisch-lexicografisch gesorteerd, waarbij typen die meer dan een keer bleken voor te komen slechts een maal werden opgenomen, terwijl tegelijkertijd de woordfrequenties werden opgeteld. Tevens werd bij ieder type opgeteld in hoeveel documenten het was aangetroffen.

De manier waarop in de eerste fase te werk werd gegaan bij het splitsen van de teksten van documenten in afzonderlijke woordtokens verdient nog enige aandacht. Het algoritme dat daarvoor werd gebruikt luidt (in woorden) als volgt:

1. lees een regel uit het document
2. zoek de eerste spatie; alles tot aan die spatie is het eerste woord
3. maak van iedere letter van dit woord een hoofdletter; verwijder alle tekens die geen letter (A..Z) of cijfer (0..9) zijn uit het woord
4. wis bij lange woorden alle tekens na het 30e teken

5. ga na of het woord al eerder voorkwam in het document. Zo ja, verhoog de frequentie met 1, zo nee, voeg het woord toe aan de lijst.
6. verwijder het nu verwerkte woord uit de ingelezen regel. Ga als de regel nog meer woorden bevat terug naar 2. Ga anders, wanneer nog niet alle regels van het document gelezen zijn, terug naar 1.

Gevolgen van het gebruik van dit algoritme zijn onder meer:

- het verschil tussen hoofd- en kleine letters vervalt.
- wanneer een corpus woorden zou bevatten die aan het einde van een regel zijn afgebroken, zouden deze niet op de juiste manier worden verwerkt. Bij het programmeren van de lees-procedures was echter al duidelijk geworden dat de corpora geen afgekorte woorden bevatten. Regels bleken altijd te worden afgebroken bij een spatie.
- Twee woorden met een verbindingsstreepje (-) of een schuine streep (/) ertussen worden aan elkaar geplakt, waarbij de streep wordt verwijderd. Twee woorden die in de tekst bij elkaar horen, maar waartussen een spatie staat worden daarentegen afzonderlijk verwerkt, waarbij het verband tussen de woorden verloren gaat.
- Zeer lange woorden worden afgekapt na het dertigste letterteken. Dit is gedaan om de geheugenruimte die nodig is voor het opslaan en sorteren van de woordenlijsten te beperken. Op deze lange woorden zal hieronder worden teruggekomen.
- De woorden worden opgeslagen in precies dezelfde vorm als waarin ze in de tekst voorkomen. Er wordt dus geen poging gedaan om woorden tot een stamvorm of lemma te herleiden, onder andere vanwege de te verwachten semantische problemen (zie ook hoofdstuk 1).

Gegevens met betrekking tot het aantal woorden (tokens) en woordtypen in elk van de drie corpora zijn te vinden in de voorgaande drie paragrafen, in de tabellen 3-2, 3-4 en 3-5.

3.6.2 Lange woorden

Door het gebruikte algoritme voor het vervaardigen van woordenlijsten werden zoals hierboven al aangegeven woorden langer dan 30 lettertekens 'afgekapt'. Natuurlijk zou het beter zijn om dit achterwege te laten, maar dan zouden problemen zijn ontstaan bij het samenstellen en sorteren van de lijsten, in verband met de hoeveelheid beschikbaar werkgeheugen. Een mogelijk nadeel van deze werkwijze is natuurlijk dat woordtekens waarvan de eerste 30 tekens overeenkomen, maar die voorbij dat punt van elkaar verschillen ten onrechte als verschijningsvormen van een en hetzelfde woordtype worden aangemerkt. Dit heeft zich inderdaad enkele keren voorgedaan. In de meeste gevallen ging het dan om het enkelvoud en het meervoud van hetzelfde woord hetgeen niet al te ernstig lijkt. In enkele gevallen (8 maal in het corpus jurisprudentieteksten en 15 maal in het corpus wetgevingsteksten) werden door het afkappen echter wel verschillende woorden aan elkaar gelijk gemaakt, zoals bijvoorbeeld in het geval van de woorden *aansprakelijkheidsverzekeringsovereenkomst* en *aansprakelijkheidsverzekeringspolis* (corpus jurisprudentie, de woorden komen 5x respectievelijk 4x voor in het corpus), en van de woorden *natriumcarbonaat-waterstofcarbonaatbuffer* en *natriumcarbonaat-water-*

stofcarbonaatbufferoplossing[15] (corpus wetgeving, de woorden komen beiden 1x voor in het corpus). In totaal werden bij de diverse corpora zo'n 934 woordtypen ingekort. Dat lijken er op het eerste gezicht nog aardig wat, maar in geen van de drie corpora is het percentage ingekorte typen hoger dan 0,25%. Het totale aantal woord*tokens* van meer dan 30 letters ligt zelfs bij geen van de drie corpora boven de 0,03% (en bij de corpora jurisprudentie en algemeen Nederlands zelfs nog een factor 10 lager). Het gaat dus om relatief zeer weinig woorden. Bijlage 2 bevat een lijst van alle woorden uit de drie corpora van meer dan 30 lettertekens. Woorden die uit een samenvoeging van twee of meer andere woorden bestaan zijn niet in de lijst opgenomen wanneer tussen de samenstellende woorden in het oorspronkelijke corpus verbindingsstreepjes of schuine strepen staan. Slechts enkele lange woorden, zoals *aansprakelijkheidsverzekeringen, arbeidsongeschiktheidsuitkering* en *arbeidsongeschiktheidsverzekering* komen in alle drie de corpora voor.

3.6.3 Het sorteren op woordfrequentie

Met behulp van de alfabetische woordenlijsten van ieder corpus, waarvan de samenstelling in 3.6.1 werd beschreven, werden vervolgens met behulp van een sorteerprogramma per corpus twee op frequentie gesorteerde lijsten samengesteld. Deze lijsten bevatten dus exact dezelfde gegevens als de alfabetische lijsten, alleen waren de woordtypen op een andere volgorde geplaatst. In de eerste lijst werden de woordtypen gesorteerd op het aantal documenten binnen het corpus waarin zij voorkwamen, en in de tweede lijst op de totale frequentie waarmee het woordtype voorkwam in het gehele corpus. Om van de resultaten alvast een eerste indruk te geven zijn in tabel 3-6 voor alle drie corpora de eerste 75 typen uit de op frequentie gesorteerde lijsten weergegeven. Bij de twee juridische corpora zijn sommige woorden gemarkeerd. De betekenis daarvan is als volgt:

- woorden die **vet** zijn afgedrukt hebben in het betreffende corpus duidelijk een hoger rangnummer (dat wil zeggen, ze komen vaker voor) dan in de andere twee corpora.
- woorden die ***vet en cursief*** zijn afgedrukt (dat zijn er maar een paar) komen in het betreffende corpus duidelijk minder vaak voor dan in de andere twee corpora.

De markeringen die in de tabel zijn aangebracht hebben nog een zeer voorlopig karakter; ze zijn gebaseerd op eerste waarnemingen van alleen de 'kop' van de frequentielijsten (de woorden met de hoogste frequenties) van de drie corpora. In de volgende drie hoofdstukken zullen de corpora afzonderlijk verder worden geanalyseerd, waarna in hoofdstuk 7 het woordgebruik in alledrie zal worden vergeleken. Daarbij zal dan duidelijk worden of de hier waargenomen zaken passen binnen een grote lijn van overeenkomsten en verschillen tussen de drie corpora.

---

[15] Met name in het corpus wetgevingsteksten komt een aanmerkelijk aantal namen van chemische stoffen en verbindingen voor, met name afkomstig uit allerlei KB's waarin voorschriften voor de chemische industrie of voor de produktie van bepaalde goederen worden gegeven. Deze zijn in de lijst van woorden langer dan 30 tekens sterk vertegenwoordigd.

Rangnr	Wetgeving	Jurisprudentie	Alg. Nederlands
1	DE	DE	DE
2	VAN	VAN	VAN
3	HET	HET	HET
4	IN	IN	EEN
5	EN	DAT	EN
6	EEN	EN	IN
7	**OF**	EEN	IS
8	VOOR	TE	DAT
9	OP	IS	TE
10	IS	OP	OP
11	MET	NIET	ZIJN
12	TE	**HEEFT**	DIE
13	ART	AAN	VOOR
14	AAN	DOOR	MET
15	**LID**	ZIJN	NIET
16	**ARTIKEL**	MET	ALS
17	WORDEN	VOOR	AAN
18	DIE	DIE	OOK
19	**1**	TOT	OF
20	WORDT	ALS	WORDEN
21	BIJ	BIJ	ER
22	ZIJN	WORDEN	DOOR
23	DOOR	ART	DAN
24	***DAT***	DEZE	DEZE
25	TOT	OF	BIJ
26	**2**	**HOF**	OM
27	***NIET***	DIT	MAAR
28	DEZE	**RB**	TOT
29	ALS	OOK	WORDT
30	**WET**	TEN	HIJ
31	**EERSTE**	**ZOU**	KAN
32	**MINISTER**	KAN	DIT
33	TEN	OM	ZICH
34	**BEDOELD**	UIT	MEN
35	DAN	GEEN	NAAR
36	INDIEN	NAAR	ZE
37	DER	DAN	UIT
38	**3**	1	KUNNEN
39	ONDER	WAS	OVER
40	DIT	HAAR	HEEFT
41	KAN	**BEROEP**	IK
42	A	ZICH	MEER
43	**ONZE**	ONDER	WAS
44	**GEMEENTE**	HIJ	WAT
45	B	HEBBEN	NOG
46	**BESLUIT**	ZIJ	WE
47	**TOEPASSING**	**TEGEN**	HEBBEN
48	**4**	WORDT	ZIJ
49	TWEEDE	2	ZO
50	**STB**	**MIDDEL**	MOET
51	HEEFT	TER	WEL
52	WELKE	DER	HUN
53	**NA**	LID	HAAR
54	**BEPAALDE**	**CASSATIE**	ANDERE
55	**DEN**	F	GEEN
56	**BEDOELDE**	**VONNIS**	JE
57	TER	KUNNEN	AL
58	HEBBEN	WELKE	ZAL
59	UIT	HAD	ZOU
60	C	MOET	TUSSEN
61	HAND	***IK***	U
62	**5**	OVER	VEEL
63	**RAAD**	**RECHT**	NU
64	ZIJ	ER	WERD
65	ANDERE	3	MOETEN
66	NAAR	A	ALLEEN
67	**I**	ZAL	ZOALS
68	HIJ	**GEVAL**	ONDER
69	BETREKKING	**GROND**	EERSTE
70	**II**	EERSTE	EIGEN
71	**KRACHTENS**	O	MAKEN
72	ZAL	**ARREST**	ECHTER
73	MEER	P	WAAR
74	**MOET**	HEM	HAD
75	OVER	NOG	NA

Tabel 3-6 De 75 meest voorkomende woordtypen uit de drie corpora

HOOFDSTUK 4 - DE BELANGRIJKSTE KWANTITATIEVE EIGENSCHAPPEN VAN HET CORPUS WETGEVINGSTEKSTEN

4.1 Inleiding

In dit hoofdstuk zullen de voor dit onderzoek van belang zijnde kwantitatieve aspecten van het corpus wetgevingsteksten, waarvan de samenstelling in het vorige hoofdstuk is beschreven, worden geanalyseerd. Daarbij zal het steeds gaan om eigenschappen van het corpus in zijn geheel; aan specifieke kenmerken van bepaalde groepen documenten uit het volledige corpus wetgevingsteksten zal in hoofdstuk 8 enige aandacht worden besteed. Vergelijking met de andere twee corpora (jurisprudentie en algemeen Nederlands) vindt plaats in hoofdstuk 7, nadat deze corpora in de komende twee hoofdstukken eerst afzonderlijk zijn onderzocht.

4.2 Absolute en relatieve woordfrequenties

De voornaamste basisgegevens betreffende het woordgebruik in een corpus zijn te vinden in een frequentielijst, waarin alle in dat corpus voorkomende woordtypen met hun absolute en hun relatieve frequentie worden vermeld. De relatieve frequentie is daarbij gelijk aan de absolute frequentie gedeeld door het aantal tokens waaruit het corpus bestaat.

Woordfrequentielijsten gerangschikt naar frequentie, het meest voorkomende woord het eerst, het minst voorkomende het laatst, bevatten gewoonlijk ook de *rangnummers* van de woordtypen. Deze rangnummers kunnen als volgt worden toegekend:

- het meest voorkomende woordtype krijgt rangnummer 1, het op één na meest voorkomende rangnummer 2, het volgende rangnummer 3 etc.;
- wanneer twee of meer woordtypen even vaak voorkomen delen ze hetzelfde rangnummer, waarna evenzoveel rangnummers worden overgeslagen alvorens de typen die volgen hun nummer krijgen. Anders gesteld: de typen die volgen na een aantal typen met dezelfde frequentie krijgen het rangnummer dat ze ook zouden krijgen wanneer de daaraan voorafgaande typen *niet* dezelfde frequentie zouden hebben.

Het rangnummer geeft duidelijk aan welke plaats een bepaald type in de frequentielijst van een corpus heeft, en kan daarnaast een rol spelen bij het trekken van vergelijkingen tussen corpora en bij het grafisch weergeven en vergelijken van frequentiegegevens (bijvoorbeeld van de frequenties van alle typen, zoals verderop in deze paragraaf).

Tegenwoordig worden voorts in frequentielijsten meestal ook gegevens met betrekking tot de *spreiding* van de woordtypen over het corpus opgenomen. Als maat voor de spreiding van woordtypen wordt in dit geval gehanteerd het percentage van de documenten uit het corpus waarin het type voorkomt. Martin (1990, p. 1318) noemt ook andere maten voor spreiding,

zoals bijvoorbeeld de 'Usage Coefficient' die werd ontwikkeld door Juilland[1] en de daarin door Carroll aangebrachte verbeteringen[2]. De subcorpora waarvoor met deze maten de spreiding kan worden berekend moeten echter aan zekere eisen voldoen. In het geval van de coëfficiënt van Juilland dienen zij ongeveer dezelfde omvang (in woordtokens) te hebben, in het geval van die van Carroll dienen zij een bepaalde minimum-omvang te hebben (dit laatste om de waarschijnlijkheid van het voorkomen van een type in een subcorpus met voldoende nauwkeurigheid te kunnen vaststellen). Aan geen van beide eisen kan met betrekking tot het corpus wetgevingsteksten worden voldaan, tenminste wanneer we de spreiding over de 18803 *documenten* van het corpus zouden willen bepalen. Deze documenten zijn namelijk allerminst gelijk van omvang, terwijl er ook documenten van zeer geringe omvang - vanaf 4 typen - voorkomen (zie hoofdstuk 3). Daarom zal bij het vervaardigen van de frequentielijsten in dit geval geen gebruik worden gemaakt van dergelijke meer geavanceerde spreidingsmaten[3].

Het bovenstaande in aanmerking nemend, bevatten de frequentielijsten die in het kader van dit onderzoeksproject worden opgesteld de volgende gegevens:
- het rangnummer
- het woordtype zelf
- de absolute frequentie (het aantal malen dat het type voorkomt in het gehele corpus)
- de relatieve frequentie (absolute frequentie gedeeld door aantal tokens in het corpus, maal 100%)
- het spreidingspercentage, ofwel het percentage van alle documenten in het corpus waarin het type te vinden is.

Een kopie van de eerste pagina van de op deze wijze ingedeelde frequentielijst van het corpus wetgevingsteksten wordt hier ter illustratie opgenomen (zie tabel 4-1).

De volledige frequentielijst van het corpus wetgevingsteksten is omvangrijk. Het corpus telt 143156 woordtypen, zodat de lijst op papier al ruim 954 pagina's zou beslaan, wanneer we ervan uitgaan dat 150 typen op een pagina kunnen worden afgedrukt. Ook wanneer alle woorden met frequentie 1 (de 'hapax legomena') worden weggelaten, resteren nog altijd meer dan 530 pagina's. Dit geldt voor de andere corpora evenzeer.

---

1 Juilland and Chang-Rodriguez (1964), Juilland *et al.* (1970).
Binnen deze coëfficiënt wordt in feite de woordfrequentie gecorrigeerd voor de spreiding van het type over in het corpus aanwezige subcorpora, zodanig dat wanneer twee woorden dezelfde frequentie hebben, het woord dat in ieder subcorpus even vaak voorkomt een hogere rang krijgt dan het woord dat ongelijk verdeeld is over de subcorpora. Als maat voor de spreiding wordt de variatie-coëfficiënt - de standaarddeviatie van de frequenties in de subcorpora gedeeld door de gemiddelde frequentie van het type in deze subcorpora - gebruikt.

2 Carroll (1970, p 61-65).
Carroll meet onder andere de spreiding over subcorpora op een andere manier, waardoor met name verschil in omvang tussen deze subcorpora een veel geringere rol speelt dan bij Juilland. De door Carroll voorgestelde coëfficiënt levert met name meer realistische waarden op voor woordtypen die zeer ongelijk over de subcorpora zijn verdeeld.

3 Bij de andere twee corpora zal dat evenmin het geval zijn, aangezien ook de documenten in het corpus jurisprudentieteksten niet aan bovengenoemde eisen voldoen, en het niet gewenst is om voor het (vergelijkings)corpus algemeen Nederlands andere frequentielijsten op te stellen dan voor de overige twee corpora.

Rangnr.	Woordtype	Abs. freq.	Rel. freq.	In % doc.
1	DE	899333	7,0218%	98,21%
2	VAN	759046	5,9265%	98,79%
3	HET	437151	3,4132%	95,44%
4	N	316151	2,4685%	94,82%
5	EN	285753	2,2311%	94,13%
6	EEN	236772	1,8487%	85,03%
7	OF	157635	1,2308%	79,06%
8	VOOR	140172	1,0944%	83,18%
9	OP	129700	1,0127%	83,41%
10	IS	115722	0,9035%	80,69%
11	MET	114068	0,8906%	85,03%
12	TE	108198	0,8448%	76,56%
13	ART	101843	0,7952%	82,09%
14	AAN	98641	0,7702%	78,99%
15	LID	96108	0,7504%	66,98%
16	ARTIKEL	92743	0,7241%	70,87%
17	WORDEN	92634	0,7233%	80,42%
18	DIE	91737	0,7163%	73,78%
19	1	86409	0,6747%	83,88%
20	WORDT	85716	0,6693%	74,58%
21	BIJ	85185	0,6651%	78,48%
22	ZIJN	85099	0,6644%	69,48%
23	DOOR	82302	0,6426%	69,13%
24	DAT	79173	0,6182%	73,56%
25	TOT	75861	0,5923%	70,90%
26	2	71135	0,5554%	80,56%
27	NIET	59673	0,4659%	62,37%
28	DEZE	59344	0,4633%	64,89%
29	ALS	56037	0,4375%	63,87%
30	WET	53805	0,4201%	49,76%
31	EERSTE	52625	0,4124%	58,17%
32	MINISTER	52053	0,4064%	55,75%
33	TEN	51294	0,4005%	59,97%
34	BEDOELD	46458	0,3627%	54,01%
35	DAN	46330	0,3617%	54,27%
36	INDIEN	45425	0,3547%	52,70%
37	DER	45375	0,3543%	50,19%
38	3	44630	0,3485%	68,25%
39	ONDER	41964	0,3276%	54,91%
40	DIT	37869	0,2957%	52,96%
41	KAN	37854	0,2956%	53,71%
42	A	37705	0,2944%	52,92%
43	ONZE	36202	0,2827%	38,62%
44	GEMEENTE	35580	0,2778%	14,38%
45	B	34734	0,2712%	51,01%
46	BESLUIT	34240	0,2673%	39,38%
47	TOEPASSING	32289	0,2521%	48,01%
48	4	29428	0,2298%	57,49%
49	TWEEDE	28855	0,2253%	47,35%
50	STB	27163	0,2121%	37,93%
51	HEEFT	26615	0,2078%	41,01%
52	WELKE	25900	0,2022%	41,84%
53	NA	25639	0,2002%	46,51%
54	BEPAALDE	25407	0,1984%	42,91%
55	DEN	25255	0,1972%	10,89%
56	BEDOELDE	24786	0,1935%	38,95%
57	TER	23536	0,1838%	42,69%
58	HEBBEN	23325	0,1821%	49,26%
59	UIT	23103	0,1804%	41,00%
60	C	22676	0,1771%	42,04%
61	HAND	22040	0,1721%	15,95%
62	5	21918	0,1711%	48,05%
63	RAAD	21487	0,1678%	31,17%
64	ZIJ	21016	0,1641%	37,42%
65	ANDERE	20277	0,1583%	39,11%
66	NAAR	19604	0,1531%	33,86%
67	I	19234	0,1502%	26,47%
68	HIJ	18875	0,1474%	26,60%
69	BETREKKING	18764	0,1465%	34,65%
70	II	18535	0,1447%	22,58%
71	KRACHTENS	18462	0,1441%	31,99%
72	ZAL	17892	0,1397%	39,55%
73	MEER	17316	0,1352%	32,39%
74	MOET	17247	0,1347%	24,95%
75	OVER	17161	0,1340%	31,81%

Rangnr.	Woordtype	Abs. freq.	Rel. freq.	In % doc.
76	F	16944	0,1323%	28,49%
77	6	16760	0,1309%	40,53%
78	KUNNEN	16703	0,1304%	35,19%
79	ALGEMENE	16466	0,1286%	30,65%
80	ZAKEN	15796	0,1233%	30,18%
81	ARTIKELEN	15670	0,1223%	34,53%
82	DERDE	15338	0,1198%	32,90%
83	DAG	14794	0,1155%	30,38%
84	BINNEN	14727	0,1150%	28,59%
85	ENZ	14426	0,1126%	24,35%
86	WEL	14346	0,1120%	30,94%
87	GEEN	14079	0,1099%	31,42%
88	BEDRAG	13741	0,1073%	18,39%
89	NR	13471	0,1052%	15,74%
90	OM	13441	0,1049%	28,94%
91	WAAROP	13385	0,1045%	30,29%
92	DATUM	13322	0,1040%	20,45%
93	ZOVER	13291	0,1038%	29,04%
94	BESTJUR	13234	0,1033%	21,30%
95	AANZIEN	13057	0,1019%	30,00%
96	JAAR	13041	0,1018%	22,23%
97	D	12920	0,1009%	29,08%
98	7	12879	0,1006%	34,98%
99	HEM	12809	0,1000%	24,84%
100	INGEVOLGE	12604	0,0984%	26,58%
101	LEDEN	12523	0,0978%	18,69%
102	WIJ	12497	0,0976%	27,06%
103	WAARIN	12402	0,0968%	33,65%
104	GEMEENTEN	12281	0,0959%	9,68%
105	GRENS	12236	0,0955%	3,55%
106	ZICH	12127	0,0947%	26,99%
107	VERSTAAN	12045	0,0940%	35,20%
108	10	11689	0,0913%	29,64%
109	DAARVAN	11606	0,0906%	30,57%
110	ALLE	11603	0,0906%	31,27%
111	8	11282	0,0881%	31,70%
112	BEPALINGEN	11100	0,0867%	29,51%
113	PLAATS	11014	0,0860%	25,81%
114	GROND	10920	0,0853%	24,37%
115	ALSMEDE	10862	0,0848%	29,56%
116	SECTIE	10813	0,0844%	2,31%
117	WAARVAN	10749	0,0839%	29,86%
118	MOETEN	10516	0,0821%	18,35%
119	HUN	10481	0,0818%	22,17%
120	VERZOEK	10437	0,0815%	19,34%
121	JANUARI	10253	0,0801%	18,92%
122	STELLEN	10209	0,0797%	27,02%
123	WERKING	10163	0,0794%	31,95%
124	BEHANDELING	9568	0,0780%	18,71%
125	PER	9974	0,0779%	16,55%
126	STATENGENERAAL	9932	0,0775%	13,71%
127	COMMISSIE	9911	0,0774%	10,08%
128	UITVOERING	9899	0,0773%	32,23%
129	STATE	9894	0,0773%	22,15%
130	DIENST	9809	0,0766%	13,15%
131	HOOFDSTUK	9772	0,0763%	27,11%
132	E	9745	0,0761%	22,79%
133	GEBIED	9682	0,0756%	10,71%
134	BESCHIKKING	9590	0,0749%	17,37%
135	TUSSEN	9523	0,0744%	15,86%
136	BEKEND	9518	0,0743%	11,05%
137	INGANG	9513	0,0743%	26,12%
138	GENOEMDE	9436	0,0737%	23,13%
139	REGELING	9433	0,0737%	17,72%
140	PERSONEN	9393	0,0733%	18,02%
141	GEVAL	9334	0,0729%	23,90%
142	9	9326	0,0728%	27,41%
143	TIJDSTIP	9322	0,0728%	20,49%
144	ADVIES	9292	0,0726%	16,74%
145	VOORSCHRIFTEN	9274	0,0724%	19,81%
146	STAATSBLAD	9252	0,0722%	26,55%
147	GEGEVEN	9231	0,0721%	29,06%
148	VORIGE	9166	0,0716%	19,49%
149	RECHT	9135	0,0713%	15,26%
150	WIJZE	9093	0,0710%	24,61%

Tabel 4-1 Eerste pagina van de frequentielijst van het corpus wetgevingsteksten

Het is daarom zaak na te gaan welk deel van de frequentielijsten de interessantste informatie-ubevat, zowel met het oog op de bestudering van de afzonderlijke corpora als op het maken van onderlinge vergelijkingen. Een eerste gedachte is dat het bovenste gedeelte van iedere frequentielijst (de 'kop', met de meest voorkomende woordtypen) waarschijnlijk de meeste aandacht verdient. Dit deel heeft in het algemeen het meest 'robuuste' karakter, in die zin dat wèlke typen erin voorkomen - met name door de hogere woordfrequenties - veel minder aan fluctuatie onderhevig is bij toe- of afnemende corpusgrootte, hetgeen het vergelijken van corpora eenvoudiger maakt. Zoals aan het eind van het vorige hoofdstuk bleek, bieden zelfs de eerste 75 woordtypen uit de corpora al perspectieven voor onderlinge vergelijking. Punt van overweging is verder dat de 'staart' (het onderste gedeelte met de woorden die 1 of 2 keer voorkomen) zeer lang is - het corpus wetgevingsteksten bevat 63303 hapax legomena - en zich moeilijk laat inkorten op basis van inhoudelijke criteria of frequentiegegevens. Typen die slechts 1 keer voorkomen in een enkel document uit het corpus bieden ook niet echt een stabiele basis voor de vergelijking met andere corpora. Dit alles leidt tot de conclusie dat het waarschijnlijk het beste is om met name de aandacht te richten op - een gedeelte van - de 'kop' van de frequentielijsten. De vraag is dan nog: welk gedeelte, tot en met welk rangnummer, komt hiervoor dan in aanmerking?

De Mulder en Oskamp gingen op dit punt in eerste instantie uit van de relatieve frequentie van de typen in de lijst[4]; zij bestudeerden alleen de typen die 0,01% of meer van de afzonderlijke corpora uitmaakten (in de meeste gevallen ging het daarbij om zo'n 1000 woordtypen). Verder maakten zij gebruik van Zipf-karakteristieken[5], welke in het geval van de door hen bestudeerde strafrechtelijke corpora een 'kop' vertoonden die zich uitstrekte over de eerste 50 tot 70 rangnummers, en een 'middenstuk' dat ongeveer tot rangnummer 300 doorliep. Zij bestudeerden met name de kop, en verbonden aan het feit dat dit gedeelte in het geval van de strafrechtelijke corpora korter was dan bij de gebruikte vergelijkingscorpora enkele (voorzichtige) conclusies met betrekking tot het karakter van juridisch taalgebruik (een kleiner aantal woorden wordt vaker gebruikt). Hoewel met name de waarde van Zipf-karakteristieken enigszins ter discussie staat (zie hoofdstuk 2, paragraaf 3), representeert het kenmerk 'kop', ontleend aan deze karakteristieken, in ieder geval gegevens uit het corpus zelf (in dit geval: rangnummers en relatieve frequenties), wat te verkiezen lijkt boven het aanleggen van criteria als 'de eerste X typen' of 'alle frequenties boven 1000'. In deze lijn doordenkend ligt het daarom voor de hand om na te gaan of uit gegevens, afkomstig uit het corpus wetgevingsteksten, afgeleid kan worden welk gedeelte van de frequentielijst van dit corpus met name interessante informatie bevat c.q. nadere bestudering verdient, zonder nochtans gebruik te maken van de Zipf-grootheid van woordfrequentie vermenigvuldigd met rangorde.

---

4 De Mulder en Oskamp (1979, p. 36).

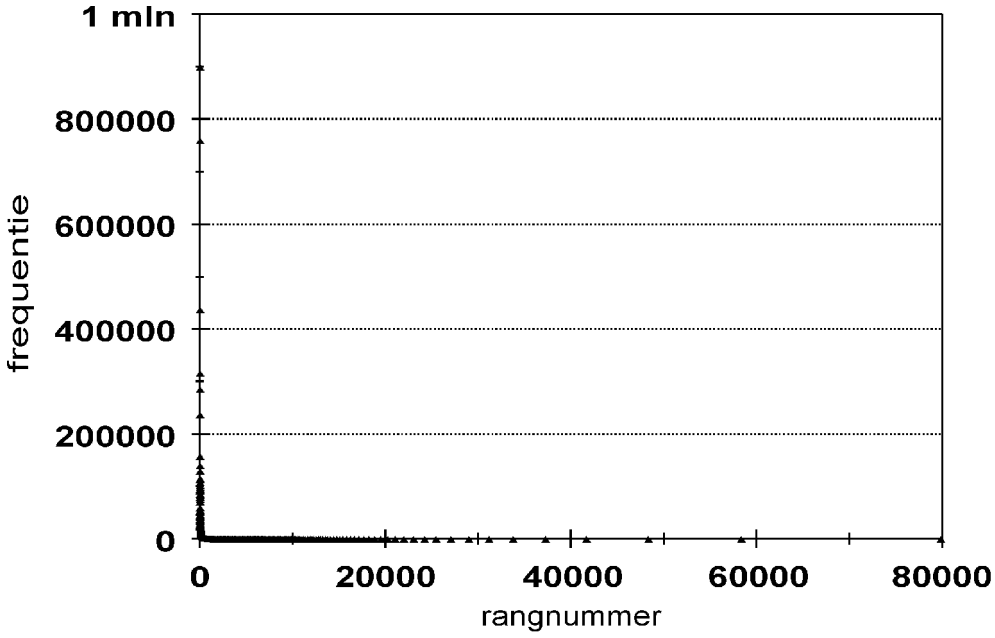
5 Met Zipf-karakteristieken worden hier bedoeld grafieken waarin het rangnummer van woordtypen is afgezet tegen het produkt van rangnummer en relatieve frequentie. Wanneer de 'Wet van Zipf' voor een bepaald tekstgedeelte voor 100% zou opgaan zou de Zipf-karakteristiek van dat tekstgedeelte er uitzien als een horizontale rechte lijn. Zie ook hoofdstuk 2, paragraaf 3.

Allereerst kan daarbij gedacht worden aan gegevens uit de frequentielijst zelf. Informatie die daaruit kan worden verkregen betreft met name:
1. de (absolute of relatieve) woordfrequenties, en
2. de spreiding van typen over de documenten van het corpus.

4.2.1 Woordfrequenties

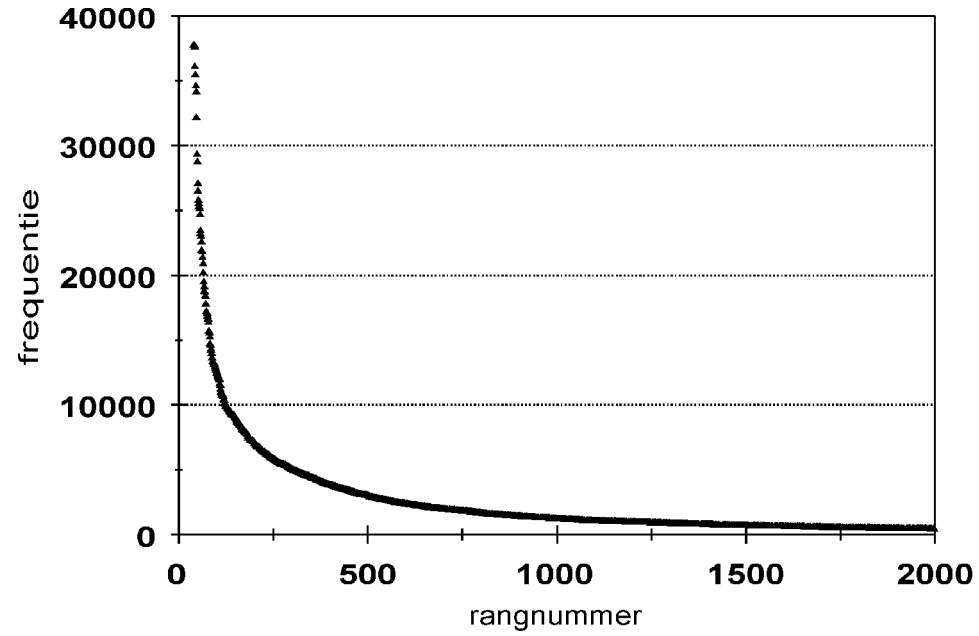
Het eerste wat opvalt wanneer we naar de *woordfrequenties* kijken, is dat deze met name in het begin zeer snel afnemen (zie tabel 4-1). Het woordtype met rangnummer 1 (DE) heeft een frequentie van 899333, terwijl het woordtype met rangnummer 20 (WORDT) nog maar 85716 keer voorkomt, een frequentie die meer dan een factor 10 lager is. Bij rangnummer 150 liggen de frequenties rond de 9000, wederom een factor 10 lager. De daling is dus in het begin zeer scherp, maar neemt ook vrij snel in omvang af. Voor de relatieve frequenties geldt hetzelfde.

De frequenties behorend bij de diverse rangnummers zijn in grafiek 4-1a in beeld gebracht. Deze grafiek laat zich echter moeilijk lezen, doordat de range van de twee variabelen erg groot is, terwijl de waarden die deze variabelen aannemen allerminst gelijkmatig over die range zijn verdeeld. Het lijkt daardoor of alle waarden op de x- en de y-as liggen.



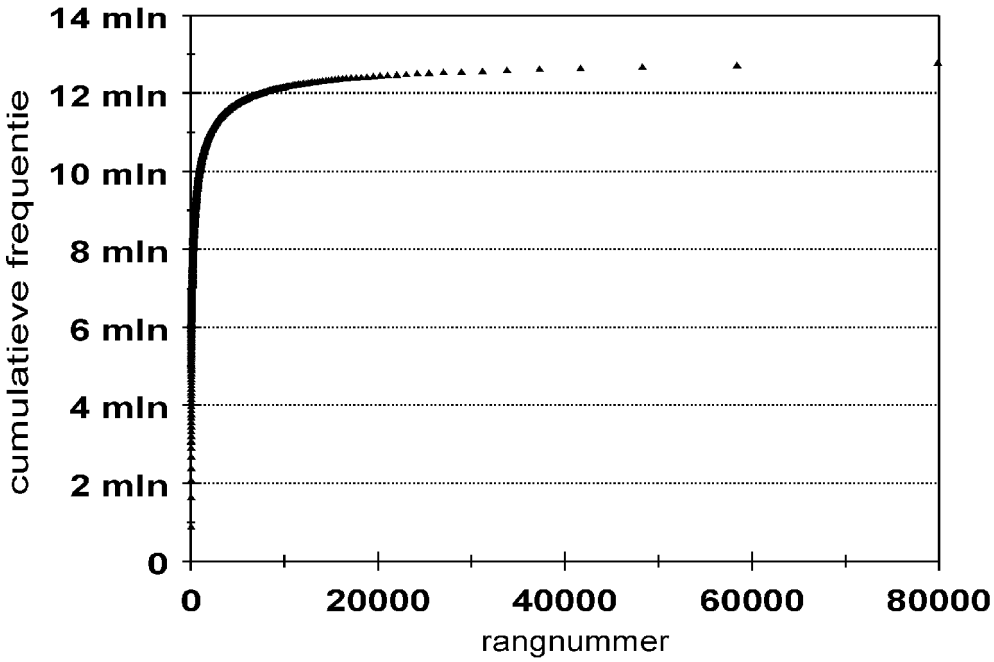
Afb. 4-1a In het corpus wetgevingsteksten voorkomende frequenties

Dat is echter niet het geval, zoals blijkt uit afbeelding 4-1b. Deze grafiek is gebaseerd op dezelfde gegevens als die in 4-1a, maar de schaalverdeling is aangepast, zodat een vergrote weergave van de punten met de laagste x- en y-waarden ontstaat. In deze grafiek is in ieder geval al goed te zien dat de frequentie bij toenemend rangnummer in het begin zeer snel, maar daarna steeds langzamer daalt. De grafiek heeft veel weg van die van een (dalende) exponentiele functie.



Afb. 4-1b Uitvergrote weergave van de laagste rangnummers uit grafiek 4-1a

Naast de frequentie bij een bepaald rangnummer, kan ook de cumulatieve frequentie worden berekend. Dit is de som van de frequenties van alle typen, vanaf de top van de frequentielijst tot en met een bepaald rangnummer. Ook dit gegeven kan weer grafisch worden weergegeven (zie afbeelding 4-2a).

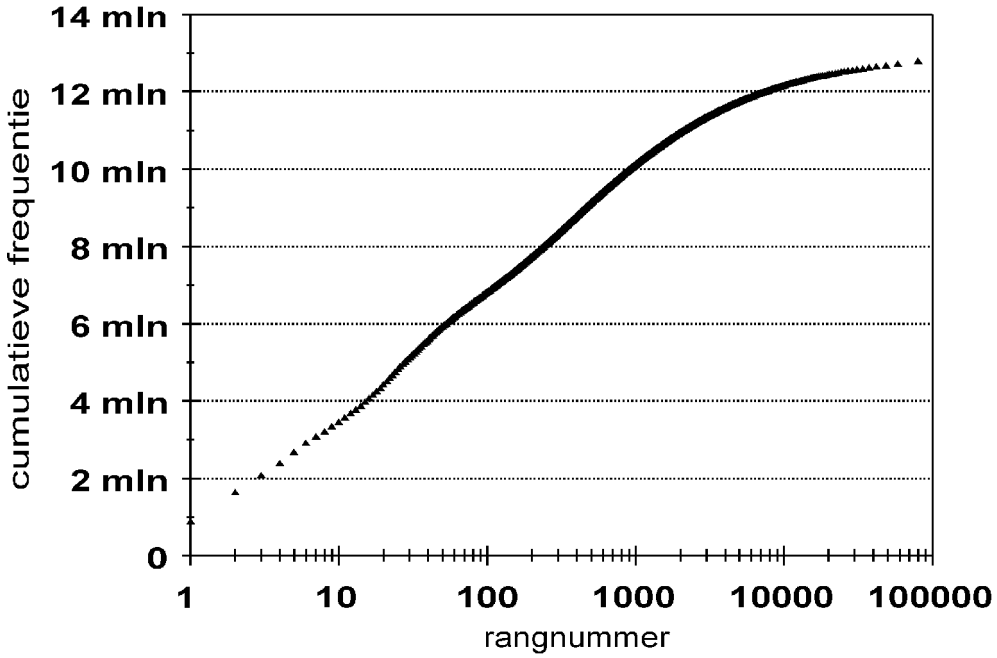


Afb. 4-2a Cumulatieve frequenties

De grafiek van de cumulatieve frequenties toont met name goed de ongelijke verdeling van de

woordtokens over de rangnummers. Al bij rangnummer 70 is ruim 50% van alle tokens vertegenwoordigd, bij rangnummer 1000 is dat 80%. Verder is duidelijk zichtbaar dat na rangnummer 30000 nog maar weinig datapunten volgen (veel typen hebben dezelfde frequentie, en dus hetzelfde rangnummer), terwijl vanaf dat punt ook het aantal tokens nog maar heel weinig toeneemt (de typen hebben allemaal een lage frequentie).

Juist vanwege de ongelijke verdeling van de rangnummers (hoe hoger de rangnummers, hoe verder ze uit elkaar liggen) is het denkbaar dat de grafiek meer informatie verschaft wanneer voor het rangnummer (dat wil zeggen, op de x-as), gebruik wordt gemaakt van een *logaritmische schaalverdeling*. Bij dit type schaalverdeling staan gelijke afstanden tussen punten in de grafiek niet langer voor even grote verschillen tussen deze punten, maar voor gelijke verhoudingen van de punten ten opzichte van elkaar. In een grafiek met logaritmische schaalverdeling kan derhalve vooral goed in beeld worden gebracht hoe meetpunten zich tot elkaar verhouden. Voor cumulatieve data is deze schaalverdeling in de meeste gevallen niet zo geschikt. Deze verschaffen in de regel op zichzelf al informatie over verhoudingen tussen gegevens (in dit geval over de mate waarin elk type bijdraagt aan het totaal aantal tokens in het corpus). Wanneer dergelijke cumulatieve data ook nog eens met een logaritmische schaalverdeling worden afgebeeld, ontstaat meestal (en ook in dit specifieke geval) niet meer dan een 'afgedempte' versie van de oorspronkelijke curve. Daarom bevat afbeelding 4-2b nogmaals de grafiek uit afbeelding 4-2a, maar nu met een logaritmische schaalverdeling voor de x-as.

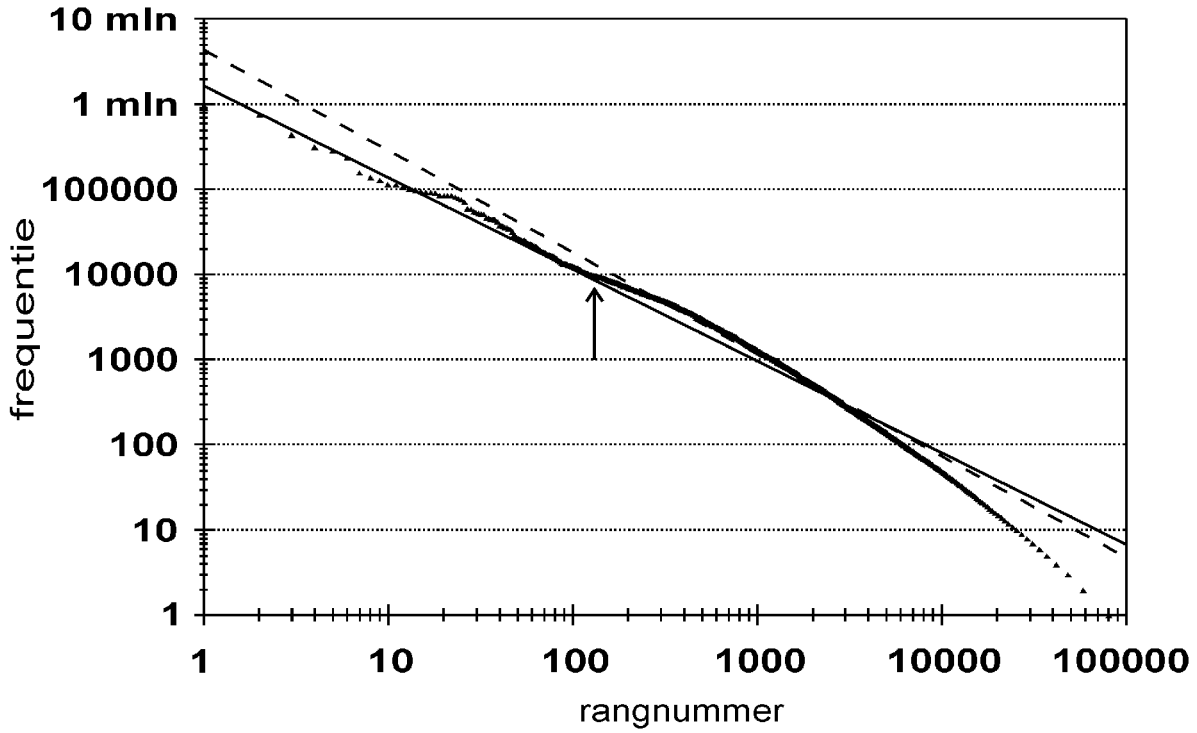


Afb. 4-2b Cumulatieve frequenties, logaritmische schaalverdeling voor het rangnummer

Wat direct opvalt bij deze versie van de grafiek is dat de 'curve' van de cumulatieve frequentie (in feite gaat het om een verzameling afzonderlijke datapunten) nu over een breed gebied vrijwel de vorm heeft van een rechte lijn. Dit gebied strekt zich uit van rangnummer 1 tot ongeveer rangnummer 800 à 1000, en omvat dus zo'n 80% van alle woordtokens in het corpus. In dit gebied bestaat er dus een praktisch constante verhouding tussen de mate waarin

rangnummer en cumulatieve frequentie toenemen. Pas boven rangnummer 1000 buigt de curve duidelijk naar beneden, een indicatie dat het aantal tokens dat vanaf dat punt wordt toegevoegd geen gelijke tred meer houdt met het rangnummer, maar relatief steeds lager wordt. Dit is op zich zeker een interessant gegeven: de rangnummers tot en met 1000 omvatten niet alleen een zeer aanzienlijk deel van alle tokens (80%), maar de cumulatieve frequentie (de som van het aantal tokens) staat daarnaast in dit gebied ook in een praktisch constante verhouding tot het rangnummer.

Natuurlijk kunnen ook de 'gewone' (niet-cumulatieve) frequenties (uit de afbeeldingen 4-1a en 4-1b) worden afgebeeld onder gebruikmaking van een logaritmische schaalverdeling. Omdat de frequenties (op de y-as) in dit geval onafhankelijke gegevens zijn (anders dan het geval was bij de cumulatieve frequenties in grafieken 4-2a en 4-2b) en aangezien ook deze een zeer ongelijke verdeling kennen (hoge frequenties liggen ver uit elkaar, lage juist heel dicht bij elkaar) is het zinvol om in dit geval ook voor de y-as een logaritmische schaalverdeling toe te passen. De datapunten in de grafiek vormen dan op het eerste gezicht een dalende (vrijwel) rechte lijn:



Afb. 4-3 In het corpus wetgevingsteksten voorkomende frequenties (logaritmische schaalverdeling)

In de grafiek zijn ook twee lijnen ingetekend (een stippellijn en een doorgetrokken lijn), waarover hieronder meer. De datapunten, in de grafiek weergegeven als driehoekjes, kennen eigenlijk alleen voor de lagere rangnummers een wat onregelmatig verloop. Dat geldt bijvoorbeeld voor de eerste 6 rangnummers (en met name voor rangnummer 1), terwijl ook rond rangnummer 25 een klein bultje zichtbaar is. Bestudering van de eerste 6 rangnummers van de frequentielijst (zie tabel 4-1) levert geen duidelijke aanwijzingen op met betrekking tot de oorzaak van de geconstateerde onregelmatigheden. Het bultje rond rangnummer 25 wordt veroorzaakt doordat daar enkele frequenties zeer dicht bij elkaar liggen (met name rangnum-

mers 20-22), waarna een en ander wordt 'goedgemaakt' door een nogal forse sprong in de frequenties tussen rangnummers 26 en 27. Bestudering van de woordtypen zelf levert geen verklaring voor dit feit op; uit tabel 3-6, aan het einde van hoofdstuk 3, blijkt wel dat het woordtype NIET, met rangnummer 27, in dit corpus relatief minder vaak voorkomt dan in de andere corpora, maar met dat feit alleen kan het geconstateerde verloop niet worden verklaard.

Afgezien van deze onregelmatigheden echter vormen de datapunten, in ieder geval tot rangnummer 100, een vrijwel rechte lijn. De eigenschappen van deze lijn, die wellicht geschikt zouden kunnen zijn voor het maken van vergelijkingen tussen de corpora, kunnen worden bepaald door het uitvoeren van een *lineaire regressie-analyse*. Dit houdt in dat, door middel van analyse van de datapunten, de vergelijking wordt bepaald van de rechte lijn die zo goed mogelijk het verband aangeeft tussen de zogenaamde onafhankelijke variabele (in dit geval het rangnummer) en de afhankelijke variabele (de woordfrequentie). De meeste datapunten zullen natuurlijk niet precies op deze lijn liggen, maar de regressieanalyse zorgt er wel voor dat de lijn zodanig wordt geconstrueerd dat de som van alle afstanden van de datapunten tot de regressielijn (de zogenaamde *residuals*) minimaal is (de methode van de *least squares*, zo genoemd omdat bij de berekening wordt gewerkt met de som van de kwadraten van alle afstanden). Nu is het in dit geval zo, dat de regressielijn op tenminste twee verschillende manieren kan worden geconstrueerd, waarbij echter een van de twee een duidelijk beter resultaat geeft. Deze twee manieren zijn:

1. De vergelijking van de regressielijn kan worden bepaald door uit te gaan van 1 punt per rangnummer. Iedere in het corpus voorkomende woordfrequentie levert dan dus, met het bijbehorende rangnummer, 1 datapunt op. De op deze datapunten (1960 stuks, aangezien in het corpus 1960 verschillende frequenties voorkomen) gebaseerde regressielijn heeft in dit geval als vergelijking: $Y = -1{,}192\ X + 6{,}635$, ofwel (in termen van de data waar het hier om gaat)

   $$\log r = -1{,}192 \log rangnr + 6{,}635 \qquad (4\text{-}1)$$

   (waarbij r staat voor de frequentie). Deze functie is in de grafiek ingetekend in de vorm van een stippellijn.

   Een probleem van methode 1 is dat deze weinig rekening houdt met de zeer ongelijke verdeling van de *woordtokens* over de verschillende rangnummers. Zo vertegenwoordigen de laagste rangnummers een zeer groot aantal tokens (rangnummer 1, het woord DE, al 899333 stuks, ruim 7% van het totale corpus), terwijl bijvoorbeeld rangnummer 2780 maar 331 tokens omvat. Het is daarom denkbaar dat een regressielijn, gebaseerd op 1 datapunt per rangnummer, niet goed aansluit bij de structuur van het woordgebruik in het corpus.

2. Om wel recht te doen aan de frequentie van ieder token, zou eigenlijk ieder datapunt net zoveel keer in aanmerking moeten worden genomen als het aantal tokens dat het omvat. Dat zou echter wel het praktische bezwaar hebben dat op die manier de regressielijn zou moeten worden berekend op basis van meer dan 12 miljoen datapunten (het aantal tokens in het corpus), wat in dit geval te hoge eisen zou stellen aan geheugen en verwerkingscapaciteit van de gebruikte computer. Om die reden is uit de totale lijst van 12 miljoen

punten een steekproef samengesteld, door telkens een datapunt op te nemen en er vervolgens 1830 over te slaan. Op die manier ontstond een lijst van bijna 7000 datapunten, een nog juist te verwerken aantal, die als het ware een afspiegeling vormen van alle frequenties, *gecorrigeerd naar hun 'gewicht'*, dat wil zeggen naar het aantal tokens dat ze vertegenwoordigen. Op basis van deze lijst van 7000 datapunten werd opnieuw de vergelijking van de regressielijn bepaald, welke dan als volgt luidt:

$$\log r = -1{,}079 \log rangnr + 6{,}226 \qquad (4\text{-}2)$$

Deze functie is in afbeelding 4-3 getekend als een doorgetrokken lijn.

Beide regressielijnen verklaren op papier (elk voor hun eigen dataset) ongeveer even goed de variantie in de waarden van de afhankelijke variabele (de frequentie): r^2, het kwadraat van de lineaire correlatiecoëfficiënt, ligt bij allebei iets boven 0,98, wat betekent dat in beide gevallen ruim 98% van de variantie door de regressielijn wordt verklaard. Wanneer we de regressielijnen bekijken, blijkt echter dat de eerste (onderbroken) lijn maar over een vrij beperkt gebied (ongeveer van rangnummer 100 tot 3000) min of meer samenvalt met de serie datapunten. Het hele gebied van rangnummer 1 tot 100, uiterst belangrijk vanwege het grote deel van het corpus dat deze typen vertegenwoordigen, wordt door de regressielijn 'gemist', doordat in dat gebied niet zo veel verschillende frequenties te vinden zijn. De tweede (doorgetrokken) regressielijn kent dit probleem niet. Met uitzondering van rangnummer 1, dat duidelijk ruim onder de lijn ligt, worden alle datapunten tot even voorbij rangnummer 100 uitstekend door de lijn bestreken. Deze regressielijn geeft dus, anders dan de eerste, goed de verhouding aan tussen rangnummer en frequentie juist in het gebied waar die verhouding, blijkens het patroon dat de datapunten vertonen, constant is. Daarom kan geconcludeerd worden dat de tweede regressielijn deze belangrijke eigenschap van het corpus het beste weergeeft, en daarom ook voor vergelijkingsdoeleinden het meest geschikt is.

De (tweede, doorgetrokken) regressielijn geeft tenslotte ook een duidelijker inzicht in wat er ná rangnummer 100 aan de hand is met de verhouding tussen rangnummer en frequentie. Vanaf een bepaald punt, na uitvergroten van het betreffende deel van de grafiek tamelijk nauwkeurig te bepalen als zijnde rangnummer 136 (zie pijl), vormen de datapunten namelijk niet langer een vrijwel rechte lijn, maar een lichte, zeer regelmatige neerwaartse kromme. Het rangnummer waar deze kromme begint ligt nog praktisch op de regressielijn, daarna volgt een gedeelte (ongeveer tot rangnummer 500) waar de frequentie wat minder snel daalt dan uit de vergelijking van de regressielijn zou volgen, waarna tenslotte de curve steeds wat sneller gaat dalen, en daarbij ook de regressielijn kruist. Alles bij elkaar kunnen bij het corpus wetgevingsteksten in de grafiek die de verhouding tussen rangnummer en frequentie weergeeft dus in ieder geval twee afzonderlijke gedeelten worden onderscheiden: een 'kop' vanaf rangnummer 1 tot ongeveer 136, en een 'staart', vanaf rangnummer 136.[6] Deze twee gedeelten waren nog niet duidelijk zichtbaar in de grafiek van de cumulatieve frequenties. Wel was in die grafiek bij de

[6] Dit is tot op zekere hoogte in overeenstemming met de bevindingen van De Mulder en Oskamp (1979, p. 48). Zij pasten een wat andere techniek toe om de kop van de frequentielijst zichtbaar te maken (Zipf-karakteristieken), echter met vergelijkbaar resultaat. Zij troffen in twee juridische corpora een kop aan die liep van rangnummer 1 tot 50 à 70, maar daarnaast ook een 'middenstuk', dat zich uitstrekte tot rangnumme 250 à 300. Een dergelijk midden-gedeelte valt in grafiek 4-3 niet duidelijk te herkennen. 250 à 300. Een dergelijk midden-gedeelte valt in grafiek 4-3 niet duidelijk te herkennen.

hogere rangnummers al hetzelfde typische gekromde verloop zichtbaar dat we nu weer in de staart aantreffen. Het exacte punt waar deze staart begint was uit grafiek 4-2b echter niet af te lezen. Voor het bepalen daarvan en van de verhouding tussen frequentie en rangnummer in met name de kop, kan de lineaire regressielijn van frequentie op rangnummer goede diensten bewijzen. De vergelijking van deze lijn kan ook de basis vormen voor het vergelijken van de structuur van het woordgebruik in de drie corpora. Voor dat doel is het nuttig om de vergelijking in een andere vorm om te zetten, zonder gebruik van logaritmen:

$$r = rangnr^{-1,079} \cdot 10^{6,226} = \frac{10^{6,226}}{rangnr^{1,079}} \qquad (4\text{-}3)$$

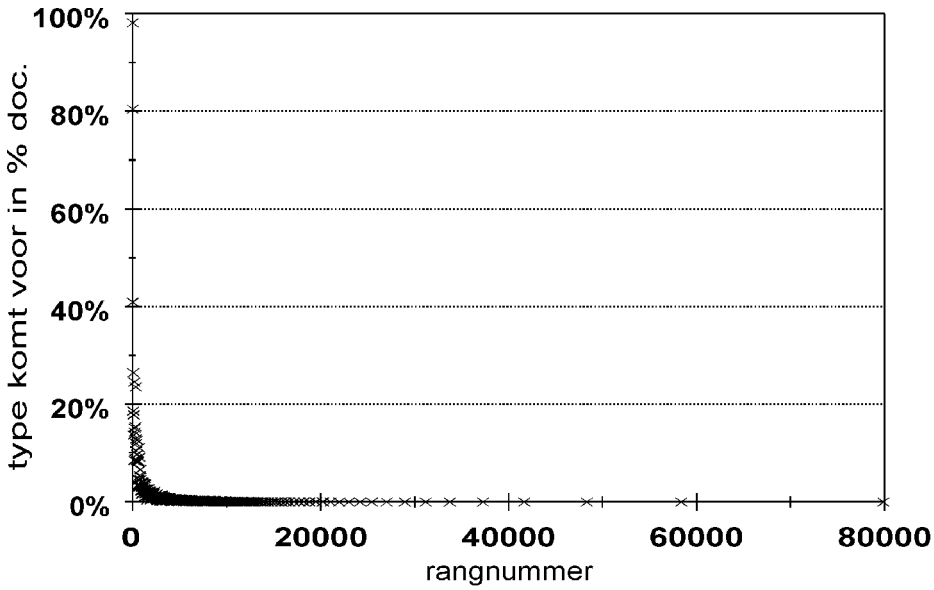
Wanneer de functionele samenhang tussen frequentie en rangnummer op deze manier wordt weergegeven, is duidelijk het verband zichtbaar met de *wet van Zipf* [7], die immers luidt dat het produkt van rangnummer en frequentie constant is. Deze wet zou alleen opgaan wanneer de regressielijn uit grafiek 4-3 een x-coëfficiënt van -1 zou hebben, en dus *een hoek van 45° zou maken* met de horizontale as. Dat is in ieder geval bij dit corpus niet zo, waardoor niet het produkt van rangnummer en frequentie, maar het produkt van rangnummer tot de macht 1,079 en frequentie een constante waarde heeft (van $10^{6,226}$). De wet van Zipf heeft dus betrekking op een bijzondere situatie, die zich echter in dit geval niet voordoet, terwijl de benadering met de regressielijn altijd kan worden toegepast, tenminste wanneer het verband tussen de logaritmen van rangnummer en frequentie voldoende lineair is.

Met betrekking tot de vraag welk deel van de frequentielijst in ieder geval bestudering verdient kan worden gesteld dat met name het geval is met de kop, het gedeelte tot even voorbij rangnummer 100, vanwege de stabiele verhouding tussen rangnummer en frequentie in dat gebied.

4.2.2. Spreiding van woordtypen

Ook de spreiding van typen over het corpus zou een criterium kunnen zijn om te bepalen welk deel van de frequentielijst de interessantste informatie bevat. Bij nadere bestudering van de percentages die aangeven in hoeveel procent van de documenten waaruit het corpus is opgebouwd de diverse woordtypen voorkomen blijkt dat tot ongeveer rangnummer 1000 de spreiding van woordtypen een sterk dalend verloop heeft: het woordtype met rangnummer 1 komt in bijna 100% van de documenten voor, het woordtype met nummer 40 nog maar in 50%, en dit percentage loopt dan snel terug tot ongeveer 5% voor de woordtypen met rangnummer 1000. Na rangnummer 1000 daalt het percentage nog maar zeer geleidelijk en langzaam. In afbeelding 4-4 wordt dit grafisch weergegeven.

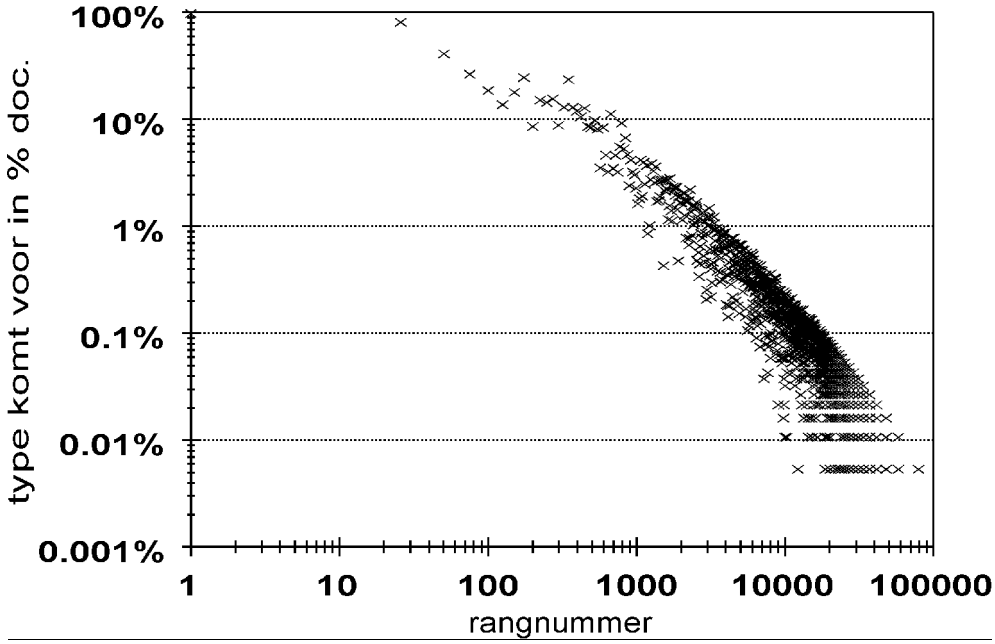
[7] Zipf (1935, p. 39 e.v.). Zie ook hoofdstuk 2, paragraaf 3.



Afb. 4-4 Spreiding van alle typen over het corpus wetgevingsteksten

Ten aanzien van deze grafiek kan het volgende worden opgemerkt:

- er is geen sprake van een lijn, maar van een puntenwolk. Dat komt doordat op de x-as nog steeds het rangnummer is afgezet, terwijl in de grafiek de spreidingspercentages van *alle afzonderlijke typen*, ook wanneer die even vaak voorkomen (en dus hetzelfde rangnummer hebben) zijn opgetekend. De spreidingspercentages van typen met hetzelfde rangnummer blijken soms aanzienlijk te verschillen; aan dit verschijnsel zal in paragraaf 4.7 aandacht worden besteed.

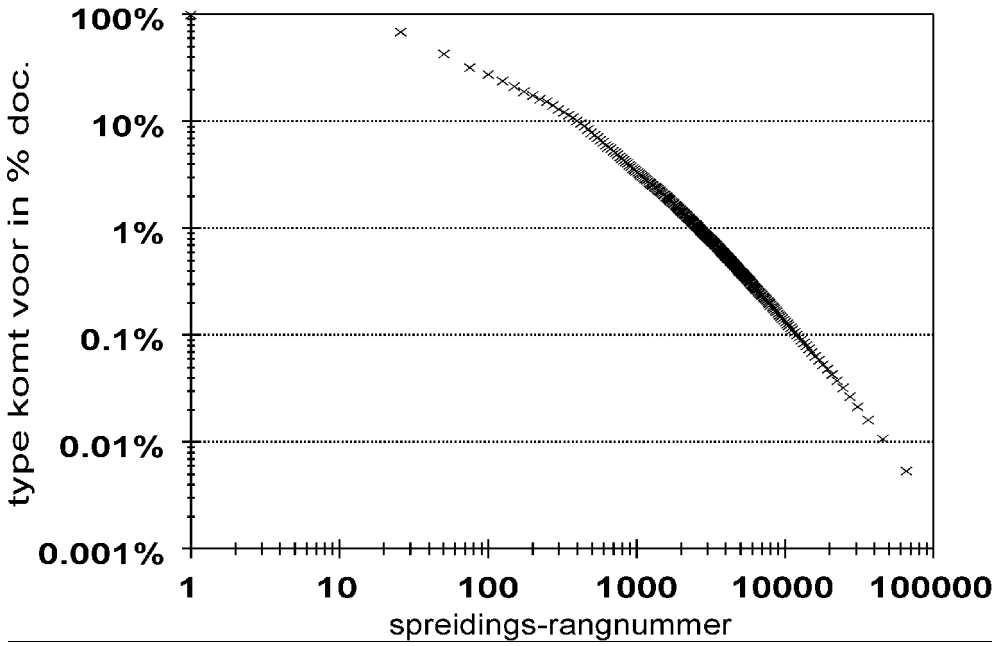


Afb. 4-5 Spreiding van alle typen over het corpus wetgevingsteksten (logaritmische schaalverdeling)

- ook deze grafiek laat zich, net als die in afbeelding 4-1a, moeilijk lezen door de grote range en ongelijke verdeling van de meetpunten. Het ligt daarom ook hier voor de hand om te kijken of een duidelijker beeld ontstaat bij toepassing van een logaritmische schaalverdeling. Zie daarvoor afbeelding 4-5

Inderdaad is die grafiek, waarin overigens om praktische redenen slechts elk 25^{e} datapunt wordt weergegeven, duidelijker zij het dat de puntenwolk niet zozeer een rechtlijnig, dan wel een licht gekromd dalend verloop kent. Nog duidelijker dan in de versie met gewone schaal-verdeling wordt in deze grafiek echter ook een vrij groot aantal 'uitbijters' zichtbaar, punten die duidelijk afwijken, in positieve of in negatieve richting, van de trend die uit de grafiek valt af te lezen. Mede daardoor vertoont de puntenwolk in het gebied tussen rangnummer 200 en 1000 een wat onregelmatig verloop.

Als laatste stap in dit proces is nog een logaritmische grafiek vervaardigd op basis van 'spreidings-rangnummers', waarbij alle woordtypen een nieuw rangnummer hebben gekregen, nu niet op basis van hun frequentie, maar op basis van het aantal documenten waarin ze voorkomen (zie afbeelding 4-6).



Afb. 4-6 Spreiding van woordtypen over het corpus wetgevingsteksten, rangnummers (x-as) op basis van de mate van spreiding

De figuur lijkt op die met de 'gewone' rangnummers, maar is veel duidelijker doordat de punten vrijwel volledig in een lijn liggen. Het uitwaaier-effect, veroorzaakt doordat typen met hetzelfde (frequentie-)rangnummer een verschillende mate van spreiding vertonen is nu verdwenen. Het algemene verloop van de twee grafieken is echter praktisch gelijk; de rangnummers waarbij sprake is van een voorkomen in tenminste 10%, 1% en 0,1% van het corpus zijn ongeveer hetzelfde. Dat de figuren zo sterk op elkaar lijken is op zich wel opvallend, omdat de wijze waarop dit corpus in documenten is gesplitst - anders dan bij de andere corpora - tamelijk arbitrair is (zie hoofdstuk 3). Verder valt nog op dat grafiek 4-6 een

lichte knik vertoont, in de buurt van de 10%-grens en ongeveer bij spreidingsrangnummer 300. De spreiding van de woordtypen gaat vanaf dat punt plotseling wat sneller afnemen.

Op basis van deze spreidingsgegevens kan niet worden gezegd dat in de frequentielijst sprake is van een duidelijk omslagpunt. Wel vertonen typen met rangnummers tussen 200 en 1000 voor wat betreft de spreiding een wat rommelig patroon, terwijl de grafiek volgens spreidingsrangnummers rond rangnummer 300 een lichte knik vertoont. Hoewel een en ander niet echt overtuigend is, verdient het op basis van deze gegevens mogelijk toch aanbeveling het primair te bestuderen deel van de frequentielijst wat ruimer te nemen dan alleen de 'kop' (ongeveer de eerste 136 rangnummers, zie de vorige subparagraaf).

Dit alles leidt tot de conclusie dat, in het geval van dit corpus, van de gegevens die direct uit de frequentielijst kunnen worden afgelezen vooral de woordfrequenties zelf enige aanwijzingen bevatten ten aanzien van de vraag welk gedeelte van deze lijst de meest interessante informatie zou kunnen opleveren. Aan de hand van deze frequenties kan een 'kop' worden onderscheiden, welke in ieder geval bestudering verdient. Voorts kan het spreidingspercentage in de overweging worden betrokken, aangezien woordtypen die breed vertegenwoordigd zijn in het corpus waarschijnlijk informatie kunnen verschaffen over het corpus als geheel. Op basis van deze spreidingsgegevens is het zinvol het te bestuderen gedeelte uit te breiden, zodat het omvat: de kop (rangnummer 1-136) en het gedeelte dat daar direct op volgt, tot en met rangnummer 300. Gezien het te verwachten belang van dit deel van de frequentielijst, zowel voor dit als voor mogelijk vervolgonderzoek, zullen dan ook ten minste deze woordtypen in bijlage 3 van dit rapport worden afgedrukt.

Afgezien van het beperkte deel van de frequentielijst dat in bijlage 3 wordt opgenomen zullen van ieder corpus natuurlijk de complete lijsten bewaard blijven en in computer-leesbare vorm - onder andere via het Internet - beschikbaar worden gesteld aan geïnteresseerden. Vanzelfsprekend zijn ook deze complete lijsten de basis voor het opstellen van bijvoorbeeld frequentieverdelingen en voor het bepalen van type/token ratio's (zie de volgende twee paragrafen).

Een conclusie die aan het einde van deze paragraaf kan worden getrokken is dat zowel voor de woordfrequenties als voor de spreidingspercentages van woordtypen het volgende geldt: wanneer deze gegevens worden afgezet in een grafiek met (dubbel) logaritmische schaalverdeling, blijken zij in vrij sterke mate een lineair verband met de rangnummers van de woordtypen te vertonen. In feite betekent dit dat er een lineair verband bestaat niet zozeer tussen de absolute waarden van rangnummer en woordfrequentie en van rangnummer en spreidingspercentage, maar tussen de *verhoudingen* in deze waarden.

4.3 Frequentieverdelingen

Het begrip frequentieverdeling werd al beschreven in hoofdstuk 2. Een frequentieverdeling is een gesorteerde lijst van frequentieklassen, waarbij voor iedere klasse wordt aangegeven hoeveel woordtypen in die klasse vallen, of anders gezegd, hoeveel woordtypen met die frequentie in het corpus voorkomen. Frequentieverdelingen kunnen nuttig zijn bij het in kaart brengen van de *structuur* van het woordgebruik in een corpus. Ze kunnen bijvoorbeeld inzicht

geven in de verhoudingen tussen de aantallen zeer vaak, gemiddeld vaak en weinig voorkomende woordtypen. Het zicht op de individuele woordtypen en hun frequentie gaat daarbij verloren; we praten nog slechts over klassen van even vaak voorkomende woordtypen.

De frequentieverdeling van een document van beperkte omvang kan dikwijls nog op compacte wijze in een tabel worden weergegeven, waarbij het aantal frequentieklassen beperkt blijft en de structuur van het woordgebruik eigenlijk al direct uit de tabel kan worden afgelezen. Zie voor een voorbeeld hiervan hoofdstuk 2, tabel 2-1. Bij de frequentieverdeling van een omvangrijk corpus ligt dat anders. Het gaat dan al gauw om enkele duizenden verschillende frequenties (corpus wetgevingsteksten 1960 stuks, corpus jurisprudentieteksten 3353 stuks, corpus algemeen Nederlands 1189 stuks), die zich een stuk moeilijker laten overzien en waaruit de structuur van het woordgebruik niet of slechts in beperkte mate valt af te lezen. Om toch zinvolle informatie te kunnen aflezen uit een dergelijke omvangrijke frequentieverdeling is in de meeste gevallen een goede optie om de frequentieklassen zodanig te kiezen dat iedere klasse niet slechts een enkele (absolute) frequentie omvat, maar een bepaald frequentie-*interval*. De totale frequentie-'range' van het corpus wetgevingsteksten loopt van 1..899333. Niet alle frequenties komen echter even vaak voor in het corpus, de lage frequenties zijn sterk vertegenwoordigd, de hoge slechts één maal of helemaal niet (de frequenties van 759046..899332 komen bijvoorbeeld niet in het corpus voor). Met andere woorden, de frequentieverdeling is buitengewoon scheef. Daarom ligt het niet erg voor de hand om het gehele frequentieinterval te verdelen in klassen van gelijke grootte: er zullen dan veel lege klassen ontstaan, tenzij de klassebreedte zeer hoog wordt gekozen, wat weer als nadeel heeft dat mogelijk veel details verloren kunnen gaan. Een beter alternatief vormt een frequentieverdeling met klassen die logaritmisch in omvang toenemen, bijvoorbeeld als volgt:

- klasse 1: 1 (10^0) frequentie
- klasse 2: 10 (10^1) frequenties
- klasse 3: 100 (10^2) frequenties

etc.

De steeds toenemende omvang van de klassen zou dan min of meer het afnemen van het aantal frequenties naarmate we het einde van de frequentieverdeling naderen kunnen compenseren. Inderdaad bleek dat voor het corpus wetgevingsteksten op te gaan; na enig experimenteren met klassebreedtes en stapgroottes kon uiteindelijk de frequentieverdeling zoals afgebeeld op de volgende bladzijde worden samengesteld.

Zoals uit de tabel blijkt is gekozen voor een indeling waarbij iedere klasse niet 10 keer zo groot is als de voorgaande, maar $10^{0,1} \approx 1,2$ keer. Er is dan sprake van een duidelijk beeld, met een goede spreiding over de klassen. Ook het aantal lege klassen (6 stuks, precies 10% dus) is voor een dergelijke scheve verdeling laag te noemen. Uit analyse blijkt dat in deze verdeling ook weinig informatie verloren gaat. Wanneer namelijk het totaal aantal tokens en het gewogen gemiddelde van alle woordfrequenties in het corpus worden geschat met behulp van de klassemiddens uit de frequentieverdeling (de middelste waarde uit het interval dat door iedere frequentieklasse wordt bestreken) worden uitkomsten verkregen die slechts 0,01% afwijken van de werkelijke waarden.

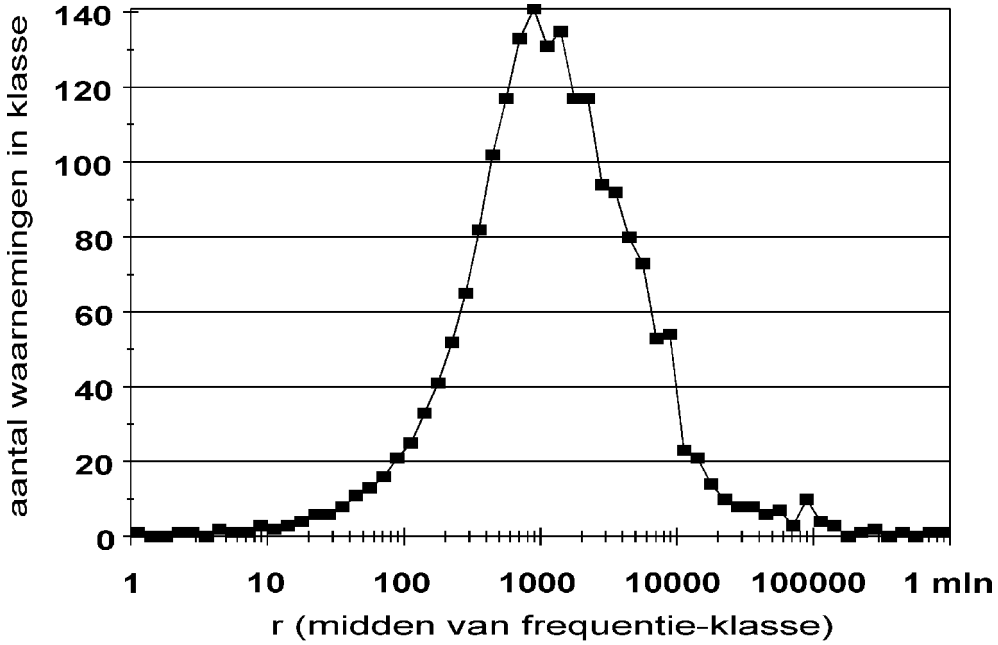
Klasse	Grenzen	Klasse-midden	Aantal freq.	Aantal typen	Aantal tokens	Klasse	Grenzen	Klassemid-den	Aantal freq.	Aantal typen	Aantal tokens
1	10^{0} - $10^{0.1}$	1.13	1	63303	63303	31	$10^{3.0}$ - $10^{3.1}$	1129.46	131	208	230889
2	$10^{0.1}$ - $10^{0.2}$	1.42	0	0	0	32	$10^{3.1}$ - $10^{3.2}$	1421.91	135	180	252294
3	$10^{0.2}$ - $10^{0.3}$	1.79	0	0	0	33	$10^{3.2}$ - $10^{3.3}$	1790.08	117	140	249652
4	$10^{0.3}$ - $10^{0.4}$	2.25	1	21462	42924	34	$10^{3.3}$ - $10^{3.4}$	2253.57	117	135	301776
5	$10^{0.4}$ - $10^{0.5}$	2.84	1	10067	30201	35	$10^{3.4}$ - $10^{3.5}$	2837.08	94	100	282644
6	$10^{0.5}$ - $10^{0.6}$	3.57	0	0	0	36	$10^{3.5}$ - $10^{3.6}$	3571.67	92	94	334952
7	$10^{0.6}$ - $10^{0.7}$	4.50	2	11048	48601	37	$10^{3.6}$ - $10^{3.7}$	4496.47	80	85	383299
8	$10^{0.7}$ - $10^{0.8}$	5.66	1	3477	20862	38	$10^{3.7}$ - $10^{3.8}$	5660.72	73	74	415163
9	$10^{0.8}$ - $10^{0.9}$	7.13	1	2618	18326	39	$10^{3.8}$ - $10^{3.9}$	7126.43	53	55	386054
10	$10^{0.9}$ - $10^{1.0}$	8.97	3	5707	50726	40	$10^{3.9}$ - $10^{4.0}$	8971.64	54	54	485034
11	$10^{1.0}$ - $10^{1.1}$	11.29	2	2443	28038	41	$10^{4.0}$ - $10^{4.1}$	11294.63	23	23	259808
12	$10^{1.1}$ - $10^{1.2}$	14.22	3	2796	38930	42	$10^{4.1}$ - $10^{4.2}$	14219.09	21	21	290371
13	$10^{1.2}$ - $10^{1.3}$	17.90	4	2539	44057	43	$10^{4.2}$ - $10^{4.3}$	17900.78	14	14	249963
14	$10^{1.3}$ - $10^{1.4}$	22.54	6	2599	57774	44	$10^{4.3}$ - $10^{4.4}$	22535.74	10	10	224164
15	$10^{1.4}$ - $10^{1.5}$	28.37	6	1842	52245	45	$10^{4.4}$ - $10^{4.5}$	28370.82	8	8	214262
16	$10^{1.5}$ - $10^{1.6}$	35.72	8	1692	59676	46	$10^{4.5}$ - $10^{4.6}$	35716.75	8	8	286473
17	$10^{1.6}$ - $10^{1.7}$	44.96	11	1682	75063	47	$10^{4.6}$ - $10^{4.7}$	44964.72	6	6	270182
18	$10^{1.7}$ - $10^{1.8}$	56.61	13	1326	75023	48	$10^{4.7}$ - $10^{4.8}$	56607.23	7	7	385031
19	$10^{1.8}$ - $10^{1.9}$	71.26	16	1246	88205	49	$10^{4.8}$ - $10^{4.9}$	71264.28	3	3	226169
20	$10^{1.9}$ - $10^{2.0}$	89.72	21	1058	94396	50	$10^{4.9}$ - $10^{5.0}$	89716.41	10	10	896574
21	$10^{2.0}$ - $10^{2.1}$	112.95	25	881	98995	51	$10^{5.0}$ - $10^{5.1}$	112946.27	4	4	439831
22	$10^{2.1}$ - $10^{2.2}$	142.19	33	771	109187	52	$10^{5.1}$ - $10^{5.2}$	142190.93	3	3	427507
23	$10^{2.2}$ - $10^{2.3}$	179.01	41	668	118436	53	$10^{5.2}$ - $10^{5.3}$	179007.78	0	0	0
24	$10^{2.3}$ - $10^{2.4}$	225.36	52	583	130186	54	$10^{5.3}$ - $10^{5.4}$	225357.44	1	1	236772
25	$10^{2.4}$ - $10^{2.5}$	283.71	65	473	132817	55	$10^{5.4}$ - $10^{5.5}$	283708.20	2	2	601904
26	$10^{2.5}$ - $10^{2.6}$	357.17	82	418	148691	56	$10^{5.5}$ - $10^{5.6}$	357167.47	0	0	0
27	$10^{2.6}$ - $10^{2.7}$	449.65	102	371	166453	57	$10^{5.6}$ - $10^{5.7}$	449647.20	1	1	437151
28	$10^{2.7}$ - $10^{2.8}$	566.07	117	343	192708	58	$10^{5.7}$ - $10^{5.8}$	566072.29	0	0	0
29	$10^{2.8}$ - $10^{2.9}$	712.64	133	251	178586	59	$10^{5.8}$ - $10^{5.9}$	712642.79	1	1	759046
30	$10^{2.9}$ - $10^{3.0}$	897.16	141	244	216952	60	$10^{5.9}$ - $10^{6.0}$	897164.12	1	1	899333

Tabel 4-2 Frequentieverdeling corpus wetgevingsteksten. De frequentieklassen nemen logaritmisch in omvang toe.

4.3.1 De aantallen frequenties per frequentieklasse

Wanneer het *aantal frequenties* in elk van de 60 klassen (kolom vier uit bovenstaande tabel) wordt afgezet tegen de klassemiddens blijkt dat deze frequenties over de klassen zijn verdeeld volgens een patroon dat gelijkt op de zogenaamde curve van de *normale verdeling*, ook wel genoemd de *kromme van Gauss* (zie afbeelding 4-7). Dit is een symmetrische, klokvormige curve, die aangeeft hoe de waarden die een bepaalde grootheid kan aannemen zijn verdeeld (dat wil zeggen, hoe vaak elke waarde voorkomt). Voor veel grootheden uit de dagelijkse praktijk, zoals bijvoorbeeld de lichaamslengte van personen, geldt dat de waarden die zij kunnen aannemen verdeeld zijn overeenkomstig deze curve van de normale verdeling. Opgemerkt kan worden dat in deze grafiek voor de x-as een logaritmische schaalverdeling is toegepast, die er voor zorgt dat alle frequentieklassen, hoewel die logaritmisch in omvang toenemen, in de grafiek een gelijke ruimte op deze as innemen. Voor de y-as (het aantal verschillende frequenties per klasse) is een gewone indeling toegepast, aangezien het hier gaat

om de aantallen verschillende frequenties *per frequentieklasse*. De klasse-indeling zorgt ervoor dat deze aantallen niet de grote verschillen vertonen die een logaritmische schaalverdeling noodzakelijk zouden maken.



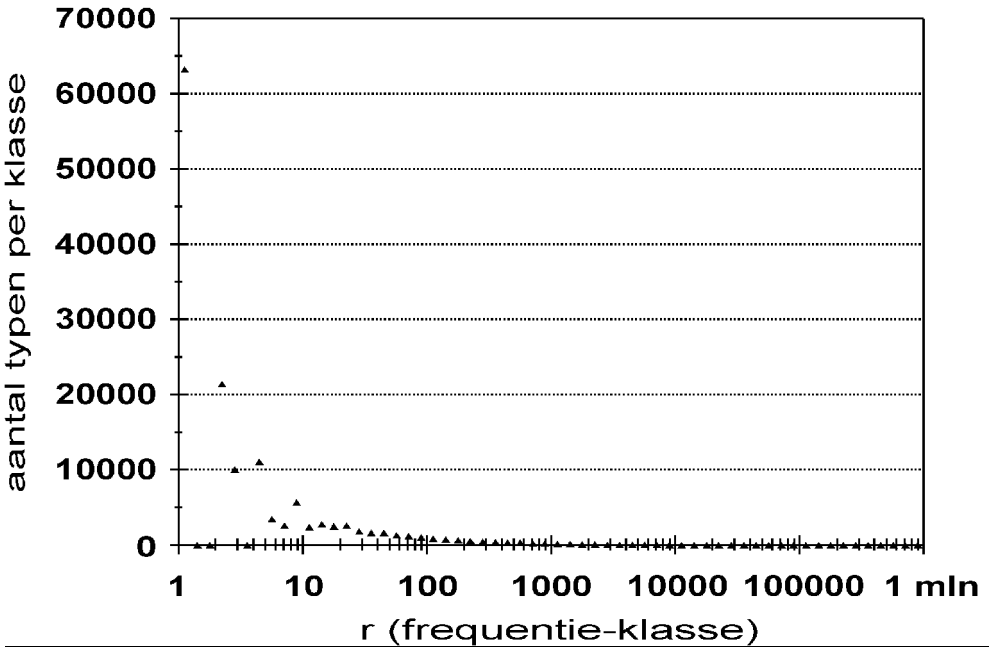
Afb. 4-7 Aantal waarnemingen (verschillende frequenties) per frequentieklasse

De grafiek toont in ieder geval de relevantie van een frequentieverdeling met logaritmisch in omvang toenemende klassen. De aantallen frequenties per klasse zijn dan volgens een regelmatig patroon verdeeld, het hoogste aantal frequenties vinden we precies in de middelste klasse. Dat de verdeling van het aantal frequenties overeenkomst vertoont met een normale verdeling lijkt in overeenstemming te zijn met bevindingen van Carroll (1967, 1970), die aangeeft dat "word frequencies or probabilities have an approximately normal distribution when scaled logarithmically"[8]. Hier is het dan interessant dat de verdeling ook grafisch zo sterk op een normale verdeling gelijkt. N.B. de aantallen frequenties (op de y-as) zijn niet logaritmisch weergegeven.

4.3.2 De aantallen woordtypen en -tokens per klasse

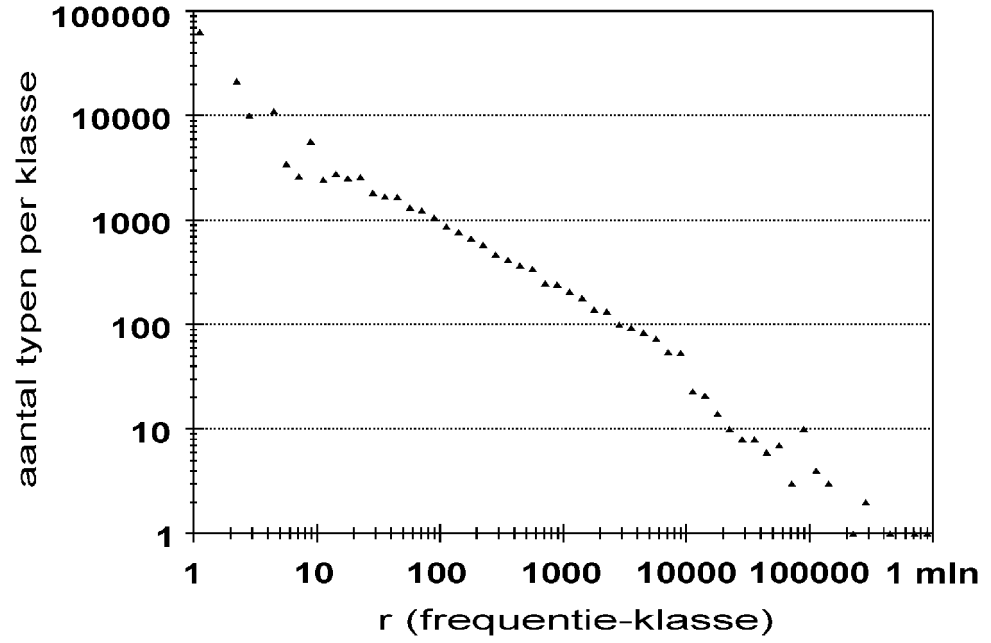
Wanneer we nu op overeenkomstige wijze de aantallen woordtypen eveneens afbeelden als functie van de klassemiddens van de frequentieklassen, ontstaat in eerste instantie geen erg duidelijk beeld:

[8] Carroll (1970, p. 65). Deze bevinding is voor hem reden om te pleiten voor het gebruik van een Standard Frequency Index (SFI) om woord-probabiliteiten (relatieve frequenties) in uit te drukken. In wezen is de SFI gelijk aan de logaritme van de relatieve frequentie van een woord. Mede gezien het feit dat relatieve frequenties in het vervolg van dit onderzoek geen essentiële rol spelen zal op deze SFI hier niet nader worden ingegaan.



Afb. 4-8a Aantal woordtypen per frequentieklasse

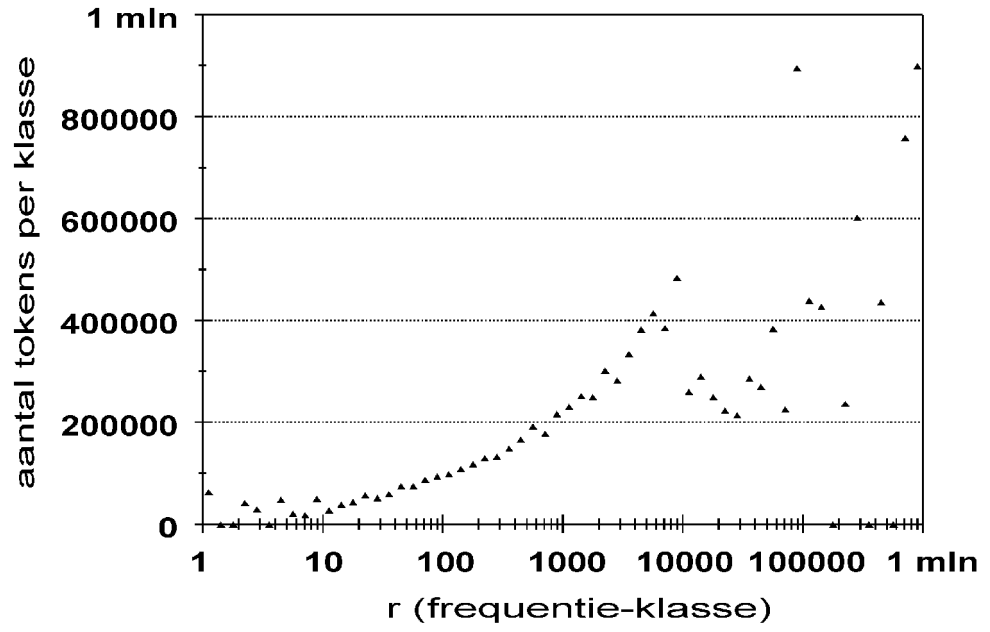
De eerste vier (niet lege) frequentieklassen, die elk meer dan 10000 woordtypen bevatten, bepalen in belangrijke mate het aanzien van deze grafiek, met name voor wat betreft de schaalverdeling op de verticale as. De aantallen typen in de overige klassen zijn daardoor niet goed meer af te lezen. Om te zien of in dit geval met behulp van de verhoudingen tussen de aantallen een duidelijker grafiek kan worden verkregen dan met behulp van de aantallen zelf is in afbeelding 4-8b nogmaals dezelfde grafiek getekend, maar nu met een logaritmische indeling voor de y-as.



Afb. 4-8b Aantal woordtypen per frequentieklasse, y-as met logaritmische schaalverdeling

Direct blijkt nu dat tussen de aantallen woordtypen veel meer samenhang bestaat dan afbeelding 4-8a deed vermoeden: de puntenwolk in afbeelding 4-8b heeft de vorm van een dalende rechte lijn. Alleen bij de lage frequenties (ongeveer tot 10) en bij frequenties boven 10000 is het verloop wat minder gelijkmatig. In deze gebieden zijn ook de lege klassen uit de frequentieverdeling te vinden (in de grafiek niet weergegeven).

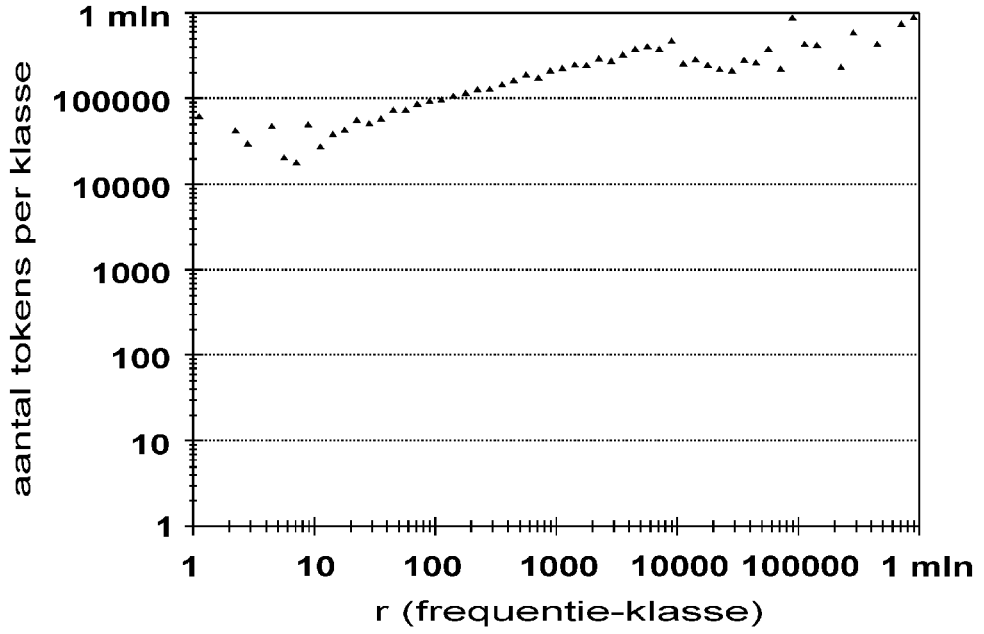
Ook de aantallen woordtokens uit tabel 4-2 kunnen worden afgebeeld als functie van de klassemiddens van de frequentieklassen (zie afbeelding 4-9a).



Afb. 4-9a Aantal woordtokens per frequentieklasse

Ook het verloop van deze aantallen kan echter uit bovenstaande grafiek niet echt duidelijk worden afgelezen. Weliswaar lijkt in het eerste gedeelte van de grafiek sprake te zijn van een behoorlijk gelijkmatige toename, maar in het tweede gedeelte (ongeveer vanaf frequentie 10000) gaat deze toename over in een op het oog bijna willekeurig patroon. Net als bij de woordtypen brengt ook hier een logaritmische schaalverdeling op de verticale as enige duidelijkheid (zie afbeelding 4-9b). De puntenwolk krijgt ook hier door de logaritmische schaalverdeling duidelijk meer samenhang, en heeft in dit geval een licht stijgend verloop. Net als bij de woordtypen zijn bij lage frequenties (ongeveer tot 10), en bij frequenties vanaf 10000 onregelmatigheden zichtbaar. Deze zijn bij de woordtokens in deze grafiek omvangrijker dan bij de woordtypen in afbeelding 4-8b.

We kunnen uit het bovenstaande toch concluderen dat de aantallen woordtokens en -typen per frequentieklasse, wanneer deze klassen logaritmisch zijn ingedeeld en de aantallen eveneens logaritmisch in een grafiek worden afgezet, over een tamelijk breed gebied een vrijwel lineair verloop kennen. Een mogelijk verklaring voor het wat minder gelijkmatige verloop bij de laagste en hoogste frequenties is dat de frequentieklassen hier slechts weinig verschillende frequenties bevatten, waardoor bijvoorbeeld een wat onregelmatige toename van het aantal typen duidelijker zichtbaar wordt.



Afb. 4-9b Aantal woordtokens per frequentieklasse, y-as met logaritmische schaalverdeling

Voor het aantal tokens speelt verder waarschijnlijk nog een rol dat de hogere frequentieklassen (zo vanaf klasse 40) vrijwel uitsluitend nog unieke frequenties bevatten (het aantal frequenties is gelijk aan het aantal typen) waardoor het onregelmatige verloop van het aantal typen versterkt wordt overgenomen.

De verwachting is dat deze frequentieverdeling door zijn regelmatige vorm, zonder uitgesproken pieken of dalen, in principe een goed uitgangspunt vormt voor vergelijking van de structuur van het woordgebruik met dat in andere corpora. Ook voor deze corpora zal daarom in de volgende hoofdstukken een zelfde frequentieverdeling worden samengesteld.

4.4 De verhouding tussen woordtokens en woordtypen

4.4.1 Token-type ratio's

Het gehele corpus wetgevingsteksten telt in totaal 12807659 woordtokens, en 143156 woordtypen, hetgeen een overall-token/typen ratio oplevert van 89,47. Dat betekent dus dat ieder type in het corpus gemiddeld bijna 90 keer wordt gebruikt. Deze token/type ratio is in hoge mate gerelateerd aan de corpusgrootte, zoals uit tabel 4-3 blijkt. In deze tabel wordt het aantal woordtokens en woordtypen vermeld in deelcorpora van steeds toenemende omvang uit het corpus wetgevingsteksten (uit het corpus worden 'random' steekproeven van telkens 1000 documenten meer genomen). De laatste regel in de tabel betreft het gehele corpus (18803 documenten). Wat daarbij opvalt is dat de token/type ratio een vrij constante toename te zien geeft; vanaf een aantal van 4000 documenten stijgt de ratio steeds met ongeveer 2 tot 4 voor

elke volgende steekproef. De procentuele toename (kolom 5 in de tabel) maakt in het begin wat grotere sprongen, wat natuurlijk veroorzaakt wordt doordat de ratio op dat punt zelf absoluut gezien nog laag is, en komt uiteindelijk uit op zo'n 2 tot 3% per 1000 toegevoegde documenten.

Aantal doc.	Typen	Tokens	Tok/Typ ratio	Toename	Log.Tok/ Typ ratio	Toename
1000	26207	685780	26,168	-	1,321	-
2000	41906	1402463	33,467	27,89%	1,330	0,68%
3000	52917	2053562	38,807	15,96%	1,336	0,49%
4000	58583	2713102	46,312	19,34%	1,349	0,97%
5000	69165	3465062	50,098	8,18%	1,351	0,14%
6000	75515	4069918	53,895	7,58%	1,355	0,28%
7000	81153	4775941	58,851	9,19%	1,360	0,41%
8000	87175	5401861	61,966	5,29%	1,363	0,17%
9000	96295	6186064	64,241	3,67%	1,363	0,00%
10000	101145	6797924	67,210	4,62%	1,365	0,17%
11000	106444	7527958	70,722	5,23%	1,368	0,20%
12000	112731	8351183	74,081	4,75%	1,370	0,16%
13000	118260	8915496	75,389	1,77%	1,370	0,00%
14000	120175	9519981	79,218	5,08%	1,374	0,27%
15000	125979	10270620	81,526	2,91%	1,375	0,07%
16000	130213	10947018	84,070	3,12%	1,376	0,11%
17000	134404	11581901	86,172	2,50%	1,377	0,08%
18000	139758	12256330	87,697	1,77%	1,378	0,02%
18803	143156	12807659	89,466	2,02%	1,379	0,07%

Tabel 4-3 Token/type ratio als functie van de corpusgrootte

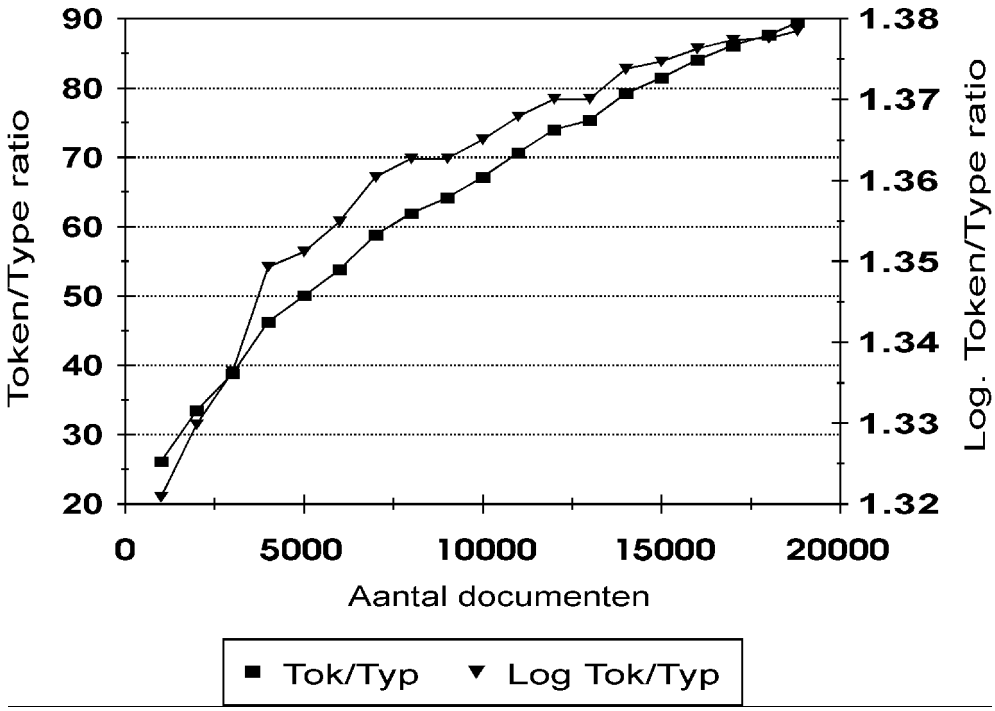
Het is in het geval van dit corpus dus niet zo dat de token/type ratio zich bij toenemende corpusgrootte steeds meer stabiliseert, en uiteindelijk constant blijft. Herdan[9] geeft aan dat dit een algemene karakteristiek is van tekstcorpora, en trekt daaruit de conclusie dat de verhouding tussen het aantal tokens en typen onvoldoende constant is om deze zonder meer te kunnen gebruiken voor bijvoorbeeld het vergelijken van schrijfstijlen. Hij geeft, onder andere met verwijzing naar werk van Chotlos en Devooght, echter aan dat dit anders ligt voor (alweer) de *logaritmische type/token* ratio, dat wil zeggen de logaritme van het aantal typen gedeeld door de logaritme van het aantal tokens. Deze verhouding zou min of meer constant moeten zijn bij toenemende corpusgrootte[10]. Voor de token/type ratio, de reciproque van de type/token ratio, zou dan natuurlijk hetzelfde moeten gelden.

---

9 Herdan (1960, p. 26).

10 Dit is de van Herdan bekende verhouding $\log V = C \log N$, ofwel $V = N^{\alpha}$, die al in hoofdstuk 2 werd vermeld. Herdan meende in feite dat op taaluitingen de zogenaamde Wet van relatieve groei van toepassing was, die tot op dat moment met name in verband was gebracht met biologische verschijnselen. Objecten waarop deze wet van toepassing is staan in een bepaalde, vaste verhouding tot elkaar, bijvoorbeeld van het type $y=bx^c$, waarbij b en c constanten voorstellen. Deze verhouding is in feite identiek aan de in een van de volgende paragrafen behandelde formule van Erikstad. Herdan meende echter dat de constante b voor taaluitingen niet van belang was; hij leidde af dat de verhouding tussen de aantallen woordtokens en -typen dient te luiden zoals hierboven aangegeven (Herdan 1960, p. 28). Zie verder paragraaf 4.4.3, hierna.

Helaas blijkt dat in het geval van dit corpus niet op te gaan. Zoals uit tabel 4-3, kolom 6 blijkt is ook de logaritmische token/type ratio zeker niet constant, maar stijgt ook deze voortdurend. Weliswaar is de range waarbinnen deze stijging plaatsvindt veel kleiner, maar wanneer we de twee ratio's grafisch weergeven en voor elk van de twee een aan hun range aangepaste schaalverdeling op de y-as aanbrengen (zie afbeelding 4-10) blijkt het patroon dat zich aftekent voor beiden sterke overeenkomsten te vertonen. Het hanteren van een logaritmische token/type ratio biedt wat dat betreft weinig voordelen. In paragraaf 4.4.3 zal duidelijk worden dat deze verhouding niettemin van groot belang is voor het verkrijgen van inzicht in de eigenschappen van een tekstcorpus, en tevens voor het onderling vergelijken van corpora.



Afb. 4-10 'Gewone' en logaritmische token/type ratio bij toenemende corpusgrootte

Het is natuurlijk interessant om na te gaan of het bovenstaande ook opgaan voor de andere twee corpora. De gegevens in tabel 4-3 zullen daartoe verderop in dit proefschrift met die van de andere corpora worden vergeleken.

4.4.2 De karakteristiek K van Yule en Herdan

Zoals al in hoofdstuk 2 werd aangegeven is de verhouding tussen het aantal woordtokens en woordtypen de afgelopen decennia meer dan eens voorwerp van onderzoek geweest. Dergelijk onderzoek resulteerde soms in de formulering van bepaalde maatgetallen of karakteristieken, die volgens de onderzoekers in kwestie in meerdere of mindere mate geschikt waren om de opbouw van teksten te beschrijven. Ook de logaritmische token/type ratio uit de voorgaande subparagraaf vormt daarvan een voorbeeld. In aansluiting daarop zal nu van enkele andere karakteristieken worden nagegaan of deze de verhoudingen in het corpus wetgevingsteksten correct omschrijven.

Als eerste komt daartoe in aanmerking de door Yule geformuleerde karakteristiek *K*, later door Herdan in een iets andere vorm geherdefinieerd en 'The Characteristic' gedoopt. De oorspronkelijke formule van Yule luidt, zoals al in hoofdstuk 2 aangegeven:

$$K = 10.000 \cdot \frac{S_2 - S_1}{S_1^{\,2}} \qquad (4\text{-}4)$$

met $S_1 = \sum r.n_r$ en $S_2 = \sum r^2.n_r$; de formule die Herdan gebruikt is

$$K = \frac{S_2}{S_1^{\,2}} = \frac{v_r^{\,2}}{N} \qquad (4\text{-}5)$$

waarbij v_r staat voor de variatie-coëfficiënt van *r*, gelijk aan $\sigma_r / \bar{r}$, de standaarddeviatie gedeeld door het gewogen gemiddelde van de in het corpus voorkomende frequenties.
Deze karakteristiek werd al in hoofdstuk 2 besproken[11], waarbij onder andere bleek dat Yule *K* met name heeft ontworpen als een maat voor de omvang van de woordenschat in een corpus. Daarvoor komt dan vooral K^{-1} (= 1/*K*) in aanmerking, aangezien *K* in feite een maat is voor de (gemiddelde) herhalingsfrequentie van de woordtypen. Hoe hoger deze herhalingsfrequentie, hoe vaker hetzelfde woordtype wordt gebruikt, en dus hoe kleiner de woordenschat. Herdan laat zien dat de karakteristiek ook op een andere manier kan worden gedefinieerd, namelijk als een speciaal soort variatie-coëfficiënt, die 'voorspelt' welke veranderingen zouden optreden in het gewogen gemiddelde van alle woordfrequenties wanneer steekproeven van verschillende omvang uit het corpus zouden worden getrokken[12]. Dit gegeven zou van belang zijn bij het onderscheiden van verschillende schrijfstijlen. Essentieel voor *K*, op welke wijze ook gedefinieerd, is dat de karakteristiek door beide onderzoekers verondersteld wordt constant te zijn voor steekproeven van alle mogelijk groottes die uit een corpus worden getrokken, waardoor het dus mogelijk zou zijn om door middel van een steekproef van beperkte omvang een typering te geven van de omvang van de woordenschat c.q. van de schrijfstijl in het gehele corpus. Bij deze veronderstelde constante waarde van *K* worden echter door andere onderzoekers kanttekeningen geplaatst[13]. Daarom is het interessant om na te gaan hoe de waarde van *K* zich ontwikkelt in steekproeven van toenemende omvang uit het corpus wetgevingsteksten. Daartoe zijn van dezelfde serie steekproeven waarvan de basisgegevens werden vermeld in tabel 4-3 frequentieverdelingen samengesteld, naar het voorbeeld van de frequentieverdeling voor het gehele corpus uit paragraaf 4.3. Op die manier was het mogelijk ook S_2 ($\sum r^2.n_r$) uit de formule van Yule en Herdan (formules 2-3 en 2-4 uit hoofdstuk 2) te berekenen; S_1 ($\sum r.n_r$) was al bekend, aangezien dit gegeven gelijk is aan het totaal aantal tokens van de steekproef. De resultaten, zowel berekend volgens de oorspronkelijke formule van Yule als volgens de iets vereenvoudigde versie van Herdan, worden weergegeven in tabel 4-4.

Het eerste dat opvalt in deze tabel is dat de verschillende berekeningsmethoden van Yule en Herdan praktisch hetzelfde resultaat opleveren, afgezien van de vermenigvuldigingsfactor van 10000 die Yule toepast. Wanneer we deze factor zouden weglaten zouden de verschillen

---

11 Zie paragraaf 2.3.

12 Herdan gebruikt, zoals in hoofdstuk 2 aangegeven, voor zijn karakteristiek het symbool v_m. De door hem gebruikte definitie sluit nauw aan bij die van Yule. Zie Herdan (1966, p. 101-102).

13 Zie bijvoorbeeld Baayen (1989, p. 75).

tussen de twee versies van *K* liggen in de orde van 1/1000000 of minder, afnemend tot 1/10000000 bij een steekproefgrootte van 14000 documenten en meer. Dat het verschil steeds kleiner wordt is eenvoudig te verklaren, aangezien de factor $1/S_1$ waarmee Yule de waarde van *K* vermindert van steeds minder belang wordt bij toenemende steekproefgrootte.

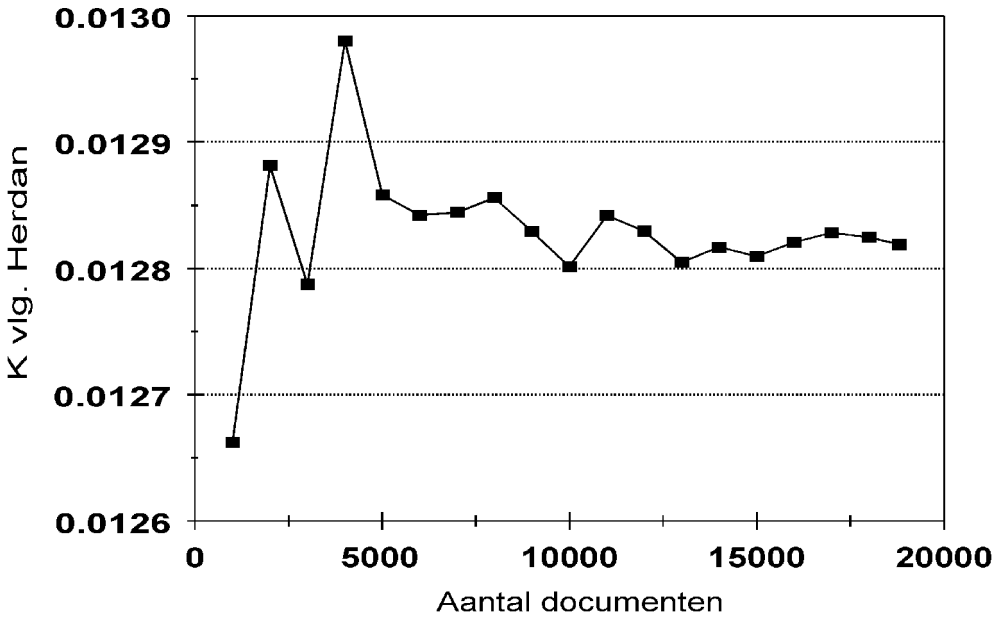
Aantal doc.	Typen	Tokens	*K* (Yule)	*K* (Herdan)
1000	26207	685780	126,609	0,0126624
2000	41906	1402463	128,809	0,0128816
3000	52917	2053562	127,874	0,0127879
4000	58583	2713102	129,802	0,0129805
5000	69165	3465062	128,581	0,0128584
6000	75515	4069918	128,424	0,0128426
7000	81153	4775941	128,445	0,0128447
8000	87175	5401861	128,563	0,0128565
9000	96295	6186064	128,297	0,0128299
10000	101145	6797924	128,016	0,0128018
11000	106444	7527958	128,422	0,0128423
12000	112731	8351183	128,300	0,0128301
13000	118260	8915496	128,052	0,0128053
14000	120175	9519981	128,171	0,0128172
15000	125979	10270620	128,100	0,0128101
16000	130213	10947018	128,208	0,0128209
17000	134404	11581901	128,283	0,0128284
18000	139758	12256330	128,250	0,0128250
18803	143156	12807659	128,193	0,0128194

Tabel 4-4 De waarde van *K* bij toenemende steekproefgrootte

Verder blijkt dat de waarde van *K* inderdaad behoorlijk stabiel blijft voor steekproeven van zeer uiteenlopende omvang uit dit corpus. De fluctuatie die optreedt in de waarde bedraagt over de hele linie minder dan 1,8%, en wanneer we alleen de steekproeven van 5000 documenten en meer in beschouwing nemen zelfs minder dan 0,4%. Dit geldt zowel voor *K* berekend volgens de methode van Yule als volgens die van Herdan. De variatie die er nog is treffen we vooral aan bij de kleinste steekproeven (tot 5000 documenten). De waarde stijgt en daalt daar twee keer, alvorens het definitieve niveau van rond 0.01285 te bereiken. Fluctuaties zijn vanaf dat punt niet groter dan ongeveer 0.00005 (zie afbeelding 4-11).

Een en ander leidt tot de volgende conclusies:

- gezien het geringe verschil in uitkomst valt de eenvoudiger berekeningsmethode voor *K* zoals die door Herdan wordt voorgesteld te prefereren boven die van Yule;
- de waarde van *K*, hoe ook berekend, is in hoge mate stabiel voor steekproeven van verschillende omvang uit het corpus wetgevingsteksten; met name bij steekproeven van 5000 documenten of meer vertoont deze waarde nog maar weinig fluctuatie.



Afb. 4-11 De waarde van *K* (Herdan) bij toenemende steekproefgrootte

Bedacht moet wel worden dat zelfs de kleinste van bovenstaande steekproeven (685780 woordtokens) al aanzienlijk groter is dan al hetgeen Yule en Herdan in beschouwing hebben genomen c.q. konden nemen, voor het berekenen van *K*[14]. In hun tijd ontbraken nu eenmaal omvangrijke, in computer-leesbare vorm opgeslagen tekstcorpora, en ook computercapaciteit was niet of in veel geringere mate voorhanden. Wanneer steekproeven van nog veel geringere omvang zouden zijn getrokken uit het corpus wetgevingsteksten zou de waarde van *K* waarschijnlijk een grotere fluctuatie hebben vertoont. De voorlopige conclusie hieruit, welke nog voor de andere corpora getoetst dient te worden, kan zijn dat voor het verkrijgen van een redelijk betrouwbare waarden van *K* een steekproef van - in absolute zin - behoorlijke afmetingen aan te bevelen is. De veel geringere mate van fluctuatie in de waarde van *K* die nog optreedt in de wat grotere steekproeven vergeleken met de kleinste doet de vraag rijzen of daarbij nu vooral de absolute grootte van steekproeven bepalend is, of dat een steekproef tenminste een bepaald percentage van het corpus moet omvatten. Hierover zal misschien meer duidelijkheid ontstaan bij de analyse van het corpus jurisprudentieteksten, dat immers een grotere omvang heeft.

Wanneer het ook voor de andere corpora mogelijk zal blijken een stabiele waarde voor *K* vast te stellen, lijkt deze karakteristiek wel geschikt als maatstaf bij het vergelijken van de drie corpora, waartoe in hoofdstuk 7 zal worden overgegaan. De waarde van *K* wordt in feite bepaald door de volledige frequentieverdeling van een corpus, en zal dus bijvoorbeeld stijgen naarmate daarin een groter aantal hoge frequenties is vertegenwoordigd[15]. Daarom wordt bij het vergelijken van de waarde van deze karakteristiek in verschillende corpora in feite een

---

[14] Yule (1944) noemt (bijvoorbeeld op blz. 86 en verder) steekproeven met een maximale omvang van 120000 tokens. Een van de grootste door Herdan gebruikte steekproeven telt 133655 tokens (Herdan 1966, p. 109).

[15] Volgens Herdan (1966, p.103) ook wanneer meer van het gemiddelde afwijkende frequenties in de verdeling voorkomen; zie ook hoofdstuk 2, paragraaf 3.

stijlelement vergeleken, namelijk een aspect van de structuur van het woordgebruik. De waarde van *K* heeft dus niet alleen betrekking op de omvang van de woordenschat, zoals Yule meende, maar op een veel complexer geheel van eigenschappen van de tekst, waarvan deze omvang er slechts één is.

4.4.3 Guiraud's R en Herdan's C

Een karakteristiek die beter geschikt zou kunnen zijn voor het meten van de omvang van de woordenschat, en in ieder geval een stabielere maat beoogt te zijn voor de verhouding tussen het aantal woordtokens en woordtypen dan de token/type ratio, is Guiraud's *R*, het quotiënt van het aantal woordtypen dat wordt aangetroffen in een steekproef uit een bepaald corpus, en de wortel uit de omvang (het aantal tokens) van die steekproef[16]. Ook deze karakteristiek werd door Guiraud verondersteld min of meer onafhankelijk te zijn van de steekproefgrootte, wat dus zou betekenen dat er een vaste en tamelijk simpele verhouding zou bestaan tussen het aantal tokens en typen in een corpus. Zoals echter al in hoofdstuk 2 bleek is *R* voor geen van de drie corpora constant bij toenemende steekproefgrootte, maar heeft de karakteristiek een voortdurend stijgend verloop, waarbij met name de curve van het corpus algemeen Nederlands duidelijk stijler verloopt, wat zou kunnen wijzen op een structureel hoger aantal typen (en dus een lagere token/type ratio) in dat corpus. Op dit laatste punt zal in de hoofdstukken 6 en 7 worden terugkomen. Een conclusie die nu al, op basis van de gegevens uit hoofdstuk 2, valt te trekken is dat de karakteristiek van *R* in principe een zelfde soort verloop kent als de token/type ratio, dat wil zeggen dat de waarde van *R*, zij het wat minder sterk dan de token/type ratio, blijft stijgen bij toenemende steekproefgrootte. *R* gedraagt zich in de praktijk dus als een wat afgedempte versie van de token/type ratio. Dit is theoretisch als volgt te verklaren:

- de token/type ratio is:

$$T_N = \frac{N}{V} \qquad (4\text{-}6)$$

- Guiraud's *R* is gedefinieerd als:

$$R = \frac{V}{\sqrt{N}} \qquad (4\text{-}7)$$

- De eerste formule ingevuld in de tweede levert op:

$$R = \frac{N}{T_N\sqrt{N}} = \frac{1}{T_N} \cdot \frac{N}{\sqrt{N}} = \frac{1}{T_N} \cdot \sqrt{N} \qquad (4\text{-}8)$$

R is dus gelijk aan de reciproque van de token/type ratio (en dus gelijk aan de *type/token* ratio, het getal dat aangeeft hoeveel 1 woordtoken gemiddeld bijdraagt aan ieder woordtype uit het corpus) vermenigvuldigd met de wortel uit de steekproefgrootte. Dat *R* blijft stijgen duidt erop dat voor alle drie de corpora deze wortel uit de steekproefgrootte over de hele linie sterker stijgt dan de token/type ratio.

---

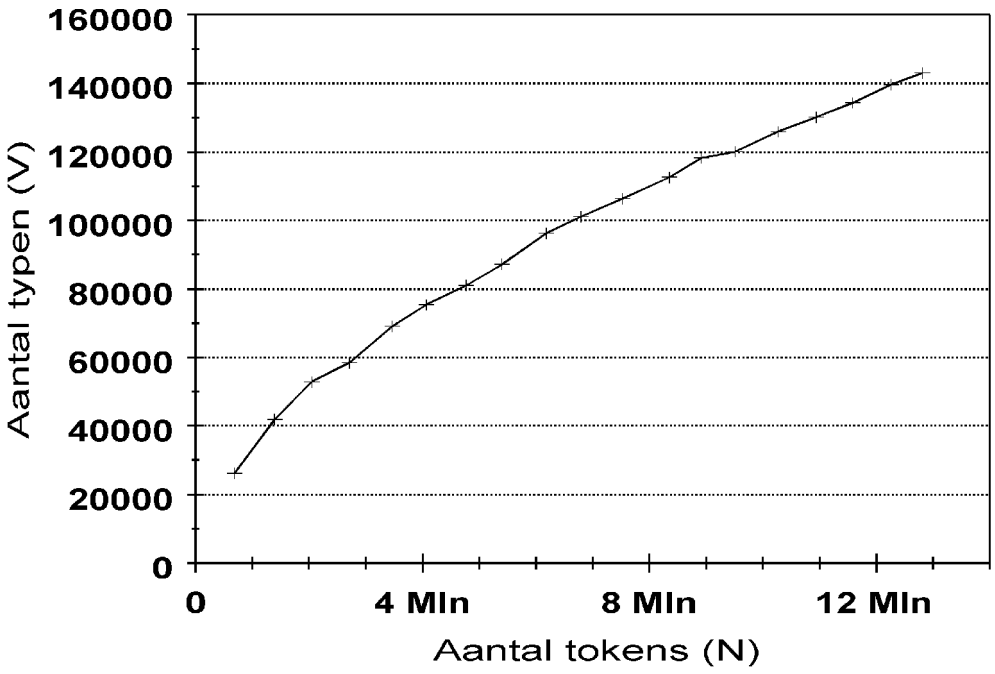
[16] Zie ook hoofdstuk 2, formule 2-5.

Aangezien *R* zich dus niet ontpopt als een constante waarmee de schrijfstijl afdoende kan worden gekarakteriseerd, maar meer als een maatgetal dat nauw samenhangt met de token/type ratio in een steekproef, ligt het voor de hand om te bezien of andere maatgetallen, die ook samenhangen met de token/type ratio, wellicht beter in staat zijn deze verhouding voor een heel corpus te karakteriseren, zodat vergelijking met andere corpora wordt vereenvoudigd.

Zoals al in hoofdstuk 2 opgemerkt komt daarvoor onder andere in aanmerking de door Herdan[17] weergegeven verhouding

$$\log V = C \log N \qquad (4\text{-}9)$$

De formule komt erop neer dat wanneer in een grafiek de logaritme van het aantal woordtypen tegen de logaritme van het aantal woordtokens wordt uitgezet een rechte lijn door de oorsprong zou moeten ontstaan met richtingscoëfficiënt *C*. Het verband tussen aantal typen en tokens kan dus volgens Herdan worden geschreven als $V=N^C$. Inderdaad ontstaat, wanneer we de steekproef-gegevens uit tabel 4-3 grafisch weergeven en de datapunten verbinden door een lijn, een beeld dat sterk doet denken aan een dergelijke exponentiële functie:

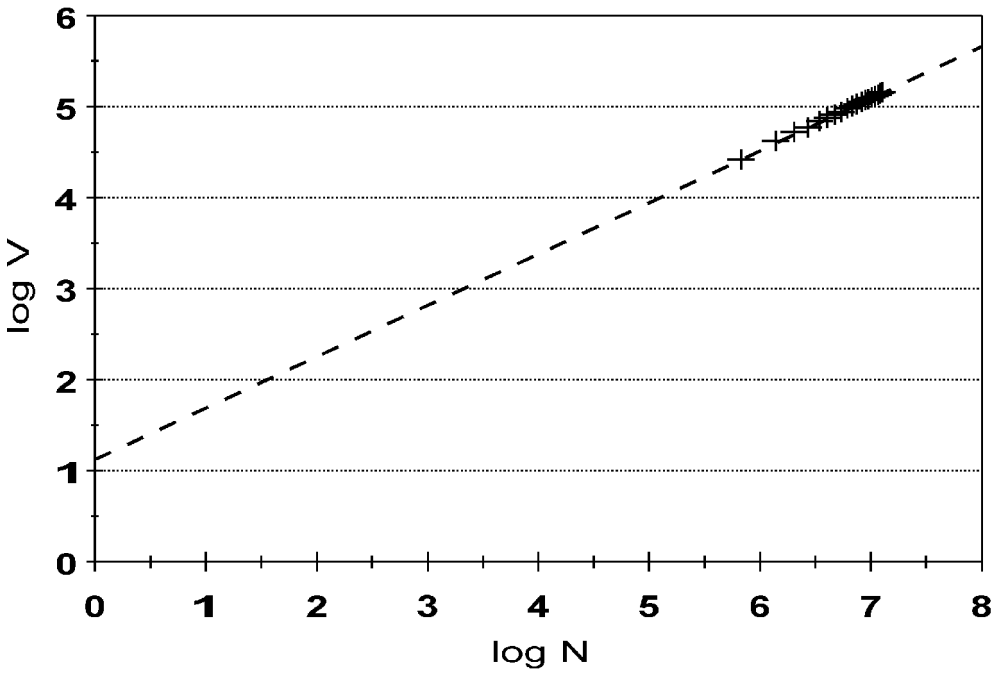


Afb. 4-12 Corpus wetgevingsteksten: het aantal typen als functie van het aantal tokens

Wanneer Herdan's bewering juist zou zijn, zou een grafiek die dezelfde gegevens bevat als die in afbeelding 4-12, maar dan met de logaritme van *N* en *V* afgezet op de x- resp. de y-as, het beeld van een rechte lijn door de oorsprong moeten vertonen. Dat gaat echter voor het corpus wetgevingsteksten in ieder geval niet op: de datapunten liggen wel

[17] Zie bijvoorbeeld Herdan (1960, p. 28).

allemaal min of meer op één (denkbeeldige) rechte lijn, maar die gaat niet door de oorsprong. Om dat aan te tonen is in afbeelding 4-13 de grafiek volgens Herdan's specificatie afgedrukt:



Afb. 4-13 Het verband tussen log *V* en log *N* (datapunten). Tevens is weergegeven de regressielijn van log *V* op log *N* (stippellijn).

De stippellijn in deze grafiek is de lineaire regressielijn van log *V* op log *N*. Deze lijn gaat niet door de oorsprong, en kan dus beschreven worden met een vergelijking van het type:

$$\log V = C \log N + B \qquad (4\text{-}10)$$

waarbij B het snijpunt met de verticale as aangeeft. Dit houdt in dat

$$V = N^{C} . 10^{B} \qquad (4\text{-}11)$$

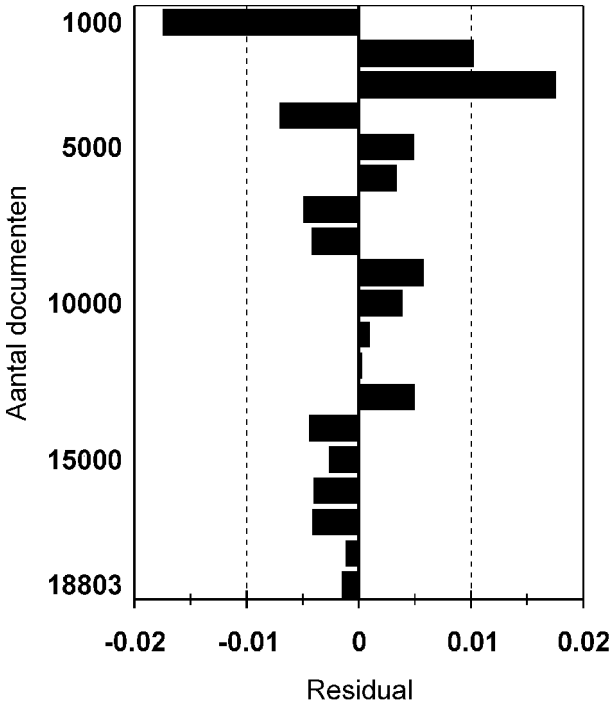
wat weer overeenkomt met de in hoofdstuk 2 reeds aangehaalde formule van Erikstad

$$V = R . N^{C} \qquad (4\text{-}12)$$

De coëfficiënt *C* blijkt bij de afgebeelde lineaire regressielijn een waarde van 0,5675 te hebben, de y-intercept *B* bedraagt 1,1237. r^2, het kwadraat van de lineaire correlatie-coëfficiënt tussen log *V* en log *N*, heeft daarbij een waarde van 0,998634, hetgeen betekent dat 99,86% van de variantie van deze twee variabelen door de regressie wordt 'verklaard', een zeer hoog percentage.

De 'residuals', de verschillen tussen de werkelijke waarden van log *V* en de waarden die uit de vergelijking van de regressielijn volgen zijn steeds gering van omvang (nergens groter dan 0,02), de aan de hand daarvan berekende standaard fout voor deze regressielijn

bedraagt slechts 0.007536. Er valt ook praktisch geen verband aan te tonen tussen de residuals onderling; de covariantie[18] bedraagt minder dan 0,00006. Ook visueel is het patroon dat de residuals vertonen willekeurig (zie afbeelding 4-14). Wel is het zo dat de eerste vier steekproeven (1000-4000 documenten) duidelijk van de rest verschillen, in die zin dat de residuals daar hoger zijn. Net als in de vorige paragraaf lijkt ook hier de boodschap duidelijk: voor het betrouwbaar bepalen van een maat voor het verband tussen aantallen tokens en typen is een bepaalde minimum steekproef-omvang nodig. Die omvang bedraagt in het geval van dit corpus kennelijk zo'n 5000 documenten c.q. rond de 3,5 miljoen woordtokens (bijna een derde van de omvang van het corpus dus).



Afb. 4-14 Residuals van de lineaire regressielijn van log *V* op log *N*

Tenslotte is nog door middel van een *t*-test berekend of het verband tussen log *V* en log *N*, zoals dat uit deze regressielijn naar voren komt, op toeval zou kunnen berusten. Deze test levert de zeer hoge waarde van ruim 111 op, bij n - 2 = 17 vrijheidsgraden. Dat betekent dat de kans op toeval hier verwaarloosbaar klein is. We kunnen daarmee dus stellen dat de lineaire regressielijn uit grafiek 4-13 voor het corpus wetgevingsteksten inderdaad het verband tussen de logaritme van het aantal woordtokens en van het aantal woordtypen in steekproeven van verschillende omvang goed weergeeft. Tevens is duidelijk dat dit verband niet correct kan worden beschreven met de eenvoudige formule van Herdan (4-9), maar dat deze dient te worden uitgebreid zoals aangegeven in (4-10). (4-10) kan worden omgewerkt tot (4-12), de formule van Erikstad uit hoofdstuk 2, waardoor nu, na de meer intuïtieve benadering in hoofdstuk 2, ook op meer 'exacte' wijze is aangetoond dat deze formule in staat moet worden geacht om het verband tussen aantallen woordtokens en -typen, in ieder geval voor wat betreft dit corpus, juist te beschrijven.

Nu dat het geval is, kan tenslotte worden bepaald welke waarde de twee constanten *R* en *C* voor dit corpus aannemen. Daarvoor wordt gebruik gemaakt van vergelijking van de lineaire regressielijn uit afb. 4-13. Deze luidt als volgt:

$$\log V = 0{,}5675 \log N + 1{,}1237 \qquad (4\text{-}13)$$

Hieruit volgt direct de waarde van *C* voor dit corpus, nl. **0,5675**. Aangezien uit (4-11) en (4-12) volgt dat $R = 10^B$, kunnen we berekenen dat *R* hier gelijk is aan **13,2979**. Ter herinnering: Erikstad berekende voor het door hem bestudeerde juridische corpus

[18] Een covariantie van 0 houdt in dat twee variabelen onafhankelijk van elkaar zijn, een hoge waarde (positief of negatief) duidt op een sterke afhankelijkheid.

waarden van respectievelijk 0,6260 en 5,2253. Extrapolatie van de waarde van *V* met behulp van (4-12) en de hier berekende waarden voor *C* en *R* levert een gemiddelde afwijking van iets meer dan 50 (absoluut) op, en een gemiddelde procentuele afwijking van 0,01%. De grootste afwijking die voorkomt bedraagt ruim 2084, de kleinste ruim 39. Al met al kan worden geconcludeerd dat formule (4-12) voor dit corpus goede resultaten oplevert, doordat het verband tussen log *V* en log *N* - zoals al kon worden vermoed in verband met het hoge percentage verklaarde variantie - inderdaad praktisch lineair is. Met de berekende waarden voor *R* en *C* kan het verband tussen het aantal woordtokens en woordtypen in dit geval dus effectief worden weergegeven. Daarom vormen deze constanten mogelijk een goede basis voor een onderlinge vergelijking van de drie corpora.

4.5 Woordlengtes

4.5.1 Inleiding

In de vorige paragrafen is onder andere de frequentielijst van het corpus wetgevingsteksten aan de orde geweest, en is aandacht geschonken aan de structuur van het voorkomen van woordtokens en -typen. In deze paragraaf zal dit eerste onderzoek van het corpus worden afgerond met een analyse van de erin voorkomende woord*lengtes*.

Natuurlijk is er meer te zeggen over het woordgebruik in dit corpus. Zo is bijvoorbeeld interessant of er woordtypen zijn aan te wijzen die kunnen gelden als specifiek voor wetgevingsteksten, of die in dit type teksten juist weinig of helemaal niet voorkomen. Om over dat soort zaken uitspraken te kunnen doen is het echter noodzakelijk om te beschikken over vergelijkingsmateriaal, in dit geval in de vorm van de frequentielijsten van de andere twee corpora. Dit materiaal zal worden gepresenteerd in de volgende twee hoofdstukken, waarna vergelijking van de corpora zal plaatsvinden in hoofdstuk 7. Daarbij kunnen dan ook specifieke kenmerken op woordniveau van elk van de corpora aan het licht komen.

4.5.2 Woordlengte verdelingen

Een eerste statistisch gegeven met betrekking tot de lengte van de woordtypen is de woordlengte-*verdeling*. Daaruit kan worden afgelezen welke woordlengtes in het corpus voorkomen, en hoe vaak. Verder blijkt eruit wat de kleinste en de grootste woordlengte is, en welke lengte het meeste voorkomt[19].

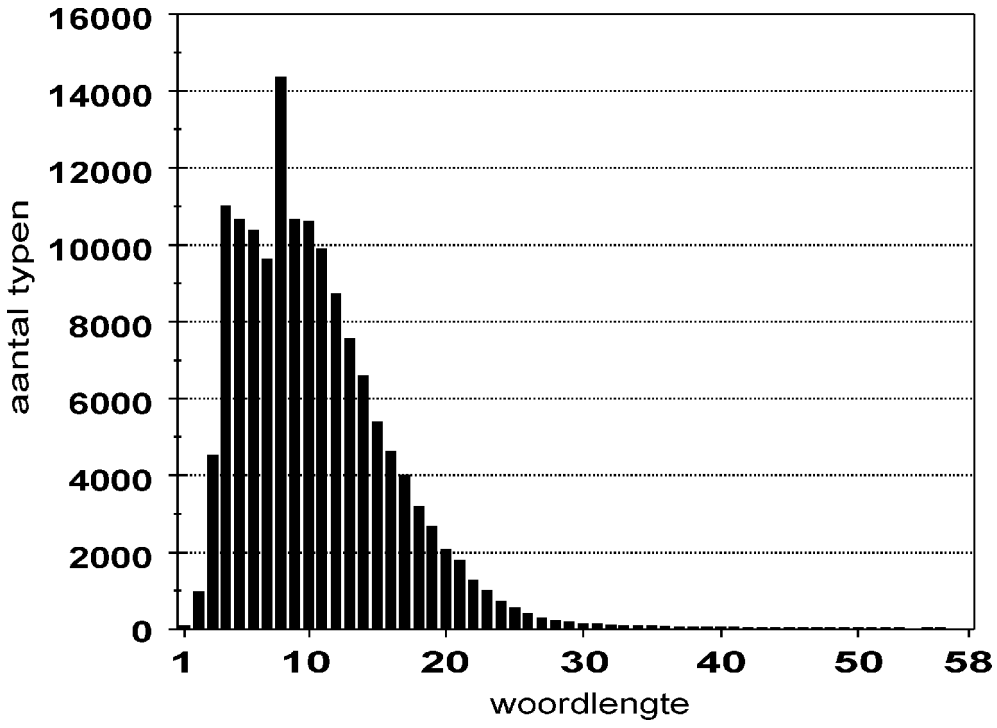
Om deze verdeling samen te stellen is opnieuw de frequentielijst geanalyseerd, waarbij is nagegaan welke woordlengtes voorkomen, hoeveel woordtypen een bepaalde, gelijke lengte hebben, en hoeveel woordtokens uit het corpus deze typen 'vertegenwoordigen'. Enige speciale aandacht was hierbij nodig voor alle woordtypen die, in de vorm waarin ze in de frequentielijst zijn opgenomen, een lengte hebben van precies dertig lettertekens.

[19] Zie voor enkele voorbeelden van woordlengte-verdelingen Martin (1976, p. 275 e.v.).

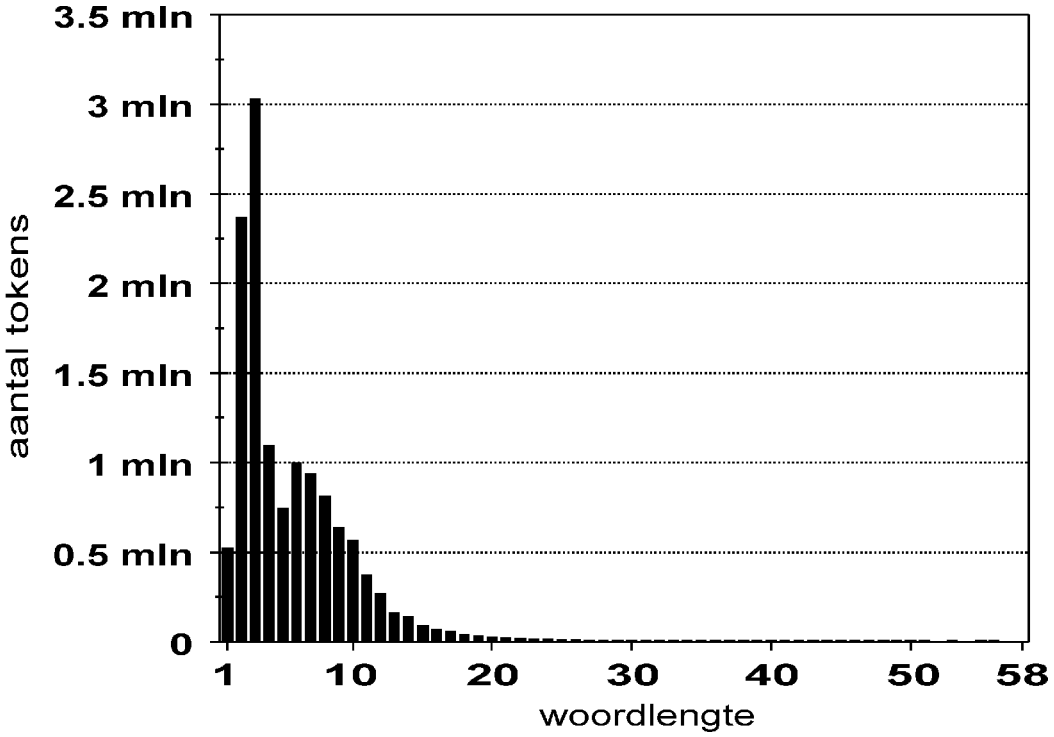
Zoals in het vorige hoofdstuk werd aangegeven, werden bij het opstellen van frequentielijsten woordtypen die uit meer dan dertig lettertekens bestonden 'afgekapt' na het 30e teken. Een aantal woorden met lengte dertig was dus oorspronkelijk langer. Van deze woordtypen (in hun originele vorm) is echter, tegelijk met het opstellen van de frequentielijst, een aparte lijst aangelegd, waarbij ook de frequentie in het corpus werd vermeld. Aan deze lijst zijn bijvoorbeeld de data in paragraaf 3.6.2 (aantallen lange woorden, percentage van het corpus dat deze vertegenwoordigen, etc.) ontleend.

De woordlengte-gegevens uit deze lijst van lange woorden zijn gebruikt om de gegevens uit de frequentielijst te corrigeren. Daartoe werd eerst het totaal aantal woordtypen bepaald waarvan de lengte in feite groter dan dertig lettertekens bleek te zijn. Opvallend was daarbij, dat soms één enkel type uit de frequentielijst bleek te staan voor twee, drie of zelfs vier typen uit de lijst met lange woorden, die slechts van elkaar verschilden ná het dertigste letterteken. Hiermee rekening houdend werd het aantal typen met (volgens de frequentielijst) lengte dertig met het juiste aantal typen verminderd, en werden de aantallen typen in de lengtecategoriën vanaf eenendertig toegevoegd. Op overeenkomstige wijze werden tenslotte ook de aantallen woordtokens per lengtecategorie gecorrigeerd.

Het aantal woordtypen per 'lengte-categorie' blijkt in dit geval een ander patroon te vertonen dan het aantal woordtokens. Daarom zijn beide woordlengte-verdelingen in aparte grafieken weergegeven:



Afb. 4-15a Verdeling van de lengtes van woordtypen



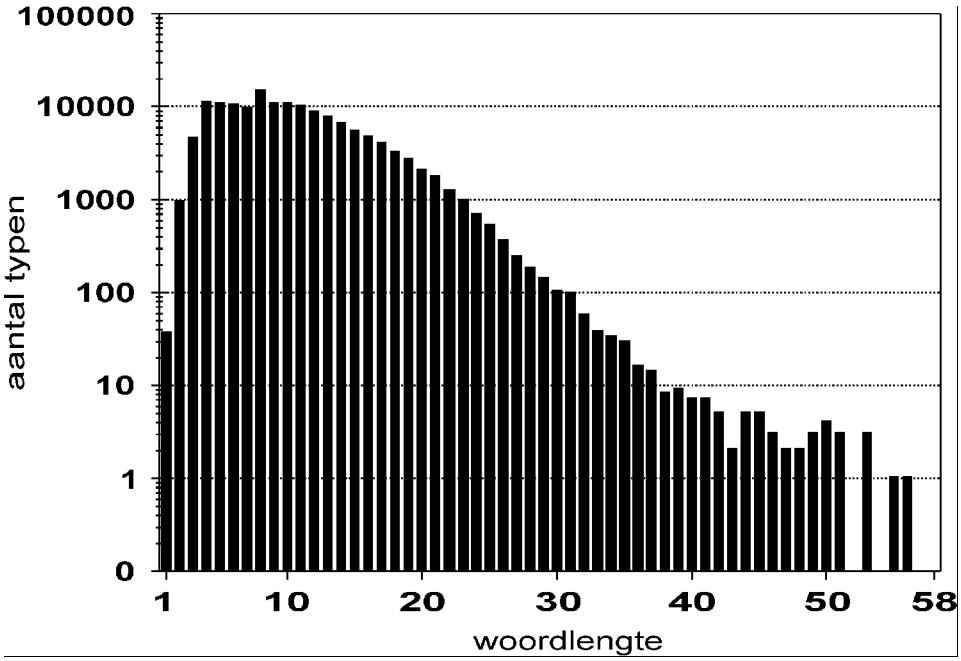
Afb. 4-15b Verdeling van de lengtes van woordtokens

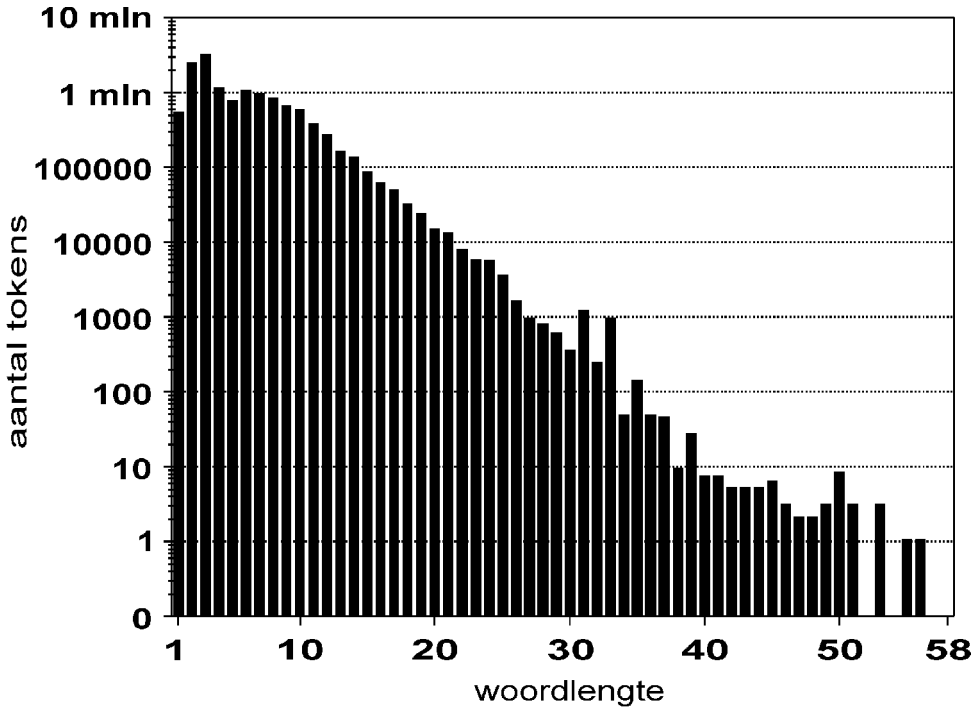
De kortste woordtypen in dit corpus bestaan slechts uit een enkel karakter, het langste type telt 56 lettertekens. Wat opvalt in deze twee grafieken is dat ze allebei een bepaalde inzinking vertonen. Bij de woordtypen bevindt dit zich bij woordlengte 7, bij de tokens bij lengte 5. De inzinking bij de woordtokens betreft mogelijk een voor het Nederlands bekend verschijnsel, waaraan bijvoorbeeld ook Battus[20] refereert. De oorzaak ligt volgens hem hierin, dat een woord altijd uit een geheel aantal lettergrepen bestaat. Een éénlettergrepig woord van 5 letters komt relatief niet zo vaak voor (woorden van bijvoorbeeld drie of vier letters zijn hier veel zwaarder vertegenwoordigd), terwijl woorden van twee lettergrepen in de meeste gevallen zes of meer letters tellen.

Met betrekking tot de woordtypen is waarschijnlijk iets soortgelijks aan de hand, hoewel het patroon hier toch iets anders is. Het lijkt wel of het corpus een enorm groot aantal typen van acht letters bevat, en dat dit ten koste gaat van het aantal typen van 5, 6 en 7, en mogelijk ook van 9 letters. Hiervoor een sluitende verklaring geven is niet eenvoudig, hoewel het niet onmogelijk is dat het verschijnsel samenhangt met het formele karakter van de (wet)teksten in dit corpus. Wanneer we namelijk de 150 meest voorkomende woordtypen bekijken (zie wederom tabel 4-1) blijkt dat praktisch alle typen van acht letters die daarin voorkomen woorden zijn die passen in zo'n wat formeler taalgebruik. Het gaat om de woorden 'minister', 'gemeente', 'bepaalde', 'bedoelde', 'algemene', 'verstaan', 'genoemde', 'regeling' en 'personen'. Wanneer deze verklaring juist is, zou met name het corpus Algemeen Nederlands op dit punt een ander patroon te zien moeten geven. In hoofdstuk 6 zal dit worden nagegaan.

[20] Battus 1983, p. 70.

Een nadeel van beide grafieken is wel dat de verdeling van woordlengtes boven de dertig (en bij de woordtokens zelfs boven de twintig) lettertekens er niet uit kan worden afgelezen. De aantallen typen resp. tokens zijn daarvoor te laag. Met behulp van een logaritmische verdeling op de y-as kan daar echter verbetering in worden gebracht:





Afb. 4-16a en b Verdeling van de lengtes van woordtypen en woordtokens; aantallen per lengte-cat. logaritmisch

De verdeling van de woordtypen blijkt nu in feite behoorlijk gelijkmatig van vorm, de piek bij lengte acht is zelfs praktisch helemaal verdwenen. Kijken we naar de lengtes boven de dertig, dan blijkt dat ook daar nog sprake is van een bijna even gelijkmatige afname van het aantal typen als daarvoor. Pas bij lengtes groter dan veertig wordt het verloop wat grilliger, waarschijnlijk doordat de aantallen typen daar absoluut gezien erg laag worden (vijf of minder). Pas bij lengtes groter dan vijftig ontstaat een tweetal lege lengte-klassen (52 en 54).

Bij de woordtokens zien we praktisch hetzelfde beeld. Enkele lengte-klassen, te weten 31, 33, 35 en 39 geven echter een piek te zien. Deze is, met name bij de eerste twee van deze vier klassen, groter van omvang dan de fluctuaties aan de 'staart' van de grafiek (en aan die van de woordtypen). In het corpus komt dus een onevenredig groot aantal woordtokens voor met een lengte van 31 en van 33 lettertekens, en - wellicht in iets mindere mate - van 35 en 39 lettertekens. Een verklaring voor dit verschijnsel kan gevonden worden door bestudering van de frequentielijst. Tabel 4-5 bevat een uittreksel uit deze lijst, met alle woordtypen die een lengte hebben van meer dan 30 lettertekens, en een frequentie hoger dan 10. De lijst bevat door deze frequentie-grenswaarde alleen die typen die afzonderlijk van substantiële invloed kunnen zijn op het aantal woordtokens in hun lengte-klasse.

Woordtype	Lengte	Frequentie	Aantal doc.
ARBEIDSONGESCHIKTHEIDSUITKERING	31	828	86
HOEVEELHEIDSAANDUIDINGENBESLUIT	31	72	17
ZIEKTEKOSTENVERZEKERINGSBEDRIJF	31	21	7
CANDIDAATNOTARISPLAATSVERVANGER	31	18	4
ASSISTENTSCHEEPSWERKTUIGKUNDIGE	31	17	3
TOEGEPASTNATUURWETENSCHAPPELIJK	31	13	4
ACCOUNTANTADMINISTRATIECONSULENT	32	77	18
PERCEELSGEDEELTENPERCEELPERCELEN	32	34	1
TANDTECHNISCHLABORATORIUMBEDRIJF	32	21	4
ARBEIDSONGESCHIKTHEIDSREGELINGEN	32	13	2
ARBEIDSONGESCHIKTHEIDSVERZEKERING	33	762	193
ARBEIDSONGESCHIKTHEIDSVOORZIENING	33	46	23
ARBEIDSONGESCHIKTHEIDSUITKERINGEN	33	18	9
ADSPIRANTBEROEPSBRANDWEEROFFICIER	33	16	1
ACCOUNTANTSADMINISTRATIECONSULENTEN	35	66	21
VORDERINGSCOMMISSARISMOTORRIJTUIGEN	35	24	1
NULPUNTSTOTALISATIEAANWIJSINRICHTING	36	14	1
KOSTWINNERSVERGOEDINGSBESLUITMILITAIREN	39	13	4

Tabel 4-5 Woordtypen met lengte > 30 en frequentie > 10

Met behulp van de gegevens uit de tabel kunnen de pieken bij lengtes 31 en 33 vlot worden verklaard. Deze komen praktisch volledig voor rekening van de typen 'arbeidsongeschiktheidsuitkering' resp. 'arbeidsongeschiktheidsverzekering', beide termen die kennelijk veel voorkomen in wetgevingsteksten (waarschijnlijk met name in wetgeving op het gebied van de sociale zekerheid). Beide klassen bevatten daarnaast een 'goede tweede', te weten 'hoeveelheidsaanduidingenbesluit' resp. 'arbeidsongeschiktheidsvoorziening'. De piek in de lengte-klasse 35 komt met name voor rekening van het type 'accountants-administratieconsulenten' (in de oorspronkelijke tekst waarschijnlijk meestal geschreven met een verbindingsstreepje). Het enige in de tabel vermelde type met lengte 39, 'kostwinnersvergoedingsbesluitmilitairen', heeft eigenlijk geen opvallend hoge frequentie. Kennelijk is in die klasse sprake van een relatief iets hoger aantal typen (inderdaad

vertoont grafiek 4-16a bij deze lengte een klein hobbeltje) gecombineerd met een wat hogere gemiddelde frequentie dan in de aangrenzende lengte-klassen.

Al met al kan gesteld worden dat de woordlengte-verdelingen voor dit corpus een behoorlijk coherent beeld vertonen, met name wanneer deze verdelingen worden weergegeven in een grafiek met een logaritmische schaalverdeling voor de aantallen typen c.q. tokens per lengteklasse. Deze grafieken bieden interessant materiaal voor vergelijking met de andere twee corpora.

4.5.3 De relatie tussen het rangnummer en de lengte van woordtypen

Zoals hierboven al werd opgemerkt, bevat het corpus wetgevingsteksten woordtypen van zeer uiteenlopende lengte. De woordtypen met de hoogste frequenties hebben veelal een geringe lengte (2 of 3, in het geval van de lidwoorden en voorzetsels die we hier bijvoorbeeld aantreffen). De langste woordtypen vinden we vooral aan de andere kant van de frequentielijst, met frequenties lager dan 10. Het gaat daarbij vooral om lange samengestelde woorden (al dan niet met verbindingsstreepjes ertussen, zoals in adspirant-beroepsbrandweerofficier) waaronder begrepen allerlei (combinaties van) telwoorden (een/negenendertigduizendtweehonderdtweeenvijftigste), en om reeksen woorden die om bepaalde redenen zijn samengevoegd, soms met tussenvoeging van leestekens of streepjes, zoals bijvoorbeeld reeksen plaatsnamen die kennelijk samen een route vertegenwoordigen (Eindhoven-Oss-Arnhem-Apeldoorn-Kampen-Joure) of reeksen stoffen die een chemische verbinding vormen (2-chloor-4-dimethylamino-6-methylppyrimidine).

Er lijkt dus een zeker verband te bestaan tussen de lengte van woordtypen en de frequentie waarmee zij voorkomen in het corpus[21]. Bovenaan de frequentielijst vinden we veel korte en weinig lange woorden, onderaan is het net andersom. Om dit verschijnsel in beeld te brengen, en een eerste indicatie te verkrijgen over de mate waarin de lengte van woordtypen toeneemt naar gelang ze een lagere positie op de frequentielijst innemen, is allereerst een grafiek vervaardigd waarin alle woordlengtes voor ieder in het corpus voorkomend woordtype-rangnummer worden weergegeven door een 'dot'[22]. Het rangnummer is hierbij afgezet op de x-as, de woordlengte op de y-as. De grafiek bleek, net als eerder in dit hoofdstuk vaak het geval was, het beste leesbaar wanneer voor de rangnummers een logaritmische schaalverdeling werd gehanteerd.

De grafiek geeft een eerste indicatie van het verband tussen rangnummer en woordlengte. Punten die opvallen zijn:

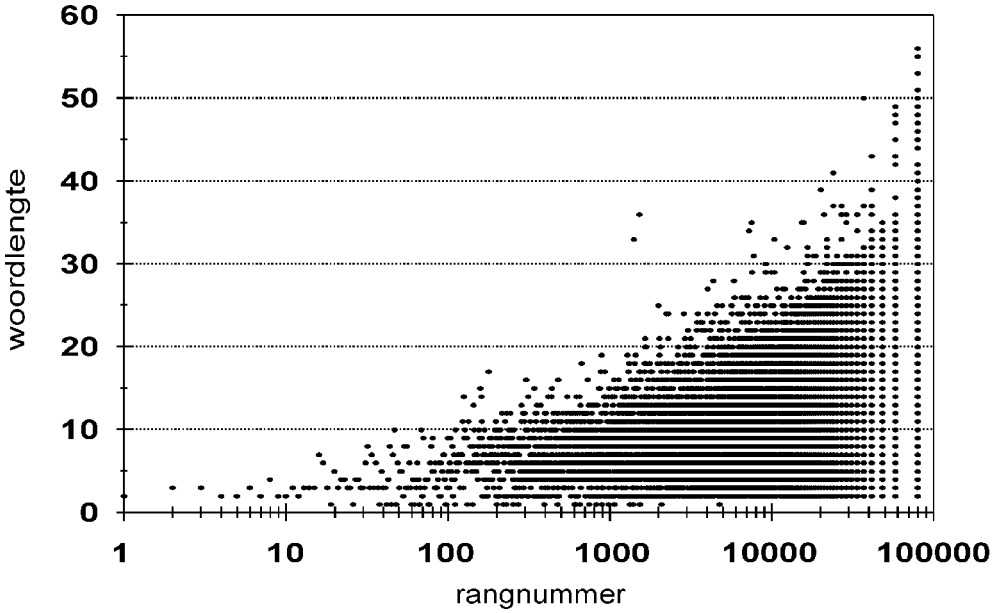
- de bovenste begrenzing van de puntenwolk vertoont een vrij constante stijgende lijn; verder lijkt de wolk op het eerste gezicht geen duidelijke richting te hebben.

---

[21] Dit verband werd al genoemd in hoofdstuk, paragraaf 3. Estoup (1916) refereerde er al aan, en ook Zipf (1935, p. 20-39) besteedde er aandacht aan. Hij benoemde het verschijnsel aanvankelijk als 'The Law of Abbreviation', en later (Zipf 1949) als een van de 'Principles of Least Effort'. Zie ook Rapoport (1982, p. 1-28).

[22] Een rangnummer kan betrekking hebben op meer dan een woordtype; het omvat immers alle woordtypen met een gelijke frequentie. Daarom kunnen bij een enkel rangnummer verschillende woordlengtes horen.

- bij nadere beschouwing blijkt echter wel dat de allerkleinste woordlengte (1) boven een bepaald rangnummer (te weten 4800) niet meer voorkomt.



Afb. 4-17 Alle woordlengtes per rangnummer

- de grafiek geeft een aantal uitbijters te zien, waarvan de twee opvallendste te vinden zijn rond rangnummer 1500. Bij nadere bestudering blijkt het om dezelfde twee woordtypen te gaan die ook in de woordlengte-verdelingen al opvielen, namelijk 'arbeidsongeschiktheidsuitkering' (rangnr. 1440, lengte 31) en 'arbeidsongeschiktheidsverzekering' (rangnr. 1535, lengte 33), beiden typen met een ongebruikelijk grote lengte voor hun rangnummer, of juister gezegd, typen die veel vaker voorkomen dan gezien hun lengte te verwachten zou zijn. Een minder duidelijke uitbijter is te vinden bij rangnummer 37277. Het blijkt daarbij te gaan om het type registeraccountant-accountantadministratieconsulent (lengte 50, oorspronkelijk voorzien van een verbindingsstreepje tussen het eerste en tweede 'accountant'), dat in totaal 5 keer in het corpus voorkomt.

Hoewel een verband tussen lengte en rangnummer niet zonder meer uit de grafiek valt af te leiden, geeft deze wel een indicatie van het *verwachtingspatroon* met betrekking tot de woordlengte. Zeer lange woorden met een relatief hoge frequentie vallen duidelijk uit de toon, zoals in het geval van de hierboven genoemde twee typen rond rangnummer 1500, zeer korte woorden (lengte 1) hebben nooit een hoger rangnummer dan 4800. Om te zien wat dit verwachtingspatroon voor gevolgen heeft voor de (gemiddelde) woordlengte op ieder punt in de frequentielijst, is vervolgens een grafiek vervaardigd waarin telkens de woordlengtes van een aantal rangnummers worden samengenomen. De gemiddelde lengte van elke serie rangnummers wordt als een enkel datapunt in de grafiek afgebeeld. Daartoe werden de volgende gegevens verzameld, uitgaande van de volledige frequentielijst van het corpus wetgevingsteksten:

- uit deze frequentielijst werden telkens clusters van 50 opeenvolgende rangnummersgenomen;

- voor elk van die clusters werd bepaald hoeveel woordtypen erin voorkwamen (sommige typen hebben immers hetzelfde rangnummer) en uit hoeveel lettertekens die in totaal waren opgebouwd; door het totale aantal lettertekens te delen door het aantal woordtypen werd de gemiddelde lengte van woordtypen voor ieder cluster verkregen.

De cluster-grootte van 50 rangnummers werd proefondervindelijk bepaald. Het werken met clusters van 50 rangnummers zorgt ervoor dat de gemiddelde woordlengte niet te zeer fluctueert, en dat het aantal datapunten niet te hoog oploopt, hetgeen de overzichtelijkheid ten goede komt.

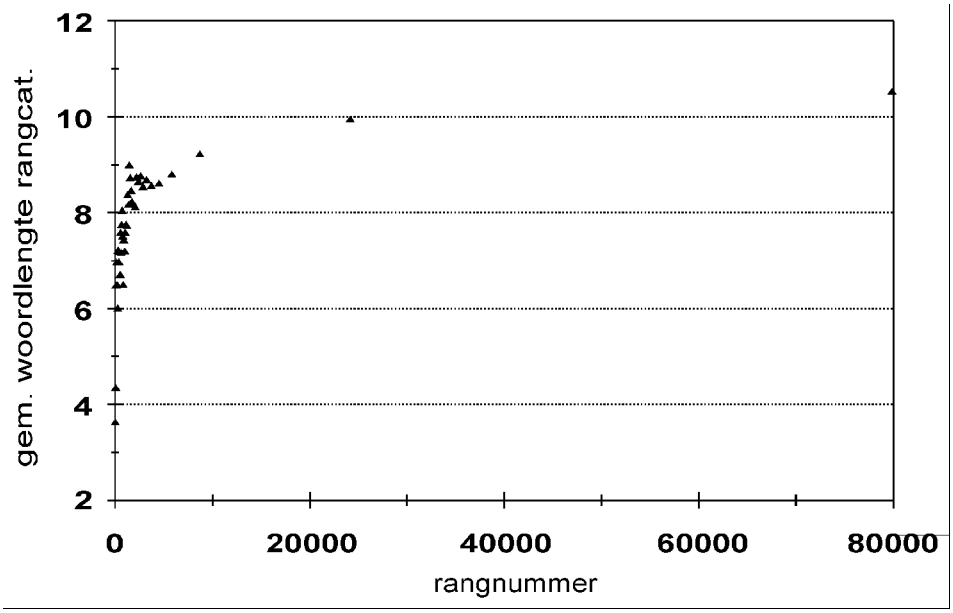
Zowel voor deze grafiek als voor de voorgaande (afbeelding 4-17) moesten weer de gegevens uit de frequentielijst worden gecorrigeerd voor het aantal 'ingekorte' woordtypen (typen die oorspronkelijk uit meer dan 30 lettertekens bestonden). Met behulp van de (gewone) frequentielijst is nagegaan welke rangnummers horen bij de ingekorte typen (aan de hand van hun frequentie), waarna de betreffende rangnummers (voor grafiek 4-17) respectievelijk clusters van rangnummers voor wat betreft de totale woordlengte werden gecorrigeerd: het aantal lettertekens dat door het afkappen was verdwenen werd nu weer bijgeteld. Al met al een redelijk ingewikkelde operatie, die voor grafiek 4-17 zonder meer noodzakelijk was om een juist totaalbeeld te krijgen, maar waarvan het effect op de gemiddelde woordlengtes in déze grafiek (4-18) uiteindelijk gering bleek te zijn: in totaal werden namelijk niet meer dan 2657 lettertekens toegevoegd op een oorspronkelijk totaal van 1479278, een toename van de lengte van alle typen samen (over het hele corpus gemeten) van minder dan 0,2%. Slechts in twee afzonderlijke clusters is sprake van een duidelijk grotere toename, te weten in de clusters met rangnummers 1389-1473 en 1474-1561 (toename in lengte met 4.2 resp. 4.4%). Ook hier gaat het weer om de twee typen 'arbeidsongeschiktheidsuitkering' en 'arbeidsongeschiktheidsverzekering'.

Met behulp van bovengenoemde gegevens is een grafiek vervaardigd waarin op de x-as het rangnummer is afgezet, en op de y-as de gemiddelde woordlengte in iedere rangcategorie, dat wil zeggen voor elk cluster van 50 in het corpus voorkomende rangnummers. Voor ieder cluster van rangnummers is als x-waarde genomen het *hoogste* in dat cluster voorkomende rangnummer.

Met betrekking tot de gemiddelde woordlengte kan men zich nog afvragen of het niet beter zou zijn om een *gewogen* gemiddelde te berekenen, waarbij ieder type even vaak meetelt als het in het corpus voorkomt. Bij controle blijkt dat echter voor dit corpus weinig verschil te maken. Alleen het allereerste cluster krijgt er een wat geringere gemiddelde lengte door (natuurlijk omdat daarin zeer veel voorkomende korte woordtypen als 'de', 'in', 'of' etc. figureren), voor de overige clusters is geen sprake van een verschil van betekenis. Om die reden is er hier voor gekozen te werken met de gemiddelde lengte van de woord*typen*. Zie voor de grafiek afbeelding 4-18.

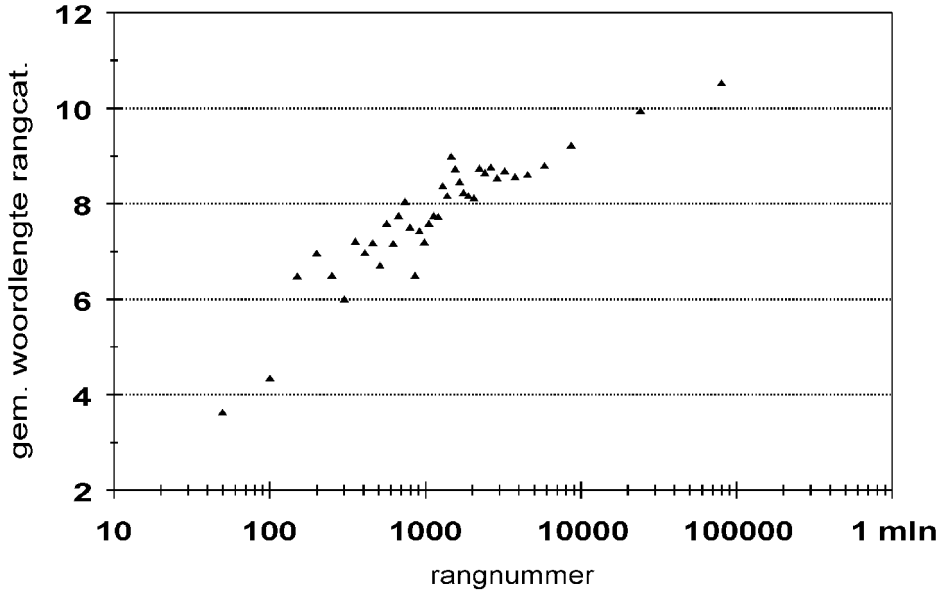
Met opzet werd in dit geval voor het rangnummer op de x-as geen logaritmische schaalverdeling gehanteerd. Hoewel de datapunten daardoor erg ongelijk verdeeld zijn, wordt wel duidelijk welke vorm de grafiek heeft, namelijk die van een typische 'groei'-functie: de gemiddelde woordlengte, gemeten over telkens 50 rangnummers, groeit aanvankelijk snel, en later steeds langzamer van ongeveer 3,5 (bij rangnummer 50) naar een maximum van 10,5 (rond rangnummer 80000). Opvallend is echter wel dat op een zeker punt in de

grafiek (analyse van de data leert dat dit ongeveer bij rangnummer 1500 is) een vrij duidelijke 'piek' zichtbaar is, waarna de grafiek even verder (rond rangnummer 4000) zijn regelmatige 'groei'-verloop herneemt.



Afb. 4-18 Gemiddelde woordlengte per categorie van 50 rangnummers

Natuurlijk kan weer worden geprobeerd om dit verschijnsel wat duidelijker in beeld te brengen door voor het rangnummer op de x-as een logaritmische schaalverdeling te gebruiken. Al eerder in dit hoofdstuk bleek dat deze schaalverdeling goed aansluit bij het karakter van dit gegeven: aan een kant van de schaal liggen rangnummers dicht bijeen terwijl zij grote aantallen woordtypen vertegenwoordigen, aan de andere kant van de schaal liggen ze ver uiteen terwijl ze typen vertegenwoordigen die maar een of twee keer voorkkomen



Afb. 4-19 Gemiddelde woordlengte per categorie van 50 rangnummers, logaritmische schaalverdeling

Het beeld dat *nu* uit de grafiek naar voren komt is dat van een consistente - zij het afnemende - stijging in de gemiddelde woordlengte, ook bij de hogere rangnummers. De puntenwolk vertoont weliswaar enige spreiding, met name in het middengebied, maar de richting van de wolk lijkt vrij duidelijk. De 'piek' rond rangnummer 1500 is nu duidelijker zichtbaar, er lijkt op dat punt inderdaad sprake te zijn van een soort van 'tijdelijk maximum'. Daarna vertoont de grafiek een lichte daling, om tenslotte (rond rangnummer 5000) weer zijn normale, stijgende verloop te hernemen. Een verklaring voor dit verschijnsel is niet eenvoudig te geven, ook niet na bestudering van de frequentielijst rond deze rangnummers. Mogelijk speelt de woordlengte-verdeling een rol. Er waren immers erg veel typen met lengte 8 (zie afbeelding 4-15a), wellicht dat juist in dit deel van de frequentielijst (waar ook de gemiddelde lengte even boven de 8 ligt) extra veel van die typen voorkomen.

Om hierover wat informatie te verkrijgen zijn in tabel 4-6 de aantallen typen met lengte 8 opgenomen per 'cluster' van 500 rangnummers, van rangnummer 1 t/m 5000. Uit de tabel blijkt dan echter dat het aantal typen met deze lengte bij rangnummer 1500 niet het hoogste is, maar pas zijn piek (65) bereikt tussen de rangnummers 2000 en 2500. Dat maakt dat het toch niet erg waarschijnlijk is dat alleen dit gegeven verantwoordelijk is voor de piek rond rangnummer 1000, hoewel het er misschien wel een rol bij speelt.

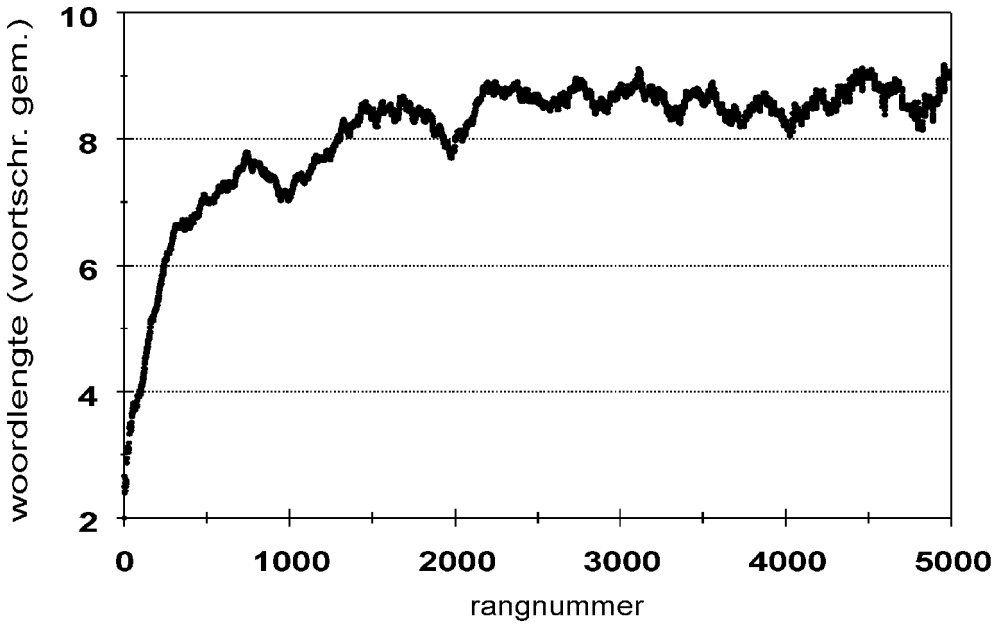
Rangnummer	Aantal typen met lengte 8
1 - 500	52
501 - 1000	53
1001 - 1500	57
1501 - 2000	63
2001 - 2500	65
2501 - 3000	62
3001 - 3500	57
3501 - 4000	62
4001 - 4500	42
4501 - 5000	55

Tabel 4-6 Aantal typen met lengte 8

Een ander opvallend punt in grafiek 4-19 is de nogal abrupte 'sprong' die deze vertoont bij rangnummer 150 (het eerste datapunt na 100). De gemiddelde woordlengte gaat daar in één keer van ongeveer 4,5 naar een waarde van rond 6,5 en blijft vanaf dat punt ook steeds
hoger dan 6. Een probleem is echter wel dat de datapunten in dit deel van de grafiek vrij ver uiteen liggen, waardoor ook het toeval hier een rol zou kunnen spelen.

Al met al geven de grafieken gebaseerd op de gemiddelde woordlengte per cluster van rangnummers wel een vrij goed beeld van de relatie tussen woordlengte en rangnummer, maar laten deze toch nog enige vragen onbeantwoord, met name voor wat betreft de lagere rangnummers (tot ± 5000). Daarom is tenslotte nog één andere techniek beproefd waarmee deze gegevens in beeld kunnen worden gebracht. Het gaat daarbij om het *voortschrijdend* gemiddelde van de woordlengtes, gemeten over telkens 200 woordtypen. Dat wil dus zeggen dat ieder datapunt in de grafiek de gemiddelde lengte vertegenwoordigt van een bepaald woordtype en van de 199 typen daarvóór. Het aantal van 200

typen is weer proefondervindelijk bepaald; bij dit aantal wordt de spreiding van datapunten zover teruggebracht dat als het ware één lijn ontstaat. In de grafiek worden alleen de eerste 5000 woordtypen in beeld gebracht, het deel van de frequentielijst waarover (na grafiek 4-18 en 4-19) nog de meeste onduidelijkheid bestaat. Meer typen zou ook een te groot aantal datapunten opleveren, aangezien een punt is opgenomen voor ieder type, en niet slechts voor elk rangnummer. Onder gebruikmaking van een gewone schaalverdeling ziet een en ander er als volgt uit:

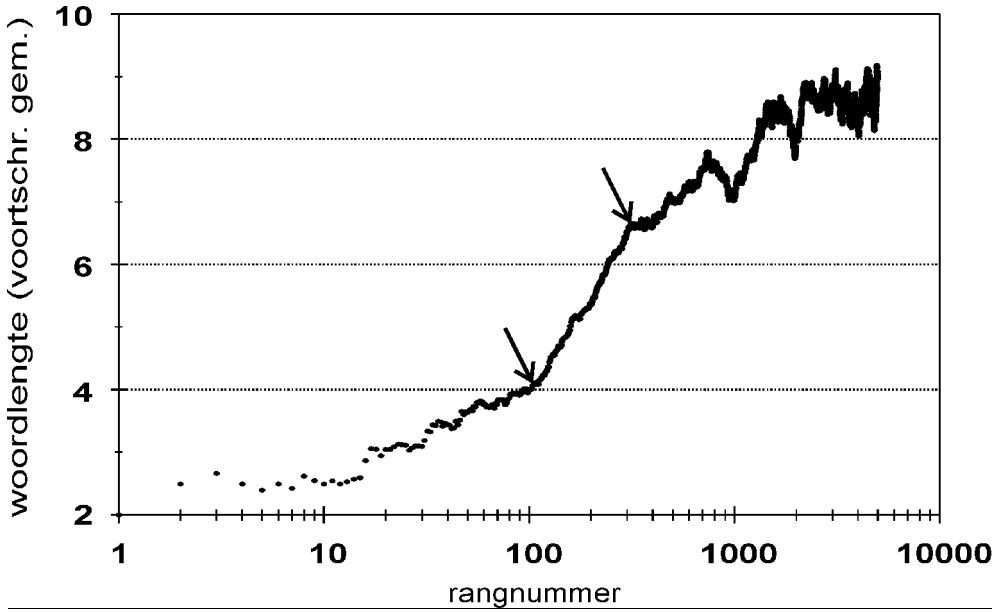


Afb. 4-20 Voortschrijdend gemiddelde van de lengte van woordtypen

De grafiek vertoont in grote lijnen dezelfde 'groei'-karakteristiek als grafiek 4-18, alleen is deze karakteristiek hier veel duidelijker waarneembaar door het grotere aantal datapunten en de gewijzigde schaalverdeling op de x-as. In de grafiek is verder een interessant verschijnsel zichtbaar. De piek rond rangnummer 1500 is duidelijk afgevlakt, wat natuurlijk wordt veroorzaakt doordat ieder datapunt nu de gemiddelde lengte van een veel groter aantal typen vertegenwoordigt. Nog opvallender is echter dat nu aan weerszijden van de piek twee duidelijke inzinkingen zichtbaar zijn, de eerste rond rangnummer 1000 en de tweede bij 2000. De gemiddelde woordlengte neemt in beide gevallen over zo'n 500 rangnummers eerst met ongeveer 1 karakter af en vervolgens weer toe. De mogelijke verklaring voor de aanwezigheid van de piek bij 1500 die hierboven werd genoemd, namelijk een oververtegenwoordiging van typen met lengte 8, wordt hier in feite onderuit gehaald nu blijkt dat de gemiddelde woordlengte juist daalt waar het aantal typen met lengte 8 stijgt (tussen rangnummer 1500 en 2000). Een andere verklaring is eigenlijk niet te vinden aan de hand van het in dit hoofdstuk gepresenteerde materiaal, noch voor de piek, noch voor de twee inzinkingen. Het is natuurlijk wel mogelijk dat het hier een kenmerkend aspect betreft van de structuur van het woordgebruik in dit corpus. In dat geval zouden de overige twee corpora op dit punt een ander beeld moeten vertonen.

Een laatste gegeven dat met behulp van het voortschrijdend gemiddelde van woordlengtes kan worden onderzocht betreft de abrupte 'sprong' in de gemiddelde woordlengte, ergens

rond rangnummer 150, van 4,5 naar 6,5 die in grafiek 4-19 zichtbaar was. Aangezien grafiek 4-19 hierover weinig uitsluitsel kan verschaffen, is hieronder weer een grafiek getekend uitgaande van dezelfde data, maar met een logaritmische schaalverdeling op de horizontale as.



Afb. 4-21 Voortschrijdend gemiddelde van de lengte van woordtypen, logaritmische schaalverdeling voor rangnummers

In deze grafiek is nog wel iets te herkennen van het patroon uit grafiek 4-19, voor wat betreft de rangnummers die in beide grafieken voorkomen (10-5000). Grafiek 4-21 is echter, met name voor rangnummers tot en met 2000, een stuk duidelijker doordat spreiding in de datapunten praktisch niet voorkomt. Verder kan worden geconstateerd dat de 'sprong' uit grafiek 4-19 hier ook zichtbaar is, zij het - als gevolg van de eigenschappen van het hier toegepaste voortschrijdend gemiddelde - op iets andere wijze. Het gedeelte van de grafiek tussen de rangnummers 100 en 300 (met pijlen aangegeven) vertoont namelijk duidelijk een veel sterkere stijging dan de rest van de grafiek. Er is hier dus inderdaad sprake van een afwijkend verloop, dat zich zelfs nog wat verder uitstrekt (namelijk tot rangnummer 300) dan uit grafiek 4-19 kon worden afgeleid.

Ná rangnummer 300 wordt het verloop van de grafiek in toenemende mate grilliger, maar dat wordt veroorzaakt door het (vaste) aantal van 200 waarnemingen waarover het voortschrijdend gemiddelde wordt berekend, in combinatie met de logaritmische schaalverdeling op de x-as. Tenslotte kan nog opgemerkt worden dat ook de inzinkingen rond de rangnummers 1000 en 2000 uit grafiek 4-20 hier goed zichtbaar zijn.

Een conclusie met betrekking tot het verband tussen rangnummer en woordlengte kan zijn dat de lengte van woordtypen gelijkmatig, maar wel steeds minder snel stijgt bij toenemend rangnummer. Wanneer het voortschrijdend gemiddelde (gemeten over 200 typen) wordt afgebeeld in een grafiek waarin voor het rangnummer een logaritmische schaalverdeling wordt gehanteerd, ontstaat voor de kop van de frequentielijst (rangnummers van 1-700) zelfs een regelmatig stijgende, vaak bijna rechte lijn, waarbij moet worden opgemerkt

dat het gedeelte tussen rangnummers 100 en 300 duidelijk wat sterker stijgt dan de overige gedeelten. De resulterende grafiek (afbeelding 4-21) vertoont ook enkele onregelmatigheden (inzinkingen rond de rangnummers 1000 en 2000) die mogelijk kunnen gelden als kenmerkend voor dit corpus wetgevingsteksten.

4.5.4 Zeer lange woordtypen

Aan het einde van deze paragraaf over woordlengtes wordt nog even teruggekomen op de zeer lange woordtypen. Van alle woordtypen die in dit corpus voorkomen is iets meer dan 4,5% langer dan 20 lettertekens, en 0,25% langer dan 30. Er zijn dus maar relatief weinig echt lange woordtypen. Wanneer we gaan kijken hoeveel procent van het corpus wordt 'beslagen' door deze typen, die dus zo'n 4,5% van de frequentielijst vertegenwoordigen, blijkt dat een nog veel geringer percentage te zijn, namelijk zo'n 0,32% (de gemiddelde relatieve frequentie van elk van deze lange typen ligt dus op ruim 0,32% / 4,5% = 7,1%). Toch ligt de veronderstelling voor de hand dat wetgevingsteksten, zoals in dit corpus aanwezig, juist relatief veel lange (en moeilijke) woorden bevatten, in ieder geval meer dan 'gewone' teksten. Dat zou betekenen dat het percentage lange woordtokens in de andere corpora wellicht nog aanzienlijk lager zou kunnen zijn. De volgende hoofdstukken zullen hierover uitsluitsel geven.

Interessant is nu nog de vraag of de lijst met woordtypen langer dan 30 lettertekens, waaraan hiervoor al werd gerefereerd, typen bevat die wellicht kunnen gelden als 'specifiek voor wetgevingsteksten'. Deze lijst is te vinden in bijlage 2, waarbij moet worden opgemerkt dat samengestelde woorden daarin alleen zijn opgenomen wanneer zij geen verbindingsstreepjes e.d. bevatten. Wanneer we deze lijst bekijken, valt op dat zeker de helft van alle daarin voorkomende typen tenminste enige relatie met onderwerpen van juridische aard hebben. In deze categorie treffen we woorden aan als 'aansprakelijkheidsverzekeringen', 'arbeidsongeschiktheidscriterium', 'basisremigratiesubsidieregeling' en 'hoeveelheidsaanduidingenbesluit'. Daarnaast zijn in de lijst veel min of meer technische termen te vinden ('ademhalingsbeschermingsapparaat', 'hartcatheterisatievoorzieningen', 'hogedrukvloeistofchromatografie', etc.), een vrij groot aantal namen van chemische verbindingen en dergelijke ('dinatriumwaterstoffosfaatoplossing', 'desoxyribonucleinezuurmoleculen') en een aantal 'uitgeschreven' getallen (zoals 'honderdnegentigduizendzeshonderdenacht' en 'negenendertigduizendtweehonderdtweeenvijftigste'). We treffen in de lijst maar twee woordsoorten aan, te weten zelfstandige naamwoorden en telwoorden. Hoewel een aantal van bovengenoemde (categoriën van) lange woorden mogelijk kenmerkend zijn voor wetgevingsteksten, is hun nut voor vergelijkingsdoeleinden beperkt, gezien het geringe percentage van het corpus dat zij vertegenwoordigen.

HOOFDSTUK 5 - DE BELANGRIJKSTE KWANTITATIEVE EIGENSCHAPPEN VAN HET CORPUS JURISPRUDENTIETEKSTEN

5.1 Inleiding

Nadat in het vorige hoofdstuk een aantal kwantitatieve analyses werd uitgevoerd op de teksten van het corpus wetgevingsteksten, is het nu de beurt aan het corpus met teksten van rechterlijke uitspraken. Dit corpus is aanzienlijk omvangrijker dan het corpus wetgevingsteksten; het telt in totaal ruim 40 miljoen woordtokens (wetgeving: ruim 12 miljoen). De analyses die zullen worden uitgevoerd zijn in grote lijnen dezelfde als die in hoofdstuk 4. Eerst zal weer aandacht worden besteed aan de frequentielijst van het corpus, en de daaruit af te lezen woordfrequentie- en spreidingsgegevens. Daarna volgt bestudering van de frequentieverdeling, van de token/type ratio en van de lengte van woordtypen. Voor elk van deze aspecten is weer getracht om eigenschappen op te sporen die mogelijk als kenmerkend voor dit corpus kunnen gelden. De grafische weergave van bepaalde gegevens vormt daarbij weer een belangrijk hulpmiddel.

5.2 Absolute en relatieve woordfrequenties

Voor het corpus jurisprudentieteksten is een frequentielijst opgesteld, volgens dezelfde richtlijnen als bij het corpus wetgevingsteksten. Naast het woordtype zelf bevat deze lijst dus weer het rangnummer, de absolute en de relatieve frequentie van het type, en het spreidingspercentage. Een kopie van de eerste pagina van de lijst is te vinden in tabel 5-1. De frequentielijst is, met 265841 woordtypen, helaas weer te omvangrijk om in zijn geheel als bijlage in dit proefschrift op te nemen. Bij het corpus wetgevingsteksten werd uiteindelijk gekozen voor de eerste 300 rangnummers, die ruimschoots de 'kop' van de frequentielijst omvatten. Voor dit corpus zal eerst worden nagegaan of de frequentielijst in grote lijnen dezelfde karakteristieken vertoont als die van de wetgevingsteksten. Indien dat inderdaad het geval is, zal ook hier worden volstaan met het opnemen van dit aantal rangnummers.

5.2.1 Woordfrequenties

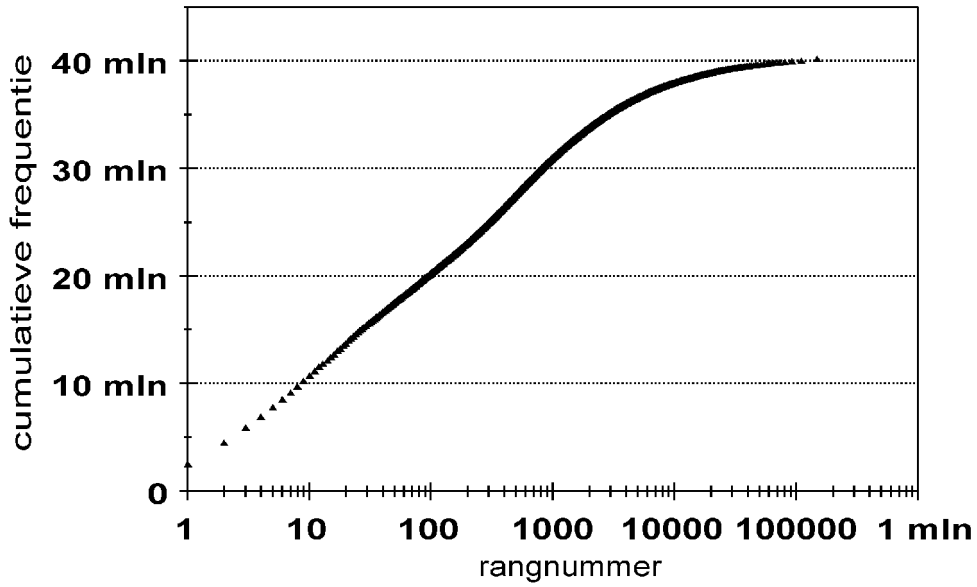
Een eerste element uit de frequentielijst dat daarbij kan worden onderzocht is weer het patroon waarmee woordfrequenties voorkomen in het corpus, ofwel het verband tussen het rangnummer van woordtypen en de frequentie waarmee zij voorkomen. Dit verband is in het vorige hoofdstuk onderzocht aan de hand van grafieken waarin de frequentie of de cumulatieve frequentie was afgezet tegen het rangnummer. Twee grafieken bleken daarbij met name zinvolle informatie op te leveren: de grafiek van de cumulatieve frequenties, met een logaritmische schaalverdeling op de horizontale as, en de grafiek van de frequenties, met een logaritmische schaalverdeling op beide assen.

Rangnr.	Woordtype	Abs. freq.	Rel. freq.	In % doc.
1	DE	2597537	6,4631%	97,57%
2	VAN	1957452	4,8705%	98,59%
3	HET	1406008	3,4984%	95,90%
4	IN	986401	2,4543%	96,23%
5	DAT	917971	2,2841%	95,02%
6	EN	707918	1,7614%	96,06%
7	EEN	649203	1,6153%	94,97%
8	TE	587032	1,4606%	97,79%
9	IS	507448	1,2626%	95,02%
10	OP	476258	1,1850%	94,69%
11	NIET	433890	1,0796%	94,56%
12	HEEFT	374749	0,9324%	92,90%
13	AAN	303584	0,7554%	92,58%
14	DOOR	301730	0,7508%	92,95%
15	ZIJN	300133	0,7468%	92,81%
16	MET	288083	0,7168%	92,26%
17	VOOR	283602	0,7056%	91,83%
18	DIE	252974	0,6294%	90,47%
19	TOT	241589	0,6011%	92,41%
20	ALS	236583	0,5887%	91,51%
21	BIJ	222475	0,5536%	91,55%
22	WORDEN	219272	0,5456%	92,99%
23	ART	197784	0,4921%	84,39%
24	DEZE	191315	0,4760%	90,10%
25	OF	187144	0,4656%	86,74%
26	HOF	174874	0,4351%	62,40%
27	DIT	149037	0,3708%	86,53%
28	RB	144535	0,3596%	74,47%
29	OOK	141002	0,3508%	85,34%
30	TEN	139606	0,3474%	87,78%
31	ZOU	124392	0,3095%	79,65%
32	KAN	117869	0,2933%	85,93%
33	OM	117551	0,2925%	83,85%
34	UIT	113662	0,2828%	83,38%
35	GEEN	109878	0,2734%	84,36%
36	NAAR	108143	0,2691%	82,33%
37	DAN	107068	0,2665%	81,68%
38	1	106313	0,2645%	80,97%
39	WAS	102549	0,2552%	74,25%
40	HAAR	101906	0,2536%	68,44%
41	BEROEP	101386	0,2523%	73,25%
42	ZICH	101142	0,2517%	83,40%
43	ONDER	95641	0,2380%	75,67%
44	HIJ	94897	0,2361%	72,90%
45	HEBBEN	91092	0,2267%	77,46%
46	ZIJ	90443	0,2250%	73,79%
47	TEGEN	88473	0,2201%	91,62%
48	WORDT	87868	0,2186%	77,74%
49	2	81905	0,2038%	77,30%
50	MIDDEL	76793	0,1911%	51,94%
51	TER	76085	0,1893%	77,72%
52	DER	75763	0,1885%	75,02%
53	LID	75586	0,1881%	62,51%
54	CASSATIE	74482	0,1853%	49,67%
55	F	73116	0,1819%	55,91%
56	VONNIS	68683	0,1709%	62,64%
57	KUNNEN	68493	0,1704%	74,25%
58	WELKE	66580	0,1657%	74,69%
59	HAD	65984	0,1642%	66,33%
60	MOET	65141	0,1621%	77,36%
61	IK	64407	0,1603%	50,30%
62	OVER	60925	0,1516%	68,44%
63	RECHT	60777	0,1512%	73,73%
64	ER	60732	0,1511%	69,80%
65	3	60440	0,1504%	70,83%
66	A	59341	0,1477%	68,20%
67	ZAL	57499	0,1431%	69,34%
68	GEVAL	57025	0,1419%	70,34%
69	GROND	57006	0,1418%	71,16%
70	EERSTE	56442	0,1404%	67,82%
71	O	56221	0,1399%	59,51%
72	ARREST	55966	0,1393%	45,83%
73	P	55786	0,1388%	47,83%
74	HEM	54346	0,1352%	65,43%
75	NOG	54028	0,1344%	69,55%

Rangnr.	Woordtype	Abs. freq.	Rel. freq.	In % doc.
76	MEER	53479	0,1331%	67,43%
77	NU	53478	0,1331%	74,96%
78	WEL	53307	0,1326%	70,19%
79	WET	52320	0,1302%	58,78%
80	NA	50989	0,1269%	67,81%
81	PP	50791	0,1264%	50,23%
82	TUSSEN	50110	0,1247%	62,45%
83	OORDEEL	49750	0,1238%	64,18%
84	NJ	48616	0,1210%	99,91%
85	VORDERING	48171	0,1199%	48,70%
86	4	46054	0,1146%	64,89%
87	BESCHIKKING	46017	0,1145%	35,23%
88	MR	45687	0,1137%	90,01%
89	GESTELD	45461	0,1131%	68,20%
90	VERDACHTE	44389	0,1104%	23,69%
91	HR	43626	0,1085%	46,76%
92	BESLISSING	43426	0,1081%	58,56%
93	OMDAT	42615	0,1060%	62,33%
94	B	42164	0,1049%	56,21%
95	MAAR	41547	0,1034%	58,15%
96	ANDERE	41311	0,1028%	61,91%
97	ZOALS	39692	0,0988%	63,24%
98	5	38859	0,0967%	61,74%
99	HOGER	38682	0,0962%	56,04%
100	MOETEN	38491	0,0958%	61,97%
101	BELANG	38355	0,0954%	57,86%
102	STAAT	38332	0,0954%	55,16%
103	GEGEVEN	38261	0,0952%	60,90%
104	WERD	37950	0,0944%	59,41%
105	VRAAG	37795	0,0940%	53,76%
106	INDIEN	37341	0,0929%	59,91%
107	BESTREDEN	37047	0,0922%	54,81%
108	ZONDER	36575	0,0910%	59,11%
109	SLECHTS	36484	0,0908%	62,48%
110	ONDERDEEL	36343	0,0904%	27,74%
111	6	36309	0,0903%	58,73%
112	ZAAK	35966	0,0895%	56,07%
113	DERHALVE	35892	0,0893%	68,12%
114	RO	35803	0,0891%	24,00%
115	ONDERHAVIGE	35279	0,0878%	59,28%
116	RAAD	35019	0,0871%	54,67%
117	OMSTANDIGHEDEN	34446	0,0857%	51,85%
118	WAARVAN	34194	0,0851%	64,00%
119	GEMEENTE	33963	0,0845%	17,35%
120	HIER	33759	0,0840%	56,73%
121	CONCLUSIE	33559	0,0835%	58,10%
122	ALTHANS	33540	0,0835%	50,08%
123	NR	33141	0,0825%	72,67%
124	DOEN	32955	0,0820%	60,16%
125	AL	32834	0,0817%	57,12%
126	REEDS	32168	0,0800%	56,60%
127	HETGEEN	31919	0,0794%	58,28%
128	S	31537	0,0785%	44,23%
129	DOCH	30841	0,0767%	59,91%
130	BEDOELD	30732	0,0765%	57,04%
131	ZIN	30531	0,0760%	50,64%
132	WAT	30490	0,0759%	53,91%
133	TWEEDE	30267	0,0753%	52,73%
134	GRIEF	29956	0,0746%	30,65%
135	HUN	29805	0,0742%	43,51%
136	DAARVAN	29216	0,0727%	57,36%
137	FEITEN	29204	0,0727%	48,30%
138	BETREKKING	28913	0,0719%	52,99%
139	SUB	28831	0,0717%	38,04%
140	ECHTER	28718	0,0715%	59,18%
141	OVEREENKOMST	28569	0,0711%	22,22%
142	BEDOELDE	28327	0,0705%	52,04%
143	GENOEMDE	28210	0,0702%	55,47%
144	ONRECHTE	27978	0,0696%	49,57%
145	IMMERS	27942	0,0695%	59,26%
146	RECHTER	27833	0,0693%	40,69%
147	FEIT	27827	0,0692%	50,37%
148	ZODAT	27723	0,0690%	61,33%
149	MEDE	27693	0,0689%	54,14%
150	BLIJKT	27535	0,0685%	56,23%

Tabel 5-1 Eerste pagina van de frequentielijst van het corpus jurisprudentieteksten

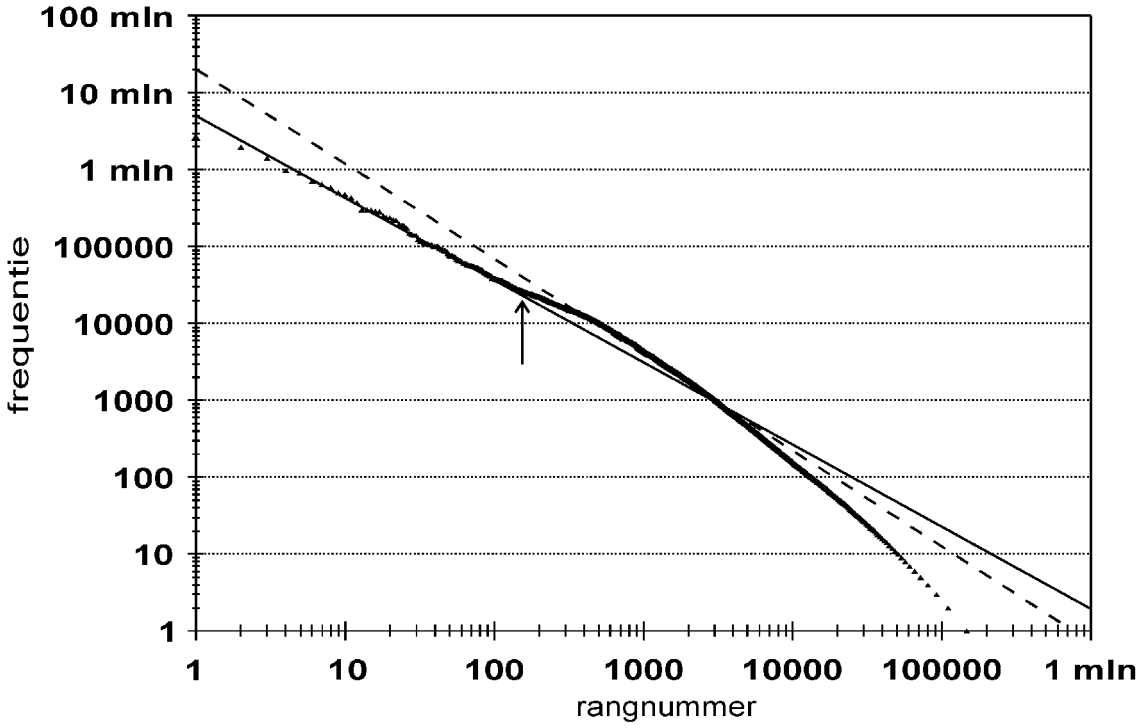
De bedoelde grafieken zijn te vinden in de afbeeldingen 4-2b en 4-3. Beide grafieken toonden een verband tussen de cumulatieve frequenties resp. de frequenties en het rangnummer dat over een bepaald bereik praktisch lineair was. Het is interessant om na te gaan of dit ook bij de jurisprudentieteksten het geval is. Daarom volgt nu eerst weer de grafiek waarin de cumulatieve frequenties zijn afgezet tegen het rangnummer (zie afbeelding 5-1):



Afb. 5-1 Cumulatieve frequenties, logaritmische schaalverdeling voor het rangnummer

Op het eerste gezicht lijkt het of in deze grafiek het 'buigpunt', het punt waar de het verband tussen rangnummer en cumulatieve frequentie niet langer min of meer lineair is, zich eerder voordoet dan bij het corpus wetgevingsteksten. Wanneer we echter kijken naar het rangnummer, dan blijkt dat ook hier het buigpunt ongeveer bij 1000 ligt, hetzelfde rangnummer waar het ook in het vorige hoofdstuk te vinden was. Tegelijk wordt dan duidelijk waarom de grafiek er toch wat anders uitziet: het corpus jurisprudentieteksten is groter, waardoor ook hogere rangnummers voorkomen (en natuurlijk ook hogere cumulatieve frequenties). De grafiek loopt daardoor aan de rechterzijde wat verder door. Dit illustreert tevens dat dergelijke gegevens, zoals absolute (cumulatieve) frequenties en rangnummers, niet goed vergelijkbaar zijn tussen corpora van verschillende omvang. Om die reden zullen dan ook in hoofdstuk 7, waar relevante kenmerken van de drie corpora worden vergeleken, methoden worden besproken en toegepast waarmee gegevens uit de corpora wél vergelijkbaar kunnen worden gemaakt, zoals het gebruik van relatieve cijfers (dat wil zeggen, gerelateerd aan de corpusgrootte), of het trekken van even grote steekproeven uit de corpora. In ieder geval is het zo dat met de ons nu ter beschikking staande gegevens nog geen gedetailleerde vergelijkingen tussen de in de twee corpora voorkomende frequenties kunnen worden getrokken (vooropgesteld dat we dat zouden willen). Wel kunnen de grafieken van de cumulatieve frequenties worden vergeleken voor wat betreft hun algemene vorm. Geconstateerd kan worden dat die voor beide corpora grote overeenkomsten vertoont. Het gebied waarover sprake is van een praktisch lineair verband tussen cumulatieve frequentie en rangnummer komt vrijwel overeen. In dit gebied is net als bij de wetgevingsteksten zo'n 80% van alle tokens te vinden.

Nu de cumulatieve frequenties voor beide corpora praktisch hetzelfde beeld vertonen, is het natuurlijk de vraag of dat voor de 'normale' frequenties, op dezelfde manier als in het vorige hoofdstuk afgezet tegen het rangnummer, ook het geval is. De betreffende grafiek is te vinden in afbeelding 5-2:



Afb. 5-2 In het corpus jurisprudentieteksten voorkomende frequenties (logaritmische schaalverdeling)

Ook bij deze grafiek valt op dat de hoogste waarde bij zowel de x-as als de y-as een factor 10 hoger ligt dan bij het corpus wetgevingsteksten het geval was. De corpusgrootte speelt daarbij waarschijnlijk weer een rol. Wanneer we van dat verschil even afzien, vertoont de grafiek in grote lijnen sterke overeenkomst met die van de wetgevingsteksten. Ook nu is bij de laagste rangnummers weer sprake van een vrijwel rechtlijnig verloop, het is zelfs zo dat de onregelmatigheden die hier bij het corpus wetgevingsteksten nog zichtbaar waren (onder andere een 'bultje' rond rangnummer 25) nu bijna geheel afwezig zijn.

Net als in het vorige hoofdstuk is op twee verschillende manieren de vergelijking uitgerekend van een lineaire regressielijn van de logaritme van de frequentie op de logaritme van het rangnummer:

1. onder gebruikmaking van 1 datapunt per *rangnummer*. De vergelijking van de regressielijn luidt dan:

$$\log r = -1{,}239 \log rangnr + 7{,}299 \qquad (5\text{-}1)$$

   Deze rechte is ingetekend in afbeelding 5-2 in de vorm van een stippellijn.

2. door (in principe) uit te gaan van 1 datapunt per *token*. Bij deze methode was het, net als bij de wetgevingsteksten, nodig om het aantal punten (meer dan 40 miljoen) te reduceren door het nemen van een steekproef. Daarbij is ernaar gestreefd ongeveer even veel datapunten 'over te houden' als bij de wetteksten (ongeveer 7000) en daarop de vergelijking van de regressielijn te baseren. Dat betekende in dit geval dat steeds een van de 40 miljoen mogelijke punten werd geselecteerd, waarna er 5750 werden overgeslagen. Onder gebruikmaking van de op die manier samengestelde verzameling datapunten is opnieuw de vergelijking van de regressielijn uitgerekend, die in dat geval luidt:

$$\log r = -1{,}069 \log rangnr + 6{,}701 \qquad (5\text{-}2)$$

   Deze functie is in afbeelding 5-2 ingetekend in de vorm van een doorgetrokken lijn.

Bij beide regressielijnen is r^2, het kwadraat van de lineaire correlatiecoëfficiënt, ongeveer even hoog (rond de 0,98), ze verklaren dus alle twee een praktisch even groot deel van de variantie in hun dataset. Net als bij de wetgevingsteksten echter, blijkt de tweede regressielijn verreweg het beste de datapunten te volgen in het belangrijke gebied van de eerste 100 rangnummers, en is om die reden, in ieder geval voor dit gebied, te prefereren boven de eerste regressielijn. Verder is het ook nu weer zo dat de datapunten vanaf een bepaald punt (in de grafiek aangegeven met een pijl) niet langer samenvallen met de (tweede) regressielijn. Door uitvergroten van het betreffende deel van de grafiek is vastgesteld dat dat punt voor dit corpus gelokaliseerd kan worden rond rangnummer 168. Het gebied tot en met dit rangnummer kan, net als in het vorige hoofdstuk, worden aangeduid als de 'kop' van de frequentielijst, en het gebied erna als de 'staart'.

Om vergelijking met de andere corpora te vergemakkelijken kan formule 5-2 weer worden opgewerkt naar een vorm zonder logaritmen:

$$r = rangnr^{-1{,}069} \cdot 10^{6{,}701} = \frac{10^{6{,}701}}{rangnr^{1{,}069}} \qquad (5\text{-}3)$$

De waarde van de constante ($10^{6{,}701}$) ligt iets hoger dan bij de wetteksten, maar het ligt voor de hand dat daarbij de corpusgrootte (en de hoogste in het corpus voorkomende frequentie) weer een rol spelen. De exponent van het rangnummer (1,069) ligt juist lager, regressielijn 5-2 verloopt dus iets vlakker dan bij de wetteksten het geval was. Met name dit laatste gegeven kon wel eens interessant zijn, alleen zal in hoofdstuk 7 moeten blijken of dit verschijnsel los staat van de corpusgrootte. Het feit dat de waarde van de exponent ook hier ongelijk is aan 1 betekent in ieder geval dat de wet van Zipf ook voor dit corpus niet opgaat[1].

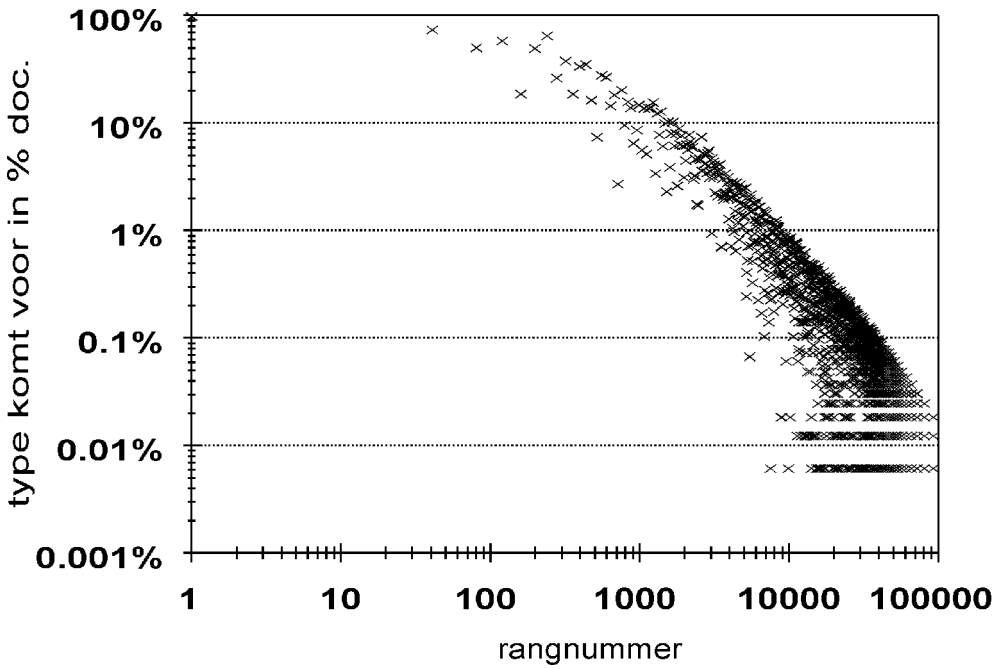
Al met al kan worden gesteld dat het verband tussen rangnummer en woordfrequentie in dit corpus sterke gelijkenis vertoont met dat in het corpus wetgevingsteksten. Net als bij dat corpus zullen we ons daarom in eerste instantie richten op bestudering van de kop van de frequentielijst, die bij dit corpus bestaat uit de rangnummers 1 tot en met (ongeveer) 168.

---

[1] Zie voor enige uitleg op dit punt ook hoofdstuk 4, paragraaf 2.

5.2.2 Spreiding van woordtypen

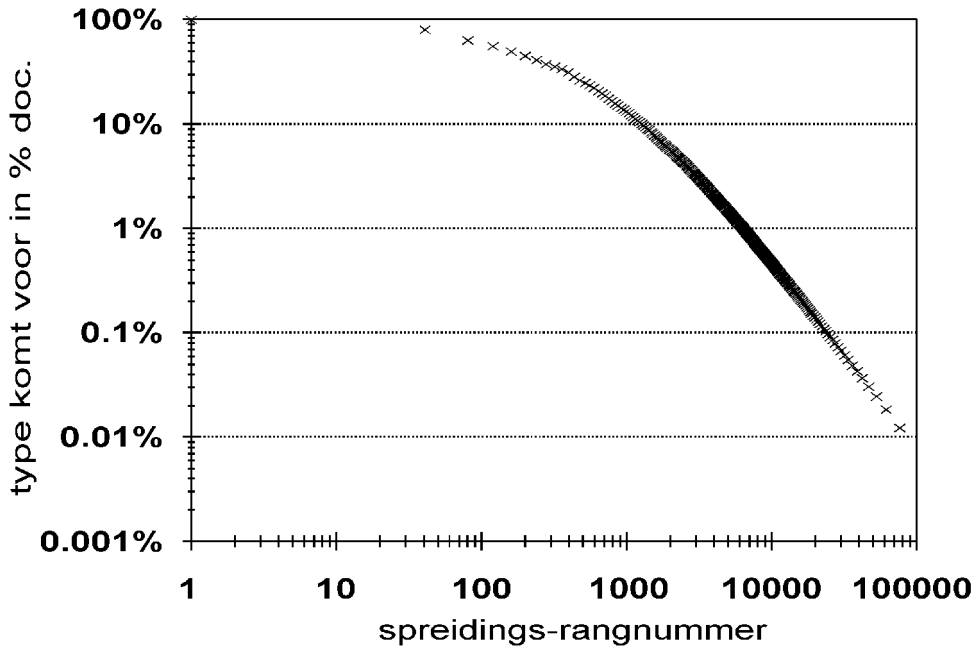
Aan de hand van de frequentielijst kan ook weer de *spreiding* van de woordtypen over het corpus worden bestudeerd. Daartoe kan weer een grafiek worden samengesteld waarin wordt aangegeven in hoeveel procent van de documenten waaruit het corpus is opgebouwd het type c.q. de typen met een bepaald rangnummer voorkomen. Bij gebruik van een logaritmische schaalverdeling voor zowel rangnummer als percentage ontstaat een grafiek die sterk lijkt op afbeelding 4-5 uit het vorige hoofdstuk:



Afb. 5-3 Spreiding van alle typen over het corpus jurisprudentieteksten (logaritmische schaalverdeling)

Ook in deze grafiek worden weer niet alle datapunten weergegeven; om praktische redenen bleek het in dit geval nodig steeds 40 punten over te slaan, hetgeen echter nauwelijks van invloed blijkt te zijn op het totaalbeeld. De puntenwolk vertoont weer het typische 'uitwaaier-effect', veroorzaakt door het steeds toenemende aantal woordtypen (met elk mogelijkerwijs een ander spreidingspercentage) per rangnummer wanneer dit rangnummer stijgt. Een globale vergelijking met grafiek 4-5 leert dat de spreidingspercentages bij de jurisprudentieteksten over de hele linie hoger liggen. Het is echter niet uit te sluiten dat de corpusgrootte daarbij weer een rol speelt. De grafiek vertoont, net als die in het vorige hoofdstuk, weer enkele uitbijters waarbij opvalt dat deze zich bijna allemaal *onder* de puntenwolk bevinden (bij de wetgevingsteksten kwamen duidelijk meer afwijkingen naar *boven* voor). Het gaat daarbij dus bijna altijd om typen met een opvallend geringere mate van spreiding.

Wanneer we de woordtypen weer sorteren op de mate van spreiding, en een 'spreidingsrangnummer' toekennen (het type dat in het hoogste aantal documenten voorkomt krijgt daarbij nummer 1, het type dat in enkele documenten minder voorkomt nummer 2, etc.), ontstaat wanneer we het spreidings-percentage afzetten tegen het spreidings-rangnummer weer een praktisch rechte lijn (zie afbeelding 5-4):



Afb. 5-4 Spreiding van woordtypen over het corpus jurisprudentieteksten, rangnummers(x-as) op basis van de mate van spreiding

Ook nu blijkt de 'lijn' c.q. 'curve' die ontstaat weer ongeveer overeen te komen met de bovenste begrenzing van de puntenwolk uit afbeelding 5-3, zoals ook bij het corpus wetgevingsteksten het geval was. Doordat deze grafiek een zoveel duidelijker beeld geeft van de voorkomende spreidingspercentages, wordt vergelijking met dat corpus ook eenvoudiger. Daarbij blijkt dan dat de spreidingspercentages voor het corpus jurisprudentieteksten niet alleen voortdurend hoger zijn, maar dat de curve ook een wat ander verloop te zien geeft, met in het begin (ongeveer tot rangnummer 1000) een veel sterkere kromming. De lichte 'knik' die in afbeelding 4-6 te zien was is hier niet aanwezig. Een voorzichtige conclusie daaruit kan zijn dat de spreiding van woordtypen in dit corpus niet slechts op enkele punten afwijkt van die in het corpus wetgevingsteksten (waarbij bijvoorbeeld slechts een beperkt aantal typen een andere spreiding vertonen), maar dat sprake is van een afwijkend verloop *over de hele linie*.

Na het bestuderen van de frequentie- en spreidingsgegevens is intussen wel duidelijk geworden dat de frequentielijst van dit corpus over het geheel genomen ongeveer dezelfde karakteristieken vertoont als die van het corpus wetgevingsteksten. Met betrekking tot het spreidingspercentage van woordtypen kan nog wel worden opgemerkt dat dit percentage hier over het geheel genomen hoger ligt, maar evenmin als bij de wetgevingsteksten een duidelijk 'omslagpunt' te zien geeft. Op basis van deze gegevens zullen we ons bij de bestudering van het corpus jurisprudentieteksten primair richten op hetzelfde gedeelte van de frequentielijst, namelijk de eerste 300 rangnummers. Dit gedeelte van de frequentielijst is ook afgedrukt in bijlage 3, volgend op de lijst van het corpus wetgevingsteksten. Ook in dit geval is een volledige frequentielijst beschikbaar voor geïnteresseerden.

5.3 Frequentieverdelingen

Klasse	Grenzen	Klasse-midden	Aantal freq.	Aantal typen	Aantal tokens	Klasse	Grenzen	Klassemid-den	Aantal freq.	Aantal typen	Aantal tokens
1	10^{0} - $10^{0.1}$	1,13	1	118150	118150	34	$10^{3.3}$ - $10^{3.4}$	2253,57	225	308	690477
2	$10^{0.1}$ - $10^{0.2}$	1,42	0	0	0	35	$10^{3.4}$ - $10^{3.5}$	2837,08	213	251	704211
3	$10^{0.2}$ - $10^{0.3}$	1,79	0	0	0	36	$10^{3.5}$ - $10^{3.6}$	3571,67	196	223	793509
4	$10^{0.3}$ - $10^{0.4}$	2,25	1	36624	73248	37	$10^{3.6}$ - $10^{3.7}$	4496,47	157	168	747188
5	$10^{0.4}$ - $10^{0.5}$	2,84	1	18377	55131	38	$10^{3.7}$ - $10^{3.8}$	5660,72	149	161	905076
6	$10^{0.5}$ - $10^{0.6}$	3,57	0	0	0	39	$10^{3.8}$ - $10^{3.9}$	7126,43	131	138	977292
7	$10^{0.6}$ - $10^{0.7}$	4,50	2	20178	89064	40	$10^{3.9}$ - $10^{4.0}$	8971,64	116	120	1072932
8	$10^{0.7}$ - $10^{0.8}$	5,66	1	6330	37980	41	$10^{4.0}$ - $10^{4.1}$	11294,63	98	99	1114477
9	$10^{0.8}$ - $10^{0.9}$	7,13	1	4890	34230	42	$10^{4.1}$ - $10^{4.2}$	14219,09	106	110	1548629
10	$10^{0.9}$ - $10^{1.0}$	8,97	3	10120	89946	43	$10^{4.2}$ - $10^{4.3}$	17900,78	74	74	1307823
11	$10^{1.0}$ - $10^{1.1}$	11,29	2	4548	52113	44	$10^{4.3}$ - $10^{4.4}$	22535,74	61	62	1393816
12	$10^{1.1}$ - $10^{1.2}$	14,22	3	5194	72293	45	$10^{4.4}$ - $10^{4.5}$	28370,82	40	40	1112415
13	$10^{1.2}$ - $10^{1.3}$	17,90	4	4844	84311	46	$10^{4.5}$ - $10^{4.6}$	35716,75	31	31	1110963
14	$10^{1.3}$ - $10^{1.4}$	22,54	6	5185	115749	47	$10^{4.6}$ - $10^{4.7}$	44964,72	15	15	678944
15	$10^{1.4}$ - $10^{1.5}$	28,37	6	3644	103343	48	$10^{4.7}$ - $10^{4.8}$	56607,23	20	20	1120898
16	$10^{1.5}$ - $10^{1.6}$	35,72	8	3404	120008	49	$10^{4.8}$ - $10^{4.9}$	71264,28	12	12	851113
17	$10^{1.6}$ - $10^{1.7}$	44,96	11	3371	151071	50	$10^{4.9}$ - $10^{5.0}$	89716,41	7	7	630319
18	$10^{1.7}$ - $10^{1.8}$	56,61	13	2793	158085	51	$10^{5.0}$ - $10^{5.1}$	112946,27	12	12	1311879
19	$10^{1.8}$ - $10^{1.9}$	71,26	16	2408	171295	52	$10^{5.1}$ - $10^{5.2}$	142190,93	4	4	574180
20	$10^{1.9}$ - $10^{2.0}$	89,72	21	2245	201194	53	$10^{5.2}$ - $10^{5.3}$	179007,78	4	4	751117
21	$10^{2.0}$ - $10^{2.1}$	112,95	25	1868	209668	54	$10^{5.3}$ - $10^{5.4}$	225357,44	4	4	919919
22	$10^{2.1}$ - $10^{2.2}$	142,19	33	1605	225608	55	$10^{5.4}$ - $10^{5.5}$	283708,20	6	6	1730106
23	$10^{2.2}$ - $10^{2.3}$	179,01	41	1354	240995	56	$10^{5.5}$ - $10^{5.6}$	357167,47	1	1	374749
24	$10^{2.3}$ - $10^{2.4}$	225,36	52	1191	266048	57	$10^{5.6}$ - $10^{5.7}$	449647,20	2	2	910148
25	$10^{2.4}$ - $10^{2.5}$	283,71	65	1033	290955	58	$10^{5.7}$ - $10^{5.8}$	566072,29	2	2	1094480
26	$10^{2.5}$ - $10^{2.6}$	357,17	82	883	314580	59	$10^{5.8}$ - $10^{5.9}$	712642,79	2	2	1357121
27	$10^{2.6}$ - $10^{2.7}$	449,65	103	773	345850	60	$10^{5.9}$ - $10^{6.0}$	897164,12	2	2	1904372
28	$10^{2.7}$ - $10^{2.8}$	566,07	128	672	377982	61	$10^{6.0}$ - $10^{6.1}$	1078044,76	0	0	0
29	$10^{2.8}$ - $10^{2.9}$	712,64	161	597	420454	62	$10^{6.1}$ - $10^{6.2}$	1421909,30	1	1	1406008
30	$10^{2.9}$ - $10^{3.0}$	897,16	183	479	428368	63	$10^{6.2}$ - $10^{6.3}$	1790077,75	1	1	1957452
31	$10^{3.0}$ - $10^{3.1}$	1129,46	212	445	498958	64	$10^{6.3}$ - $10^{6.4}$	2253574,37	0	0	0
32	$10^{3.1}$ - $10^{3.2}$	1421,91	235	406	572688	65	$10^{6.4}$ - $10^{6.5}$	2837082,05	1	1	2597537
33	$10^{3.2}$ - $10^{3.3}$	1790,08	240	349	621697						

Tabel 5-2 Frequentieverdeling corpus jurisprudentieteksten. De frequentieklassen nemen logaritmisch in omvang toe.

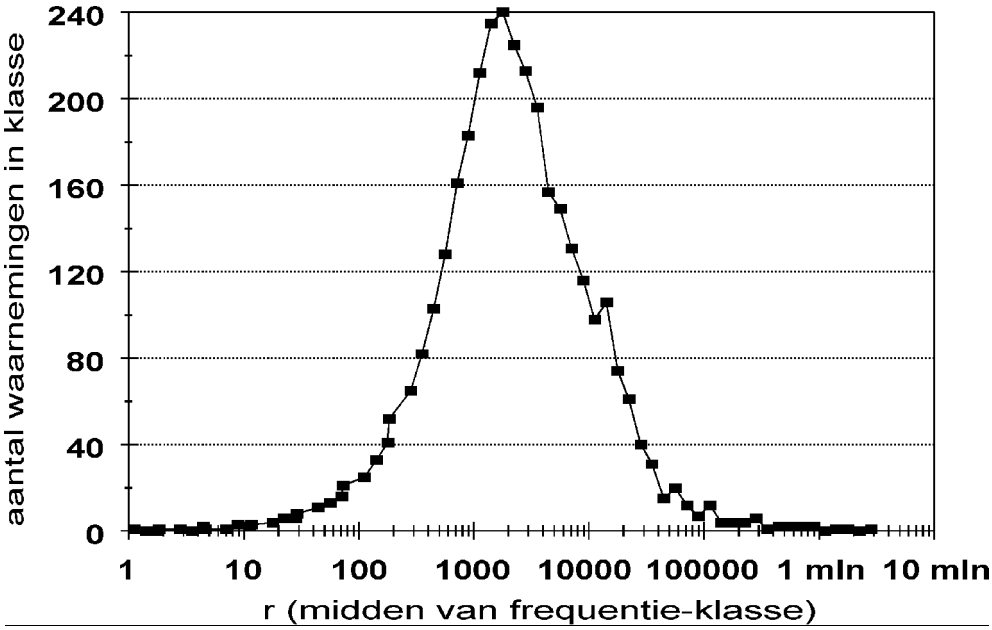
Ook voor het corpus jurisprudentieteksten is een frequentieverdeling opgesteld (zie tabel 5-2). Daarvoor is dezelfde methode toegepast als in het vorige hoofdstuk, dat wil zeggen: frequentieklassen met een logaritmisch toenemende omvang (klassebreedte), om te compenseren voor de grote verschillen in afstand tussen de afzonderlijke frequenties (de lage frequenties zijn talrijk, en liggen dicht bij elkaar, de hoge zijn schaars en liggen ver uiteen). Het aantal frequentieklassen moest in dit geval echter met 5 worden uitgebreid tot 65, om ook de hoogste frequenties die in dit (ruim drie maal zo grote) corpus voorkomen te kunnen indelen. De eerste 60 klassen komen echter helemaal overeen met die bij het corpus wetgevingsteksten.

Ook nu ontstaat op deze manier weer een evenwichtige verdeling van de aantallen verschillende frequenties, met weinig lege klassen (in dit geval 5). De indeling is bovendien, net als bij het voorgaande corpus, in hoge mate symmetrisch: het hoogste aantal

verschillende frequenties (240) treffen we aan in de middelste klasse, de overige klassen geven een gelijkmatig stijgend c.q. dalend verloop te zien. Dat ook bij deze frequentieverdeling weinig informatie verloren gaat blijkt weer wanneer we het totaal aantal tokens en de token/type ratio schatten met behulp van het produkt van de klassemiddens en de aantallen frequenties uit de tabel: de op die manier berekende waarden wijken ongeveer 0.45% af van de werkelijke waarden. Dat is weliswaar wat meer dan bij de wetgevingsteksten (daar lag het percentage op slechts 0,01%), maar het is nog steeds buitengewoon laag. In de volgende subparagrafen zal op de belangrijkste aspecten van deze frequentieverdeling worden ingegaan.

5.3.1 De aantallen frequenties per frequentieklasse

Ook bij dit corpus is het zo dat de aantallen verschillende frequenties per frequentieklasse, in een grafiek afgezet tegen de klassemiddens, een beeld vertonen dat sterk gelijkt op de curve van een normale verdeling (zie afbeelding 5-5).

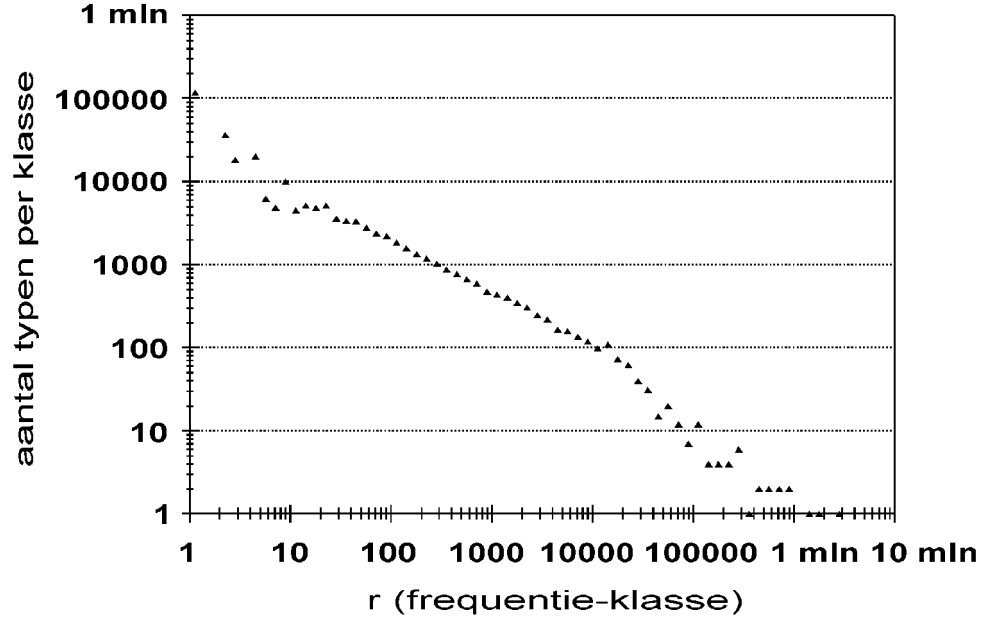


Afb. 5-5 Aantal waarnemingen (verschillende frequenties) per frequentieklasse

De grafiek lijkt, afgezien van de schaalverdeling, sterk op de overeenkomstige grafiek van het corpus wetgevingsteksten (afb. 4-7), maar is nog iets regelmatiger van vorm, met name rond de top. Opvallend is dat net als in afbeelding 4-7 enkele onregelmatigheden zichtbaar zijn in de rechter'flank' van de grafiek, waarbij vooral het in beide grafieken voorkomende bultje rond r = 14000 (om precies te zijn bij 14219,09, klasse 42) opvalt. Een reden voor dat laatste is niet eenvoudig te geven; er kan sprake zijn van toeval, het kan ook zijn dat er een hoge mate van overeenstemming bestaat tussen de frequentieverdelingen van de twee corpora. Wellicht dat vergelijking met de verdeling van het corpus algemeen Nederlands hierop nog wat meer licht kan werpen.

5.3.2 De aantallen woordtypen en -tokens per klasse

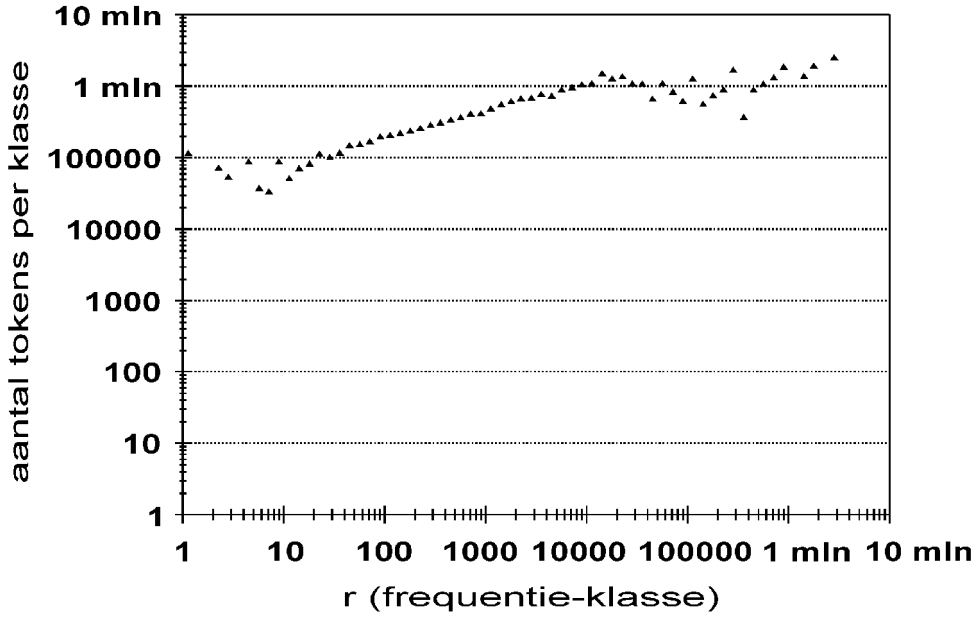
Om inzicht te krijgen in het verloop van de aantallen typen per frequentieklasse, worden deze aantallen weer afgezet tegen de klassemiddens. In hoofdstuk 4 werd al duidelijk dat het daarbij verstandig is niet alleen voor *r*, maar ook voor de aantallen typen gebruik te maken van een logaritmische schaalverdeling, om te voorkomen dat de eerste paar klassen te veel een stempel drukken op het aanzien van de hele grafiek.



Afb. 5-6 Aantal woordtypen per frequentieklasse

Deze grafiek lijkt qua verloop zeer sterk op de overeenkomstige grafiek van het corpus wetgevingsteksten. Net als in laatstgenoemde grafiek is in het middengebied (ongeveer tussen *r* = 10 en *r* = 10000) sprake van een vrijwel rechte lijn, en vertoont de grafiek een wat onregelmatiger verloop bij de laagste en de hoogste frequentieklassen (tot *r* = 10 resp. vanaf *r* = 10000). Wel is ook nu de schaal van de grafiek natuurlijk weer aangepast aan de in dit corpus geldende aantallen. Ook van dit aspect van de structuur van het woordgebruik kan dus pas nauwkeurige vergelijking plaatsvinden nadat de data onderling vergelijkbaar zijn gemaakt. Zie daarvoor hoofdstuk 7.

Datzelfde geldt eigenlijk ook voor de aantallen woordtokens per frequentieklasse. Wanneer we deze aantallen grafisch weergeven (zie afbeelding 5-7) ontstaat weer een verloop dat zeer veel gelijkenis vertoont met dat van het corpus wetgevingsteksten. Ook de onregelmatigheden in beide grafieken vertonen zeer veel overeenkomsten. Dat dit toeval zou zijn lijkt vrij onwaarschijnlijk, er is waarschijnlijk inderdaad sprake van een duidelijk verband met het verloop van de aantallen typen (dat voor beide corpora ook al veel overeenkomst vertoont), zoals ook in het vorige hoofdstuk al werd aangegeven.



Afb. 5-7 Aantallen woordtokens per frequentieklasse

Voor wat betreft de aantallen woordtypen en -tokens per frequentieklasse kan worden geconcludeerd dat deze aantallen ook in dit corpus over een bepaald gebied, ongeveer van frequentie 10 tot frequentie 10000, een vrijwel lineair verloop kennen. De curves voor beide corpora vertonen, afgezien van de schaal, veel gelijkenis. Wel lijkt de spreiding van datapunten aan de uiteinden van de grafieken bij het corpus jurisprudentieteksten iets geringer te zijn.

We kunnen voorlopig dus stellen dat, globaal genomen, de frequentieverdeling van het corpus jurisprudentieteksten als geheel op de meeste punten sterke overeenkomsten vertoont met die van het corpus wetgevingsteksten. Gedetailleerde vergelijking heeft echter pas zin wanneer zeker is dat de gegevens uit de twee corpora voor 100% onderling vergelijkbaar zijn. De voornaamste reden waarom dat nu nog niet het geval is, is dat de twee corpora sterk verschillen in omvang. Daarvoor zal in ieder geval moeten worden gecorrigeerd. Op dit punt zal in hoofdstuk 7 worden teruggekomen.

5.4 De verhouding tussen woordtokens en woordtypen

5.4.1 Token-type ratio's

Het corpus jurisprudentieteksten bestaat uit in totaal 40190212 tokens en telt 265841 typen. De token/type ratio over het hele corpus komt daarmee op 151,18, een waarde die bijna 1,7 keer zo hoog is als die van het corpus wetgevingsteksten. Dat hangt natuurlijk weer samen met de corpusgrootte. In tabel 5-3 wordt echter weer aangegeven hoe deze token/type ratio zich ontwikkelt wanneer we uit het corpus 'random' steekproeven nemen

van steeds toenemende omvang. De laatste regel in de tabel heeft weer betrekking op het hele corpus, dat 16430 documenten telt[2].

Aantal doc.	Typen	Tokens	Tok/Typ ratio	Toename	Log.Tok/ Typ ratio	Toename
1000	52512	2473097	47,096	-	1,354	-
2000	75846	4663820	61,491	30,56%	1,367	0,90%
3000	99718	7501604	75,228	22,34%	1,375	0,64%
4000	115051	9575674	83,230	10,64%	1,379	0,30%
5000	131356	12160137	92,574	11,23%	1,384	0,35%
6000	145949	14583449	99,922	7,94%	1,387	0,22%
7000	160669	16966633	105,600	5,68%	1,389	0,11%
8000	174520	19394131	111,128	5,24%	1,390	0,11%
9000	186756	21877046	117,142	5,41%	1,392	0,16%
10000	197953	24328895	122,902	4,92%	1,395	0,15%
11000	210585	26906370	127,770	3,96%	1,396	0,08%
12000	220409	29300425	132,937	4,04%	1,397	0,13%
13000	232361	31772247	136,737	2,86%	1,398	0,04%
14000	241930	34181982	141,289	3,33%	1,399	0,10%
15000	251090	36522062	145,454	2,95%	1,401	0,08%
16000	262165	39224678	149,618	2,86%	1,401	0,06%
16430	265841	40190212	151,181	1,04%	1,402	0,03%

Tabel 5-3 Token/type ratio als functie van de corpusgrootte

Ook nu blijken noch de token/type ratio, noch de logaritmische token/type ratio (de logaritme van het aantal tokens gedeeld door de logaritme van het aantal typen) constant te zijn, of naar een meer constante waarde te 'groeien', bij toenemende steekproefgrootte. Wat dat betreft vertoont dit corpus dezelfde karakteristieken als het corpus wetgevingsteksten. Ook de percentages waarmee beide ratio's toenemen zijn voor beide corpora min of meer vergelijkbaar. De groeipercentages op de onderste regel zijn niet echt maatgevend: op het laatst werden immers minder documenten toegevoegd. Interessant zijn de getallen bij steekproef nummer 5 (5000 documenten), die ongeveer hetzelfde aantal woordtokens telt als het gehele corpus wetgevingsteksten. De token/type ratio's van beide blijken heel dicht bij elkaar te liggen: 92,57 voor de steekproef, 89,47 voor de wetgevingsteksten. Ook hier geldt echter dat een echte vergelijking pas kan plaatsvinden in hoofdstuk 7, nadat de data uit de corpora geheel onderling vergelijkbaar zijn gemaakt.

5.4.2 De karakteristiek K

Net als in hoofdstuk 4 zal ook hier weer de waarde worden berekend van de karakteristiek *K* van Yule en Herdan. Tevens zal worden nagegaan of fluctuaties in de waarde van *K*

[2] Het corpus jurisprudentieteksten telt dus wat minder documenten dan het corpus wetgevingsteksten (16430 tegen 18803), maar bestaat wel uit een ruim 3 maal zo groot aantal tokens. Jurisprudentiedocumenten zijn dus gemiddeld aanmerkelijk groter dan wetgevings-documenten, zoals ook blijkt uit de cijfers in de tabellen 3-2 en 3-4 (hoofdstuk 3): ze tellen gemiddeld ruim 2446 tokens, tegen ruim 681 bij de wetgevingsteksten.

bij dit corpus net zo gering zijn als bij de wetgevingsteksten. Bij dat corpus bleek de karakteristiek bij steekproeven van 10000 documenten en meer (dat kwam neer op steekproeven van zo'n 7 miljoen tokens en groter, d.w.z. ongeveer de helft van het totale corpus) buitengewoon stabiel te zijn.

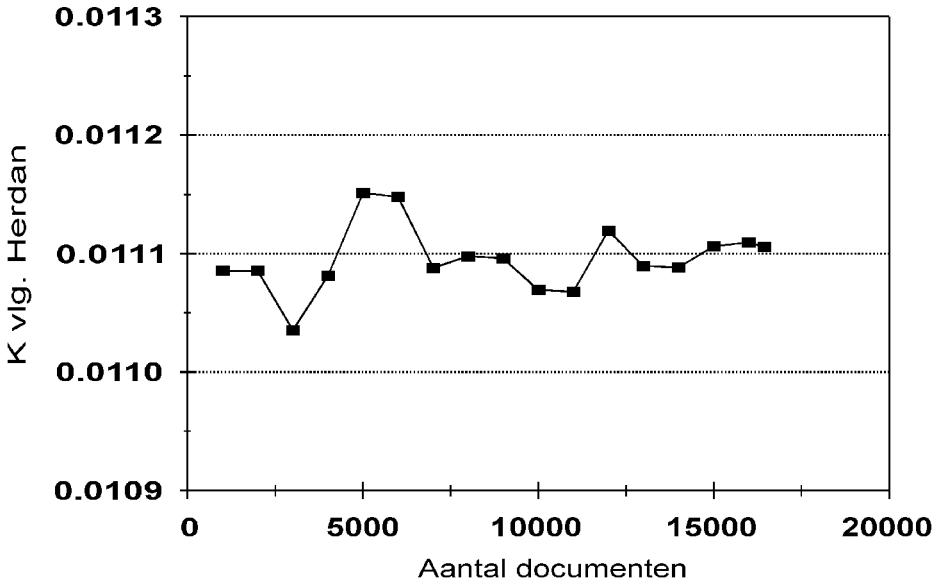
Om de waarde van *K* te berekenen voor het hele corpus en voor de in tabel 5-3 vermelde steekproeven, werd weer voor iedere steekproef (en voor het gehele corpus) een frequentieverdeling opgesteld. De met behulp daarvan berekende waarden van *K* (volgens de methoden van Yule en Herdan) zijn te vinden in tabel 5-4.

Aantal doc.	Typen	Tokens	*K* (Yule)	*K* (Herdan)
1000	52512	2473097	110,855	0,0110859
2000	75846	4663820	110,858	0,0110860
3000	99718	7501604	110,355	0,0110357
4000	115051	9575674	110,816	0,0110817
5000	131356	12160137	111,515	0,0111516
6000	145949	14583449	111,481	0,0111481
7000	160669	16966633	110,883	0,0110884
8000	174520	19394131	110,981	0,0110981
9000	186756	21877046	110,966	0,0110966
10000	197953	24328895	110,700	0,0110701
11000	210585	26906370	110,681	0,0110681
12000	220409	29300425	111,196	0,0111197
13000	232361	31772247	110,901	0,0110901
14000	241930	34181982	110,886	0,0110886
15000	251090	36522062	111,067	0,0111067
16000	262165	39224678	111,097	0,0111097
16430	265841	40190212	111,059	0,0111060

Tabel 5-4 De waarde van *K* bij toenemende steekproefgrootte

De waarden berekend volgens de methoden van Yule en Herdan blijken in dit geval steeds nog dichter bij elkaar te liggen dan bij de wetgevingsteksten het geval was. Dat komt waarschijnlijk doordat alle steekproeven ook aanmerkelijk groter zijn: ruim 2 miljoen woordtokens per stuk, tegen zo'n 600.000 bij de wetgevingsteksten. Daardoor wordt al bij de vierde steekproef een omvang van ruim 9,5 miljoen tokens bereikt. Bij die omvang werd ook bij de wetteksten geconstateerd dat de verschillen tussen de twee berekeningsmethoden zeer klein waren geworden (in de orde van 1/10000000 wanneer we afzien van de vermenigvuldigingsfactor van 10000). Ook bij dit corpus blijkt het verschil bij de vierde en alle daarop volgende steekproeven in deze orde van grootte (of nog minder) te liggen.

De waarde van *K* fluctueert bij dit corpus in het begin (bij de kleinste steekproeven) minder dan bij het corpus wetgevingsteksten. Al vrijwel direct (bij de tweede steekproef) wordt een niveau van stabiliteit bereik dat bij de wetteksten pas na vier of vijf ontstond. De grafiek waarin deze fluctuaties in beeld worden gebracht lijkt qua vorm sterk op afbeelding 4-11, met name wanneer we daarvan de eerste vier steekproeven buiten beschouwing laten:



Afb. 5-8 De waarde van *K* (Herdan) bij toenemende steekproefgrootte

Uit dit alles kunnen voorlopig de volgende conclusies worden getrokken:

- de waarde van *K* is eigenlijk voortdurend behoorlijk stabiel gebleken voor alle tot nu toe onderzochte steekproeven uit de twee corpora;
- voorzover deze waarde dan toch nog wat fluctueert, treedt dit met name op bij steekproeven van *absoluut gezien* geringere omvang; bij steekproeven van meer dan 3,5 à 4 miljoen tokens is de waarde in hoge mate stabiel.

Het lijkt er nu dus op dat met name de *absolute* omvang van de steekproef (het aantal woordtokens dat deze omvat) voor de betrouwbaarheid van *K* een belangrijker rol speelt dan de *relatieve* omvang (welk percentage van het corpus de steekproef omvat). Deze veronderstelling zal in het volgende hoofdstuk verder worden getoetst.

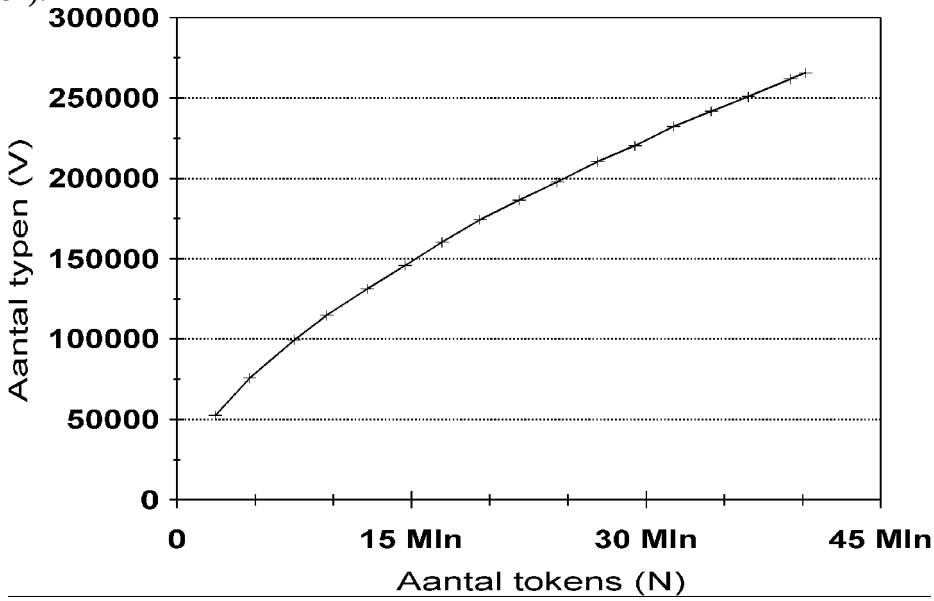
De berekende waarden van *K*, ongeveer 111 (vlg. Yule) c.q. 0,0111 (vlg. Herdan), wijken duidelijk af van de gevonden waarden voor het corpus wetgevingsteksten (ongeveer 128 c.q. 0,0128). Geïnterpreteerd volgens de theorie van Yule zou dat inhouden dat het corpus jurisprudentieteksten een *lagere* gemiddelde herhalingsfrequentie van woordtypen zou kennen, en (mogelijk) een *grotere* woordenschat (meer verschillende woordtypen) aangezien daarvoor 1/*K* maatgevend zou zijn. Hoewel aan het vergelijken van de omvang van de woordenschat in hoofdstuk 7 nog wat uitgebreider aandacht zal worden besteed, kan nu al worden vastgesteld dat steekproef nummer 5 uit tabel 5-3, die ongeveer dezelfde omvang heeft als het hele corpus wetgevingsteksten, voor beide corpora praktisch dezelfde token/type ratio te zien geeft. Wanneer we een en ander interpreteren volgens de theorie van Herdan dan zouden de verschillende waarden voor *K* wijzen op verschillen tussen de frequentieverdelingen van beide corpora. Een hogere waarde van *K* zou volgens Herdan betekenen dat een corpus meer "van het gemiddelde afwijkende" frequenties bevat[3]. Een dergelijk verschil zou zichtbaar moeten zijn in de frequentie-verdeling, en in bijvoorbeeld grafiek 5-5 (meer frequenties in de hoogste en laagste klassen).

[3] Zie voor uitleg met betrekking tot dit punt hoofdstuk 2, paragraaf 3.

Daarvan is echter in de vorige paragraaf nog niets gebleken. In hoofdstuk 7 wordt op dit punt wat meer duidelijkheid verschaft, op basis van gedetailleerder analyses op qua omvang vergelijkbaar gemaakte corpusgedeelten. Voorlopig moet echter de conclusie zijn dat de karakteristiek *K* voor beide corpora een andere waarde heeft, dat dit waarschijnlijk wel te maken heeft met verschillen tussen bepaalde aspecten van het woordgebruik, maar dat vooralsnog niet precies is aan te geven welke aspecten dat zijn.

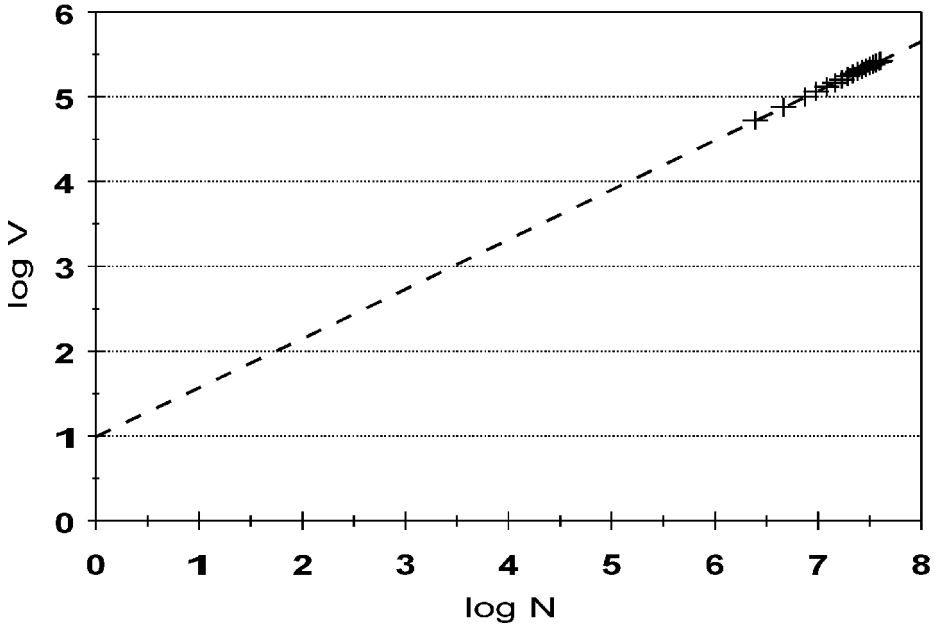
5.4.3 De constanten R en C

In het vorige hoofdstuk werd al duidelijk dat aan de constante *R* van Guiraud als op zichzelf staand statistisch maatgetal niet al te veel waarde kan worden toegekend. De constante, gedefinieerd als het quotiënt van het aantal woordtypen en de tweedemachts-wortel uit het aantal woordtokens in een steekproef, blijkt sterk afhankelijk van de omvang van die steekproef, en gedraagt zich in feite als een soort 'afgedempte' type/token ratio. Het bleek echter mogelijk een verband te leggen tussen Guiraud's *R* en de constante *C* van Herdan, zoals deze voorkomt in de formule log $V = C$ log N. Aangezien het verband tussen log *V* en log *N* bij het corpus wetgevingsteksten inderdaad vrijwel lineair bleek te zijn, konden de waarden voor de constanten *R* en *C* nauwkeurig worden bepaald. Om na te gaan of de omstandigheden in het corpus jurisprudentieteksten wat dit betreft hetzelfde zijn wordt in afbeelding 5-9 allereerst weer het verband tussen de aantallen woordtokens en -typen in beeld gebracht (de datapunten worden weer door lijnen verbonden):



Afb. 5-9 Corpus jurisprudentieteksten: het aantal woordtypen als functie van het aantal woordtokens

Het verloop van de 'curve' lijkt sterk op dat bij de wetgevingsteksten, ook hier lijkt sprake te zijn van een exponentieel verband tussen het aantal tokens en het aantal typen. Herdan's formule, die immers inhoudt dat $V=N^{c}$, lijkt daarom ook hier in beginsel geschikt om dit verband uit te drukken. Om daarover zekerheid te krijgen is in de volgende grafiek, net als bij het corpus wetgevingsteksten, de logaritme van het aantal typen afgezet tegen de logaritme van het aantal tokens.

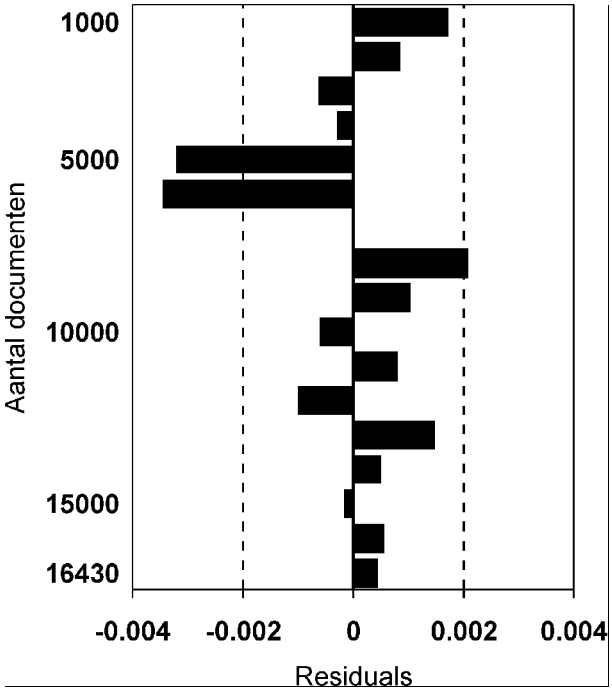


Afb. 5-10 De verhouding tussen log *V* en log *N* (datapunten), en de regressielijn van log *V* op log *N* (stippellijn)

Ook nu lijkt - op het oog - sprake te zijn van een praktisch lineair verband tussen log *V* en log *N*, maar evenmin als bij de wetgevingsteksten liggen de datapunten op een lijn door de oorsprong. De stippellijn in de grafiek is de lineaire regressielijn van log *V* op log *N*, waarvan de vergelijking luidt:

$$\log V = 0{,}5827 \log N + 0{,}9929 \qquad (5\text{-}4)$$

r^2, het kwadraat van de lineaire correlatiecoëfficiënt, bedraagt hier 0,999944, hetgeen betekent dat niet minder dan 99,9944% van de variantie van de twee variabelen door de regressie wordt 'verklaard'. De residuals zijn van dezelfde orde van grootte als bij het corpus wetgevingsteksten het geval was (nooit groter dan 0,004), de standaard fout van de regressielijn bedraagt slechts 0,00155. De covariantie tussen de residuals onderling is ook extreem laag (minder dan 0,0000022). Het patroon dat de residuals vertonen is zichtbaar gemaakt in afbeelding 5-11. Net als in het vorige hoofdstuk is sprake van een (op het oog) willekeurig patroon, een verdere indicatie dat een lineaire regressielijn inderdaad het verband tussen log V en log N goed weergeeft. Wat nog opvalt is dat steekproef 5 en 6 duidelijk hogere



Afb. 5-11 Residuals van de lineaire regressielijn van log *V* op log *N*

residuals te zien geven dan alle andere steekproeven; een reden hiervoor is niet eenvoudig te geven. Het verschijnsel dat bij het corpus wetgevingsteksten werd waargenomen, namelijk dat de kleinste steekproeven allemaal hogere residuals te zien gaven, treedt hier niet op (althans dat blijkt niet uit de grafiek). Natuurlijk is het zo dat iedere steekproef nu groter is, de eerste steekproef bevat al even veel tokens als de derde steekproef bij de wetgevingsteksten. Het zou kunnen zijn dat ook de eerste steekproef al zo omvangrijk is dat de verhouding tussen log V en log N zich heeft gestabiliseerd.

Een t-test van de regressielijn levert tenslotte de zeer hoge waarde op van ruim 519, bij n - 2 = 15 vrijheidsgraden, hetgeen inhoudt dat de kans dat de verhouding tussen log V en log N, zoals uitgedrukt in de vergelijking van de regressielijn, op toeval berust vrijwel nul is. Ook dat wijst er weer op dat het verband tussen log V en log N bij dit corpus praktisch lineair is, en dat dit verband door de hierboven afgebeelde regressielijn goed wordt weergegeven. Nu dat het geval is, kunnen met behulp van de vergelijking van die regressielijn de twee constanten uit de formule van Erikstad:

$$V = R \,.\, N^{C} \qquad (5\text{-}5)$$

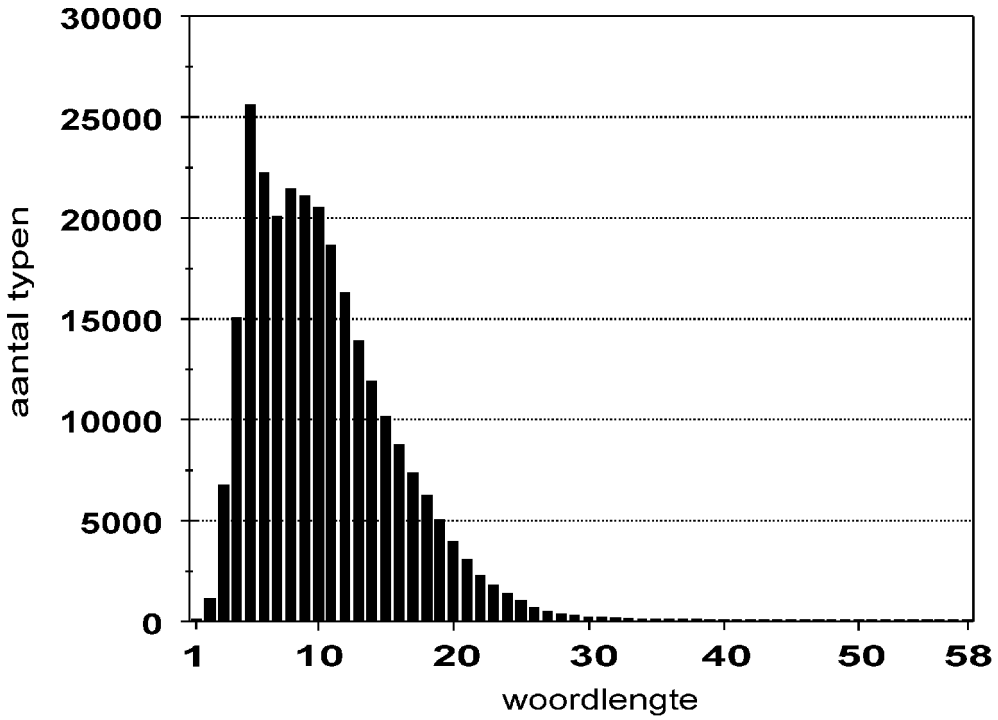
worden berekend[4]. Deze constanten hebben in dit geval de volgende waarde: C = **0,5827**, en R = $10^{0,9929}$ = **9,8383**. Deze waarden - en dan met name die voor R - wijken duidelijk af van die in het corpus wetgevingsteksten (0,5675 resp. 13,2979). Net als in het voorgaande hoofdstuk heb ik, onder gebruikmaking van bovenstaande formule en de berekende waarden voor beide constanten, voor de diverse steekproeven de waarde van V berekend uit die van N, en vervolgens het resultaat vergeleken met de werkelijke waarden. Ook nu waren de verschillen gering, gemiddeld niet meer dan 23 typen (absoluut), of 0,0006%. De grootste afwijking bij een steekproef was 1158 typen, de kleinste ruim 3. Dit alles leidt tot de veronderstelling dat de waarden van R en C in dit corpus (evenals in het corpus wetgevingsteksten) buitengewoon stabiel zijn, praktisch onafhankelijk van de steekproefgrootte en dus in feite ook van de corpusgrootte. Dat betekent dat de twee constanten een zinvol vergelijkingsinstrument vormen voor het woordgebruik in corpora, zelfs wanneer die van verschillende grootte zijn. Deze veronderstelling zal in hoofdstuk 7 nog nader worden getoetst. Voorlopig kan in ieder geval worden geconstateerd dat met name de waarde voor R bij dit corpus substantieel lager ligt; dat betekent in dit geval dat het aantal typen in kleine steekproeven bij dit corpus aanvankelijk lager ligt dan bij de wetgevingsteksten. De wat hogere waarde van C betekent echter dat het aantal typen sterker reageert op een toename van het aantal tokens; ieder toegevoegd document levert dus een wat groter aantal nieuwe typen op.

[4] Zie voor de complete afleiding hoofdstuk 4, paragraaf 4.

5.5 Woordlengtes

5.5.1 Woordlengte verdelingen

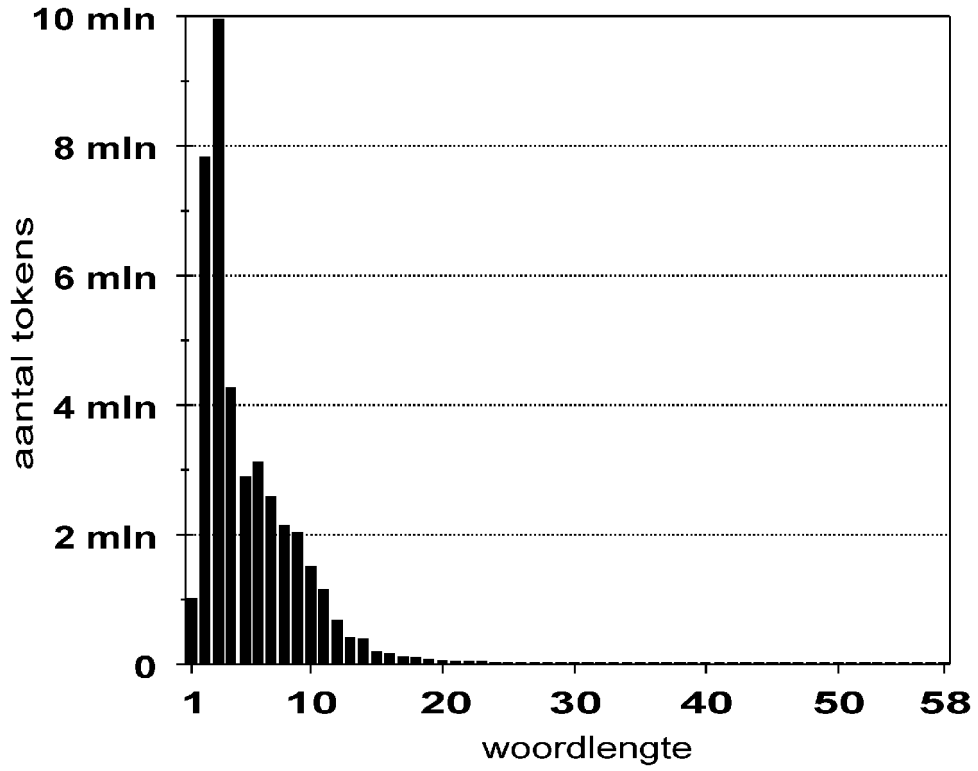
Ook voor dit corpus zijn, op overeenkomstige wijze als bij de wetgevingsteksten, woordlengte-verdelingen opgesteld, zowel voor de aantallen woordtypen als woordtokens. Ook nu moest weer rekening worden gehouden met zeer lange woordtypen, die bij het opstellen van de frequentielijst waren 'afgekapt' na het 30e karakter. De gegevens uit de frequentielijst werden weer aan de hand van een aparte lijst van 'lange woordtypen' gecorrigeerd. De grafieken voor de lengtes van woordtypen en -tokens zien er als volgt uit (zie afbeeldingen 5-12a en b):



Afb. 5-12a Verdeling van de lengtes van woordtypen

Een eerste vergelijking leert al direct dat met name de lengteverdeling van de woordtypen sterk afwijkt van die van het corpus wetgevingsteksten. Weliswaar is ook hier bij lengte 7 sprake van een 'dip', maar aan weerszijden daarvan ziet het patroon er totaal anders uit. In plaats van een piek bij lengte 4 zien we er nu een bij lengte 5, terwijl de zeer hoge piek bij lengte 8 verdwenen is, of juister gezegd, tot 'normale' proporties is teruggebracht (want 8 is nog steeds de meestvoorkomende lengte na de 'dip' bij 7). Een verklaring voor het relatief geringe aantal typen met lengte 7 is nu niet meer zo makkelijk te geven[5], aangezien er nu geen aangrenzende lengte-categorie is die duidelijk buitenproportioneel veel typen bevat.

[5] Het is, gezien dit feit, nog maar de vraag of de verklaring in het vorige hoofdstuk (veel formele woorden met lengte 8) juist was.



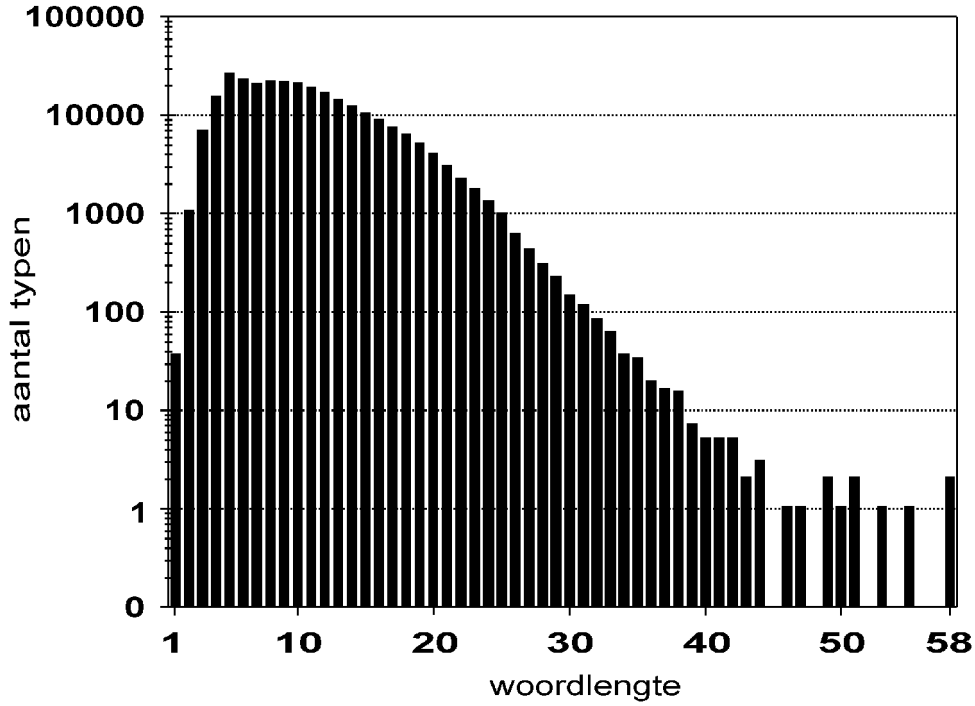
Afb. 5-12b Verdeling van de lengtes van woordtokens

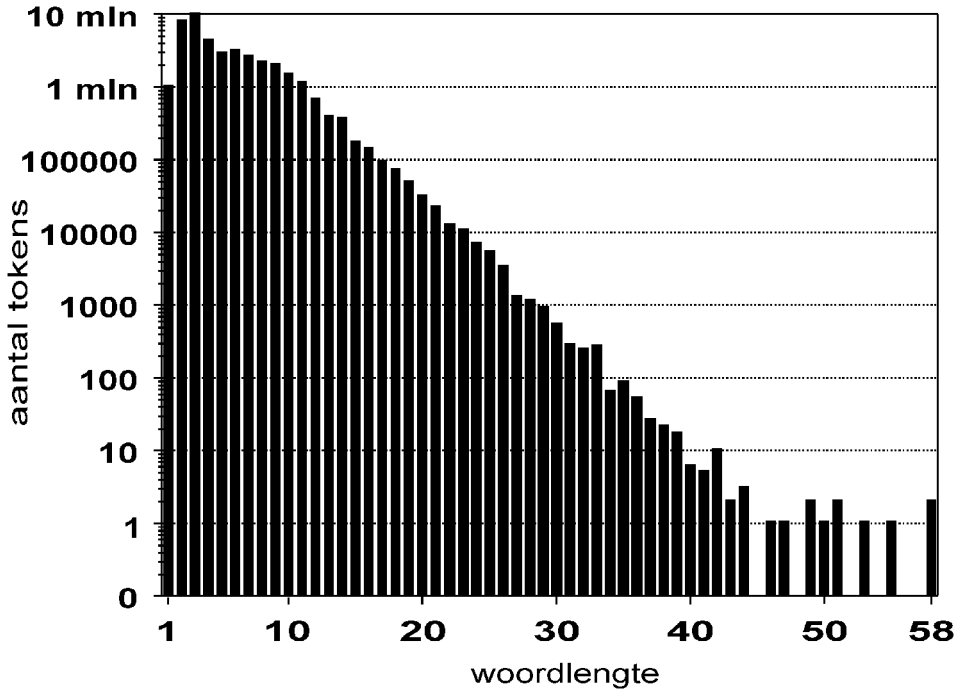
Met betrekking tot de woordtokens komt het patroon veel meer overeen met dat in het vorige hoofdstuk. De belangrijkste verschillen zijn eigenlijk dat de aantallen tokens na de piek bij lengte drie wat geleidelijker (dat wil zeggen, minder abrupt) terug lijken te lopen, en dat de inzinking bij lengte 5 minder diep is. Toch is die inzinking nog steeds duidelijk aanwezig, waardoor de verklaring uit het vorige hoofdstuk, namelijk dat het hier een algemeen verschijnsel voor de Nederlandse taal betreft (minder 5-letter woorden doordat dit net de grens tussen woorden met één en met twee lettergrepen is) aan kracht wint.

Ook nu is weer een nadeel van beide grafieken dat de verdeling van zeer lange woorden er niet uit kan worden afgelezen. Daarom is van beide ook weer een versie met een logaritmische verdeling op de y-as opgenomen (zie afbeeldingen 5-13a en b).

Voor de woordtypen blijkt dan sprake te zijn van een bijzonder gelijkmatige lengteverdeling, waarbij de piek bij 5 en het dal bij 7 toch nog goed te herkennen zijn. Bij woordlengtes boven de 40 wordt ook bij dit corpus het verloop wat grilliger, maar in mindere mate dan bij het corpus wetgevingsteksten. Verder valt op dat er meer lege lengteklassen zijn dan bij dat corpus (6 in plaats van 2), en dat er wat minder zeer lange typen lijken voor te komen, aangezien niet minder dan 5 lengte-klassen slechts een enkel type omvatten[6].

[6] De kans is natuurlijk groot dat deze twee feiten met elkaar samenhangen: doordat de typen met lengtes > 45 karakters zo dun gezaaid zijn, ontbreken er ook meer lengte-categorieën geheel.





Afb. 5-13a en b Verdeling van de lengtes van woordtypen en woordtokens; aantallen per lengte-cat. logaritmisch

Controle van de frequentielijst leert dat dit inderdaad juist is: er komen in de jurisprudentieteksten slechts 11 typen voor met een lengte van 45 karakters of meer, tegen 28 bij de wetteksten.

Voor wat betreft de woordtokens lijkt het beeld zeer sterk op dat bij de wetgevingsteksten. Er zijn in dit geval eigenlijk geen klassen die qua aantal tokens duidelijk van de algemene tendens afwijken, het aantal tokens met lengtes groter dan 45 is echter duidelijk geringer (waardoor een groter aantal lege lengte-klassen ontstaat). Inderdaad blijkt bij nadere beschouwing geen enkel type dat langer is dan 45 karakters meer dan een enkele keer in het corpus voor te komen; er zijn dus ook slechts 11 tokens van deze lengte. Bij de wetgevingsteksten waren dat er 26 (en 22 typen), een type met lengte 50 kwam vijf keer voor. Het kleine 'piekje' bij lengte 33 blijkt praktisch geheel voor rekening te komen van het type 'arbeidsongeschiktheidsverzekering', dat 137 keer voorkomt. Ditzelfde type kwam in de wetgevingsteksten 762 keer voor, en zorgde ook daar voor een (veel duidelijker) piek bij lengte 33.

Met betrekking tot de woordlengte-verdelingen kan dus worden geconstateerd dat met name de versies met logaritmische schaalverdeling voor dit corpus een nog gelijkmatiger beeld te zien geven dan bij het corpus wetgevingsteksten, praktisch zonder pieken en dalen. Met betrekking tot de woordtypen is nog opvallend dat lengte 5 in dit corpus vaker voorkomt dan lengte 4 (bij de wetteksten was het andersom), en dat lengte 8 geen opvallend hoge frequentie heeft, zoals bij de wetteksten wel het geval was. Aangezien het echter mogelijk is dat ook hier de verschillen in afmeting van de corpora een rol spelen, is het nog niet mogelijk uit deze bevindingen conclusies te trekken. Hierop zal dan ook in hoofdstuk 7 worden terugkomen.

5.5.2 De relatie tussen het rangnummer en de lengte van woordtypen.

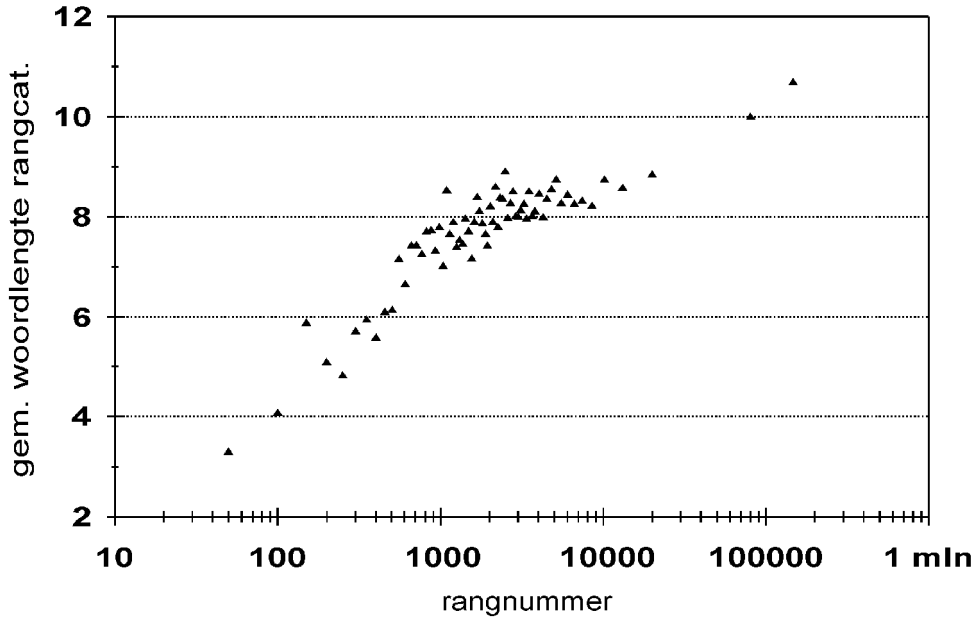
In het vorige hoofdstuk werd de relatie tussen de lengte en de positie die een woordtype inneemt op de frequentielijst onderzocht op drie verschillende manieren:

- door middel van een grafiek waarin voor ieder rangnummer alle woordlengtes waren opgetekend; de resulterende puntenwolk gaf niet veel meer dan een indicatie c.q. verwachtingspatroon van de richting waarin de gemiddelde woordlengte zich beweegt;
- door de gemiddelde lengte te berekenen van alle typen in elkaar opvolgende 'clusters' van 50 rangnummers; en tenslotte
- door het voortschrijdend gemiddelde de berekenen van de lengte van de eerste 5000 typen uit de frequentielijst; deze methode was vanwege de grote hoeveelheid data niet geschikt voor de volledige frequentielijst.

Vanwege het wat globale karakter van de eerste methode, die wel geschikt is om een eerste indruk te krijgen van het verband tussen rangnummer en woordlengte, maar geen exacte gegevens verschaft zullen voor het corpus jurisprudentieteksten (en ook voor het corpus algemeen Nederlands, in het volgende hoofdstuk) alleen de laatste twee methoden worden toegepast.

Van die twee methoden komt weer als eerste de cluster-methode aan bod. Bij die methode wordt dus de gemiddelde woordlengte bepaald van alle typen in elkaar opvolgende clusters van 50 rangnummers. De berekende waarden kunnen worden weergegeven in een grafiek met op de x-as het rangnummer (ieder cluster heeft als x-waarde het hoogste erin

voorkomende rangnummer) en op de y-as de gemiddelde woordlengte in het cluster. Voor het corpus wetgevingsteksten werd van de betreffende grafiek zowel een versie getoond waarin voor het rangnummer een normale schaalverdeling was gebruikt, als een versie met voor de x-as een logaritmische schaalverdeling[7]. De versie met normale schaalverdeling toonde wel het 'groei'patroon van de gemiddelde woordlengte, maar was verder toch niet echt duidelijk (de datapunten lagen erg dicht bij elkaar). Ook voor dit corpus blijkt dat nadeel te bestaan; de grafiek die ontstaat is daardoor voor vergelijkingsdoeleinden nauwelijks geschikt. Daarom wordt hier volstaan met alleen de versie met logaritmische schaalverdeling.

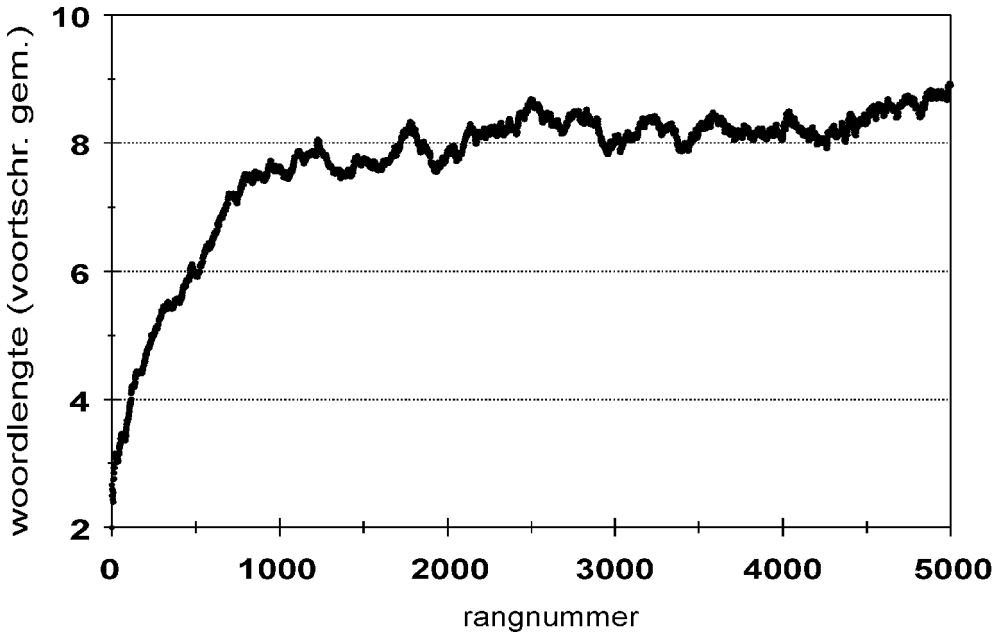


Afb. 5-14 Gemiddelde woordlengte per categorie van 50 rangnummers

In grote lijnen komt het patroon in deze grafiek overeen met dat bij de wetgevingsteksten. Er is sprake van een consistente, maar afnemende stijging in de gemiddelde woordlengte. De mate van spreiding van de datapunten lijkt hier wat geringer te zijn, kennelijk fluctueert de gemiddelde lengte per cluster wat minder dan bij de wetteksten. Verder lijkt de 'sprong' in gemiddelde lengte bij rangnummer 150, die in het vorige hoofdstuk werd waargenomen, ook hier op het eerste gezicht aanwezig te zijn, maar de clusters daarna geven weer waarden die geheel in lijn zijn met de rest van de grafiek, zodat de kans dat hier sprake is van toeval wat groter is dan bij het corpus wetgevingsteksten. Tenslotte geeft ook deze grafiek een piek of tijdelijk maximum zien, alleen ligt het nu niet rond rangnummer 1500, maar bij 2500. Net als bij de wetteksten bereikt de gemiddelde woordlengte daar een waarde van rond de 9, om daarna weer te dalen. In het vorige hoofdstuk werd nog nagegaan of daarvoor een verklaring kon worden gevonden in de woordlengte-verdeling (veel typen met lengte 8), maar echt overtuigend bleek dat toch niet te zijn. Ook voor dit corpus is een dergelijke verklaring eigenlijk niet te geven. Aangenomen moet worden dat het hier een specifiek kenmerk van dit corpus betreft, vergelijking moet uitwijzen of het corpus algemeen Nederlands het verschijnsel ook kent.

[7] Zie paragraaf 4.5.3.

Om tenslotte, net als in het vorige hoofdstuk, de situatie met betrekking tot de lengte van woordtypen in het bovenste deel van de frequentielijst (tot rangnummer 5000) in nog wat meer detail te kunnen bestuderen, zal ook nu de methode van het voortschrijdend gemiddelde worden toegepast. Deze methode houdt in dat een grafiek wordt samengesteld waarin voor ieder type een datapunt wordt opgenomen. Dit datapunt vertegenwoordigt de gemiddelde woordlengte van het type in kwestie, en van de 199 typen daarvóór. Het aantal van 200 typen is gelijk aan dat in het vorige hoofdstuk. Weergegeven met een normale schaalverdeling ziet de grafiek er voor dit corpus als volgt uit:

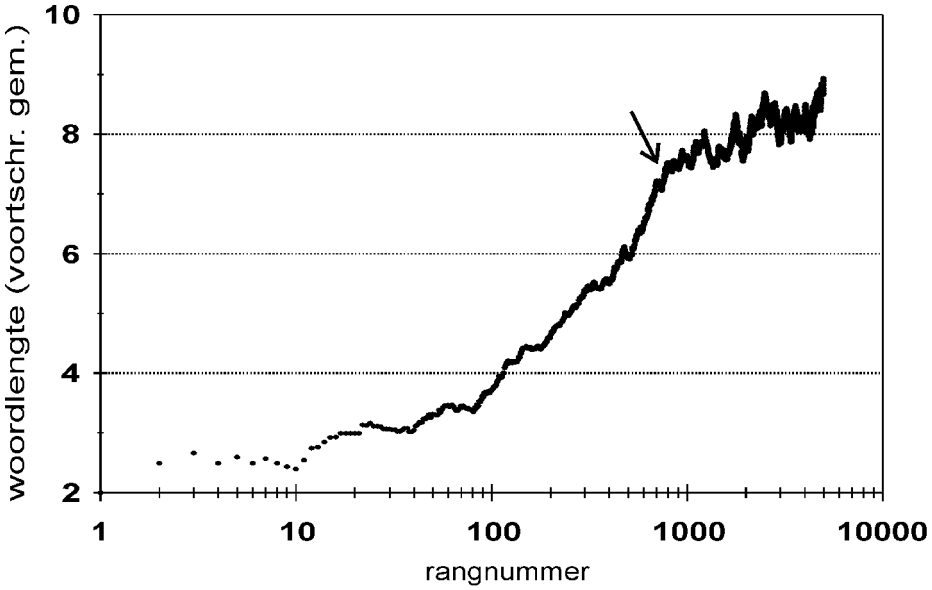


Afb. 5-15 Voortschrijdend gemiddelde van de lengte van woordtypen

Gedetailleerde vergelijking van deze grafiek met de overeenkomstige grafiek uit het vorige hoofdstuk (afbeelding 4-20) is op dit moment natuurlijk niet goed mogelijk, aangezien het verschil in corpusgrootte weer een rol zou kunnen spelen. In globale zin vallen wel de volgende zaken op:

- de gemiddelde woordlengte stijgt in het begin duidelijk minder snel; een lengte van 6 wordt hier pas bereikt bij rangnummer 500, terwijl dat bij de wetgevingsteksten al rond rangnummer 250 gebeurde.
- net als bij de wetgevingsteksten gaat de curve zich ongeveer vanaf rangnummer 2000 min of meer evenwijdig aan de x-as bewegen; de woordlengte waarmee dit 'horizontale' deel van de curve correspondeert is voor de jurisprudentieteksten echter duidelijk wat lager.
- de twee 'inzinkingen' die de curve bij de wetgevingsteksten vertoonde (ongeveer bij rangnummers 1000 en 2000) zijn hier niet of nauwelijks waar te nemen.

Ook is van bovenstaande grafiek ook een versie met logaritmische schaalverdeling op de x-as gemaakt. De voornaamste reden daarvoor is dat dan het verschijnsel van het minder snel stijgen van de gemiddelde woordlengte, tussen rangnummers 1 en 1000, beter bestudeerd kan worden.



Afb. 5-16 Voortschrijdend gemiddelde van de lengte van woordtypen, logaritmische schaalverdeling voor rangnummers

Deze grafiek geeft inderdaad een wat ander beeld te zien dan de overeenkomstige grafiek van de wetgevingsteksten (afbeelding 4-21). Met name valt op dat de curve tot even voor rangnummer 1000 veel 'holler' verloopt, terwijl er bij de wetgevingsteksten van rangnummer 100 tot ongeveer 300 juist sprake was van een praktisch rechtlijnig verloop. Ongeveer bij rangnummer 750 geeft deze grafiek een duidelijke 'knik' te zien (pijl). Iets dergelijks is bij de wetteksten eigenlijk alleen te zien bij rangnummer 300.

Al met al kan geconstateerd worden dat binnen het corpus jurisprudentieteksten sprake is van een verband tussen rangnummer en woordlengte dat op het eerste gezicht enigszins afwijkt van dat in het corpus wetgevingsteksten. Wanneer we de gemiddelde woordlengte bepalen over clusters van rangnummers, valt op dat minder spreiding in waarden ontstaat dan wanneer we hetzelfde doen voor de wetgevingsteksten. Het voornaamste verschil wordt echter zichtbaar wanneer het voortschrijdend gemiddelde van de woordlengte wordt bepaald. Dan blijkt namelijk dat dit gemiddelde bij het corpus wetgevingsteksten aanmerkelijk sneller stijgt dan bij de jurisprudentieteksten, en dat deze stijging op optreedt volgens een wat ander patroon. In hoeverre de geconstateerde verschillen los staan van het verschil in corpusgrootte zal in hoofdstuk 7 moeten blijken.

5.5.3 Zeer lange woordtypen

Het percentage lange woordtypen is bij het corpus jurisprudentieteksten wat lager dan bij het corpus wetgevingsteksten. 4,22% van alle woordtypen in het corpus is langer dan 20 lettertekens, en 0,15% langer dan 30 (wetteksten: 4,52% resp. 0,24%). Dit relatief geringere aantal typen blijkt ook een beperkter deel van het corpus uit te maken, namelijk 0,16% (wetteksten: 0,32%). De gemiddelde frequentie van deze typen is 5,82, de relatieve gemiddelde frequentie bedraagt bijna 3,9% (0,16% / 4,22%). Dit laatste percentage bedroeg bij de wetgevingsteksten 7,1%, elk van de lange typen wordt in dat corpus dus

gemiddeld bijna twee keer zo vaak herhaald. Merk op dat de waarde van de constante K voor dit corpus jurisprudentieteksten lager lag dan bij de wetteksten, wat volgens Yule ook een lagere gemiddelde herhalingsfrequentie van woordtypen betekent (zie paragraaf 5.4.2). Het gedrag van de lange woordtypen is hiermee dus in overeenstemming. Een verklaring voor het feit dat jurisprudentieteksten wel een flink aantal lange, ingewikkelde woorden bevatten, maar toch niet zoveel als teksten van wetgeving zou kunnen zijn dat rechterlijke uitspraken naast juridisch jargon ook altijd verwijzingen naar de casuspositie bevatten, en soms ook een weergave van verklaringen van bijvoorbeeld procespartijen of getuigen. Dergelijke teksten zorgen er wellicht voor dat jurisprudentieteksten wat meer het karakter van 'gewone' teksten benaderen.

Met betrekking tot de lijst van zeer lange woordtypen, die te vinden is in bijlage 2, kan nog het volgende worden opgemerkt. De lijst bevat, net als die van het corpus wetgevingsteksten, een groot aantal typen die te maken hebben met onderwerpen van juridische aard: 'bedrijfsaansprakelijkheidsverzekeringen', 'doelmatigheidsoverwegingen', 'handelsvertegenwoordigingsovereenkomst', etc. Het aantal technische termen is veel geringer dan bij de wetgevingsteksten, in feite gaat het daarbij om minder dan 10 typen. Voorbeelden zijn: 'bestrijdingsmiddelenbewaarplaats', 'gasverwarmingsinstallatiebedrijf', 'oliestookverwarmingsinstallatie', etc. Ook namen van chemische verbindingen komen bijna niet voor (een uitzondering is 'diaethylstilboestroldipropionaat'), maar getalsnamen zijn weer ruim vertegenwoordigd ('drieendertigduizenddriehonderddrieendertig', 'zesmiljoennegenhonderdzesennegentigduizendvijfhonderdenelf'). Opvallend zijn ook enkele Duitse woorden (het Duits kent wellicht nog in ruimere mate dan het Nederlands de mogelijkheid om lange samengestelde woorden te gebruiken): 'Unbedenklichkeitsbescheinigungen', 'Unfallversicherungsgesellschaft', 'Vollstreckbarkeitsbescheinigung'. Naast (veel) zelfstandige naamwoorden en telwoorden treffen we ook een enkel bijvoeglijk naamwoord aan ('alleenvertegenwoordigingsbevoegd'). Ook nu kan geconstateerd worden dat de zeer lange woorden, hoewel wellicht voor een deel specifiek voor jurisprudentieteksten, vanwege hun lage frequentie nauwelijks bruikbaar zijn voor vergelijkingsdoeleinden.

HOOFDSTUK 6 - DE BELANGRIJKSTE KWANTITATIEVE EIGENSCHAPPEN VAN HET CORPUS ALGEMEEN NEDERLANDS

6.1 Inleiding

In dit hoofdstuk zal het vergelijkingscorpus met algemeen (hedendaags) Nederlandse teksten worden bestudeerd. Dit corpus is minder omvangrijk dan de twee juridische corpora, het omvat ruim 4,6 miljoen woordtokens. Deze - verhoudingsgewijs - wat beperkte omvang is een rechtstreeks gevolg van het streven om een corpus samen te stellen dat met name zeer geschikt is voor vergelijkingsdoeleinden. Het instituut voor Nederlandse Lexicologie had namelijk nog wel meer teksten in electronische vorm voorradig, maar daarbij ging het om veel 'ruwer' materiaal, teksten die nog niet volledig gecorrigeerd en/of geclassificeerd waren. Besloten werd om van het gebruik daarvan af te zien, aangezien dan mogelijk een corpus zou ontstaan met een aantal gebreken. Door typefouten in de teksten zouden mogelijk aanzienlijk meer 'hapax legomena' (woorden die maar een keer voorkomen) in het corpus terecht kunnen komen, terwijl het feit dat de teksten nog niet in categorieën ingedeeld waren gevolgen zou kunnen hebben voor de 'representativiteit' (hoewel dat eigenlijk geen juiste term is in dit verband, zie hoofdstuk 3 paragraaf 5).

Van dit corpus zullen weer dezelfde analyses worden gemaakt als in de voorgaande twee hoofdstukken. Eerst dus weer de frequentielijst, vervolgens de frequentieverdeling, de analyse van token/type ratio's en van de voorkomende woordlengtes. Ook nu is het onderzoek er weer op gericht om met name die eigenschappen op te sporen die als kenmerkend voor het corpus kunnen gelden, en daardoor misschien uitgangspunt kunnen zijn voor vergelijking met de andere twee corpora.

6.2 Absolute en relatieve woordfrequenties

Ook voor dit corpus werd weer een frequentielijst opgesteld, met frequentie- en spreidingsgegevens van alle in het corpus voorkomende woordtypen. Tabel 6-1 bevat een kopie van de eerste bladzijde van de lijst. Van de voorgaande twee hoofdstukken werden steeds alleen de woordtypen met rangnummers 1 t/m 300 opgenomen in de bijlagen, achterin dit proefschrift. Aangezien het corpus dat in dit hoofdstuk wordt behandeld vooral gebruikt zal worden als 'referentiecorpus', aan de hand waarvan het woordgebruik in de andere twee corpora kan worden vergeleken en beoordeeld, ligt het niet erg voor de hand om van dat aantal woordtypen af te wijken. Wel is het van belang om ook nu de in het corpus voorkomende frequenties en de spreiding van woordtypen in kaart te brengen, zodat deze karakteristieken kunnen worden vergeleken met die voor de andere twee corpora.

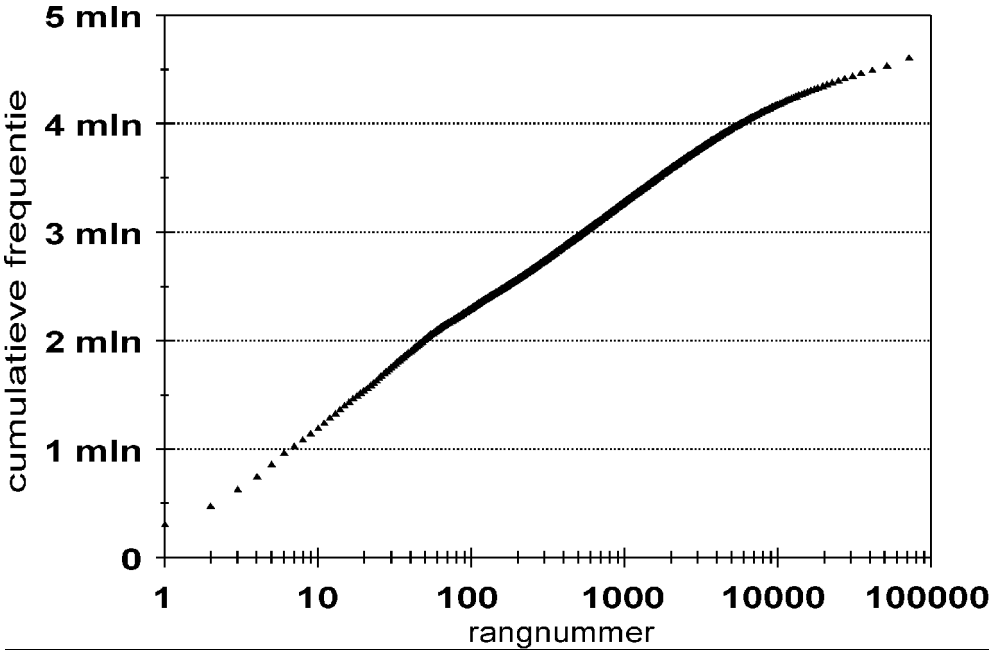
Rangnr.	Woordtype	Abs. freq.	Rel. freq.	In % doc.
1	DE	313273	6,7860%	100,00%
2	VAN	170425	3,6917%	100,00%
3	HET	151236	3,2760%	100,00%
4	EEN	118626	2,5696%	100,00%
5	EN	114921	2,4894%	100,00%
6	IN	106200	2,3005%	100,00%
7	IS	60319	1,3066%	99,01%
8	DAT	60126	1,3024%	100,00%
9	TE	59392	1,2865%	100,00%
10	OP	48776	1,0565%	100,00%
11	ZIJN	47271	1,0240%	100,00%
12	DIE	46569	1,0088%	100,00%
13	VOOR	39504	0,8557%	100,00%
14	MET	37319	0,8084%	100,00%
15	NIET	37158	0,8049%	100,00%
16	ALS	32809	0,7107%	99,75%
17	AAN	29217	0,6329%	99,75%
18	OOK	26930	0,5833%	99,01%
19	OF	25273	0,5475%	98,28%
20	WORDEN	24959	0,5407%	95,57%
21	ER	23426	0,5074%	99,51%
22	DOOR	23344	0,5057%	98,77%
23	DAN	22915	0,4964%	99,51%
24	DEZE	22529	0,4880%	96,55%
25	BIJ	22482	0,4870%	99,01%
26	OM	22162	0,4801%	99,26%
27	MAAR	21567	0,4672%	99,75%
28	TOT	19420	0,4207%	99,26%
29	WORDT	17954	0,3889%	89,90%
30	HIJ	17949	0,3888%	91,38%
31	KAN	17538	0,3799%	95,07%
32	DIT	17312	0,3750%	97,04%
33	ZICH	16772	0,3633%	98,77%
34	MEN	15814	0,3426%	86,70%
35	NAAR	13760	0,2981%	98,03%
36	ZE	13752	0,2979%	85,22%
37	UIT	13408	0,2904%	99,01%
38	KUNNEN	12787	0,2770%	94,33%
39	OVER	12752	0,2762%	98,28%
40	HEEFT	12609	0,2731%	95,57%
41	IK	12371	0,2680%	72,91%
42	MEER	12350	0,2675%	98,52%
43	WAS	12339	0,2673%	90,64%
44	WAT	12178	0,2638%	96,80%
45	NOG	11647	0,2523%	98,52%
46	WE	11144	0,2414%	81,77%
47	HEBBEN	10757	0,2330%	96,55%
48	ZIJ	10341	0,2240%	89,66%
49	ZO	9938	0,2153%	98,52%
50	MOET	9741	0,2110%	92,61%
51	WEL	9697	0,2101%	98,77%
52	HUN	9494	0,2057%	91,87%
53	HAAR	9203	0,1994%	83,74%
54	ANDERE	8816	0,1910%	95,32%
55	GEEN	8656	0,1875%	97,54%
56	JE	8655	0,1875%	50,99%
57	AL	8545	0,1851%	95,07%
58	ZAL	8057	0,1745%	90,39%
59	ZOU	7307	0,1583%	94,33%
60	TUSSEN	7264	0,1574%	89,90%
61	U	7188	0,1557%	35,22%
62	VEEL	7033	0,1523%	91,87%
63	NU	6551	0,1419%	92,36%
64	WERD	6212	0,1346%	87,93%
65	MOETEN	6016	0,1303%	87,93%
66	ALLEEN	5855	0,1268%	91,13%
67	ZOALS	5782	0,1252%	87,93%
68	ONDER	5510	0,1194%	91,63%
69	EERSTE	5400	0,1170%	93,10%
70	EIGEN	5191	0,1124%	85,22%
71	MAKEN	5148	0,1115%	88,42%
72	ECHTER	5119	0,1109%	82,51%
73	WAAR	5118	0,1109%	92,12%
74	HAD	5095	0,1104%	75,37%
75	NA	5089	0,1102%	92,36%

Rangnr.	Woordtype	Abs. freq.	Rel. freq.	In % doc.
76	TEN	4961	0,1075%	84,73%
77	OMDAT	4848	0,1050%	89,41%
78	KOMEN	4737	0,1026%	90,64%
79	ALLE	4716	0,1022%	87,44%
80	HOE	4697	0,1017%	87,68%
81	GOED	4675	0,1013%	83,74%
82	GROTE	4649	0,1007%	88,92%
83	DUS	4632	0,1003%	84,48%
84	TEGEN	4631	0,1003%	86,21%
85	WIJ	4628	0,1003%	66,26%
86	GAAN	4600	0,0996%	91,87%
87	HEM	4594	0,0995%	71,92%
88	ONS	4545	0,0985%	80,05%
89	WEER	4483	0,0971%	87,68%
90	TOCH	4480	0,0970%	88,67%
91	HIER	4444	0,0963%	89,66%
92	NIEUWE	4436	0,0961%	82,02%
93	GAAT	4422	0,0958%	84,48%
94	WANNEER	4370	0,0947%	70,44%
95	TIJD	4368	0,0946%	88,18%
96	DAAR	4340	0,0940%	87,93%
97	VOORAL	4199	0,0910%	81,03%
98	MOGELIJK	4188	0,0907%	84,73%
99	PLAATS	4173	0,0904%	89,16%
100	STAAT	4068	0,0881%	83,25%
101	DOEN	4051	0,0878%	84,98%
102	TWEE	4014	0,0869%	88,18%
103	ZELF	3938	0,0853%	85,47%
104	KOMT	3915	0,0848%	84,98%
105	MENSEN	3835	0,0831%	73,65%
106	AANTAL	3808	0,0825%	74,14%
107	BINNEN	3803	0,0824%	83,00%
108	STEEDS	3755	0,0813%	81,28%
109	WAARIN	3738	0,0810%	84,98%
110	WIL	3635	0,0787%	83,50%
111	WAREN	3603	0,0780%	78,33%
112	VAAK	3602	0,0780%	73,15%
113	ELKAAR	3518	0,0762%	81,77%
114	IETS	3505	0,0759%	79,31%
115	LEVEN	3489	0,0756%	71,92%
116	BEPAALDE	3485	0,0755%	73,15%
117	GEVEN	3458	0,0749%	83,00%
118	ZONDER	3352	0,0726%	85,96%
119	ZEER	3328	0,0721%	76,85%
120	ÉÉN	3326	0,0720%	83,50%
121	VERSCHILLENDE	3279	0,0710%	75,12%
122	LATEN	3150	0,0682%	78,08%
123	JAREN	3140	0,0680%	78,33%
124	TOE	3083	0,0668%	87,44%
125	ZIEN	3065	0,0664%	81,28%
126	JAAR	3045	0,0660%	74,63%
127	GEVAL	3029	0,0656%	79,80%
128	WELKE	3020	0,0654%	71,43%
129	AF	3012	0,0652%	88,92%
130	ZULLEN	2974	0,0644%	75,62%
131	DER	2954	0,0640%	65,52%
132	MINDER	2940	0,0637%	74,63%
133	ALTIJD	2900	0,0628%	79,06%
134	EENS	2894	0,0627%	81,77%
135	MIJN	2860	0,0620%	56,16%
136	MENS	2814	0,0610%	51,23%
137	LAATSTE	2810	0,0609%	86,45%
138	TOEN	2803	0,0607%	68,97%
139	VRAAG	2798	0,0606%	79,06%
140	SLECHTS	2777	0,0602%	75,86%
141	WIJZE	2762	0,0598%	74,14%
142	DUIDELIJK	2745	0,0595%	80,05%
143	ONZE	2734	0,0592%	69,70%
144	ZELFS	2711	0,0587%	80,30%
145	BELANG	2575	0,0558%	66,50%
146	WERK	2521	0,0546%	68,97%
147	ONDERZOEK	2511	0,0544%	53,94%
148	BIJVOORBEELD	2502	0,0542%	59,36%
149	TWEEDE	2475	0,0536%	78,33%
150	BAL	2449	0,0530%	6,65%

Tabel 6-1 Eerste pagina van de frequentielijst van het corpus algemeen Nederlands

6.2.1 Woordfrequenties

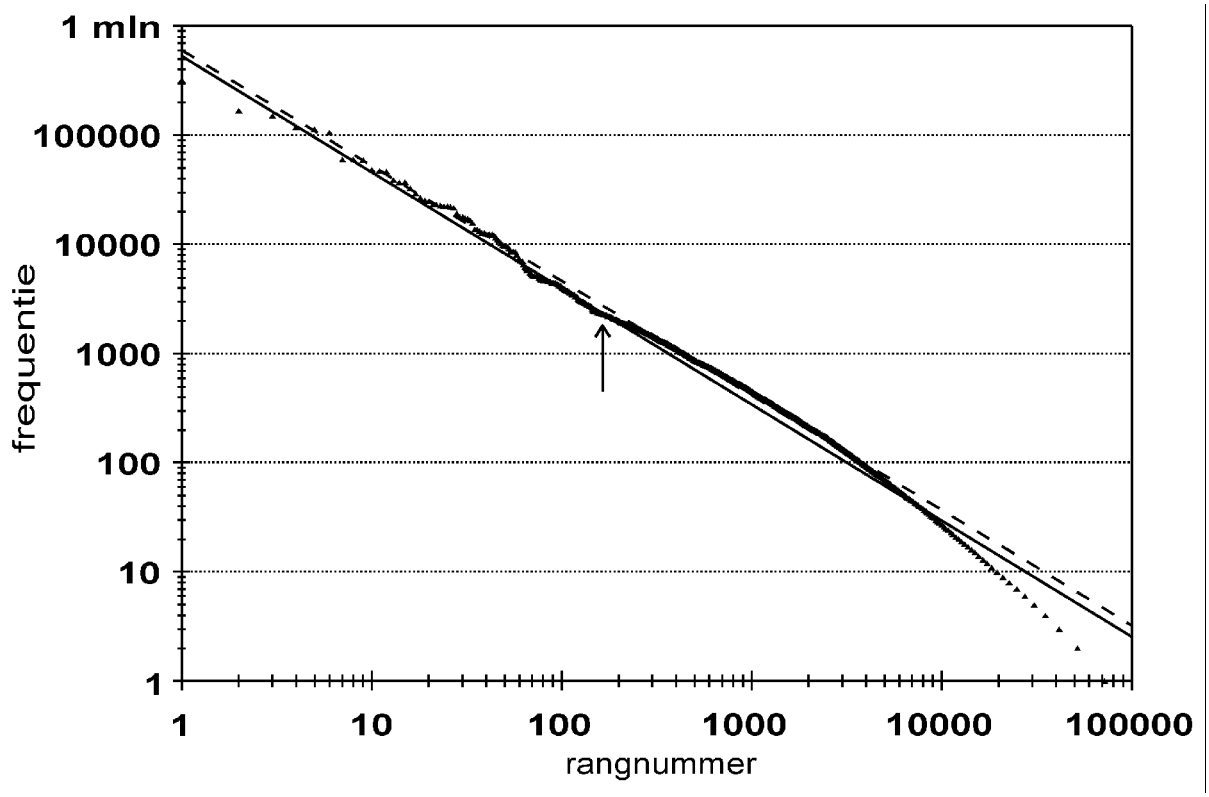
Allereerst komt daarbij weer aan de orde het verband tussen rangnummer en woordfrequentie. Om dit verband te onderzoeken zijn net als in het vorige hoofdstuk weer grafieken vervaardigd van de cumulatieve frequenties, afgezet tegen het rangnummer, en van de 'normale' frequenties, eveneens afgezet tegen het rangnummer. In deze grafieken is net als in het vorige hoofdstuk voor de horizontale as resp. voor beide assen weer gebruik gemaakt van een logaritmische schaalverdeling. Eerst zullen we weer kijken naar de cumulatieve frequenties:



Afb. 6-1 Cumulatieve frequenties, logaritmische schaalverdeling voor het rangnummer

Een globale vergelijking van deze grafiek met de overeenkomstige grafieken van de andere twee corpora (4-2b en 5-1) leert dat de curve in dit geval over een grotere lengte dan bij de andere corpora de vorm heeft van een rechte. De kromming aan het einde treedt pas op in de buurt van rangnummer 5000, en is ook minder sterk. Verder is het zo (hoewel dat niet direct zichtbaar is in de grafiek) dat de rangnummers 1 t/m 1000 nu geen 80% (zoals bij de juridische corpora), maar slechts iets meer dan 70% van alle tokens omvatten. Voor 80% van de tokens moeten we nu tot en met rangnummer 2500 gaan. Dat betekent dus een lagere token/type ratio, in ieder geval in dit gedeelte van de frequentielijst.

De cumulatieve frequenties vertonen dus een wat ander beeld dan bij de juridische corpora. Voor de 'normale' (niet-cumulatieve) frequenties blijk dat in zekere zin ook op te gaan. Deze worden afgebeeld in grafiek 6-2 (zie de volgende bladzijde). Vooral in het gebied van de laagste rangnummers (ongeveer tot 20) vertonen de datapunten wat meer spreiding, ze vormen minder één lijn en vallen minder goed samen met de ingetekende regressielijnen (waarover zodadelijk meer). Het aantal onregelmatigheden is nog wat groter dan bij het corpus wetgevingsteksten, dat op zijn beurt weer een onregelmatiger verloop te zien gaf dan het corpus jurisprudentieteksten. De absolute frequentie daalt bijvoorbeeld van rangnummer 6 tot



Afb. 6-2 In het corpus algemeen Nederlands voorkomende frequenties (logaritmische schaalverdeling)

rangnummer 7 in één keer van 106200 tot 60319, en in relatieve zin van 2,3% tot 1,3%, een aanmerkelijke daling. Een dergelijke sterke daling vinden we slechts op twee andere plaatsen: in ditzelfde corpus, en wel tussen rangnummer 1 en 2, en in het corpus wetgevingsteksten, tussen rangnummer 2 en 3. Gegeven het feit dat bij alle drie corpora de daling in frequentie gemiddeld steeds geringer wordt bij oplopend rangnummer, is deze daling bij rangnummer 6 toch wel opmerkelijk te noemen. Een verklaring is echter niet eenvoudig te geven, tenminste niet aan de hand van de typen in de frequentielijst.

Naarmate het rangnummer stijgt wordt het aantal onregelmatigheden geringer, waarna de curve ongeveer vanaf rangnummer 80 over een kort gebied (korter dan bij de andere corpora) praktisch de vorm van een rechte lijn heeft. Daarna treedt weer het verschijnsel op dat we al kennen uit de vorige twee hoofdstukken: vanaf een bepaald punt, in dit geval rond rangnummer 155 (zie pijl), is niet langer sprake van een al dan niet onregelmatig rechtlijnig verloop, maar vormen de datapunten een lichte, neerwaartse kromme welke zeer regelmatig van vorm is. Het gedeelte tot aan dit punt zal net als bij de andere corpora weer worden beschouwd als de kop van de frequentielijst, welke bij dit corpus dus met name aan het begin een wat groter aantal onregelmatigheden te zien geeft.

Om het verband tussen rangnummer en frequentie in de kop van de frequentielijst weer te geven kan weer gebruik worden gemaakt van een lineaire regressielijn. De vergelijking van deze lijn kan volgens twee methoden worden bepaald. De eindresultaten van beide methoden liggen in dit geval echter dichter bij elkaar dan bij de andere corpora:

1. wanneer gebruik wordt gemaakt van 1 datapunt per *rangnummer* luidt de vergelijking van de regressielijn:

$$\log r = -1{,}054 \log rangnr + 5{,}777 \qquad (6\text{-}1)$$

   Deze rechte is in de grafiek ingetekend in de vorm van een stippellijn.

2. wanneer wordt uitgegaan van 1 datapunt per *token* ziet deze vergelijking er als volgt uit:

$$\log r = -1{,}065 \log rangnr + 5{,}732 \qquad (6\text{-}2)$$

   Deze functie is ingetekend in de vorm van een doorgetrokken lijn. Vermeld moet nog worden dat het ook bij dit kleinste corpus nodig was om uit het totale aantal tokens een steekproef te nemen. Het aantal datapunten zou anders met ruim 4 miljoen toch nog veel te groot zijn geworden. Door in dit geval na ieder datapunt de 660 volgende punten over te slaan ontstond net als bij de andere corpora een serie van ongeveer 7000 datapunten, aan de hand waarvan de vergelijking van de regressielijn werd bepaald.

Bij beide regressielijnen hoort weer een ongeveer even hoge r^2 (het kwadraat van de lineaire correlatiecoëfficiënt), in dit geval zelfs van meer dan 0,99. Dat betekent dus dat ze allebei ongeveer een even groot deel van de variantie in hun datanet verklaren. Opvallend is nu echter dat in feite tussen beide regressielijnen slechts een zeer gering verschil bestaat. De tweede lijn daalt slechts een fractie sneller (de x-coëfficiënt is 0,011 lager), wat met het blote oog bijna niet te zien is, terwijl ook de y-intercepts dicht bij elkaar in de buurt liggen. Het resultaat is dat beide lijnen over het gehele bereik praktisch parallel lopen, waarbij regressielijn 1 steeds boven regressielijn 2 blijft. Dat de methode volgens welke de vergelijking van de regressielijn wordt berekend hier veel minder verschil maakt vindt zijn oorzaak mogelijk weer in de geringere omvang van het corpus, en in de lagere token/type ratio (in ieder geval in het eerste deel van de frequentielijst). Het aantal tokens per rangnummer is vanwege deze twee factoren lager, waardoor hun aantal ook minder 'gewicht in de schaal legt', en hun invloed bij het berekenen van de vergelijking van de regressielijn geringer is. Toch is er iets voor te zeggen om ook nu weer te kiezen voor de tweede regressielijn. Deze volgt namelijk beter de datapunten, juist in het gebied waar deze min of meer een rechte vormen (van rangnummer 80 tot 155). Daarbij ligt ook het punt waar de curve de vorm van een kromme krijgt praktisch op deze tweede regressielijn (en duidelijk onder de eerste). Daarom zal ik, onder andere voor het maken van vergelijkingen met de andere corpora, uitgaan van de tweede regressielijn. De vergelijking van die lijn kan worden geschreven als:

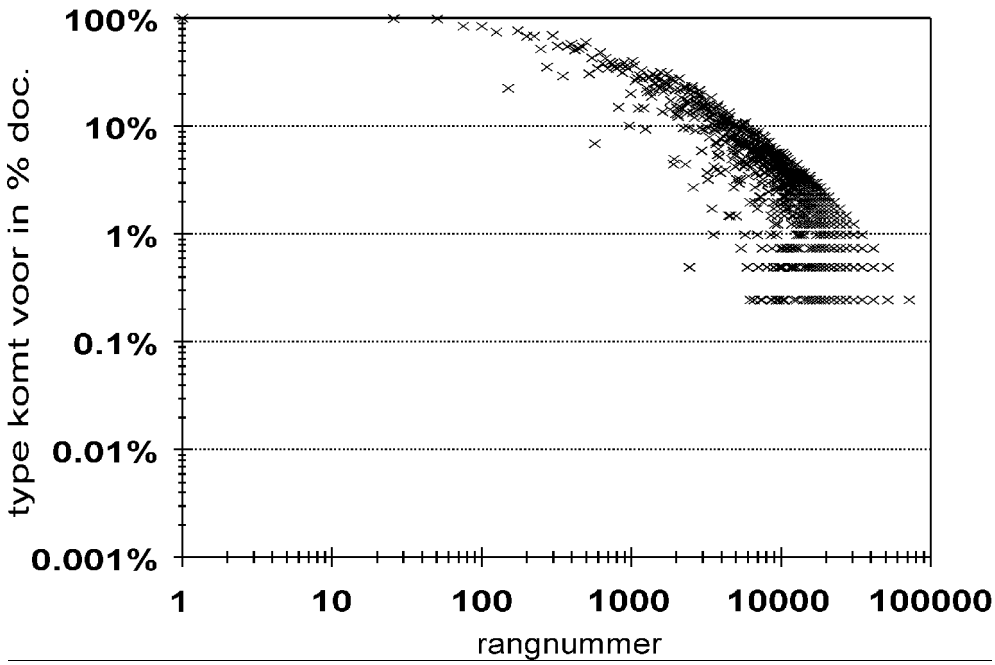
$$r = rangnr^{-1{,}065} \cdot 10^{5{,}732} = \frac{10^{5{,}732}}{rangnr^{1{,}065}} \qquad (6\text{-}3)$$

De lagere waarde van de constante $10^{5{,}732}$ vindt zijn oorzaak waarschijnlijk in de geringere omvang van het corpus. Opvallend is echter dat de waarde van de exponent van het rangnummer (de x-coëfficiënt uit de vergelijking van de regressielijn) bijna gelijk is aan die bij het corpus jurisprudentieteksten (1,069). Dit betekent tevens dat ook voor dit corpus de wet van

Zipf niet opgaat (de exponent is ongelijk aan 1). De conclusie kan zijn dat het verband tussen rangnummer en frequentie in dit corpus sterk lijkt op dat in de andere corpora. Het enige duidelijke verschilpunt is de lagere token/type ratio, als gevolg waarvan onder andere de twee verschillende methoden voor het uitvoeren van een regressieanalyse resultaten opleveren die veel minder uiteenlopen dan bij de andere corpora. Van doorslaggevend belang is dit punt echter niet, we zullen ons daarom net als bij de andere corpora in eerste instantie richten op bestudering van de kop van de frequentielijst, dat wil zeggen op de rangnummers 1 tot (ongeveer) 155.

6.2.2 Spreiding van woordtypen

Het corpus algemeen Nederlands bestaat uit in totaal 406 documenten. Het spreidingspercentage dat in de frequentielijst staat vermeld geeft weer aan in hoeveel van die documenten een bepaald woordtype voorkomt. De spreidingspercentages per rangnummer worden hieronder grafisch weergegeven. Net als in de voorgaande hoofdstukken ontstaat het duidelijkste beeld wanneer zowel voor rangnummer als spreidingspercentage een logaritmische schaalverdeling wordt gebruikt.



Afb. 6-3 Spreiding van alle typen over het corpus algemeen Nederlands (logaritmische schaalverdeling)

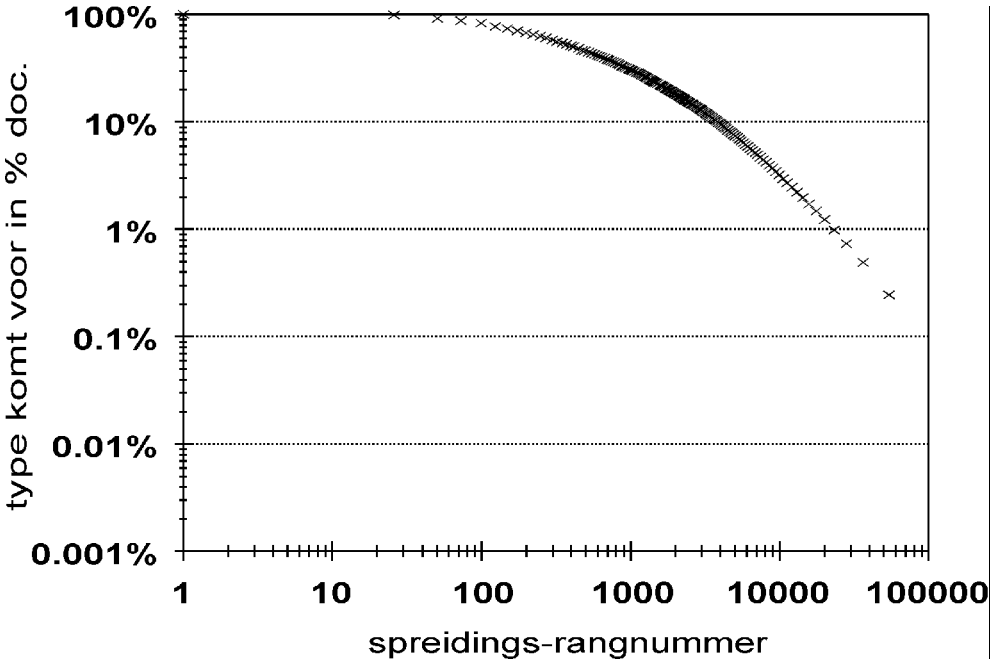
Ook nu was het om technische redenen niet mogelijk om ieder datapunt (d.w.z. het spreidingspercentage van ieder woordtype) in de grafiek op te nemen, maar moest worden volstaan met 1 op elke 25 punten. De grafiek wordt daar weliswaar wat minder nauwkeurig door, maar biedt nog steeds (of eigenlijk: juist om die reden) een goed totaalbeeld. De puntenwolk waaiert naar het einde toe uit, net als bij de andere twee corpora, doordat ieder rangnummer daar een steeds groter aantal woordtypen met gelijke frequentie, maar met verschillende spreidingspercentages omvat. Wanneer we afbeelding 6-3 bekijken valt op dat de spreidings-

percentages in dit corpus ruimschoots hoger liggen dan in beide andere corpora. De omvang van het corpus, en ook de afmetingen van de documenten kan daarop echter van invloed zijn. Dat blijkt bijvoorbeeld al wanneer we kijken naar het *laagste* spreidingspercentage dat voorkomt. Een type komt nooit in minder dan 1 document voor, en 1/406 levert een hoger percentage op dan 1/18803 (het aantal wetgevings-documenten). Daardoor bevat de grafiek geen datapunten bij spreidingspercentages lager dan 0,25%. Verder valt nog op dat de grafiek enkele duidelijke uitbijters bevat. Het gaat daarbij steeds om typen met een veel geringere spreiding dan men zou verwachten op grond van hun rangnummer. De grafiek toont een drietal van deze punten, het blijkt te gaan om UW (rangnr. 151, in 22,4% van de documenten), BABY (rangnr. 576, in 6,9% van de documenten) en om de typen SUZANNE en SWING (beiden rangnr. 2465, in 0,49% van de documenten). Bestudering van de frequentielijst levert nog een aantal andere typen op met een opvallend geringe mate van spreiding in verhouding tot hun rangnummer (de grafiek bevat immers niet alle datapunten). Zie voor een opsomming van enkele voorbeelden daarvan tabel 6-2.

Rang-nummer	Woordtype	In % docu-menten
56	JE	50,99%
61	U	35,22%
150	BAL	6,65%
518	DUIVEN	2,96%
570	SPELER	2,46%
809	EUTHANASIE	2,46%
1055	VOORLICHTER	0,74%
1109	HOK	0,99%
1584	VOORLICHTERS	0,74%
1986	DUIF	0,74%
2050	ANMB	0,49%
2067	LOMBOK	0,74%
2254	BETONBUURT	0,49%
2465	SUZANNE	0,49%
2465	SWING	0,49%

Tabel 6-1 Typen met opvallend lagere spreiding dan andere typen met vergelijkbaar rangnummer

Ook bij dit corpus kunnen de typen weer worden gesorteerd op hun mate van spreiding, waarna een 'spreidings-rangnummer' kan worden toegekend. Wanneer de spreidingspercentages dan worden afgezet tegen dit nieuwe rangnummer ontstaat net als bij de andere corpora een enkelvoudige 'curve', die min of meer overeen komt met de bovenste begrenzing van de puntenwolk uit afbeelding 6-3. Deze curve heeft een duidelijk ander verloop dan in beide eerder behandelde corpora het geval was (zie afbeelding 6-4). Zij verloopt aanvankelijk vlakker, en vertoont dan een sterkere kromming (geen knik, zoals bij het corpus wetgevingsteksten). Nader onderzoek wijst uit dat de oorzaak daarvan niet alleen gezocht kan worden in het geringere aantal documenten en het daaraan gekoppelde hogere minimale spreidingspercentage: ook wanneer de schaalverdeling op de y-as wordt aangepast blijft het eerste deel van de grafiek een duidelijk sterkere kromming vertonen. Het lijkt erop dat dit corpus relatief veel 'algemene' woordtypen bevat, die in een flink aantal documenten terugkomen. Aan overeenkomsten en verschillen met de andere twee corpora zal in het volgende hoofdstuk nog enige aandacht worden besteed.



Afb. 6-4 Spreiding van woordtypen over het corpus algemeen Nederlands, rangnummers (x-as) op basis van de mate van spreiding

De spreiding van woordtypen is bij dit corpus dus weer enigszins anders dan bij de juridische corpora. Het totaalbeeld dat naar voren komt uit frequentie- en spreidingsgegevens is echter dat de corpora op deze punten naast verschillen toch ook duidelijke overeenkomsten vertonen. Er is dan ook, na bestudering van dit derde corpus, geen reden om af te wijken van de gehanteerde werkwijze bij de andere corpora. We zullen ons daarom weer vooral richten op het eerste deel van de frequentielijst (de eerste 300 rangnummers). Dit deel van de frequentielijst van het corpus algemeen Nederlands is te vinden in het laatste gedeelte van bijlage 3. Voor geïnteresseerden zij verder nog vermeld dat ook in dit geval de volledige lijst is op te vragen bij de auteur.

6.3 Frequentieverdelingen

Een frequentieverdeling voor het corpus algemeen Nederlands, op dezelfde manier opgebouwd als in de voorgaande twee hoofdstukken, is te vinden in tabel 6-3 (zie de volgende bladzijde). De logaritmisch in omvang toenemende frequentieklassen in deze tabel bieden weer een tegenwicht voor de ongelijke spreiding van de verschillende frequenties. Voor dit corpus kon met 55 frequentieklassen worden volstaan (in verband met de hoogste in het corpus voorkomende frequentie), terwijl voor de wetgevingsteksten 60 en voor de jurisprudentieteksten 65 klassen nodig waren. Dit houdt weer verband met de omvang van de corpora.

Klasse	Grenzen	Klasse-midden	Aantal freq.	Aantal typen	Aantal tokens	Klasse	Grenzen	Klassemid-den	Aantal freq.	Aantal typen	Aantal tokens
1	10^{0} - $10^{0.1}$	1,13	1	70713	70713	29	$10^{2.8}$ - $10^{2.9}$	712,64	98	161	114101
2	$10^{0.1}$ - $10^{0.2}$	1,42	0	0	0	30	$10^{2.9}$ - $10^{3.0}$	897,16	95	122	107770
3	$10^{0.2}$ - $10^{0.3}$	1,79	0	0	0	31	$10^{3.0}$ - $10^{3.1}$	1129,46	76	87	97821
4	$10^{0.3}$ - $10^{0.4}$	2,25	1	20315	40630	32	$10^{3.1}$ - $10^{3.2}$	1421,91	75	82	115205
5	$10^{0.4}$ - $10^{0.5}$	2,84	1	10302	30906	33	$10^{3.2}$ - $10^{3.3}$	1790,08	66	73	129522
6	$10^{0.5}$ - $10^{0.6}$	3,57	0	0	0	34	$10^{3.3}$ - $10^{3.4}$	2253,57	52	54	120961
7	$10^{0.6}$ - $10^{0.7}$	4,50	2	10817	47644	35	$10^{3.4}$ - $10^{3.5}$	2837,08	25	25	72116
8	$10^{0.7}$ - $10^{0.8}$	5,66	1	3313	19878	36	$10^{3.5}$ - $10^{3.6}$	3571,67	19	19	68372
9	$10^{0.8}$ - $10^{0.9}$	7,13	1	2496	17472	37	$10^{3.6}$ - $10^{3.7}$	4496,47	27	27	120949
10	$10^{0.9}$ - $10^{1.0}$	8,97	3	5306	47013	38	$10^{3.7}$ - $10^{3.8}$	5660,72	12	12	65535
11	$10^{1.0}$ - $10^{1.1}$	11,29	2	2220	25402	39	$10^{3.8}$ - $10^{3.9}$	7126,43	5	5	35343
12	$10^{1.1}$ - $10^{1.2}$	14,22	3	2444	34010	40	$10^{3.9}$ - $10^{4.0}$	8971,64	10	10	90802
13	$10^{1.2}$ - $10^{1.3}$	17,90	4	2185	37948	41	$10^{4.0}$ - $10^{4.1}$	11294,63	8	8	93127
14	$10^{1.3}$ - $10^{1.4}$	22,54	6	2193	48763	42	$10^{4.1}$ - $10^{4.2}$	14219,09	7	7	94882
15	$10^{1.4}$ - $10^{1.5}$	28,37	6	1431	40417	43	$10^{4.2}$ - $10^{4.3}$	17900,78	6	6	106945
16	$10^{1.5}$ - $10^{1.6}$	35,72	8	1344	47448	44	$10^{4.3}$ - $10^{4.4}$	22535,74	8	8	183384
17	$10^{1.6}$ - $10^{1.7}$	44,96	11	1348	60029	45	$10^{4.4}$ - $10^{4.5}$	28370,82	3	3	81420
18	$10^{1.7}$ - $10^{1.8}$	56,61	13	1026	57901	46	$10^{4.5}$ - $10^{4.6}$	35716,75	4	4	146790
19	$10^{1.8}$ - $10^{1.9}$	71,26	16	859	61162	47	$10^{4.6}$ - $10^{4.7}$	44964,72	3	3	142616
20	$10^{1.9}$ - $10^{2.0}$	89,72	21	780	69626	48	$10^{4.7}$ - $10^{4.8}$	56607,23	3	3	179837
21	$10^{2.0}$ - $10^{2.1}$	112,95	25	606	68253	49	$10^{4.8}$ - $10^{4.9}$	71264,28	0	0	0
22	$10^{2.1}$ - $10^{2.2}$	142,19	33	534	75188	50	$10^{4.9}$ - $10^{5.0}$	89716,41	0	0	0
23	$10^{2.2}$ - $10^{2.3}$	179,01	41	498	88921	51	$10^{5.0}$ - $10^{5.1}$	112946,27	3	3	339747
24	$10^{2.3}$ - $10^{2.4}$	225,36	52	402	89341	52	$10^{5.1}$ - $10^{5.2}$	142190,93	1	1	151236
25	$10^{2.4}$ - $10^{2.5}$	283,71	65	347	97701	53	$10^{5.2}$ - $10^{5.3}$	179007,78	1	1	170425
26	$10^{2.5}$ - $10^{2.6}$	357,17	80	273	97123	54	$10^{5.3}$ - $10^{5.4}$	225357,44	0	0	0
27	$10^{2.6}$ - $10^{2.7}$	449,65	88	210	93522	55	$10^{5.4}$ - $10^{5.5}$	283708,20	1	1	313273
28	$10^{2.7}$ - $10^{2.8}$	566,07	97	192	107265						

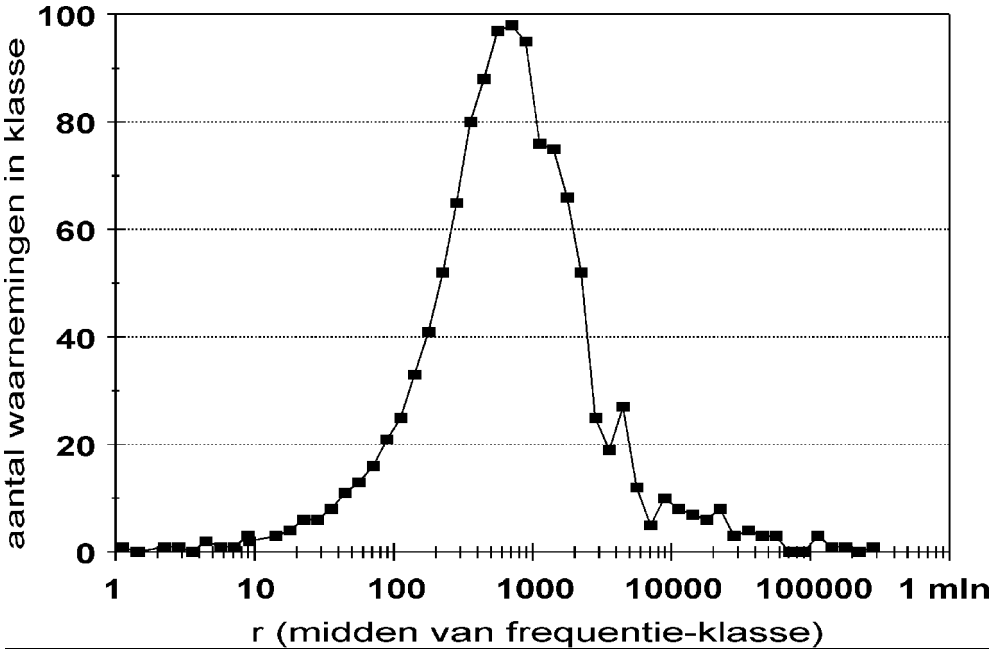
Tabel 6-2 Frequentieverdeling corpus algemeen Nederlands. De frequentieklassen nemen logaritmisch in omvang toe.

Het aantal lege klassen in deze frequentieverdeling - 6 stuks - ligt in dezelfde orde van grootte als bij de andere twee corpora. Hoewel dit aantal relatief gezien iets hoger is, aangezien het totaal aantal klassen nu geringer is, kan het nog steeds laag worden genoemd, zeker gezien de scheve verdeling van de verschillende frequenties. De symmetrie van de verdeling is ook wat minder dan bij de voorgaande twee corpora; het hoogste aantal verschillende frequenties (98) treffen we niet in de middelste klasse aan, maar een klasse daar voorbij (klasse 29). De toe- c.q. afname van dit aantal verloopt daarentegen juist weer wat regelmatiger (zie ook de volgende subparagraaf). Wanneer met behulp van de cijfers uit tabel 6-3 het totaal aantal tokens en de totale token/type ratio worden geschat door telkens klassemidden en aantal typen per klasse te vermenigvuldigen, ontstaat net als bij de andere corpora een resultaat dat zeer dicht bij de werkelijkheid ligt: de afwijking is niet hoger dan zo'n 0,43%. Hieronder zal in nog wat meer detail op enkele punten met betrekking tot deze frequentieverdeling worden ingegaan.

6.3.1 De aantallen frequenties per frequentieklasse

Wanneer de aantallen verschillende frequenties per klasse in een grafiek worden uitgezet, en de punten worden verbonden door lijnen, ontstaat weer een beeld dat sterk gelijkt op dat bij de

andere corpora. Opnieuw zijn onregelmatigheden vrijwel uitsluitend te vinden in het rechtergedeelte van de grafiek (dat wil zeggen, rechts van de top). Ook deze curve vertoont in dat gedeelte een 'bultje', en wel bij r = 4496,47 (klasse 37). In dit geval is het echter zo dat het gehele verloop van de curve vanaf dat punt duidelijk onregelmatiger wordt. Wanneer we de frequentieverdeling nog eens goed bestuderen, kan daaruit een mogelijke verklaring worden afgeleid: vanaf de klasse waar het 'bultje' optreedt is sprake van frequentieklassen met alleen nog maar unieke frequenties (dat wil zeggen, iedere frequentie in de klasse komt nog maar voor bij een enkel type). Dat zou er misschien de reden van kunnen zijn dat bij deze klassen de fluctuatie in de aantallen verschillende frequenties groter is dan bij de klassen daarvoor, waarin verschillende typen met gelijke frequentie vertegenwoordigd zijn. Wanneer we de frequentielijst van de andere twee corpora bestuderen, kunnen we vaststellen dat dit verschijnsel ook daar mogelijk een rol speelt: onregelmatigheden, zij het geringer van omvang, ontstaan met name vanaf het punt waar de frequentieklassen alleen nog maar unieke frequenties bevatten. Het is dus mogelijk dat we hier niet met een specifiek kenmerk van een of meer corpora van doen hebben, maar met een verschijnsel dat rechtstreeks gerelateerd is aan de gehanteerde methode voor het opstellen van een frequentieverdeling.

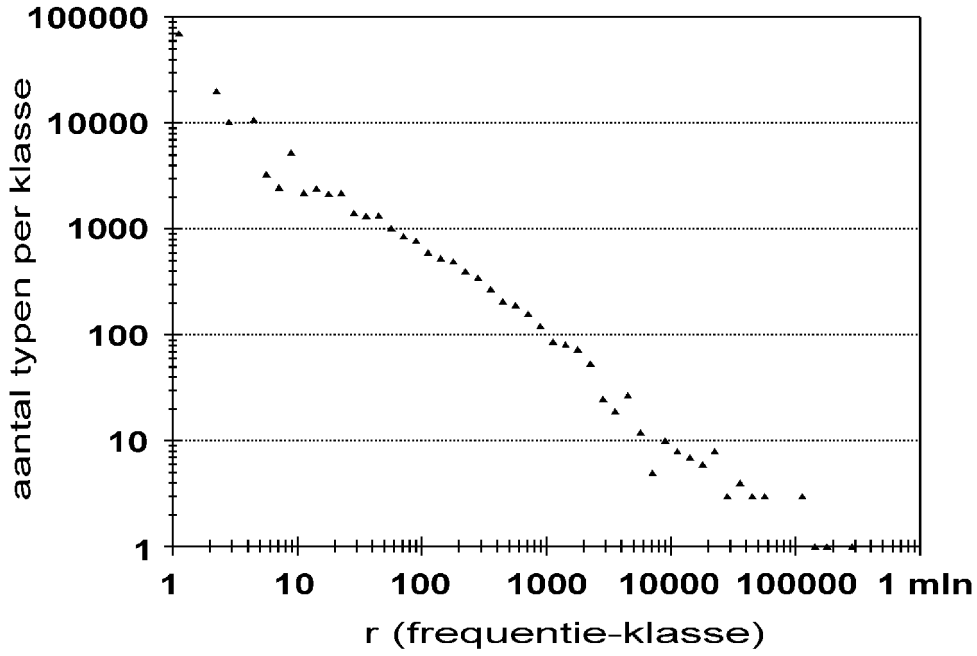


Afb. 6-5 Aantal waarnemingen (verschillende frequenties) per frequentieklasse

6.3.2 De aantallen woordtypen en -tokens per klasse

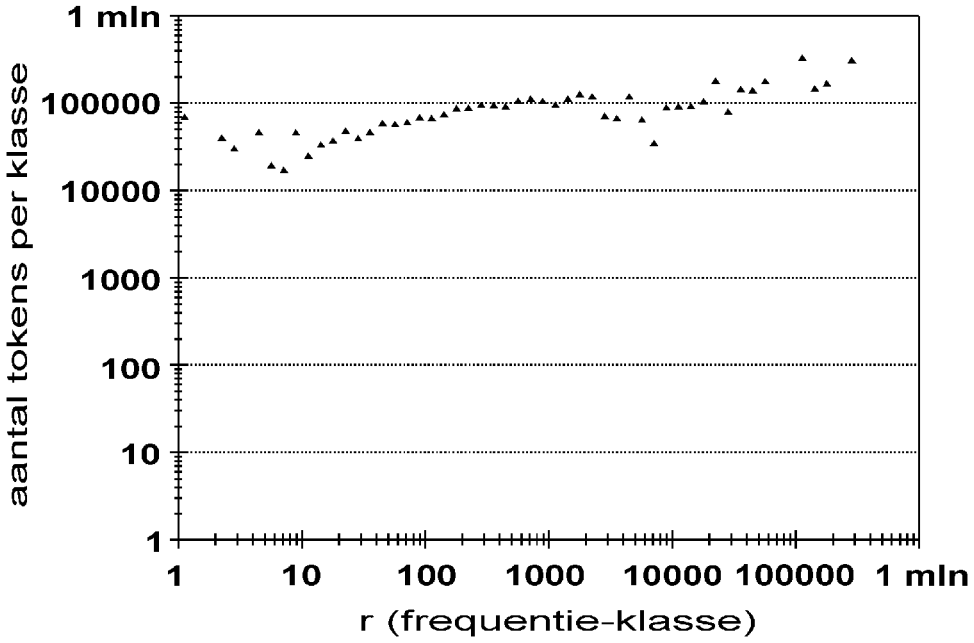
Wanneer we de aantallen woordtypen in iedere frequentieklasse (kolom 5 uit tabel 6-3) afzetten tegen de klassemiddens, ontstaat een beeld dat (afgezien van de schaal) zeer sterk gelijkt op dat bij de andere twee corpora (afbeeldingen 4-8b en 5-6). Wel is nu de frequentierange waarover sprake is van een vrijwel lineair verband wat korter: ongeveer van r = 10 tot r = 2500 in plaats van tot r = 10000. Frappant zijn echter de overeenkomsten tussen de eerste 10 frequentieklassen van alle corpora. Het patroon dat de datapunten daar vertonen lijkt in eerste instantie nogal onregelmatig, maar komt bij alle corpora vrijwel helemaal overeen. Dat

geldt echter niet voor de onregelmatigheden bij de hoogste frequentieklassen (in dit geval boven *r* = 2500); het corpus algemeen Nederlands vertoont daar duidelijk meer spreiding in de datapunten dan bijvoorbeeld het geval is bij het corpus jurisprudentieteksten. Het zou heel goed kunnen dat dit verband houdt met het wat grotere aantal onregelmatigheden in de *aantallen frequenties* per klasse in dit corpus, zoals beschreven in de vorige subparagraaf.



Afb. 6-6 Aantal woordtypen per frequentieklasse

Zoals te verwachten valt, gezien de overeenkomsten in het verloop van de aantallen typen, biedt ook de grafiek van de aantallen woordtokens een vertrouwde aanblik (zie afbeelding 6-7). De spreiding van de datapunten aan beide uiteinde van de 'curve' is bij dit corpus duidelijk groter dan bij de andere corpora. Een verband met de hierboven geconstateerde onregelmatigheden in de aantallen verschillende frequenties en de aantallen woordtypen per klasse ligt daarbij voor de hand. Voor al deze aspecten geldt verder dat ook de corpusgrootte van invloed zou kunnen zijn (het corpus algemeen Nederlands is immers een stuk kleiner dan de andere twee corpora), evenals de duidelijk lagere token/type ratio.



Afb. 6-7 Aantal woordtokens per frequentieklasse

Net als in het vorige hoofdstuk, kan aan het einde van deze paragraaf over de frequentieverdeling worden geconstateerd dat de corpora op veel punten sterke overeenkomsten vertonen. De verschillen die er zijn zullen in het volgende hoofdstuk nader onder de loupe worden genomen, nadat de corpora zoveel mogelijk onderling vergelijkbaar zijn gemaakt.

6.4 De verhouding tussen woordtokens en woordtypen

6.4.1 Token-type ratio's

Het corpus algemeen Nederlands bevat in totaal 4616455 woordtokens, en 142879 woordtypen. Dat betekent dat de token/type ratio, gemeten over het hele corpus, 32,31 bedraagt, een waarde ver onder die bij zowel de wetteksten (89,47) als bij de jurisprudentieteksten (151,18). Hoewel de corpusgrootte hier natuurlijk weer een rol speelt, is het corpus op dit punt toch structureel anders dan de andere twee corpora. Dat blijkt wel wanneer we uit het corpus weer steekproeven nemen, en die op bepaalde punten vergelijken met de in de voorgaande hoofdstukken samengestelde steekproeven uit de andere corpora.

N.B.: werden bij de andere twee corpora telkens steekproeven genomen bestaande uit een veelvoud van 1000 documenten, bij dit corpus is er voor gekozen steekproeven te trekken bestaande uit veelvouden van 25 documenten. De reden hiervoor is dat het aantal documenten in dit corpus veel lager is (406, tegen 18803 resp. 16430 bij de andere corpora), terwijl de gemiddelde documentgrootte hoger ligt (11370 tokens, tegen 681 resp. 2446). Bij een toename van de steekproefgrootte met telkens 25 documenten ontstaat een voldoende aantal 'stappen', waardoor nog goed kan worden waargenomen hoe bijvoorbeeld de token/type ratio zich ontwikkelt. Bedacht moet natuurlijk wel worden dat, ondanks het feit dat het aantal

stappen ongeveer even groot is als bij de andere corpora, bij elk van de corpora (gemiddeld) verschillende aantallen tokens worden toegevoegd: ruim 680000 bij de wetteksten, ruim 2,4 miljoen bij de jurisprudentieteksten, en ruim 280000 in dit corpus. De verhoudingen tussen die aantallen weerspiegelen weer min of meer de verschillen in afmeting tussen de corpora.

Aantal doc.	Typen	Tokens	Tok/Typ ratio	Toename	Log.Tok/ Typ ratio	Toename
25	21689	277093	12,776	-	1,255	-
50	42996	639126	14,865	16,35%	1,253	-0,17%
75	50218	879150	17,507	17,77%	1,264	0,92%
100	60436	1236180	20,454	16,84%	1,274	0,77%
125	63936	1270416	19,870	-2,86%	1,270	-0,32%
150	73182	1485967	20,305	2,19%	1,269	-0,10%
175	77595	1810616	23,334	14,92%	1,280	0,86%
200	94009	2342689	24,920	6,80%	1,281	0,08%
225	97338	2444341	25,112	0,77%	1,281	-0,01%
250	105136	2821593	26,838	6,87%	1,285	0,30%
275	114519	3209625	28,027	4,43%	1,286	0,13%
300	118838	3342708	28,128	0,36%	1,286	-0,05%
325	123393	3625269	29,380	4,45%	1,288	0,22%
350	127688	3921172	30,709	4,52%	1,291	0,23%
375	131738	4126597	31,324	2,00%	1,292	0,07%
400	142273	4586407	32,237	2,91%	1,293	0,04%
406	142879	4616455	32,310	0,23%	1,293	0,01%

Tabel 6-3 Token/type ratio als functie van de corpusgrootte

De token/type ratio groeit ook bij dit corpus vrijwel voortdurend, alleen bij de vijfde steekproef treedt even een lichte daling op. Het groeipercentage ligt aan het einde op ongeveer hetzelfde niveau als bij de andere corpora: rond de 3%. Het groeipercentage bij 406 documenten doet niet echt mee, er werden daar slechts 6 documenten toegevoegd. Iets anders ligt dat met de logaritmische token/type ratio, die niet bepaald regelmatig toeneemt en in een vijftal gevallen zelfs een daling te zien geeft. Een en ander wordt veroorzaakt door de nogal hevige *fluctuatie* die met name in het begin optreedt in het toenamepercentage. Ook hierbij speelt het (vergeleken met de andere corpora) geringe volume aan woordtokens in de steekproeven een rol, hetgeen kan worden aangetoond door de toenamepercentages te berekenen over telkens twee steekproeven (dus: van 25 naar 75 documenten, van 75 naar 125 documenten, etc.). Dan ontstaan voor de logaritmische token/type ratio toenamepercentages van achtereenvolgens 1,69%, -0,42%, 0,95%, 0,29%, 0,08%, 0,45% en 0,12%, waarbij dus nog maar in één geval sprake is van een daling, en verder van toenames die in dezelfde lijn liggen als bij bijvoorbeeld de wetgevingsteksten.

6.4.2 De karakteristiek K

De steekproef-gegevens uit tabel 6-4 zijn weer gebruikt om de waarde van de karakteristiek *K* van Yule en Herdan te berekenen, en om na te gaan in hoeverre deze waarde stabiel is. Na het opstellen van frequentieverdelingen van elk van de steekproeven werden de twee varianten van *K* berekend met behulp van formules 4-4 en 4-5 (zie hoofdstuk 4).

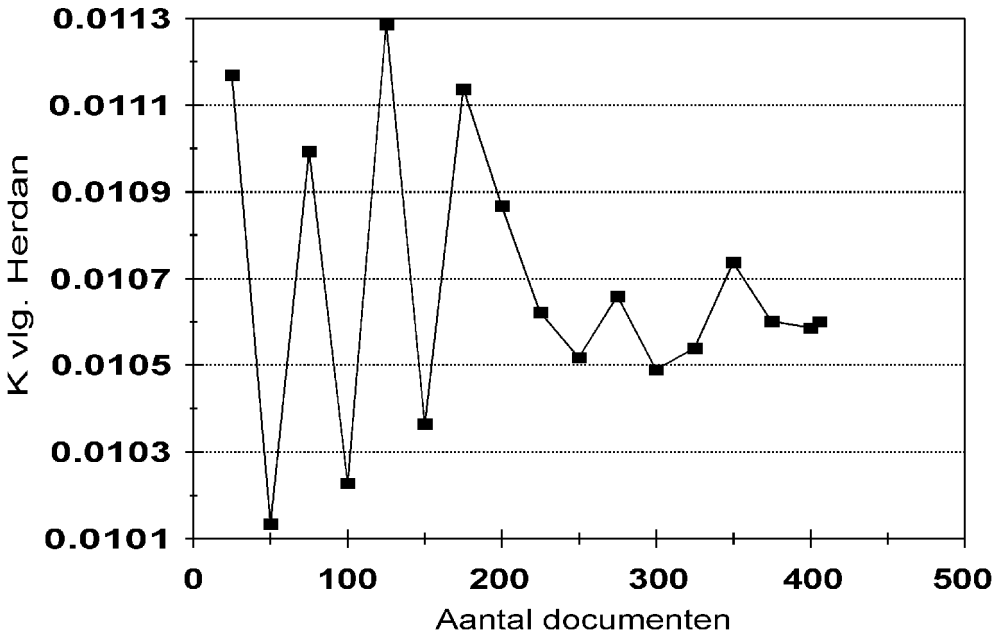
Aantal doc.	Typen	Tokens	*K* (Yule)	*K* (Herdan)
25	21689	277093	111,660	0,0111696
50	42996	639126	101,325	0,0101341
75	50218	879150	109,942	0,0109954
100	60436	1236180	102,276	0,0102284
125	63936	1270416	112,864	0,0112872
150	73182	1485967	103,648	0,0103655
175	77595	1810616	111,378	0,0111384
200	94009	2342689	108,676	0,0108681
225	97338	2444341	106,227	0,0106231
250	105136	2821593	105,183	0,0105187
275	114519	3209625	106,602	0,0106605
300	118838	3342708	104,898	0,0104901
325	123393	3625269	105,398	0,0105401
350	127688	3921172	107,371	0,0107373
375	131738	4126597	106,020	0,0106022
400	142273	4586407	105,868	0,0105871
406	142879	4616455	106,015	0,0106017

Tabel 6-4 De waarde van *K* bij toenemende steekproefgrootte

De waarde van *K* (volgens beide berekeningsmethoden) blijkt in het begin nog flink te fluctueren. In de tweede helft van de serie steekproeven treedt echter een stabilisering op (ongeveer vanaf 225 documenten). Opnieuw zien we dus dat kennelijk een bepaalde minimum steekproef-omvang (in dit geval zo'n 2,5 miljoen tokens) nodig is om de waarde van *K* redelijk nauwkeurig te kunnen vaststellen. Bij de andere corpora was dat ook het geval, alleen was toen sprake van omvangrijker steekproeven, waardoor de benodigde minimum-omvang al na twee tot vijf steekproeven werd bereikt. De waarde van *K* werd daardoor na een geringer aantal steekproeven - maar bij ongeveer hetzelfde aantal woordtokens - voldoende stabiel.

De verschillen tussen de berekeningswijzen van Yule en Herdan blijken ook in dit geval gering, hoewel over de hele linie net iets groter dan bij de andere corpora. Uiteindelijk (bij de laatste paar steekproeven) ligt het verschil rond de 2/10000000 (tegen 1/10000000 bij de andere corpora). Het gaat in ieder geval om een verwaarloosbaar klein verschil, bij steekproeven van een omvang waarbij *K* redelijk nauwkeurig berekend kan worden. Nogmaals kan daarom de voorkeur worden gegeven aan de iets eenvoudiger berekeningsmethode van Herdan.

De steeds geringere fluctuatie van *K* wordt in grafiek 6-8 goed in beeld gebracht (zie de volgende pagina).



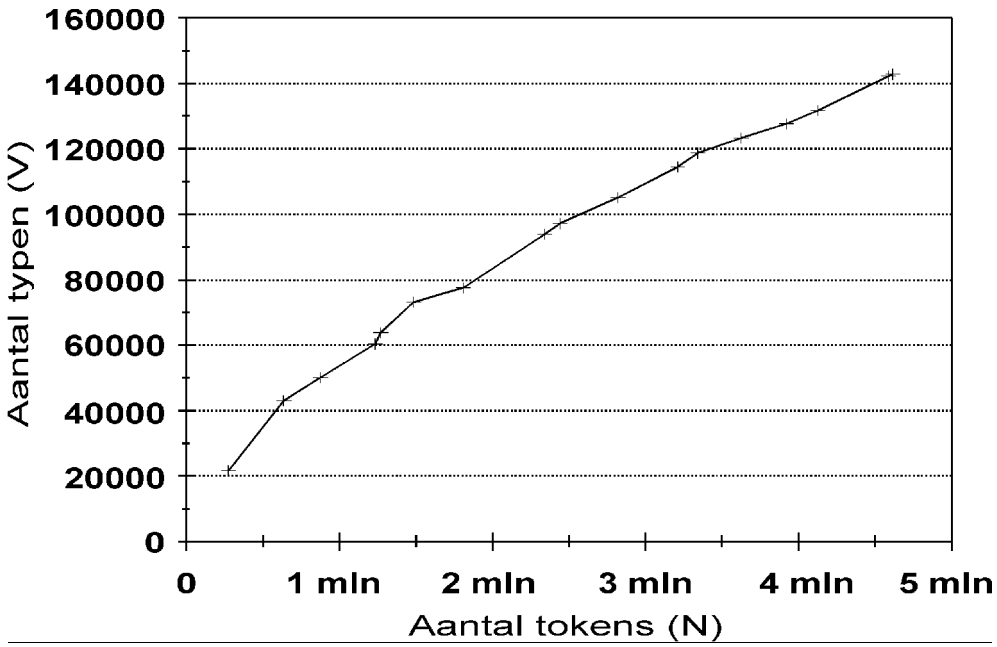
Afb. 6-8 De waarde van *K* (Herdan) bij toenemende steekproefgrootte

De veronderstelling, geuit in het vorige hoofdstuk, dat met name de absolute omvang van een steekproef bepalend is voor de nauwkeurigheid waarmee *K* wordt berekend vindt hier steun. Een niveau van stabiliteit dat vergelijkbaar is met hetgeen werd bereikt bij de andere twee corpora ontstaat ook nu bij een steekproefgrootte van rond de 4 miljoen woordtokens. Dat komt in dit geval echter neer op bijna het volledige corpus, terwijl een dergelijk aantal tokens slechts zo'n 10% uitmaakt van het corpus jurisprudentieteksten, en ruim 30% van het corpus wetgevingsteksten. Dat houdt dus in dat voor de hier onderzochte corpora zeer uiteenlopende percentages van elk corpus nodig waren voor een betrouwbare berekening, maar dat het aantal woordtokens dat nodig was steeds in dezelfde orde van grootte lag.

De waarde van *K* ligt uiteindelijk voor het corpus algemeen Nederlands rond de 106 (vlg. Yule) c.q. 0.0106 (vlg. Herdan). Dat is duidelijk lager dan bij de andere corpora (0.0128 resp. 0.0111). In dit geval is, anders dan bij de juridische corpora, wel sprake van een duidelijk lagere token/type ratio, en dus van een lagere gemiddelde herhalingsfrequentie van woordtypen. Aangezien de beide juridische corpora echter ook duidelijk verschillende waarden voor *K* te zien gaven, maar toch een vergelijkbare token/type ratio bleken te hebben, kan ook nu niet worden uitgesloten dat nog andere factoren de waarde van *K* mede bepalen, zoals bepaalde elementen uit de frequentieverdeling. Of dat het geval is zal in het volgende hoofdstuk moeten blijken.

6.4.3 De constanten R en C

Net als in de voorgaande hoofdstukken zal nu worden getracht de waarde van de twee constanten *R* en *C* voor dit corpus vast te stellen[1]. Van belang daarbij is of het verband tussen aantallen woordtokens en -typen in grote lijnen vergelijkbaar is met dat in de andere corpora. Om na te gaan of dat zo is worden de aantallen typen, ontleend aan de steekproeven in tabel 6-4, allereerst weer afgezet tegen de aantallen tokens. Dit levert, wanneer de datapunten worden verbonden door lijnen, de volgende grafiek op:



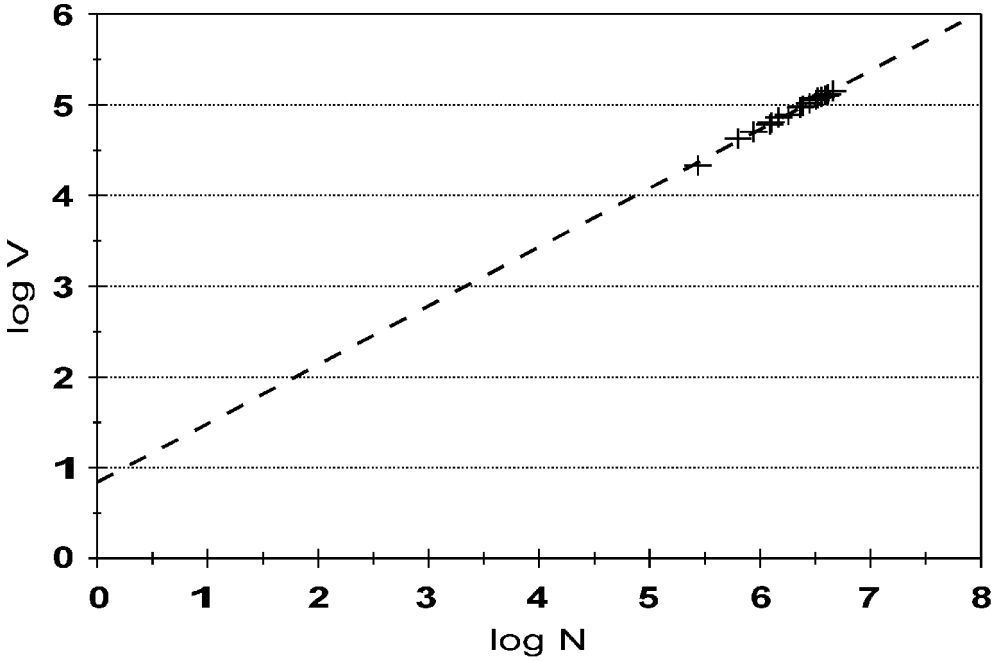
Afb. 6-9 Corpus algemeen Nederlands: het aantal woordtypen als functie van het aantal woordtokens

De 'curve' verloopt wat minder vloeiend dan bij de andere corpora, waarschijnlijk als gevolg van de geringere omvang van elk van de steekproeven. Ondanks dat kan echter worden gesteld dat de algehele vorm nog steeds, net als bij de andere corpora, doet denken aan een exponentieel verband tussen de twee grootheden. De formule $\log V = C \log N$ (om te vormen tot $V = N^c$) zou daarom ook in dit geval in beginsel in staat moeten zijn om dat verband uit te drukken. Of dat inderdaad het geval is kan weer worden nagegaan door de logaritme van het aantal typen af te zetten tegen de logaritme van het aantal tokens. Dat zou dan weer moeten resulteren in een serie datapunten die allemaal ongeveer op één rechte lijn liggen. Om dat laatste te verifiëren is tevens de lineaire regressielijn van log *V* op log *N* bepaald met behulp van de 'least squares' methode. De vergelijking van deze regressielijn luidt:

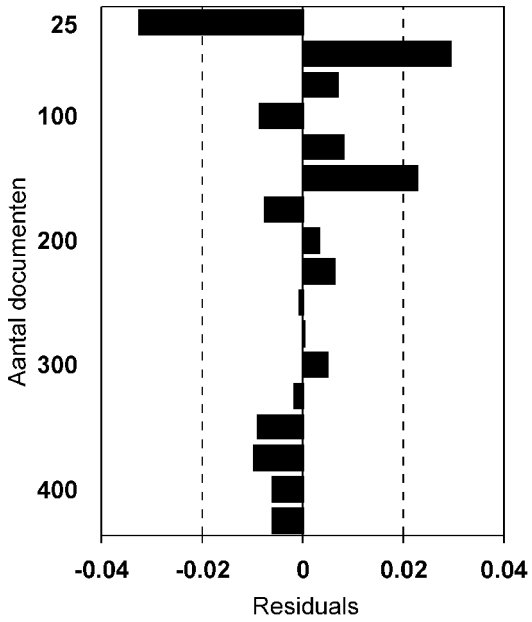
$$\log V = 0{,}6484 \log N + 0{,}8396 \qquad (6\text{-}4)$$

Deze lijn is weer in de grafiek ingetekend in de vorm van een stippellijn. Een en ander levert het volgende plaatje op (zie de volgende bladzijde):

[1] Zie voor een beschrijving van de betekenis van deze constanten hoofdstuk 4, paragraaf 4.4.3.



Afb. 6-10 De verhouding tussen log *V* en log *N* (datapunten) en de regressielijn van log *V* op log *N* (stippellijn)



Afb. 6-11 Residuals van de lineaire regressielijn van log *V* op log *N*

Omdat de datapunten dicht bij elkaar liggen is op het oog eigenlijk niet goed te bepalen of de 'fit' van de regressielijn net zo goed is als bij de andere corpora. Het allereerste datapunt (behorend bij de eerste steekproef) lijkt in ieder geval duidelijk onder de lijn te liggen, maar dat hoeft op zich nog niet veel te betekenen. Om zinvolle uitspraken over de kwaliteit van de regressielijn te doen zijn nadere gegevens erover onmisbaar. Om te beginnen weer r^2, het kwadraat van de lineaire correlatiecoëfficiënt. Deze bedraagt 0,996162, de regressielijn verklaart dus 99,6162% van de variantie. Dat is iets minder dan bij de andere corpora, maar toch nog bijzonder hoog. Van belang zijn ook de residuals, de afstand van ieder datapunt tot de regressielijn. Ook deze zijn wat hoger dan bij de andere corpora, maar komen toch nergens boven 0.04 (absoluut) uit. Aan de hand van de residuals kan de standaard fout worden berekend, deze bedraagt 0.0141 (ongeveer twee keer zo hoog als bij het corpus wetgevingsteksten). Voor de covariantie tussen de residuals geldt eigenlijk hetzelfde: met een waarde van iets onder 0,0002 is deze weliswaar wat hoger dan bij de andere corpora, maar nog steeds ruimschoots laag genoeg om vertrouwen in de berekende regressielijn te houden. Belangrijk is vooral dat de residuals net als bij de andere corpora een willekeurig patroon vertonen, dat wil zeggen, een patroon vrij van systematische afwijkingen. Wel kan worden vastgesteld dat de twee hoogste residuals te vinden zijn bij de eerste twee steekproeven. Dat was echter bij het

corpus wetgevingsteksten ook het geval, en geconcludeerd werd toen dat dit er waarschijnlijk slechts op duidt dat voor het enigermate betrouwbaar bepalen van de verhouding tussen aantallen tokens en typen steekproeven van een zekere minimum omvang nodig zijn. Dat de residuals zich na de eerste steekproeven stabiliseren op een veel lager niveau (met een enkele uitschieter bij steekproef 6) geeft alle reden om aan te nemen dat de lineaire regressielijn het verband tussen de waarden van (de logaritme van) *V* en *N* goed benadert. Ook een *t*-test met betrekking tot de waarde van de berekende coëfficiënt voor log *N* bevestigt dit nog eens. Deze levert namelijk een waarde op van ruim 62 (15 vrijheidsgraden), wat er op wijst dat de kans dat het gevonden verband op toeval berust weer praktisch nihil is.

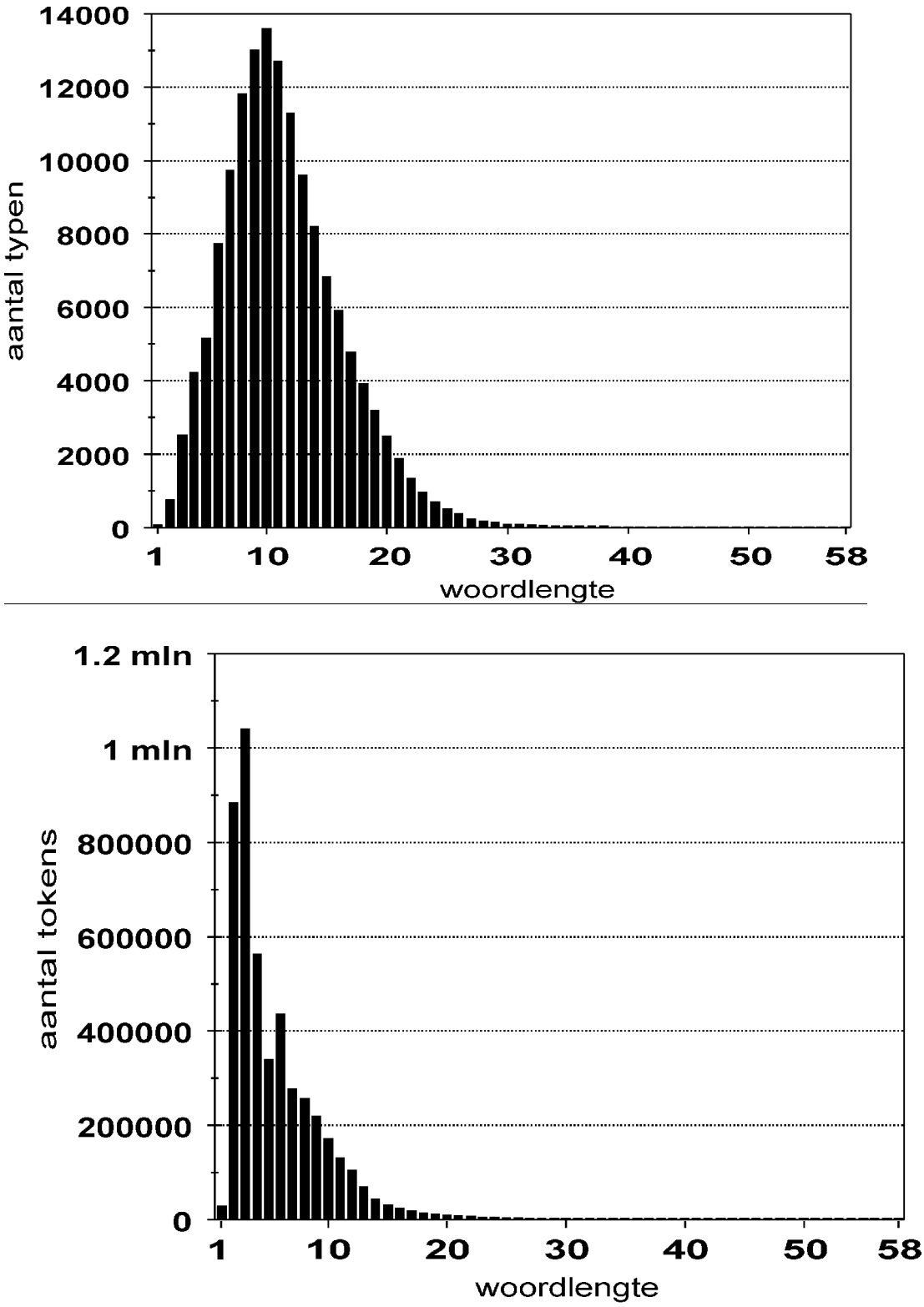
De conclusie kan zijn dat de residuals bij de regressielijn in dit corpus weliswaar wat hoger liggen dan in de andere corpora, maar dat niets erop wijst dat tussen de logaritmen van de aantallen tokens en typen geen lineair verband bestaat. Dat betekent dat we ook nu weer de twee constanten *R* en *C* op betrouwbare wijze kunnen berekenen. De waarde van *C*, **0,6484**, volgt direct uit de vergelijking van de regressielijn. *R* is gelijk aan $10^{0,8396}$ = **6,9118**. Deze waarden wijken allebei af van die in de twee andere corpora; de waarde van *C* is hoger, die van *R* lager dan bij de wet- en jurisprudentieteksten. Dat betekent net als bij de jurisprudentieteksten dat het aantal typen aanvankelijk (in kleine steekproeven) lager zal zijn dan in even grote steekproeven uit de andere corpora, maar dat dit aantal - door de hogere waarde van de exponent *C* - sterker dan bij de twee juridische corpora afhankelijk is van het aantal tokens, en dus bij toename van de steekproefgrootte sneller zal stijgen. Op deze verschillen en hun gevolgen zal in het volgende hoofdstuk nog worden teruggekomen.

Volledigheidshalve kan tenslotte nog worden vermeld dat ook voor de serie steekproeven uit dit corpus is nagegaan of de met behulp van *R* en *C* berekende waarden voor *V* overeenkomen met de werkelijkheid. De afwijkingen bleken wel wat hoger te liggen dan bij de andere corpora, maar lagen gemiddeld toch niet boven de 146 absoluut (ongeveer 0,05%). Bij steekproef 6 (150 documenten), die ook een hoge residual had, was de afwijking het hoogst: ruim 3700 typen (7,8%). Al met al toch nog wel een goed resultaat, gegeven de wat grotere spreiding van de datapunten en de wat hogere residuals dan bij de andere corpora.

6.5 Woordlengtes

6.5.1 Woordlengte verdelingen

Om de lengte van de woorden in het corpus algemeen Nederlands te kunnen analyseren zijn, net als bij de andere corpora, lengteverdelingen opgesteld voor zowel de woordtypen als de woordtokens. Woordlengte-gegevens uit de frequentielijst werden weer, voorzover het typen betrof die waren 'afgekapt' na het 30^{e} karakter, gecorrigeerd aan de hand van een speciaal daartoe opgestelde lijst met de langste woordtypen uit het corpus. De woordlengte verdelingen worden hieronder grafisch weergegeven.



Afb. 6-13a en b Verdeling van de lengtes van woordtypen en woordtokens

Wanneer we de woordlengte-verdelingen bekijken, valt direct op dat met name die van de woordtypen er heel anders uitziet dan bij de andere corpora het geval was. In plaats van een piek, gevolgd door een dal en weer een piek, vertoont deze grafiek een fraaie

regelmatige stijging naar één enkele piek bij lengte 10, gevolgd door een al even gelijkmatige daling. De klassen rond de piek (6 tot 9 en 11 tot 14) zien er zelfs praktisch volmaakt symmetrisch uit. Het is opvallend dat we deze symmetrische lengteverdeling juist aantreffen bij het kleinste corpus, waarin bijvoorbeeld onregelmatigheden als gevolg van toevallige omstandigheden in theorie talrijker zouden kunnen zijn. De mogelijkheid bestaat natuurlijk ook altijd (en is voor dit corpus ook groter dan voor de andere twee) dat juist de regelmatige vorm van de verdeling op toeval berust. Dat laatste lijkt echter niet erg waarschijnlijk, bijvoorbeeld gezien het feit dat een ander gedeelte van de grafiek, ongeveer van lengte 12 tot 30, praktisch hetzelfde is opgebouwd als de overeenkomstige delen van de grafieken van de andere corpora.

Verder kan worden opgemerkt dat de twee juridische corpora, hoe verschillend van omvang ook, voor wat betreft de vorm van de lengteverdeling van de woordtypen (zie grafieken 4-15a en 5-12a) duidelijk meer overeenkomst met *elkaar* vertonen dan met het corpus algemeen Nederlands. Met name de lengteklassen 4, 5 en 6 lijken bij deze corpora 'over-vertegenwoordigd' te zijn, wanneer we grafiek 6-12a ernaast leggen. Om na te gaan of daarvoor een oorzaak is aan te wijzen zijn voor alle drie corpora aparte lijsten gemaakt van de typen met deze lengtes, en zijn deze vervolgens vergeleken. Daaruit komt in de eerste plaats naar voren dat bij de juridische corpora een veel groter percentage woordtypen 4, 5 of 6 karakters lang is: 22,3% en 23,6% bij respectievelijk de wetteksten en de jurisprudentieteksten, tegen 11,9% bij het corpus algemeen Nederlands. Bij nauwkeurige beschouwing blijkt verder dat tussen lengte-4-tot-6-typelijsten van de juridische corpora enerzijds en het corpus algemeen Nederlands anderzijds in ieder geval één belangrijk verschil is aan te geven: de juridische corpora bevatten een veel hoger percentage *getallen*. Onder een getal wordt in dit verband verstaan: een woordtype waarvan in ieder geval het eerste karakter uit een cijfer (0 - 9) bestaat; ook 5A, 12DE en 1279BIS zijn dan dus getallen. Van alle typen van 4-6 karakters is bij de juridische corpora 59,7% resp. 49,9% een getal, en bij het corpus algemeen Nederlands niet meer dan 9,1%. Ter vergelijking: de volledige typelijsten van de corpora bevatten respectievelijk 21,7%, 14,8% en 2,1% getallen. Bij elk van de corpora zijn getallen dus sterk oververtegenwoordigd in de lengteklassen 4, 5 en 6. Een verklaring daarvoor zou met name bij de juridische corpora kunnen zijn dat veel van deze getallen nummers van wetsartikelen zijn; deze tellen in de praktijk 1 tot 4 cijfers (meer komt zelden voor), eventueel gevolgd door een of meer letters: 161BIS, 1638AA, etc. Een andere categorie vormen wellicht de nummers van artikelen uit het in 1992 ingevoerde Burgerlijk Wetboek. Deze zijn in de jaren voorafgaand aan de invoering steeds aangeduid met een combinatie van getallen, gescheiden door punten: 3.1.1.1, 6.1.7.16, etc. Dergelijke cijfercombinaties zijn in de typelijsten terug te vinden als één getal, zonder punten: 3111 en 61716.

Om te verifiëren of de getallen in de genoemde lengtecategorieën inderdaad een punt van onderscheid vormen, is nagegaan welke typen van 4 tot 6 karakters *ontbreken* in het corpus algemeen Nederlands. Voor de wetteksten gaat het daarbij om 25355 typen, waarvan 17997 stuks (70,9%) getallen zijn, en voor de jurisprudentieteksten om 52159 typen, waarvan 30141 getallen (57,8%). Zoals hierboven vermeld was van het totaal aantal typen van 4 tot 6 karakters 59,7% respectievelijk 49,9% een getal, zodat we kunnen concluderen dat, voor beide corpora, getallen oververtegenwoordigd zijn onder de typen die ontbreken in het corpus algemeen Nederlands. De meeste getallen hebben een

vrij lage frequentie, wat zou verklaren waarom de lengteverdeling van de woordtokens wel veel gelijkenis vertoont bij alle drie corpora.

De aanwezigheid van meer getallen in de juridische corpora zou de geconstateerde verschillen in de woordlengte-verdelingen van de corpora dus (wellicht gedeeltelijk) kunnen verklaren. Het is natuurlijk mogelijk dat er daarnaast nog andere dingen aan de hand zijn. Het is bijvoorbeeld heel goed mogelijk dat in juridische teksten vaker een wat langer woord wordt gebruikt op plaatsen waar in een gewone tekst een kort woord zou staan: *indien* in plaats van *als*, *welke* in plaats van *die*, *bedoelde* in plaats van *deze* of *die*, *teneinde* in plaats van *om*, enzovoort[2]. Dit 'inruilen van woordtypen' zou ertoe kunnen leiden dat sommige lengtecategorieën voor de woordtypen bij juridische teksten wat oververtegenwoordigd raken, terwijl de lengteverdeling van de woord*tokens* (zie hieronder) weinig verschil tussen de corpora te zien geeft[3]. Het probleem met dergelijke verschuivingen in het woordgebruik is echter dat zij in de meeste gevallen niet goed kunnen worden aangetoond louter aan de hand van kwantitatieve verschillen tussen de corpora. Er is daarnaast ook een inhoudelijke analyse van de tekst nodig, waaruit blijkt welke woordtypen anders worden gebruikt, en welke gevolgen dat verder nog kan hebben, bijvoorbeeld op semantisch niveau. Een dergelijke analyse ligt echter buiten de reikwijdte van dit onderzoek, waardoor dit soort verklaringen voor de verschillen in woordlengte-verdelingen helaas op deze plaats niet kunnen worden geverifieerd. Met betrekking tot het grotere percentage getallen in de juridische corpora moet verder nog worden opgemerkt dat nog wel moet worden nagegaan (in hoofdstuk 7) in hoeverre de verschillen in corpusgrootte daarop nog van invloed zijn.

Met betrekking tot de lengteverdeling van woordtokens valt, zoals al opgemerkt, weinig verschil waar te nemen tussen dit corpus en de twee andere. Lengte 3 komt wederom het meeste voor, bij lengte 5 vertoont de verdeling weer een inzinking. De in de voorgaande hoofdstukken genoemde verklaring daarvoor - lengte 5 is net de grens tussen woorden met een en met twee lettergrepen - is daarom ook hier waarschijnlijk van toepassing.

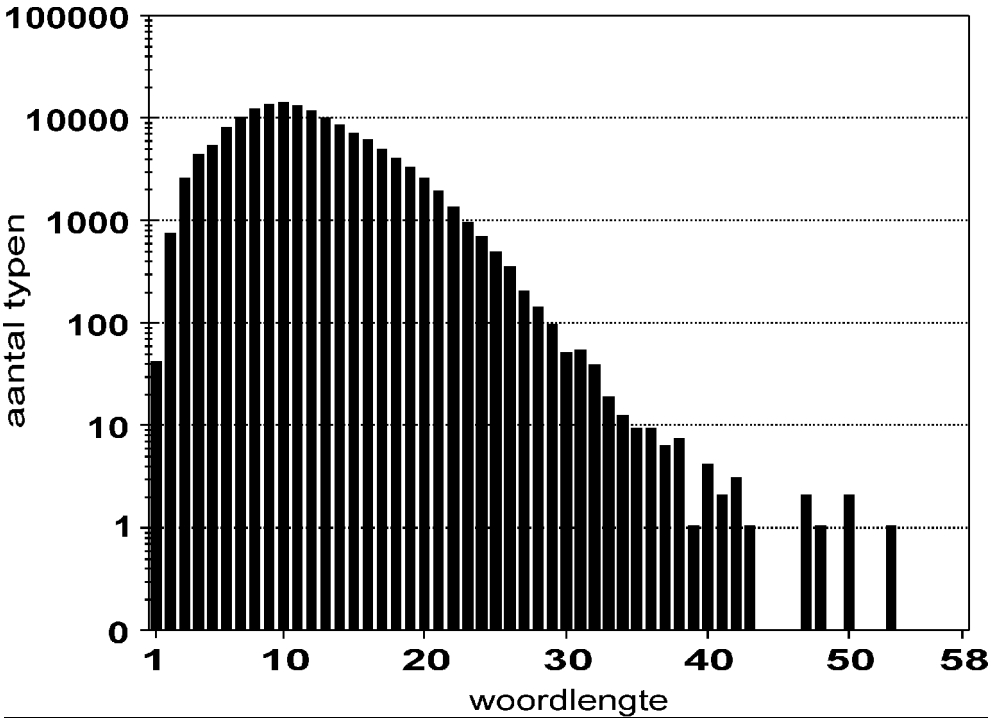
Hieronder zijn beide woordlengte-verdelingen nogmaals opgenomen, maar nu met logaritmische schaalverdeling op de y-as. Daardoor wordt ook de verdeling van de langste woorden zichtbaar. Afbeelding 6-13a (zie de volgende bladzijde) toont een zeer gelijkmatige lengte-verdeling voor de woordtypen (de symmetrische opbouw rond lengte 10 is nog goed zichtbaar), die voor de langere woordtypen nauwelijks een ander beeld laat zien dan bij de andere corpora. Wel wordt duidelijk dat dit corpus minder zeer lange typen (lengte > 45 karakters) bevat dan de andere twee (6, tegen 11 bij de jurisprudentieteksten en 22 bij de wetteksten). Er ontstaan daardoor meer lege lengte-klassen, terwijl de niet-lege klassen steeds maar een of twee typen bevatten. Verder valt nog op dat lengte 39 in dit corpus duidelijk ondervertegenwoordigd is, met slechts een enkel type. Een verklaring

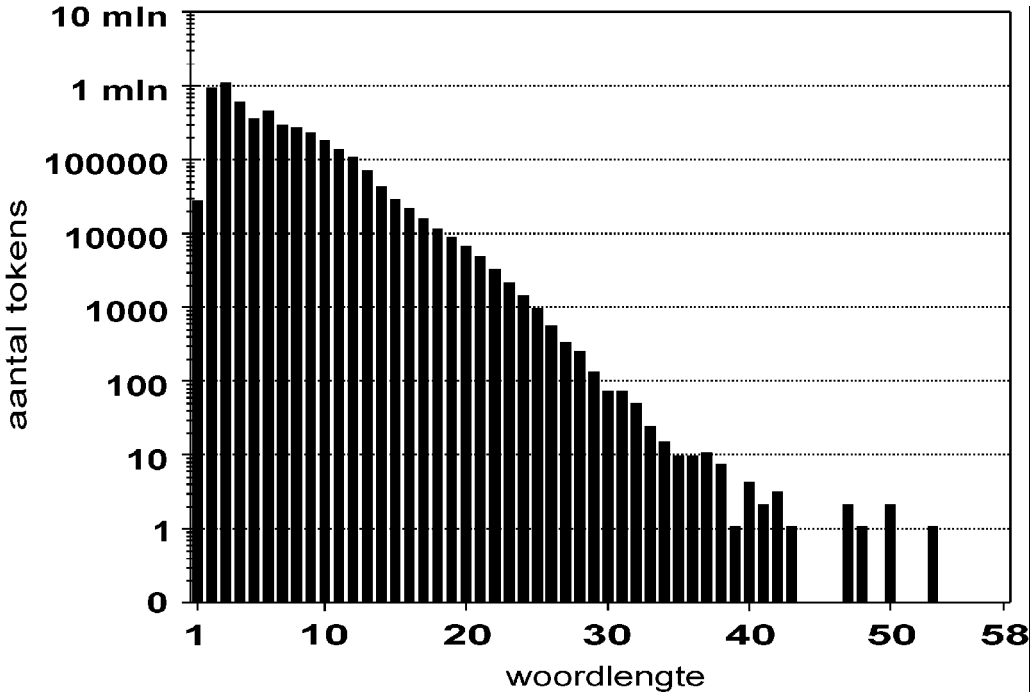
---

[2] Deze neiging om, als er van een woord twee synoniemen bestaan, altijd het zwaarste en meest omslachtige te kiezen wordt bijvoorbeeld genoemd door Reinsma en Reinsma (1976, p. 870). Ook Van Eemeren *et al.* (1991, p. 169) geven voorbeelden van, wat zij noemen, plechtstatige en verouderde woorden in rechterlijke uitspraken, waarvoor in de meeste gevallen een of enkele kortere vervangingswoorden zouden volstaan.

[3] Voorwaarde daarbij is natuurlijk wel dat niet te veel zeer frequente woordtypen (zoals lidwoorden en veel van de voorzetsels) worden 'ingeruild' voor iets langers; dat zou ook in de verdeling van de woordtokens zichtbaar zijn.

daarvoor is eigenlijk niet te geven.





Afb. 6-13a en b Verdeling van de lengtes van woordtypen en woordtokens; aantallen per lengte-cat. logaritmisch

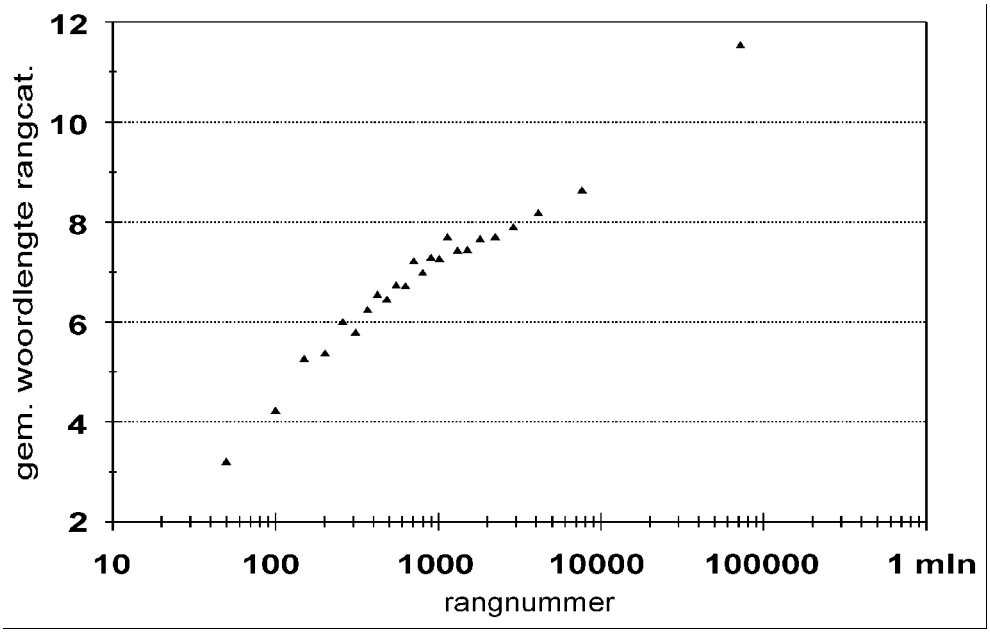
Van de verdeling van de tokens kan gezegd worden dat deze een bijzonder gelijkmatig aflopende lijn vertoont. Het verloop bij de hoogste lengte-klassen lijkt veel op dat bij het corpus jurisprudentieteksten: wat onregelmatig, en vrij veel lege klassen, waarschijnlijk

doordat het (vanaf lengte 38) alleen nog maar om unieke typen gaat
(bij de wetsteksten was het aantal unieke typen wat lager).

Al met al kan worden geconstateerd dat de logaritmische lengteverdelingen in dit geval weinig nieuwe inzichten scheppen. Dat geldt echter niet voor de gewone lengteverdelingen, met name niet voor de lengteverdeling van de woordtypen. Deze vertoont bij dit corpus een zeer gelijkmatige en symmetrische opbouw, duidelijk verschillend van die bij de twee juridische corpora. Een mogelijke verklaring voor dit geconstateerde verschil werd hierboven gegeven. Wel zal in het volgende hoofdstuk nog moeten worden gecontroleerd of de verschillen in omvang tussen de corpora hierop nog van invloed zijn.

6.5.2 De relatie tussen het rangnummer en de lengte van woordtypen

Om de relatie tussen het rangnummer en de lengte van woordtypen te onderzoeken zal voor dit corpus (net als bij de jurisprudentieteksten) gebruik worden gemaakt van de 'cluster-methode', waarbij telkens de gemiddelde woordlengte wordt bepaald in opeenvolgende clusters van 50 rangnummers, en van de methode van het voortschrijdend gemiddelde. Deze twee methoden bleken bij de andere corpora het best geschikt om het bedoelde verband in beeld te brengen. Begonnen zal weer worden met de clustermethode. Daarbij wordt de frequentielijst van het corpus verdeeld in in totaal 24 clusters van 50 rangnummers, en wordt de gemiddelde lengte bepaald van alle woordtokens in ieder cluster. Deze gemiddelde lengte varieert van 3,22 karakters in het eerste tot 11,55 karakters in het laatste cluster.



Afb. 6-14 Gemiddelde woordlengte per categorie van 50 rangnummers

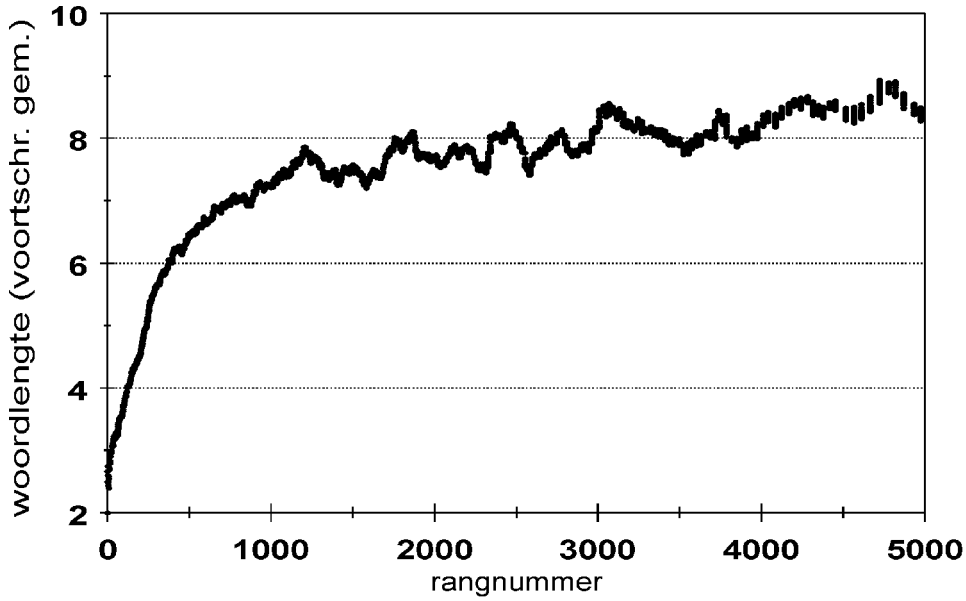
De berekende gemiddelde lengtes zijn weer in een grafiek afgezet tegen het hoogste in elk cluster voorkomende rangnummer. Voor de x-as is opnieuw gebruik gemaakt van een logaritmische schaalverdeling, die compenseert voor de ongelijke spreiding van de rangnummers: aan het begin van de frequentielijst liggen deze dicht opeen, en omvatten

slechts een of enkele tokens, aan het einde (bij de lage frequenties) liggen ze ver uiteen, en omvatten een groot aantal woordtokens.

Het eerste dat opvalt met betrekking tot grafiek 6-14 is dat het aantal datapunten duidelijk geringer is dan bij de andere twee corpora. Dat is natuurlijk een gevolg van de geringere omvang, en dus het geringere aantal rangnummers, van dit corpus. De datapunten die er zijn vertonen verder een opvallend geringe mate van spreiding, ze vormen samen een vrij duidelijke kromme (de gemiddelde woordlengte neemt steeds minder snel toe). Eén datapunt vormt daarop een uitzondering: het hoogste, behorend bij de laatste 'rangcategorie'. De gemiddelde woordlengte in dit cluster bedraagt 11,55, hoger dan de gemiddelde lengte in elk van de clusters van de twee andere corpora. De stijging ten opzichte van de clusters daarvoor is onevenredig hoog. Een dergelijke sterke stijging kan niet alleen het gevolg zijn van het feit dat de langste woordtypen (van bijvoorbeeld meer dan 45 karakters) allemaal in dit cluster te vinden zijn; daarvoor zijn die typen te gering in aantal, terwijl ze allemaal ook maar één keer voorkomen (een gering aantal tokens dus). Er moet daarnaast sprake zijn van een hoog aantal woordtokens van 'gemiddelde' lengte (20 tot 30 karakters), die samen het gemiddelde op het waargenomen hoge peil brengen. Tenslotte kan worden opgemerkt dat grafiek 6-14, anders dan bij de andere corpora het geval was, eigenlijk geen opvallende 'sprongen' of 'pieken' vertoont. Het zou dus kunnen zijn dat de onregelmatigheden in de grafieken van de juridische corpora inderdaad een voor dat soort teksten kenmerkend patroon vormen. Daaraan zal in hoofdstuk 7 in wat meer detail aandacht worden geschonken.

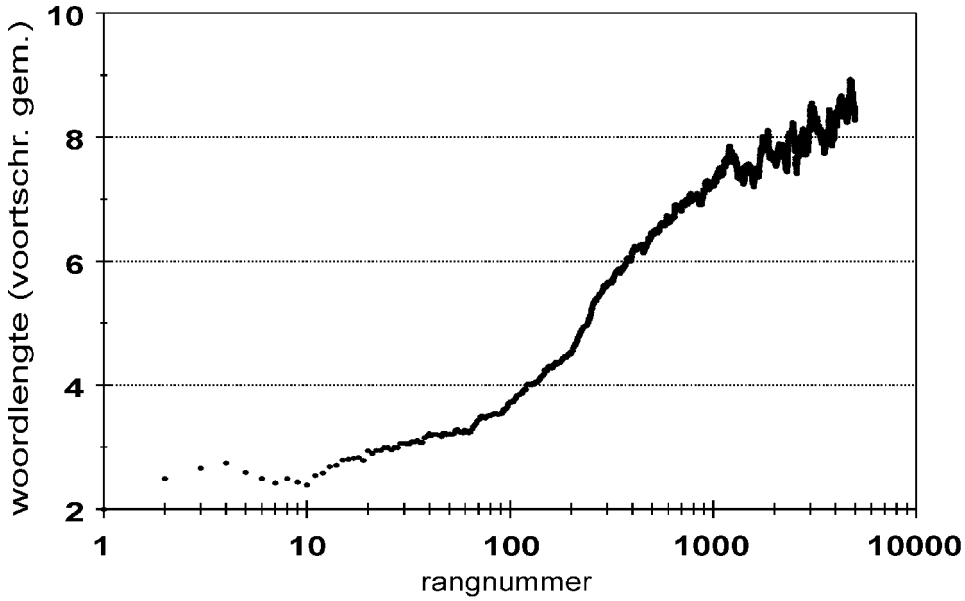
De tweede techniek waarmee het verband tussen rangnummer en woordlengte zal worden bestudeerd is weer de analyse van het voortschrijdend gemiddelde van de woordlengte. Daarbij geldt weer als beperking dat alleen de eerste 5000 rangnummers worden bestudeerd, gedeeltelijk om technische redenen (het aantal datapunten zou te hoog worden) en gedeeltelijk omdat de gemiddelde woordlengte in dat deel van de frequentielijst de sterkste ontwikkeling doormaakt.

De betreffende grafiek is te vinden in afbeelding 6-15. Ieder datapunt in deze grafiek vertegenwoordigt weer de gemiddelde lengte van een woordtype en van de 199 typen die daaraan vooraf gaan. Voor beide assen is in eerste instantie weer een normale schaalverdeling gebruikt. Bij globale vergelijking met de andere corpora (grafieken 4-20 en 5-15) kan worden geconstateerd dat het patroon dat de stijging van de woordlengte te zien geeft voor dit corpus vrij veel weg heeft van dat bij het corpus jurisprudentieteksten, alhoewel de snelle stijging tot ongeveer rangnummer 250 weer meer doet denken aan het overeenkomstige gedeelte van de grafiek van de wetgevingsteksten. Hoewel de drie grafieken ook een aantal verschillen vertonen, is het voor de bestudering daarvan een gedetailleerde analyse nodig. Deze is te vinden in het volgende hoofdstuk, waar tevens maatregelen zijn getroffen om te voorkomen dat ook het verschil in corpusgrootte weer invloed uitoefent.



Afb. 6-15 Voortschrijdend gemiddelde van de lengte van woordtypen

Om voor het corpus algemeen Nederlands met name het eerste deel van de grafiek, ongeveer tot rangnummer 1000, in meer detail te kunnen bestuderen kan op de x-as weer een logaritmische schaalverdeling worden gebruikt. De grafiek ziet er dan als volgt uit:



Afb. 6-16 Voortschrijdend gemiddelde van de lengte van woordtypen, logaritmische schaalverdeling voor rangnummers

Ook deze grafiek heeft zowel aspecten die aan de overeenkomstige grafiek van het corpus jurisprudentieteksten doen denken, als aspecten die doen denken aan de grafiek van het corpus wetgevingsteksten. De algemene vorm komt weer het meeste overeen met die bij de jurisprudentieteksten, maar de mate van stijging van rangnummer 1 tot 100 en van rangnummer 500 tot 1000 (ongeveer) lijkt meer op de situatie bij de wetgevingsteksten.

Het aantal onregelmatigheden dat de 'curve' bevat is bij dit corpus het kleinst van allemaal. Er is ook eigenlijk nergens sprake van een 'knik' in het verloop, wat bij de andere corpora wel het geval was (zij het op verschillende plaatsen).

Samenvattend kan gesteld worden dat het verband tussen rangnummer en gemiddelde woordlengte voor het corpus algemeen Nederlands bepaalde overeenkomsten lijkt te vertonen met dat bij de twee andere corpora. De overeenkomsten met het corpus jurisprudentieteksten lijken echter wat groter dan die met het corpus wetgevingsteksten, vooral wanneer we kijken naar het voortschrijdend gemiddelde van de woordlengte. In het volgende hoofdstuk zal hierop nog worden teruggekomen.

6.5.3 Zeer lange woordtypen

Men zou misschien in eerste instantie verwachten dat het percentage lange woordtypen in een corpus met algemeen Nederlandse teksten aanzienlijk lager ligt dan bij juridische tekstcorpora. In dat licht is het opvallend dat het percentage woordtypen langer dan 20 karakters voor dit corpus 4,32% bedraagt, zelf nog iets hoger dan dat bij de jurisprudentieteksten (4,22%) en maar weinig onder het percentage bij de wetteksten (4,53%). De lange woorden vertegenwoordigen 0,29% van het corpus, duidelijk meer dan bij de jurisprudentieteksten (0,16%) en maar iets minder dan bij de wetgevingsteksten (0,32%). De gemiddelde relatieve frequentie per type bedraagt 0,29% / 4,32% = 6,68%, een waarde die ook weer boven die bij de jurisprudentieteksten (3,85%) en dicht bij die bij de wetteksten (7,1%) ligt. Ook wanneer we ons beperken tot uitsluitend typen van meer dan 30 karakters verandert deze situatie maar weinig: weliswaar daalt het percentage typen dan tot 0,12% (juridische corpora: 0,24% resp. 0,16%), maar het relatieve aantal tokens blijft met 0,004% hoger dan bij de jurisprudentieteksten (0,003%), alhoewel niet zo hoog als bij de wetteksten (0,02%). Ook de gemiddelde relatieve frequentie heeft dan een waarde die ergens tussen die bij de andere corpora inligt: 3,78%, tegen 8,46% resp. 1,73%. Het op het eerste gezicht wat merkwaardige feit doet zich dus voor dat wanneer we vergelijken op basis van relatieve cijfers, het corpus algemeen Nederlands voor wat betreft de lange woorden een middenpositie tussen beide juridische corpora inneemt. Ook dit punt zal echter in het volgende hoofdstuk worden geverifieerd aan de hand van qua omvang vergelijkbaar gemaakte corpora. Het is namelijk zo dat relatieve frequentie-cijfers nogal eens problemen geven, met name bij de allerlaagste frequenties. Een frequentie van 1 in een klein corpus levert immers een hoger percentage op dan dezelfde frequentie in een groter corpus. Op dit probleem zal aan het begin van het volgende hoofdstuk worden terugkomen.

De lijst met zeer lange woordtypen in bijlage 2 weerspiegelt goed het algemene karakter van dit corpus, de meeste typen hebben inderdaad een 'algemeen' karakter. Zo treffen we woorden aan als 'projectontwikkelingsmaatschappijen' en 'samenlevingsverantwoordelijkheid', die in allerlei teksttypen zouden kunnen voorkomen. Enkele typen zijn in zekere mate 'technisch', zoals 'haemagglutinatieremmingsreactie', 'konstituentenstructuurgrammatika', en misschien ook 'huwelijksvruchtbaarheidscijfers' en 'marktverzadigingsverschijnselen'. Er zijn ook enkele typen met een zekere juridische connotatie, zoals 'aansprakelijkheidsverzekeringen', 'arbeidsongeschiktheidsuitkering' en 'vennootschapsbelastingplichtige'. Een vreemde eend in de bijt vormt 'neemdanwatdruppelsvoordezenuwenvandereform',

een type dat kennelijk is gevormd doordat in een tekst de samenstellende woorden waren verbonden door streepjes of iets dergelijks. Het voorkomen van dergelijke typen is dus een rechtstreeks gevolg van de gekozen techniek om woorden uit een tekst te isoleren. Nogmaals kan naar voren worden gebracht dat deze lange woorden slechts een zeer geringe fractie vormen van het totale aantal woordtokens en -typen in de corpora, waardoor ze bij veel kwantitatieve analyses nauwelijks een rol spelen.

HOOFDSTUK 7 - EEN VERGELIJKING VAN KWANTITATIEVE EIGENSCHAPPEN VAN DE DRIE CORPORA

7.1 Inleiding

Nadat in de voorgaande drie hoofdstukken elk van de corpora afzonderlijk werd bestudeerd, waarbij slechts in beperkte mate vergelijking met de andere corpora plaatsvond, ligt op dat aspect in dit zevende hoofdstuk juist de nadruk. De analyses die in dit hoofdstuk worden beschreven en de conclusies die aan de hand daarvan kunnen worden getrokken kunnen dan ook gerekend worden tot de kern van dit onderzoek naar juridisch woordgebruik.

Alvorens over te gaan tot de beschrijving van gedetailleerde vergelijkingen tussen de corpora zal in paragraaf 7.2 allereerst opnieuw een probleem aan de orde worden gesteld dat ook in de voorgaande hoofdstukken diverse malen ter sprake kwam, namelijk het verschil in afmeting tussen de drie corpora en de gevolgen die dit kan hebben voor het vergelijken van kwantitatieve eigenschappen. Besproken zal worden hoe kan worden voorkomen dat dit verschil in corpusgrootte invloed heeft op resultaten van metingen. Uiteindelijk wordt daartoe uit de twee grootste corpora een aselecte steekproef genomen van dezelfde afmetingen als het kleinste corpus. Nadat op deze wijze als het ware nieuwe corpora met gereduceerde afmetingen zijn samengesteld, is het natuurlijk wel zaak om te controleren of bepaalde constanten c.q. maatgetallen waarvan de waarde *niet* afhankelijk zou moeten zijn van de corpusgrootte, in de gereduceerde corpora ook inderdaad dezelfde waarde hebben als in de volledige. Bevindingen op dit gebied staan vermeld in paragraaf 7.3.

In paragraaf 7.4 zullen vervolgens, aan de hand van de gereduceerde corpora, die kwantitatieve eigenschappen worden vergeleken waarop de corpusgrootte mogelijk wel van invloed is. Het betreft bijvoorbeeld bepaalde aspecten van de frequentieverdeling en van token-type ratio's, en zaken die samenhangen met de lengte van woorden. Bij dit vergelijkingsproces zal vanzelfsprekend ook gebruik worden gemaakt van gegevens die in de voorgaande hoofdstukken werden gepresenteerd.

Aan de woordtypen met de hoogste rangnummers zal apart aandacht worden geschonken in paragraaf 7.5. Al in hoofdstuk 4 werd immers duidelijk dat in de frequentielijst van de corpora door bestudering van de relatie tussen rangnummer en frequentie een 'kop' kan worden onderscheiden die bij ieder corpus ruwweg de eerste 150 rangnummers blijkt te omvatten. Het woordgebruik in dit deel van de drie frequentielijsten is hier in detail vergeleken. Daarbij is in de eerste plaats een gecombineerde lijst van woordtypen samengesteld, waarbij voor elk type wordt vermeld in welke corpora het voorkomt en hoe vaak. Daarnaast zijn ook lijsten vervaardigd gebaseerd op de verschillen tussen de typelijsten van de corpora. Het hoofdstuk wordt afgesloten met een aantal conclusies ten aanzien van de overeenkomsten en verschillen in de bestudeerde aspecten van het woordgebruik in de drie corpora.

7.2 Het compenseren van de verschillen in afmeting tussen de corpora

Tussen de drie corpora die in dit onderzoek centraal staan bestaan aanzienlijke verschillen in afmetingen. De verklaring daarvoor is dat bij de twee juridische corpora uitgegaan is van een zo ruim mogelijke opzet, waarbij al het beschikbare tekstmateriaal in de betreffende categorie (wetgeving resp. jurisprudentie) werd opgenomen. Deze corpora bevatten daarom ook de teksten van *alle* wet- en regelgeving die op het moment dat de corpora werden samengesteld in Nederland in werking was respectievelijk *alle* op dat moment in elektronische vorm beschikbare jurisprudentie uit een bepaalde bron (het tijdschrift Nederlandse Jurisprudentie). Bij het vergelijkingscorpus met algemeen Nederlandse teksten speelden vooral praktische overwegingen een rol. De samenstelling van dit corpus werd voornamelijk overgelaten aan het Instituut voor Nederlandse Lexicologie, waarbij echter wel een beoordeling van de inhoud plaatsvond, zie hoofdstuk 3. Het INL beschikte over tekstmateriaal in elektronische vorm met de gewenste eigenschappen (een breed spectrum aan bronnen en onderwerpen, teksten die grondig waren nagelezen en gecorrigeerd), maar de omvang daarvan was beperkt tot ongeveer 4,6 miljoen woordtokens. De drie corpora die aldus ontstonden liepen in afmetingen uiteen van 40 miljoen woordtokens bij de jurisprudentieteksten tot ruim 12 miljoen woordtokens bij de wetgevingsteksten en 4,6 miljoen woordtokens bij de algemeen Nederlandse teksten.

Deze verschillen in afmetingen bleken in de voorgaande hoofdstukken al een belemmering te vormen bij het vergelijken van diverse aspecten van het woordgebruik, zoals bijvoorbeeld de verhouding tussen rangnummer en frequentie en tussen aantallen woordtokens en woordtypen, en de lengte van woordtokens en -typen. In dit hoofdstuk, waarin het vergelijken van de corpora centraal staat, dient er daarom in de eerste plaats zorg voor worden gedragen dat de afmetingen van de corpora geen invloed meer hebben op metingen. Twee voor de hand liggende manieren om daarvoor te zorgen zijn:

1. uitsluitend gebruik maken van relatieve waarden. In het geval van de woordfrequentie zou dat bijvoorbeeld inhouden dat uitsluitend wordt gewerkt met de relatieve frequentie, berekend door de absolute frequentie te delen door het aantal woordtokens in het corpus.
2. de twee langste corpora inkorten door daaruit een steekproef te nemen. Deze steekproef zou dan dezelfde afmetingen moeten hebben als het derde en tevens kleinste corpus.

Aan beide methoden zijn voor- en nadelen verbonden. Deze worden beschreven in de volgende twee subparagrafen.

7.2.1 Het werken met relatieve waarden

Wanneer gewerkt wordt met datasets van verschillende afmetingen, waarvan bepaalde eigenschappen moeten worden vergeleken, is in het algemeen de eenvoudigste oplossing om geen gebruik te maken van absolute maar van relatieve waarden. Relatief wil in dit verband zeggen: gerelateerd aan de afmetingen van de dataset. In het geval van tekstcorpora zou dat bijvoorbeeld betekenen dat gewerkt wordt met relatieve frequenties, berekend door de absolute frequentie te delen door het aantal woordtokens in het corpus

en het resultaat eventueel te vermenigvuldigen met 100%. Voor corpora waarvan de afmetingen niet al te veel verschillen kan dit goede resultaten geven. Wanneer de afmetingen wel sterk uiteenlopen, zoals bij de corpora uit dit onderzoek, kunnen echter problemen optreden. Een reden daarvoor is dat de variabelen waar het om gaat (frequentie, woordlengte, etc.) geen zogenaamde *continue*, maar *discrete* variabelen zijn. Het verschil tussen deze twee is kort gezegd dat de mogelijke waarden van continue variabelen een *continuüm* vormen, voor elke twee waarden is het mogelijk een derde waarde te vinden die daar precies tussenin ligt[1]. Bij discrete variabelen is dat echter niet het geval. Zo is het bijvoorbeeld niet denkbaar dat een corpus 5½ keer een bepaald woordtype bevat; het woordtype komt 5 *of* 6 keer voor, een tussenweg is niet mogelijk. Een gevolg daarvan wordt bijvoorbeeld zichtbaar wanneer we kijken naar de *laagste woordfrequentie* die voorkomt in een corpus. Een woord komt altijd tenminste 1 keer voor in een corpus, 1 is dus de laagste frequentie die we in elk van de corpora aantreffen. Wanneer we deze laagste *absolute* frequentie delen door het aantal tokens in de respectievelijke corpora, ontstaan echter laagste *relatieve* frequenties die sterk uiteenlopen:

Corpus	Laagste absolute frequentie	Aantal tokens	Laagste relatieve frequentie
Wetgeving	1	12.807.659	0,0000078%
Jurisprudentie	1	40.190.212	0,0000025%
Algemeen Nederlands	1	4.616.455	0,0000216%

Tabel 7-1 Laagste absolute en relatieve frequenties

De afmetingen van de corpora weerspiegelen zich dus (onder andere) duidelijk in de laagste relatieve frequenties. Dit heeft bijvoorbeeld tot gevolg dat een grafiek die het verband tussen rangnummer en relatieve frequentie toont voor de grootste corpora verder doorloopt. Er kan zo een situatie ontstaan waarin vergelijking nog steeds lastig is.

Een ander nadeel van het gebruik van relatieve cijfers is dat deze geen oplossing bieden wanneer tussen bepaalde eigenschappen van de corpora systematische verschillen bestaan die zijn terug te voeren op het verschil in omvang. Een voorbeeld hiervan is de verhouding tussen de aantallen woordtokens en -typen. Wanneer deze verhouding wordt berekend over elk van de (volledige) corpora ontstaan waarden die behoorlijk uiteen lopen, zoals in de voorgaande hoofdstukken duidelijk werd. Deze verschillen blijven bestaan, ook wanneer gebruik wordt gemaakt van relatieve frequenties, waardoor het vergelijken van eigenschappen van de corpora wordt bemoeilijkt.

7.2.2 Datareductie door het trekken van steekproeven

Een andere manier om datasets van verschillende afmetingen goed vergelijkbaar te maken is het trekken van aselecte steekproeven van gelijke afmetingen uit de datasets. In dit

[1] De Mulder (1984, p. 32).

geval zou de meest praktische uitvoering daarvan zijn het trekken van steekproeven uit de twee juridische corpora, waarbij de afmetingen van iedere steekproef zo dicht mogelijk de afmetingen van het corpus algemeen Nederlands benaderen. Het meest in het oog springende voordeel van deze methode is waarschijnlijk dat na het trekken van de steekproeven alle gegevens, zowel absolute als relatieve, met betrekking tot ieder corpus direct met die van de andere vergelijkbaar zijn. Het belangrijkste nadeel bij het gebruik van relatieve cijfers, namelijk dat de afmetingen van het corpus nog steeds enigszins een rol blijven spelen, treden bij deze methode van datareductie niet op.

Toch heeft ook deze methode enkele nadelen. In de eerste plaats introduceert het trekken van steekproeven *fouten*. Een deel van de data in de volledige corpora wordt immers genegeerd. Daarom is zorgvuldigheid vereist, onder andere waar het betreft het kiezen van een *methode* om de steekproef samen te stellen. Van belang is daarbij bijvoorbeeld of in de corpora een bepaald patroon aanwezig is (dat zou bijvoorbeeld het geval zijn wanneer alle documenten over een bepaald onderwerp en met een specifiek woordgebruik bij elkaar zouden staan, aan het begin of het einde van het corpus), waarmee bij het trekken van de steekproef rekening moet worden gehouden[2]. Voorzover bekend is dat bij de twee juridische corpora echter niet het geval. Wel is de volgorde van documenten in het corpus jurisprudentieteksten grotendeels chronologisch, maar dat hoeft geen probleem te zijn. Aangezien een specifiek patroon waarmee rekening moet worden gehouden ontbreekt vormt het trekken van een willekeurige steekproef, bestaande uit een aantal willekeurige documenten uit elk van de corpora, in dit geval waarschijnlijk de beste mogelijkheid.

Een ander, meer praktisch nadeel van het gebruik van steekproeven is dat daarbij in feite sprake is van afzonderlijke, nieuwe corpora die apart moeten worden opgeslagen en verwerkt. Het is daarbij onder andere nodig om nieuwe frequentieverdelingen en woordlengte-verdelingen te maken. Ondanks deze nadelen vormt het trekken van steekproeven uit de juridische corpora in dit geval de te prefereren methode om de corpora optimaal vergelijkbaar te maken. Het alternatief, uitsluitend gebruik maken van relatieve cijfers, levert hier te veel problemen op, mede door het grote verschil in omvang tussen de corpora. Besloten is daarom om voor de methode van steekproeven te kiezen. In de rest van deze paragraaf zal een korte beschrijving worden gegeven van de wijze waarop deze steekproeven precies tot stand zijn gekomen en zal een opsomming worden gegeven van hun belangrijkste eigenschappen.

Het samenstellen van steekproeven uit de corpora komt in feite neer op het vervaardigen van nieuwe frequentielijsten, niet gebaseerd op *alle* documenten in het corpus, maar op een *selectie* daaruit. In principe kan daarbij dezelfde procedure worden gevolgd als bij de volledige corpora, alleen moeten nu niet alle documenten systematisch worden verwerkt, maar moet een keuze worden gemaakt. Aangezien het hier gaat om een willekeurige steekproef, moeten telkens *willekeurige* documenten worden geselecteerd. Elk document mag daarbij slechts één keer worden verwerkt, er is dus sprake van trekkingen *zonder teruglegging*. Voor dit doel werd het computerprogramma voor het maken van frequentielijsten uitgebreid met een eenvoudige routine die telkens met behulp van een 'random

[2] Zie hierover bijvoorbeeld Ehrenberg (1982, p. 93-107).

number generator' een willekeurig documentnummer genereert, waarbij in dezelfde 'run' nooit twee maal hetzelfde nummer kan voorkomen. Door gebruik te maken van deze routine worden de documenten uit een corpus dus in willekeurige volgorde verwerkt. Om ervoor te zorgen dat de steekproef de juiste afmetingen krijgt, is tevens een eindvoorwaarde ingebouwd die ervoor zorgt dat het programma stopt zodra na het verwerken van een document het aantal tokens in de steekproef het gewenste aantal heeft bereikt. Aangezien alleen gewerkt wordt met *volledige documenten* zal de steekproef die resulteert uiteindelijk meestal iets groter zijn dan de opgegeven grenswaarde. Dat zal in de praktijk naar verwachting geen problemen opleveren, behalve wanneer het zou gaan om een zeer kleine steekproef.

Nadat op deze wijze nieuwe woordenlijsten waren samengesteld, werden deze gesorteerd op woordfrequentie. Ook werden net als bij de volledige corpora enkele statistische gegevens met betrekking tot (in dit geval) de steekproef en de documenten die daarvan deel uitmaken opgeslagen. Deze worden weergegeven in tabel 7-2. Ter vergelijking worden tevens de gegevens van het corpus algemeen Nederlands vermeld.

	Wetgeving (gereduceerd)	Jurisprudentie (gereduceerd)	Algemeen Nederlands
Corpus:			
Grootte in bytes	29296096	28172050	30889416
Totaal aantal tokens	4619141	4617676	4616455
Totaal aantal typen	78574	74827	142879
Aantal documenten	6780	1992	406
Grootste document:			
Grootte in bytes	127860	172235	656836
Aantal tokens	20820	27736	99194
Aantal typen	1763	2979	10215
Kleinste document:			
Grootte in bytes	43	63	3179
Aantal tokens	7	11	502
Aantal typen	6	10	215
Gemiddeld:			
Grootte in bytes	4320,96	14142,60	76082,31
Aantal tokens	681,29	2318,11	11370,58
Aantal typen	210,36	581,64	2020,66
Standaarddeviatie:			
Grootte in bytes	4556,47	2504,75	115955,98
Aantal tokens	898,05	2428,96	17631,14
Aantal typen	164,13	396,24	1869,37

Tabel 7-2 Statistische gegevens gereduceerde corpora

Uit deze statistische gegevens blijkt dat de beide steekproeven in grote lijnen dezelfde karakteristieken vertonen als de volledige corpora. Zo zijn de gemiddelde documentgrootte in tokens voor de wetgevingsteksten vrijwel identiek (681,29, tegen 681,15 voor het volledige corpus) en liggen die voor de jurisprudentieteksten zeer dicht bij elkaar (2318,11 tegen 2446,15). Hetzelfde geldt voor het gemiddeld aantal woordtypen. Enig

verschil is wel te constateren bij de gegevens met betrekking tot de grootste en kleinste documenten: de volledige corpora bevatten, niet geheel onlogisch, wat grotere respectievelijk kleinere documenten dan beide steekproeven. Het aantal tokens in de gehele steekproef is in beide gevallen iets hoger (2689 respectievelijk 1221 stuks) dan in het corpus algemeen Nederlands (doordat ze bestaan uit gehele documenten), maar deze verschillen zijn te verwaarlozen op de totale omvang van ruim 4,5 miljoen tokens. Opvallend is wel dat beide gereduceerde corpora slechts ongeveer half zoveel woord*typen* bevatten als het corpus algemeen Nederlands. Dit verschil zal onder andere voorwerp van studie zijn in paragraaf 7.4. Ten aanzien van de twee steekproeven kan worden gesteld dat deze in ieder geval voor wat betreft hun samenstelling een goede weergave te vormen van de desbetreffende corpora. De vraag is wel of hetzelfde ook gezegd kan worden met betrekking tot de structuur van het woordgebruik. Om daarin enig inzicht te krijgen zal in de volgende paragraaf worden nagegaan of enkele 'linguïstische constanten' die in de voorgaande drie hoofdstukken aan de orde kwamen voor beide steekproeven evenzeer gelden als voor de volledige corpora.

7.3 Linguïstische constanten

7.3.1 De karakteristiek K van Yule en Herdan

In de voorgaande hoofdstukken kwam bij de analyse van de drie corpora onder andere aan de orde de karakteristiek *K*, gedefinieerd door Yule en Herdan. Dit gegeven wordt afgeleid uit de *frequentieverdeling*. Het zou onder andere een indicatie zijn voor de gemiddelde herhalingsfrequentie van woordtypen. Belangrijk is dat voor elk van de corpora kon worden vastgesteld dat de waarde van *K* inderdaad redelijk constant blijft bij steekproeven van uiteenlopende omvang uit een en hetzelfde corpus. Gegeven dat feit ligt het voor de hand om ook voor de steekproeven die in de voorgaande paragraaf werden beschreven na te gaan of de waarde van *K* overeenkomt met die van het volledige corpus.

Om daarover zekerheid te krijgen werd van de twee steekproeven een frequentieverdeling gemaakt, en werd de waarde van *K* bepaald volgens formules 4-4 en 4-5 (zie hoofdstuk 4). Het resultaat is te vinden in tabel 7-3.

Corpus	Aantal doc.	Typen	Tokens	*K* (Yule)	*K* (Herdan)
Wetgeving (volledig)	18803	143156	12807659	128,193	0,0128194
Wetgeving (gereduceerd)	6780	78574	4619141	128,921	0,0128923
Jurisprudentie (volledig)	16430	265841	40190212	111,059	0,0111060
Jurisprudentie (gereduceerd)	1992	74827	4617676	111,798	0,0111799

Tabel 7-3 De waarde van *K* voor de volledige corpora en voor beide steekproeven

De waarde van *K* (vlg. Herdan) bedroeg bij het volledige corpus *wetgevingsteksten* 0,0128194; bij steekproeven van uiteenlopende grootte lag de waarde tussen 0,0126624

en 0,0129805, gemiddeld 0,0128287, met een standaarddeviatie van 0,0000574. De berekende waarde voor de steekproef ligt minder dan 1,3 maal de standaarddeviatie boven zowel de gemiddelde waarde als de waarde voor het hele corpus, dat wil zeggen net iets meer dan de gemiddelde afwijking van de andere steekproeven. De waarden liggen echter zo dicht bij elkaar dat de overeenkomst tussen de waarde van *K* in de steekproef en die in het volledige corpus nog hoog kan worden genoemd.

Bij het volledige corpus jurisprudentieteksten had *K* een waarde van 0,0111060; de range bij diverse steekproeven (waarvan de meeste groter waren dan de hier bestudeerde) liep daar van 0,0110357 tot 0,0111516, gemiddeld 0,0110959, met een standaarddeviatie van 0,0000279. De range was bij dit corpus dus duidelijk kleiner, evenals de standaarddeviatie, waarschijnlijk doordat het gehele corpus en de meeste steekproeven daaruit veel omvangrijker waren (meer tokens omvatten) dan bij de wetgevingsteksten. De waarde van *K* zoals die in tabel 7-3 staat vermeld ligt ongeveer 3 maal de standaarddeviatie boven de gemiddelde waarde uit hoofdstuk 5. Wanneer we echter de kleinere range in aanmerking nemen, en tevens het feit dat de steekproef uit dit hoofdstuk slechts zo'n 10% van het oorspronkelijke corpus omvat kan de overeenkomst tussen de waarde van *K* in steekproef en volledig corpus toch nog als heel behoorlijk gelden.

De conclusie met betrekking tot de karakteristiek *K* kan zijn dat constateringen uit de voorgaande drie hoofdstukken, namelijk dat de waarde van deze karakteristiek praktisch onafhankelijk is van de corpus- c.q. steekproefgrootte, hier worden bevestigd. De verschillen tussen de waarden van *K* die betrekking hebben op de twee steekproeven uit dit hoofdstuk en op de volledige corpora zijn gering, waarbij de steekproef uit het corpus wetgevingsteksten nog duidelijk beter scoort dan die uit het corpus jurisprudentieteksten. Voorzover het deze karakteristiek betreft kunnen de beide steekproeven dan ook gelden als representatief voor de respectievelijke volledige corpora.

7.3.2 De verhouding tussen aantallen tokens en typen

Een belangrijk gegeven bij de beschrijving van de drie corpora vormde in de voorgaande hoofdstukken het verband tussen het aantal woordtokens en woordtypen. De door Erikstad[3] gedefinieerde verhouding $V = R \,.\, N^{\frac{1}{C}}$ bleek bij elk van de corpora geschikt om dit verband weer te geven voor het volledige corpus, en voor steekproeven van uiteenlopende afmetingen daaruit. De waarde van de twee constanten *R* en *C* werd daarbij voor ieder corpus door middel van regressieanalyse bepaald.

Het opnieuw bepalen van de waarde van *R* en *C*, maar nu voor de steekproeven uit de corpora, is een procedure die niet zozeer voor de hand ligt. Daarvoor zou namelijk net als voor de volledige corpora weer regressieanalyse nodig zijn, waarbij - analoog aan de methode die in de voorgaande hoofdstukken werd gevolgd - de twee steekproeven in nog kleinere deelcorpora zouden moeten worden geknipt. De omvang van die deelcorpora zou dan echter nog maar gering zijn, waardoor de foutmarge bij de regressieanalyse waarschijnlijk zou toenemen; ook in de voorgaande hoofdstukken was bij de kleinste

[3] Erikstad (1980, p. 223).

steekproeven immers sprake van de hoogste residuals. Wel kunnen de waarden van de twee constanten worden *toegepast* op beide steekproeven. De aantallen tokens en typen waaruit deze bestaan zijn immers bekend, waardoor kan worden nagegaan of de verhouding tussen *V* en *N* die voor het volledige corpus geldt, voor deze steekproeven ook opgaat. Van belang is daarbij vooral de vraag of met behulp van de formule van Erikstad en onder gebruikmaking van de eerder berekende waarden voor *R* en *C* het aantal typen in de twee nieuwe steekproeven correct kan worden berekend uit het aantal tokens dat ze bevatten. Deze berekening is voor beide steekproeven uitgevoerd, de resultaten zijn te vinden in tabel 7-4.

Gereduceerd Corpus	Typen (*V*)	Tokens (*N*)	*R*	*C*	Berekend aantal typen	Verschil in %
Wetgeving	78574	4619141	13,2979	0,5675	80523,3	2,5%
Jurisprudentie	74827	4617676	9,8383	0,5827	75210,7	0,5%

Tabel 7-4 Aantal typen in de twee steekproeven, berekend met behulp van de formule van Erikstad

Zoals uit de tabel blijkt is vooral bij het corpus jurisprudentieteksten het verschil tussen de uitkomst van de formule van Erikstad en het werkelijke aantal typen buitengewoon klein (383,7 typen of 0,5%). Bij het corpus wetgevingsteksten is het verschil wat groter (1949,3 typen of 2,5%). Dit resultaat is enigszins opvallend, aangezien de situatie precies omgekeerd was met betrekking tot de constante *K* (zie de vorige subparagraaf). Toch is ook de afwijking bij het corpus wetgevingsteksten niet buitenproportioneel groot. Ook in hoofdstuk 4 werd immers, nadat de waarden voor de constanten *R* en *C* waren vastgesteld, geconstateerd dat de aantallen typen in de daar gehanteerde steekproeven maximaal zo'n 2000 (bijna 4% van het aantal in de betreffende steekproef) afweken van de berekende waarden.

De conclusie aan het einde van deze paragraaf is dan ook dat voorzover op basis van de hier besproken 'linguïstische constanten' kan worden vastgesteld er sterke overeenkomsten bestaan tussen de structuur van het woordgebruik in de twee steekproeven uit de juridische corpora die in dit hoofdstuk zijn samengesteld, en de volledige corpora. Aangezien dat het geval is lijken de steekproeven geschikt om te worden gebruikt voor het maken van gedetailleerde vergelijkingen tussen de corpora. Aangezien elk van de steekproeven op bepaalde punten ook zekere afwijkingen vertoonde ten opzichte van het volledige corpus, zal bij deze vergelijkingen echter toch af en toe worden teruggegrepen op bevindingen met betrekking tot de volledige corpora, die in dat geval als referentie kunnen dienen.

Met betrekking tot de in deze paragraaf besproken linguïstische constanten kan worden vastgesteld dat zowel de karakteristiek *K* als de twee constanten *R* en *C* ook van toepassing zijn op willekeurige steekproeven uit het corpus waarvoor ze gelden, zelfs als die steekproeven maar een zeer beperkt deel van het corpus (zo'n 10% van de oorspronkelijke omvang) omvatten. Dit onderstreept de kracht van deze constanten bij het in kaart brengen van de eigenschappen van corpora.

7.4 Andere overeenkomsten en verschillen tussen de corpora

In de nu volgende paragraaf zal in detail het woordgebruik in de gereduceerde juridische corpora en in het corpus algemeen Nederlands worden vergeleken. Daarbij zullen de diverse aspecten van het woordgebruik aan de orde worden gesteld in dezelfde volgorde als in de voorgaande drie hoofdstukken. Eerst zal dus aandacht worden besteed aan de frequentielijsten en de aspecten die daar direct mee samenhangen, daarna volgen de frequentieverdelingen, vervolgens zullen (in het verlengde van de voorgaande paragraaf) enkele opmerkingen worden gemaakt over de token/type ratio's en tenslotte zullen aspecten die samenhangen met woordlengtes worden vergeleken.

7.4.1 Cumulatieve frequenties

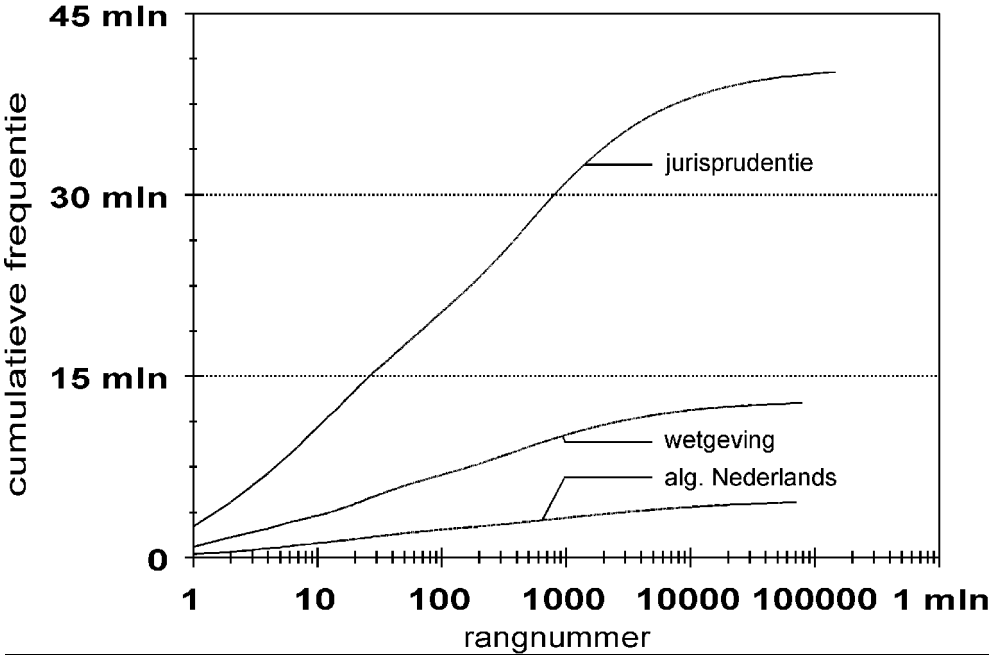
In de voorgaande hoofdstukken bleek het vergelijken van bepaalde aspecten die samenhangen met de frequentielijst nogal eens problemen op te leveren, vanwege de verschillen in afmetingen tussen de corpora. Tot die aspecten behoorden bijvoorbeeld het patroon dat de cumulatieve frequenties vertoonden, en het verband tussen het rangnummer en de frequentie van de typen in een corpus. In deze paragraaf zullen bedoelde vergelijkingen wel worden gemaakt, onder gebruikmaking van de gereduceerde juridische corpora en het (volledige) corpus algemeen Nederlands. Eén opvallend verschil tussen de frequentielijsten werd eerder in dit hoofdstuk al genoemd: het aantal woordtypen in de juridische corpora is ruwweg maar half zo groot als in het corpus algemeen Nederlands. Zoals zal blijken heeft dit consequenties voor verschillende aspecten van de structuur van het woordgebruik.

Bij het vergelijken van de frequentielijsten komen allereerst aan de orde de *cumulatieve frequenties*. Zoals in de voorgaande hoofdstukken werd vastgesteld vertoont de 'curve' van de cumulatieve frequenties (de som van de frequenties van alle typen tot en met een bepaald rangnummer), wanneer deze worden afgezet tegen het rangnummer en voor dit rangnummer een logaritmische schaalverdeling wordt gebruikt, over een groot gebied bijna de vorm van een (stijgende) rechte lijn. Verder was bij elk corpus sprake van een zich wijzigend verloop bij de hoogste rangnummers. Elk van de curves vertoonde in dat gebied een zekere kromming, welke bij de twee juridische corpora echter duidelijk sterker was dan bij het corpus algemeen Nederlands (zie de afbeeldingen 4-2b, 5-1 en 6-1). Tot dusver was niet duidelijk of de afmetingen van de corpora daarbij een rol speelden. De curves van de volledige corpora zijn ook nogal lastig direct met elkaar te vergelijken, vanwege het grote schaalverschil. Wanneer deze drie curves dan ook in één grafiek worden afgebeeld is onder andere de kromming in de curve van het corpus wetgevingsteksten nauwelijks meer te zien doordat de verticale schaal van deze grafiek tot ongeveer een derde is gereduceerd (zie afbeelding 7-1a).

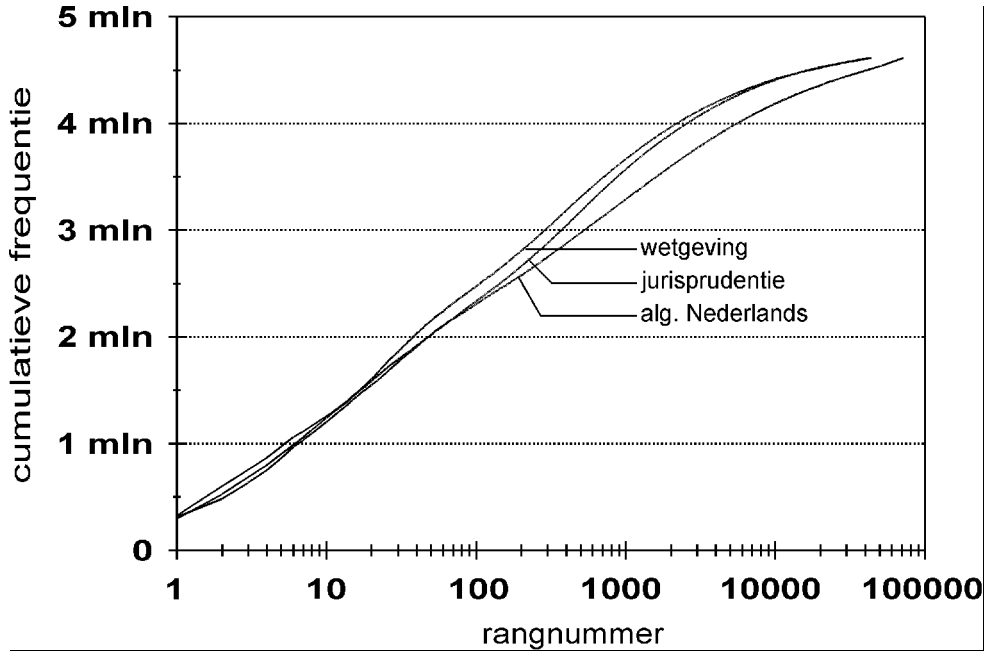
De cumulatieve frequentie is dan ook bij uitstek een gegeven dat het beste kan worden vergeleken aan de hand van corpora die voor wat betreft hun afmetingen zoveel mogelijk aan elkaar gelijk zijn gemaakt, aangezien dan de schaal op de verticale as voor alle corpora gelijk is. In grafiek 7-1b zijn daarom de cumulatieve frequentie-curves van de twee in omvang gereduceerde juridische corpora en die van het corpus algemeen Neder-

lands samengebracht. Aangezien de curves elkaar dicht benaderen is er dit keer voor

gekozen om niet de afzonderlijke datapunten af te beelden, maar alle punten door middel van een lijn te verbinden. Het verloop van de afzonderlijke curves is daardoor beter zichtbaar.



Afb. 7-1a De cumulatieve frequentie-curves van de drie volledige corpora zijn lastig te vergelijken door het grote verschil in omvang tussen de corpora



Afb. 7-1b Cumulatieve frequentiegrafiek van de gereduceerde juridische corpora en het corpus algemeen Nederlands

Met betrekking tot grafiek 7-1b is allereerst van belang in hoeverre de curve van de twee gereduceerde juridische corpora overeenkomt met die van de respectievelijke volledige

corpora. Vanzelfsprekend gaat het daarbij om overeenkomst in vorm, aangezien de schaal bij de gereduceerde corpora afwijkt van die bij de volledige. Voor wat betreft het corpus jurisprudentieteksten kan de vorm van de curve in 7-1b goed worden vergeleken met die in 7-1a. Geconstateerd kan dan worden dat beide curves veel gelijkenis vertonen. In beide gevallen ligt het begin van het 'gekromde' gedeelte ongeveer bij rangnummer 1000, de curve van het volledige corpus loopt aan de rechterzijde echter wat verder door als gevolg van het grotere aantal woordtypen (en dus hogere rangnummers) in dat corpus[4]. De curve van het gereduceerde corpus wetgevingsteksten laat zich niet zo goed vergelijken met de overeenkomstige curve uit 7-1a (het bereik op de verticale as is bij die curve veel geringer). Wanneer echter afbeelding 4-2b wordt opgeslagen blijkt ook hier sprake te zijn van een sterke overeenkomst in vorm.

Wanneer we de curves in grafiek 7-1b met elkaar vergelijken kan naar aanleiding daarvan in eerste instantie het volgende worden opgemerkt:

- Bij de twee juridische corpora is nog steeds sprake van een sterkere kromming bij de hoogste rangnummers dan bij het corpus algemeen Nederlands.

- Bij de laagste rangnummers (1-10) vallen de drie curves niet samen. Welke lijn precies bij welk corpus hoort is in grafiek 7-1b wat lastig te zien, ook al doordat geen gebruik kon worden gemaakt van verschillende kleuren. Een en ander is echter zo goed mogelijk aangegeven in grafiek 7-1c, die een uitvergroting voor de eerste 100 rangnummers bevat.

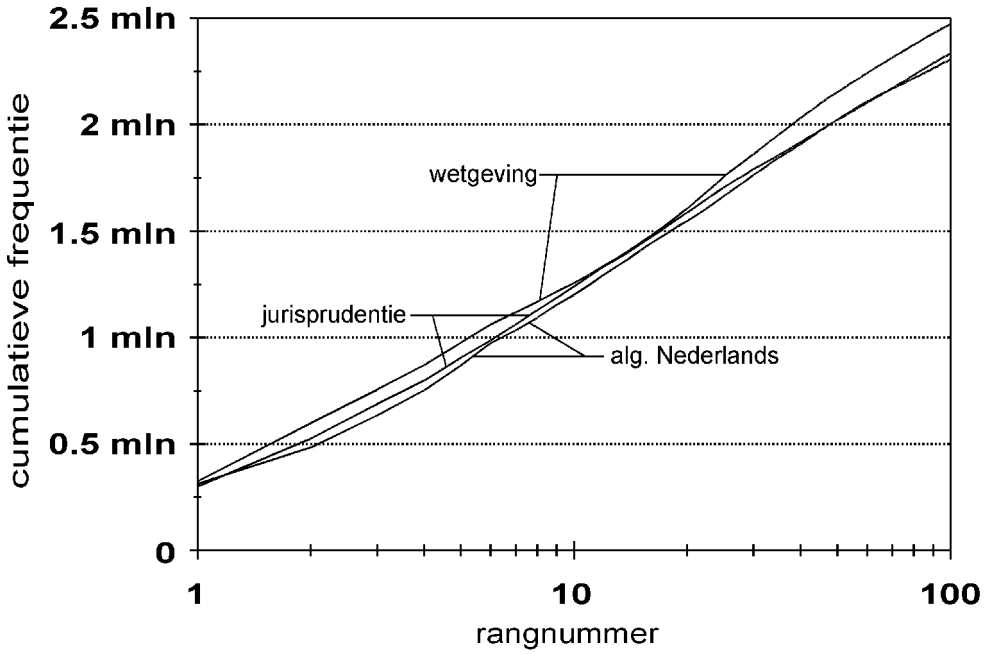
- Tussen de rangnummers 10 en 100 vallen aanvankelijk alle drie en later twee van de curves vrijwel samen. Daarna zijn de drie lijnen over een bepaald interval afzonderlijk te volgen, totdat op een bepaald punt de lijnen van de twee juridische corpora weer samenkomen (en dat ook blijven). Op het oog bevindt dit punt zich nog voor rangnummer 10000, in werkelijkheid blijkt het even voorbij rangnummer 14000 te liggen.

Aangezien grafiek 7-1b zoals vermeld niet voldoende duidelijk is in het gebied van de laagste rangnummers, wordt dat gedeelte hierna nogmaals afzonderlijk weergegeven (zie afbeelding 7-1c). Hoewel ook in die versie de lijnen elkaar soms zeer dicht naderen, kan toch worden vastgesteld dat ook bij de laagste rangnummers de curves dezelfde 'volgorde' hebben: bovenaan de curve van het corpus wetgevingsteksten, met kennelijk gemiddeld een groter aantal tokens per rangnummer dan de andere twee corpora. Onderaan de curve van het corpus algemeen Nederland, en daartussenin die van de jurisprudentieteksten. De totaalindruk die daardoor uit de twee grafieken naar voren komt is als volgt:

- De eerste tien rangnummers omvatten bij de juridische corpora en vooral bij de wetteksten een groter aantal woordtokens, dat wil zeggen, de tien meest voorkomende typen hebben bij de juridische corpora gemiddeld een hogere frequentie.

---

[4] Eenzelfde verschil werd in hoofdstuk 5 al geconstateerd bij een globale vergelijking van de curves van het corpus jurisprudentieteksten en het corpus wetgevingsteksten.



Afb. 7-1c Cumulatieve frequentiegrafiek, alleen eerste 100 rangnummers

- Na rangnummer 10 is er even een situatie waarin de drie corpora voor wat betreft de cumulatieve frequenties zeer dicht bij elkaar liggen. De curve van het corpus wetgevingsteksten lijkt hier een lichte 'dip' te vertonen.

- Daarna gaan voor beide juridische corpora de cumulatieve frequenties vanaf een zeker punt sterker stijgen dat het geval is bij het corpus algemeen Nederlands. Bij de wetteksten ligt dit punt rond rangnummer 20, bij de jurisprudentieteksten rond rangnummer 100. Het corpus jurisprudentieteksten maakt de wat latere start goed met een hoger stijgpercentage, waardoor de cumulatieve frequentie rond rangnummer 5000 toch gelijk is aan die bij het corpus wetgevingsteksten. De curves van de twee juridische corpora vallen vanaf dat punt samen.

De conclusie uit dit alles is als volgt: de aantallen woordtokens zijn bij de twee juridische corpora anders verdeeld over de rangnummers (lees: de woordtypen) dan bij het corpus algemeen Nederlands. Al bij de laagste rangnummers is sprake van een zeker verschil, waarbij ieder woordtype in de juridische corpora een groter aantal tokens omvat (een hogere frequentie heeft). Bij het corpus wetgevingsteksten gaat de cumulatieve frequentie vervolgens vanaf rangnummer 20 nog wat sterker stijgen; ditzelfde gebeurt bij het corpus jurisprudentieteksten rond rangnummer 100. Dit heeft tot gevolg dat bij de juridische corpora een wat groter percentage van alle woordtokens te vinden is bij de lagere rangnummers. Zo omvatten de eerste 100 rangnummers van het corpus wetgevingsteksten 53,58% van de woordtokens, tegen precies 50% bij het corpus algemeen Nederlands. Bij rangnummer 1000 is dit verschil opgelopen tot 8,1% (79,35% tegen 71,25%). Bij het corpus jurisprudentieteksten is sprake van een analoge situatie, het verschil bij rangnummer 1000 bedraagt daar 6,04% in het voordeel van dit corpus. Bij rangnummer 10000 is het verschil al wat kleiner, namelijk ongeveer 5% bij beide juridische corpora. Doordat

meer tokens te vinden zijn bij de lagere rangnummers is het logisch dat bij de juridische corpora bij de hogere rangnummers sprake is van een sterkere kromming in de grafiek, dat wil zeggen van een relatief geringere toename van het aantal woordtokens. Het lijkt er dus op dat bij de juridische corpora de *meestgebruikte woordtypen* elk ook *vaker worden gebruikt*. Eenzelfde resultaat vonden De Mulder en Oskamp ten aanzien van het woordgebruik in de wetboeken van Strafrecht en van Strafvordering[5].

7.4.2 Het verband tussen rangnummer en frequentie

In de voorgaande hoofdstukken is steeds vrij uitgebreid aandacht geschonken aan het verband tussen het rangnummer van woordtypen en hun frequentie. Onder andere werd daarbij in een grafiek de frequentie afgezet tegen het rangnummer, waarbij voor beide assen gebruik werd gemaakt van een logaritmische schaalverdeling. Het bleek in dat geval mogelijk te zijn om het verband tussen beide gegevens in een bepaald interval, de 'kop' van de frequentielijst, te benaderen door middel van een regressierechte. Twee verschillende methoden voor het bepalen van de vergelijking van deze regressielijn werden uitgeprobeerd, waarbij voor elk van de corpora het beste resultaat werd bereikt wanneer datapunten werden gewogen naar gelang het aantal *tokens* dat ze omvatten. Het aan de hand van de vergelijking van deze regressielijn vastgestelde verband tussen rangnummer en frequentie bleek voor elk van de corpora enigszins verschillend te liggen. Voor het (volledige) corpus wetgevingsteksten was de vergelijking van deze regressielijn:

$$\log r = -1{,}079 \log rangnr + 6{,}226 \qquad (7\text{-}1)$$

waarbij r staat voor de frequentie van een type. Voor de jurisprudentieteksten was dit

$$\log r = -1{,}069 \log rangnr + 6{,}701 \qquad (7\text{-}2)$$

en voor het corpus algemeen Nederlands

$$\log r = -1{,}065 \log rangnr + 5{,}732 \qquad (7\text{-}3)$$

Zonder gebruik van logaritmen ziet het verband tussen rangnummer en frequentie er voor de corpora dus als volgt uit: voor het corpus wetgevingsteksten

$$r = \frac{10^{6{,}226}}{rangnr^{1{,}079}} \qquad (7\text{-}4)$$

voor het corpus jurisprudentieteksten:

$$r = \frac{10^{6{,}701}}{rangnr^{1{,}069}} \qquad (7\text{-}5)$$

en voor het corpus algemeen Nederlands

$$r = \frac{10^{5{,}732}}{rangnr^{1{,}065}} \qquad (7\text{-}6)$$

Op het eerste gezicht lijkt de constante in de teller van elk van de breuken hier de voornaamste verschilfactor, deze is bij het corpus wetgevingsteksten ruim 3 keer zo hoog als bij het corpus algemeen Nederlands, en bij het corpus jurisprudentieteksten bijna 10

[5] De Mulder (1984, p. 150).

keer zo hoog. Ook in de noemer van de breuken kunnen echter aanzienlijke verschillen ontstaan, ondanks het feit dat de exponenten voor de rangnummers dicht bij elkaar liggen. Zo is de waarde van de noemer bij het corpus wetgevingsteksten bij rangnummer 100 nog maar ongeveer 9 hoger dan bij het corpus algemeen Nederlands (een verschil van zo'n 7%), maar dit verschil loopt op tot rond de 37000 (17,5%) bij rangnummer 100000. Uiteindelijk gaat het natuurlijk om het gezamenlijk effect van beide parameters. Dit effect komt hierop neer, dat bij gelijk rangnummer de frequentie in de juridische corpora aanmerkelijk hoger zal liggen dan in het corpus algemeen Nederlands; bij het corpus wetgevingsteksten is dit iets minder dan een factor 3, bij het corpus jurisprudentieteksten ongeveer een factor 9. Vergelijk ook de grafieken 4-3, 5-2 en 6-2. Aangezien deze verhouding tussen de frequenties (bij gelijk rangnummer), 3 : 9 : 1, ongeveer gelijk is aan de verhouding tussen de aantallen tokens waaruit de volledige corpora bestaan, ligt de veronderstelling voor de hand dat de afmetingen van de corpora op de bovenstaande gegevens van invloed zijn. Om na te gaan in hoeverre dat inderdaad het geval is zal ook voor de twee gereduceerde corpora eenzelfde analyse worden uitgevoerd, waarna de resultaten met die voor het volledige corpus zullen worden vergeleken. Pas dan zal de vraag aan de orde worden gesteld in hoeverre op dit punt, de verhouding tussen rangnummer en frequentie, verschillen tussen de corpora kunnen worden vastgesteld. Om die vraag te beantwoorden zullen de twee gereduceerde corpora worden vergeleken met het corpus algemeen Nederlands.

In de eerste plaats dus aandacht voor het verband tussen bovengenoemde constanten en de afmetingen van de corpora. Teneinde hierover uitspraken te kunnen doen is voor beide gereduceerde corpora precies dezelfde regressie-analyse uitgevoerd als voor de volledige corpora. Daarbij is, evenals bij de volledige corpora, weer gebruik gemaakt van twee methoden om de vergelijking van de regressielijn vast te stellen. Eén van deze methoden leverde betere resultaten op, maar juist om na te gaan of de eerder gevonden verschillen ook bij de gereduceerde corpora aanwezig zijn is ook de andere methode hier toegepast. De resultaten zijn daarbij als volgt:

1. Corpus wetgevingsteksten
Wanneer hier gebruik wordt gemaakt van de eerste methode, waarbij *per rangnummer* 1 datapunt wordt toegekend, luidt de vergelijking van de regressielijn:

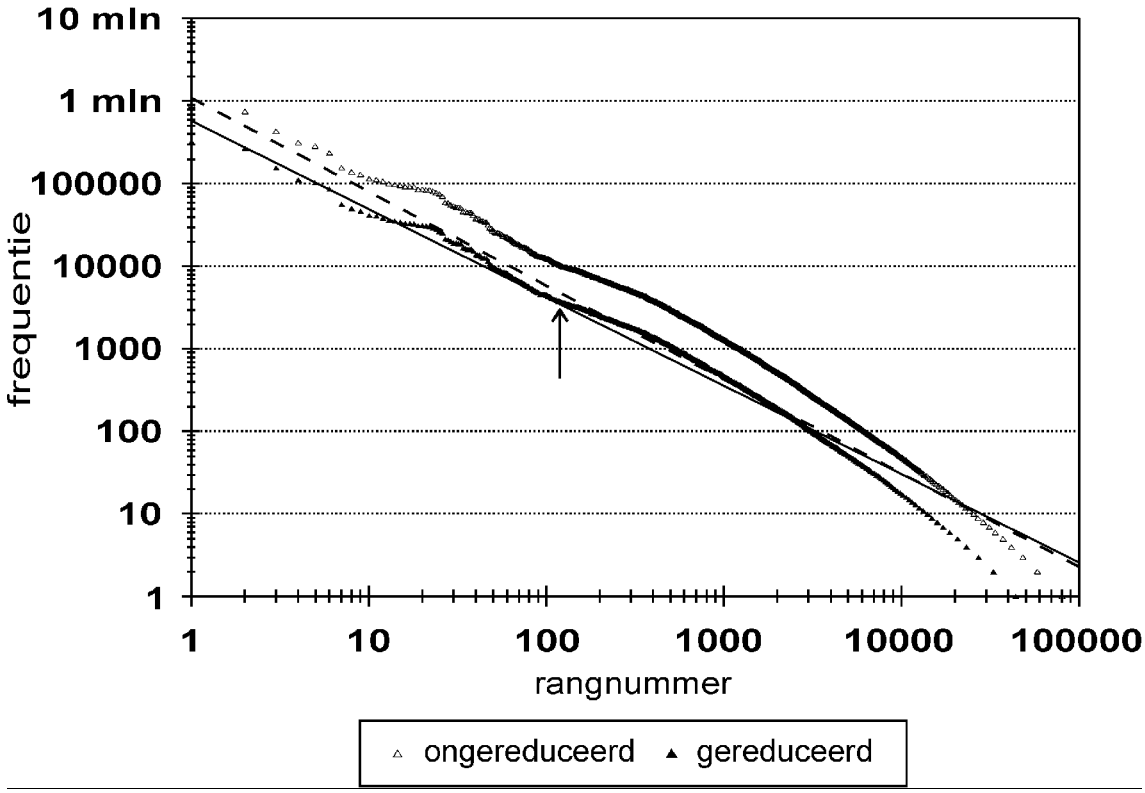
$$\log r = -1{,}137 \log rangnr + 6{,}043 \qquad (7\text{-}7)$$

Wanneer de tweede methode wordt toegepast (1 datapunt per *woordtoken*) luidt de vergelijking:

$$\log r = -1{,}071 \log rangnr + 5{,}772 \qquad (7\text{-}8)$$

Beide regressielijnen zijn weer ingetekend in een grafiek waarin de frequenties zijn afgezet tegen de rangnummers. De lijn volgens methode 1 (de bovenste) is gestippeld getekend, die volgens methode 2 is doorgetrokken. Om te controleren of de verhouding tussen rangnummer en frequentie bij het gereduceerde corpus wetgevingsteksten ongeveer hetzelfde verloop te zien geeft als bij het volledige zijn in dit geval ook de datapunten die horen bij het volledige corpus afgebeeld. De datapunten voor het gereduceerde corpus liggen daarbij zoals te verwachten was onder die van het volledige; alle frequenties zijn door het geringere totaal aantal woordtokens lager. Voor het overige blijkt de overeenkomst in

vorm tussen beide reeksen hoog te zijn; zelfs de onregelmatigheden, bijvoorbeeld tussen rangnummer 10 en 100, zijn bij het gereduceerde corpus vrijwel gelijk aan die in het volledige.



Afb. 7-2a Corpus wetgevingsteksten: verband tussen rangnummer en frequentie.
Ter vergelijking zijn ook de datapunten voor het volledige corpus opgenomen.

De regressielijnen die horen bij het volledige corpus zijn hier niet ingetekend (de grafiek zou daardoor mogelijk wat onduidelijk worden), maar vergelijking met afbeelding 4-3 leert dat ook deze bij het gereduceerde corpus ongeveer hetzelfde verlopen. Een klein verschil is te vinden in de plaats waar de datapunten de vorm van een kromme krijgen (zie pijl). Dit punt, dat in de voorgaande hoofdstukken is beschouwd als markering voor het einde van de 'kop' van de frequentielijst, lag bij het volledige corpus bij rangnummer 136, en hier bij 120. Dat betekent wel dat het deel van de frequentielijst dat tot nu toe is beschouwd als de 'kop' enigszins *afhankelijk is van de steekproefgrootte*. Op dit punt zal later in dit hoofdstuk nog worden terugkomen.

Voor wat betreft de regressielijnen voor het gereduceerde corpus kan worden vastgesteld dat de tweede methode ook nu weer het beste resultaat geeft, een lijn die tot een bepaald punt (rangnummer 120) de datapunten goed volgt. Dat betekent dat de tweede vergelijking (7-8) weer het meest van toepassing is, waardoor het verband tussen rangnummer en frequentie voor deze steekproef het beste kan worden uitgedrukt als

$$r = rangnr^{-1,071} \cdot 10^{5,772} = \frac{10^{5,772}}{rangnr^{1,071}} \qquad (7\text{-}9)$$

Wanneer we dit vergelijken met de waarden voor het volledige corpus, blijkt met name de constante in de teller flink te zijn teruggelopen (van $10^{6,226}$ tot $10^{5,772}$). Deze constante hangt dus kennelijk in belangrijke mate samen met de corpusgrootte. De exponent van het rangnummer is slechts in zeer geringe mate afgenomen (van 1,079 tot 1,071), hoewel het effect daarvan bij wat hogere rangnummers toch nog aanzienlijk is (een afname van de noemer met bijna 10% bij rangnummer 40000).

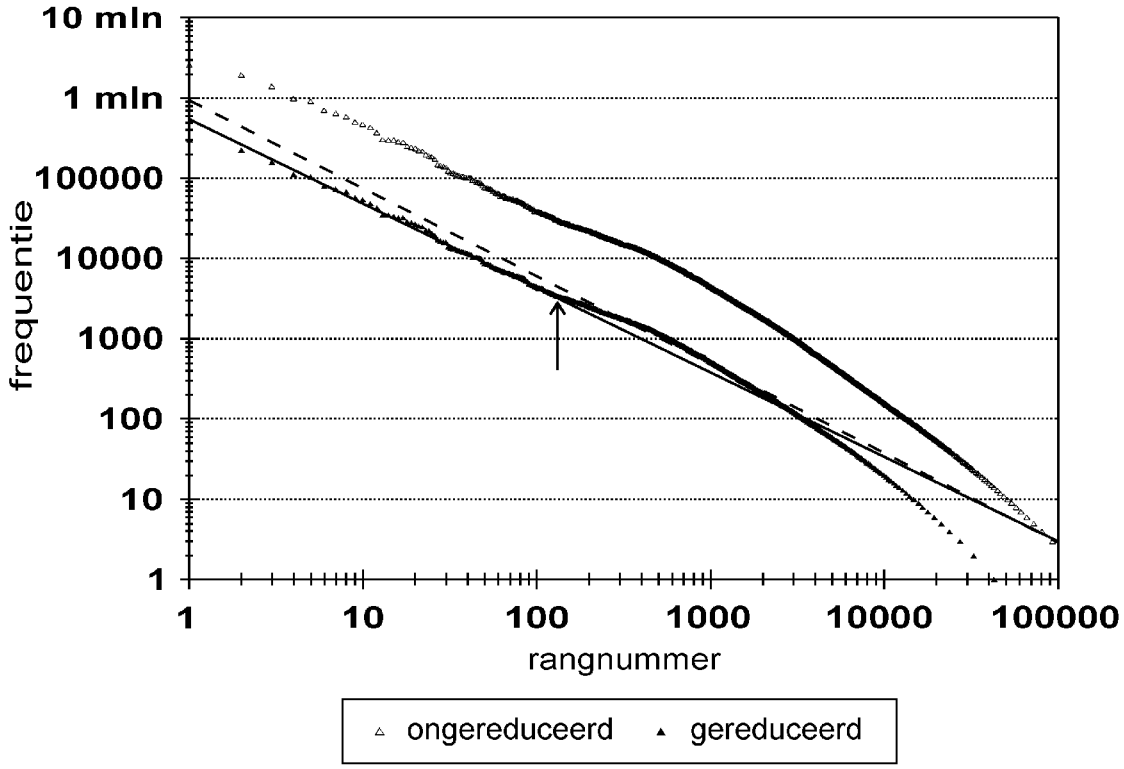
2. Corpus jurisprudentieteksten
Bij gebruik van de eerste methode (1 datapunt *per rangnummer*) is de vergelijking van de regressielijn:

$$\log r = -1{,}103 \log rangnr + 5{,}974 \qquad (7\text{-}10)$$

Bij de tweede methode (1 datapunt per *woordtoken*) is dit:

$$\log r = -1{,}051 \log rangnr + 5{,}735 \qquad (7\text{-}11)$$

Deze twee lijnen zijn weer ingetekend in een grafiek waarin de frequenties zijn afgezet tegen de rangnummers (7-10 als stippellijn, 7-11 als doorgetrokken lijn). Ook in dit geval zijn de datapunten die horen bij het volledige corpus weer ter vergelijking opgenomen.



Afb. 7-2b Corpus jurisprudentieteksten: verband tussen rangnummer en frequentie.
Ter vergelijking zijn ook de datapunten van het volledige corpus opgenomen.

Deze grafiek vertoont evenals 7-2a, met name voor wat betreft de datapunten, veel overeenkomst met die voor het volledige corpus. Bij vergelijking met afbeelding 5-2 is

echter een verschil te constateren met betrekking tot een van beide regressielijnen, namelijk die welke gestippeld is weergegeven (7-10). Deze lijn kruist nu niet de andere regressielijn in de buurt van rangnummer 5000, maar pas veel later, rond 50000. De doorgetrokken regressielijn, berekend volgens de methode 'een datapunt per woordtoken' verloopt wel ongeveer hetzelfde als bij het volledige corpus, een extra argument in het voordeel van deze methode. Wel ligt net als bij de wetteksten het punt waar de datapunten van de lijn gaan afwijken bij een wat lager rangnummer, namelijk bij 133 in plaats van bij 168, zodat ook in dit geval sprake is van een kortere 'kop' van de frequentielijst. Doordat regressielijn 7-11 het verband tussen rangnummer en frequentie weer het beste weergeeft, kan dit verband worden geschreven als

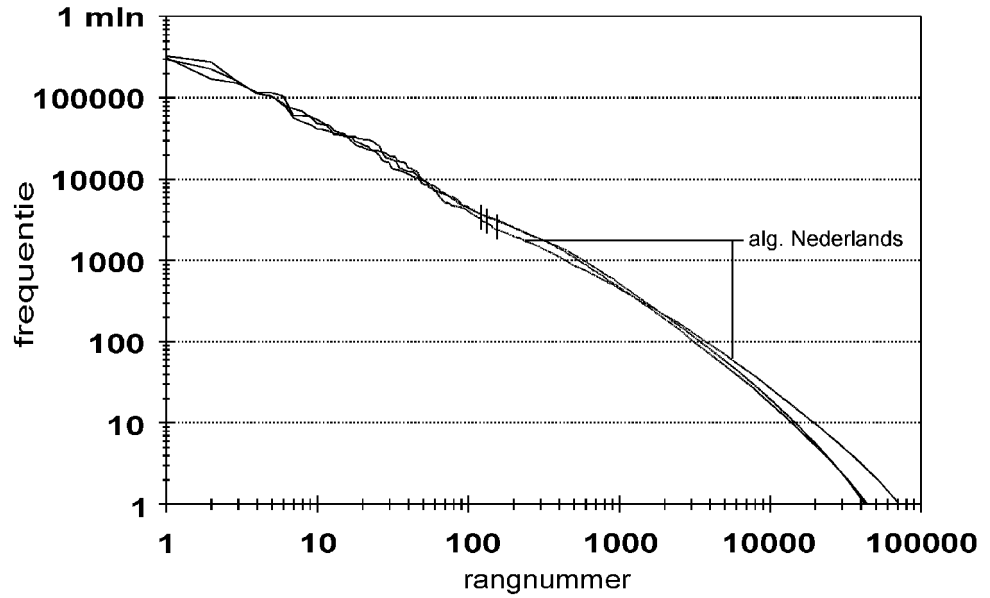
$$r = rangnr^{-1,051} \cdot 10^{5,735} = \frac{10^{5,735}}{rangnr^{1,051}} \qquad (7\text{-}12)$$

De constante in de teller ligt bijna een factor 10 lager dan bij het volledige corpus ($10^{5,735}$ in plaats van $10^{6,701}$). Dat hangt waarschijnlijk hiermee samen, dat deze steekproef een geringer deel uitmaakt van het volledige corpus (namelijk iets meer dan 10%) dan de steekproef uit het corpus wetgevingsteksten doet (die omvat ongeveer 30% van het volledige corpus). De relatief geringere omvang van de steekproef zou misschien ook een verklaring kunnen zijn voor de sterkere afname van de exponent voor het rangnummer dan bij de wetteksten (van 1,069 tot 1,051).

Aan de hand van deze gegevens kan de eerste vraag, naar de rol die de corpusgrootte speelt bij het met behulp van regressieanalyse vastgestelde verband tussen rangnummer en frequentie, beantwoord worden. Het blijkt inderdaad zo te zijn dat de afmetingen van de corpora op deze verhouding van invloed zijn. Deze invloed is het sterkst bij de *constante* die van de verhouding deel uitmaakt; deze blijkt in het geval van het corpus jurisprudentieteksten met een factor 10 te zijn afgenomen, ongeveer evenveel als het corpus in omvang is gereduceerd. Ook de *exponent* van het rangnummer blijkt echter te variëren met de corpusgrootte, zij het in veel geringere mate. De conclusie die hieruit getrokken kan worden is dat de met behulp van regressieanalyse bepaalde verhouding tussen rangnummer en frequentie geen stabiele verhouding is, maar dat deze afhankelijk blijkt te zijn van de corpusgrootte. Dat betekent dat het vergelijken van deze verhouding voor verschillende corpora alleen zin heeft wanneer die corpora dezelfde afmetingen hebben. Met behulp van de regressielijn van (log) frequentie op (log) rangnummer werd in de voorgaande hoofdstukken ook bepaald welk deel van de frequentielijst als de 'kop' kan gelden. Ook de afmetingen van deze kop blijken nu afhankelijk te zijn van de corpusgrootte, zij het niet zo sterk. De afmetingen van de kop nemen bij het gereduceerde corpus wetgevingsteksten af van 136 tot 120, en bij het gereduceerde corpus jurisprudentieteksten van 168 naar 133. De afname is duidelijk groter bij het corpus jurisprudentieteksten, waar ook een sterkere reductie van het aantal tokens heeft plaatsgevonden. Bij het corpus algemeen Nederlands is sprake van een langere kop dan bij de gereduceerde juridische corpora; deze omvat namelijk 155 rangnummers.

Nu hierover duidelijkheid bestaat, zal vervolgens aandacht besteed worden aan het tweede punt dat hierboven werd genoemd, namelijk het vergelijken van de verhouding tussen frequentie en rangnummer voor de drie corpora. Gezien de conclusie die hierboven werd getrokken over het verband tussen deze verhouding en de corpusgrootte zullen daarbij alleen de twee gereduceerde corpora worden vergeleken met het corpus

algemeen Nederlands. Een grafiek is hiervoor weer het meest geëigende middel. Wanneer we de data voor de drie corpora in één grafiek samenbrengen blijkt echter dat de 'curves' zo dicht bij elkaar liggen dat de gegevens voor elk afzonderlijk corpus bijna niet meer zijn waar te nemen. Dit is zelfs nog het geval wanneer we niet meer de afzonderlijke datapunten afbeelden, maar alleen een lijn die deze verbindt (zie afbeelding 7-3). De drie curves verlopen in het begin (ongeveer tot rangnummer 100) enigszins grillig, en kruisen elkaar voortdurend. Vlak voorbij dat punt is door middel van drie verticale lijntjes het punt aangegeven vanaf waar elk van de curves begint af te wijken van de bijbehorende regressielijn. Van links naar rechts horen de lijntjes bij het corpus wetgevingsteksten, jurisprudentieteksten en algemeen Nederlands.

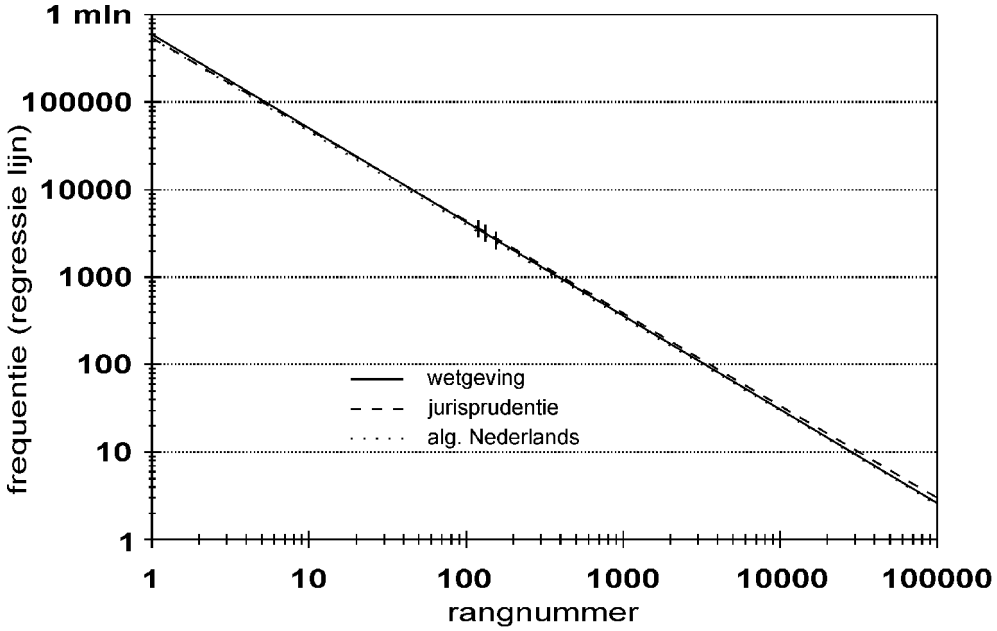


Afb. 7-3 De verhouding tussen frequentie en rangnummer voor de drie corpora in beeld gebracht

Rechts van deze punten is er één curve die een wat afwijkend verloop heeft. Bij nader onderzoek blijkt dit de curve van het corpus algemeen Nederlands te zijn. Deze daalt in eerste instantie wat sterker, kruist vervolgens de curves van de juridische corpora (die nauwelijks van elkaar te onderscheiden zijn), en vertoont voor het overige een wat minder sterke neerwaartse kromming. Dat betekent dat in dit corpus de frequenties aanvankelijk wat sneller afnemen (ongeveer van rangnummer 100 tot 1000), maar dat bij de hoogste rangnummers het omgekeerde het geval is (de frequenties in de juridische corpora dalen sneller). De snellere daling bij de hoogste rangnummers van de curves van de juridische corpora is in overeenstemming met de bevindingen bij de *cumulatieve* frequenties. De verklaring is dezelfde: bij de juridische corpora zijn de frequenties bij de *lagere* rangnummers gemiddeld wat *hoger*, waardoor aan het eind minder woordtokens 'over zijn' (elk van de corpora bevat nu immers even veel tokens) en de frequenties wat sneller teruglopen.

Bij de laagste rangnummers gaan de grafieken van de drie corpora zoals vermeld ongeveer gelijk op. Het is dan ook niet verwonderlijk dat de drie regressielijnen, die vooral het verband tussen rangnummer en frequentie in de *kop van de frequentielijst*

weergeven, zeer dicht bij elkaar liggen. Dat blijkt ook wanneer we deze lijnen samen in één grafiek tekenen (zie afbeelding 7-4). De vergelijkingen van deze regressielijnen vertonen zeer veel overeenkomst, zowel voor wat betreft de x-coëfficiënt (-1,071, -1,051 en -1,065) als de y-intercept (5,772, 5,735 en 5,732). Het enige punt dat misschien nog als opvallend kan gelden is dat de lijn van het corpus jurisprudentieteksten een iets geringere daling vertoont dan die van de andere twee corpora.



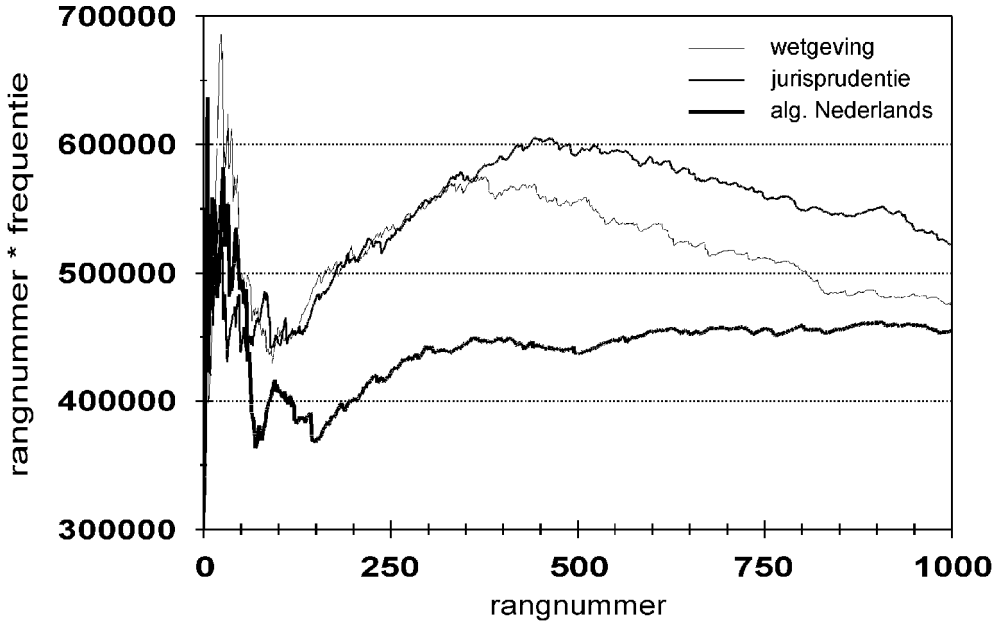
Afb. 7-4 Regressielijn van frequentie op rangnummer voor de drie corpora (gereduceerde data)

Een vergelijking met de bevindingen van De Mulder en Oskamp

De verhouding tussen rangnummer en frequentie, zoals in beeld gebracht in grafieken 7-3 en 7-4, lijkt voor de drie corpora bij de laagste rangnummers dus slechts geringe verschillen te vertonen. Bij het onderzoek van De Mulder en Oskamp[6] was sprake van duidelijker zichtbare verschillen tussen de door hen bestudeerde strafrechtelijke corpora en het vergelijkingscorpus algemeen Nederlands (waarvoor zij het Eindhovens corpus gebruikten). Zij maakten gebruik van zogenaamde 'Zipf-karakteristieken' om de frequenties in de kop van de frequentielijsten met elkaar te vergelijken. Deze karakteristieken zijn gebaseerd op de al in hoofdstuk 2 besproken 'wet van Zipf', die inhoudt dat bij natuurlijke talen het produkt van rangnummer en frequentie van de woordtypen constant is. Zonder een uitspraak te doen over de juistheid van deze wet gebruikten De Mulder en Oskamp haar met name als een analyse-instrument, waarmee juist *afwijkingen* in het verloop van de frequenties zichtbaar kunnen worden gemaakt. In de grafieken die zij samenstelden was op de horizontale as (net als in dit proefschrift steeds het geval is) het rangnummer afgezet, en op de verticale as het produkt van rangnummer en frequentie (*rangnr * r*).

[6] Zie De Mulder en Oskamp (1979) en De Mulder (1984, p.135 e.v.).

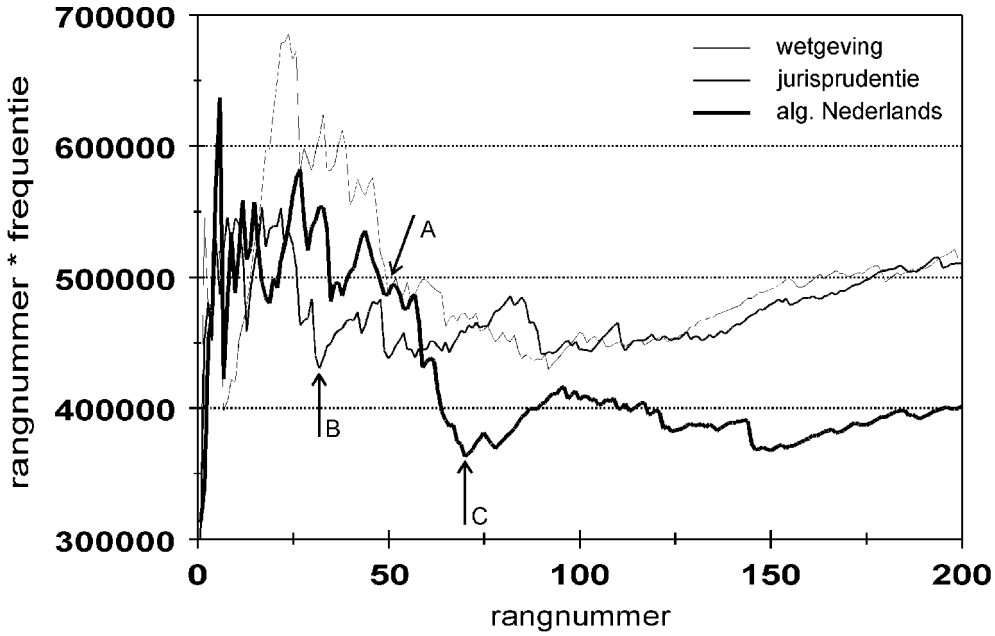
Wanneer de wet van Zipf voor een corpus volledig zou opgaan, zou dit dus moeten resulteren in een 'Zipf-karakteristiek' bestaande uit een rechte lijn evenwijdig aan de horizontale as. Wanneer voor de gereduceerde juridische corpora en voor het corpus algemeen Nederlands echter zulke karakteristieken worden getekend ontstaan, evenals trouwens bij De Mulder en Oskamp, curves die zeker bij de laagste rangnummers een nogal grillig verloop hebben (zie afbeelding 7-5a). Wanneer deze curves verder wordt doorgetrokken (hier is alleen het gedeelte tot rangnummer 1000 afgebeeld) blijkt dat ze ook bij hogere rangnummers geen horizontaal, maar een licht dalend verloop hebben. Dat is ook logisch, gezien wat intussen bekend is over de verhouding tussen rangnummer en frequentie in deze corpora. Al in hoofdstuk 4 (paragraaf 4.2.1) werd erop gewezen dat alleen wanneer de regressielijn van log *r* op log *rangnr* een x-coëfficiënt van -1 heeft (en dus een hoek van 45° met de horizontale as maakt) het produkt van rangnummer en frequentie constant kan zijn. De Zipf-karakteristieken laten overigens *wel* duidelijke verschillen zien tussen de drie corpora. Zo ligt de curve van het corpus algemeen Nederlands vanaf ongeveer rangnummer 50 duidelijk onder de curves van de juridische corpora (wat zou duiden op lagere woordfrequenties in dit deel van de frequentielijst). De curves van de juridische corpora verlopen aanvankelijk ongeveer gelijk, maar ongeveer bij rangnummer 300 gaat de curve van de wetgevingsteksten dalen, terwijl die van de jurisprudentieteksten nog even verder stijgt.



Afb. 7-5a 'Zipf-karakteristieken' van de gereduceerde juridische corpora en het corpus algemeen Nederlands (eerste 1000 rangnummers)

De Mulder en Oskamp gebruikten de Zipf-karakteristieken ook om vast te stellen welk gedeelte van de frequentielijst van een corpus als 'kop' kon gelden. Daartoe werd door hen het eerste gedeelte van de karakteristieken (tot en met rangnummer 200 à 300) verder uitvergroot. Zowel bij de door hen onderzochte corpora als bij de drie corpora die in dit proefschrift centraal staan wordt dan de 'piek' bij de laagste rangnummers duidelijker zichtbaar. In grafiek 7-5b is dit voor de twee gereduceerde juridische corpora en het corpus algemeen Nederlands in beeld gebracht. De Mulder en Oskamp beschouwden het

punt waar de piek eindigt, en de curve weer een min of meer horizontaal of licht stijgend verloop krijgt als de kop van de frequentielijst. De woordtypen in dit deel van de lijst betitelden zij als 'zeer vaak voorkomend'. Het eerst wat oplopende een daarna weer dalende gedeelte na de piek, dat zich bij de door hen onderzochte corpora ongeveer uitstrekte tot rangnummer 250 à 300 (en bij de hier onderzochte corpora ongeveer tot rangnummer 1000, zie afbeelding 7-5a) beschouwden zij als het 'middenstuk' van de frequentielijst, met de 'vaak' voorkomende woordtypen.



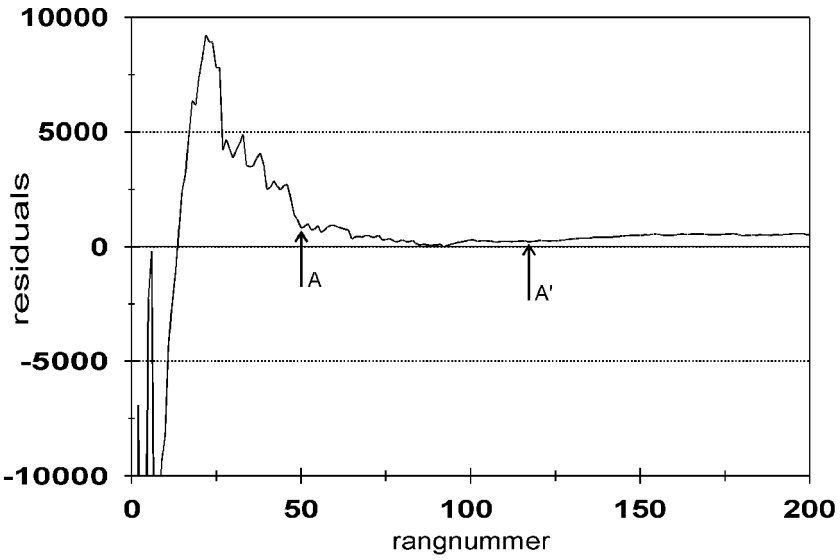
Afb. 7-5b 'Zipf-karakteristieken' als in 7-5a, eerste gedeelte uitvergroot

Tot nu toe is steeds als de kop van de frequentielijst beschouwd het gedeelte waar de verhouding tussen rangnummer en frequentie (min of meer) overeenkomt met de lineaire regressielijn van log frequentie op log rangnummer (zie het voorgaande gedeelte). Dit deel van de frequentielijst omvat (afhankelijk van het corpus) de eerste 120 tot 160 rangnummers. Wanneer we echter volgens de criteria van De Mulder en Oskamp op zoek gaan naar het rangnummer waar de kop van de frequentielijst van de drie corpora zou moeten ophouden, moeten we voor het corpus wetgevingsteksten waarschijnlijk uitgaan van het punt dat gemarkeerd is als punt A, wat zou betekenen dat de 'echte' kop met 'zeer vaak' voorkomende woordtypen (opnieuw volgens de definitie van De Mulder en Oskamp) slechts de eerste 50 rangnummers zou omvatten. Voor het corpus jurisprudentieteksten zou deze 'echte kop' waarschijnlijk lopen tot punt B (rangnummer 1 t/m 32), en bij het corpus algemeen Nederlands tot punt C (rangnummer 1 t/m 70). Wat hier tot nu toe steeds is beschouwd als de kop zou dus bij De Mulder en Oskamp ook een deel van het 'middenstuk' van de frequentielijst omvatten.

Aangezien dus sprake is van verschillende methoden om vast te stellen welk gedeelte kan gelden als de kop van de frequentielijst en beide methoden een ander resultaat opleveren, dient de vraag te worden beantwoord van welke methode in dit geval het beste kan worden uitgegaan. De methode van De Mulder en Oskamp is volledig gebaseerd op de Zipf-karakteristieken. Bij de beoordeling van deze karakteristieken is *niet* van belang dat bekend is dat de 'wet van Zipf', waarop zij zijn gebaseerd, niet als een absoluut geldende

wet kan worden beschouwd[7]. Van een dergelijke absolute 'wetmatigheid' is bijvoorbeeld ook bij gebruik van logaritmen geen sprake. Wel kan als een methodologisch nadeel van deze karakteristieken worden beschouwd dat op de verticale as een 'samengesteld' gegeven is afgezet (het produkt van rangnummer en frequentie) terwijl op de horizontale as het rangnummer is afgezet, een van de componenten van dat samengestelde gegeven. Een verband tussen deze gegevens is zo als het ware al ingebouwd. Van belang is dan ook of de kop van de frequentielijsten welke volgt uit de Zipf-karakteristieken ook op een andere manier kan worden aangetoond.

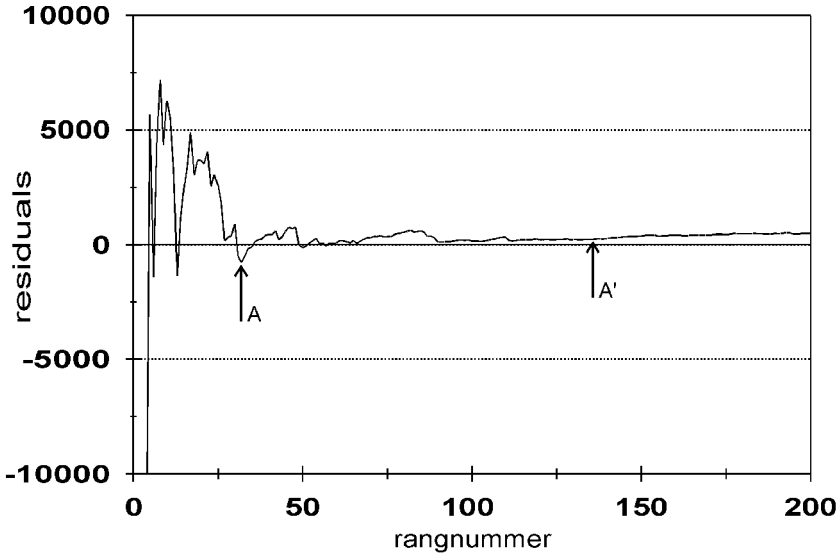
Dit blijkt inderdaad het geval te zijn, wanneer gebruik wordt gemaakt van de eerder besproken *regressielijn* van log frequentie op log rangnummer[8]. Zoals bijvoorbeeld uit de afbeeldingen 7-2a en 7-2b blijkt, volgt deze regressielijn de datapunten (die aangeven wat de werkelijke frequentie behorend bij een bepaald rangnummer is) niet volledig, de eerste datapunten liggen wat onder de lijn en vervolgens bevindt zich een aantal punten boven de regressielijn. Pas daarna volgt een serie datapunten die praktisch samenvallen met de regressielijn. Dit patroon van *afwijkingen* ten opzichte van de regressielijn kan in beeld worden gebracht door de *residuals* van de regressielijn in een grafiek weer te geven. Gezien het vrij grote aantal punten is daarbij gekozen voor een lijngrafiek, welke in dit geval - omwille van de duidelijkheid - voor ieder corpus afzonderlijk is opgesteld; zie de afbeeldingen 7-6a, b en c. Deze residuals, die zoals uit de grafieken blijkt overigens niet voor alle corpora in dezelfde orde van grootte liggen, vertonen een patroon dat doet denken aan dat in de Zipf-karakteristieken. Wanneer in deze grafieken het punt wordt gemarkeerd waar de piek van positieve residuals ophoudt (het punt A in onderstaande grafieken) blijkt dit punt overeen te komen met het 'einde van de kop' zoals dat blijkt uit de Zipf-karakteristieken.



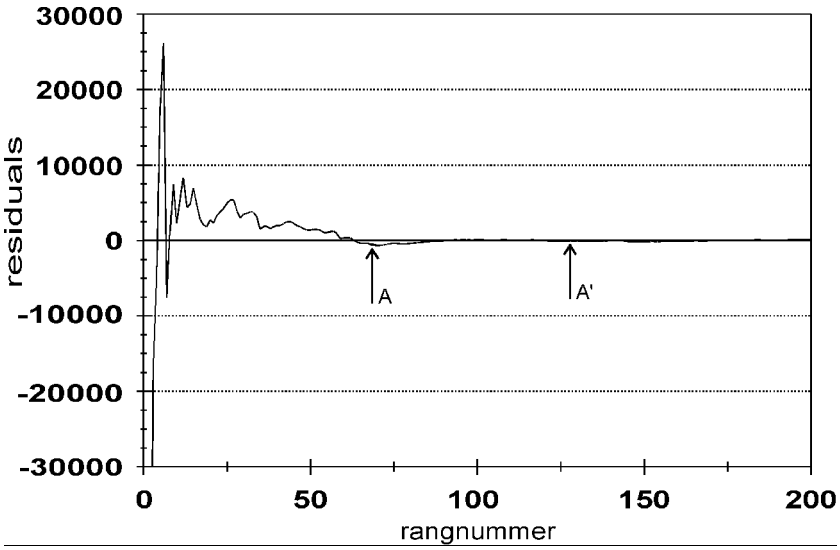
Afb. 7-6a Corpus wetgevingsteksten (gereduceerd): residuals van de regressielijn van log frequentie op log rangnummer.

[7] De Mulder en Oskamp geven zelf ook expliciet aan geen uitspraak te willen doen over de juistheid van de 'wet van Zipf', noch deze te willen toetsen (De Mulder 1984, p. 142).

[8] Het gaat hierbij om de regressielijn gebaseerd op aantallen woordtokens, waarvan voor alle corpora is vastgesteld dat deze het verband tussen (log) frequentie en (log) rangnummer het beste weergeeft.



Afb. 7-6b Corpus jurisprudentieteksten (gereduceerd): residuals van de regressielijn van log frequentie op log rangnummer.



Afb. 7-6c Corpus algemeen Nederlands: residuals van de regressielijn van log frequentie op log rangnummer.

N.B. het bedoelde punt is bij het corpus wetgevingsteksten het eenduidigst vast te stellen, terwijl bij het corpus algemeen Nederlands de piek minder stijl lijkt, waardoor het gekozen 'eindpunt' daar te veel naar rechts lijkt te liggen. Dit laatste is voornamelijk het gevolg van de andere schaal op de verticale as van de grafiek, noodzakelijk door de sterke positieve 'uitschieter' bij de laagste rangnummers. In de drie grafieken is tenslotte ook het einde van de kop volgens de eerder gehanteerde definitie (de plaats waar de datapunten gaan afwijken van de regressielijn) aangegeven; dit is het punt A'.

De door De Mulder en Oskamp beschreven kop van de frequentielijst blijkt dus een gegeven te zijn dat ook los van de - voor sommigen misschien wat discutabele - Zipf-karakteristieken kan worden vastgesteld, zij het niet bij ieder corpus even eenduidig. *Deze*

('echte') kop blijkt, net als in het onderzoek van De Mulder en Oskamp het geval was, bij de juridische corpora aanzienlijk korter te zijn dan bij het corpus algemeen Nederlands (50 respectievelijk 32 rangnummers tegenover 70 rangnummers). Wanneer we kijken naar het percentage van alle woordtokens die in dit deel van de frequentielijst vertegenwoordigd zijn dan blijken dat er voor het corpus wetgevingsteksten en het corpus algemeen Nederlands ongeveer evenveel te zijn, namelijk 46,6% respectievelijk 47%, bij het corpus jurisprudentieteksten iets minder, namelijk 39,4% (het aantal rangnummers in de 'echte kop' is bij dat corpus natuurlijk minder, na 50 rangnummers blijkt het percentage woordtokens daar echter eveneens ruim 46% te bedragen). Wanneer we er evenals De Mulder en Oskamp van uitgaan dat in de 'echte kop' de woordtypen te vinden zijn die 'zeer vaak' voorkomen, lijkt het feit dat deze 'echte kop' bij de juridische corpora korter is dat bij het corpus algemeen Nederlands hun conclusie te bevestigen dat in juridische teksten (in dit geval ook in jurisprudentieteksten) een geringer aantal woordtypen 'zeer vaak' wordt gebruikt dan in algemeen Nederlandse teksten. In ieder geval kan worden vastgesteld dat in de hier beschreven 'echte kop' van de drie corpora een hoog percentage van alle woordtokens waaruit de corpora bestaan (tot 47%) is vertegenwoordigd. Dit percentage is zelfs niet eens veel lager dan in het langere gedeelte van de frequentielijst dat in de voorgaande hoofdstukken steeds als kop is beschouwd (ongeveer de eerste 150 rangnummers), waar het op zo'n 52% tot 56% ligt. Verder zijn er aanwijzingen dat de omvang van de 'echte' kop - anders dan die van de 'gewone' kop - niet of slechts in zeer geringe mate afhankelijk is van de steekproefgrootte, aangezien onderzoek leert dat de residuals van de volledige corpora voor dit deel van de frequentielijst praktisch hetzelfde patroon vertonen als die van de gereduceerde corpora. Tenslotte kan nog opgemerkt worden dat een punt dat zowel met behulp van de Zipf-karakteristieken als met behulp van de residuals is vast te stellen is dat bij het corpus wetgevingsteksten niet alleen sprake is van een kortere 'echte' kop, maar ook van hogere frequenties binnen die kop. De aanwezigheid van deze hogere frequenties was al af te leiden uit de cumulatieve frequentiegrafieken (zie de afbeeldingen 7-1a, b en c). Voor het corpus jurisprudentieteksten blijkt uit beide grafiektypen *niet* van hogere frequenties binnen de kop (in vergelijking tot het corpus algemeen Nederlands).

Samenvatting en conclusies met betrekking tot de verhouding rangnummer - frequentie

Wanneer de curves van log frequentie afgezet tegen log rangnummer voor de drie corpora worden vergeleken kunnen aan de hand daarvan in eerste instantie geen duidelijke verschillen tussen het frequentieverloop over de eerste 100 rangnummers (de 100 meest voorkomende woordtypen) worden aangetoond. Ook de regressielijnen die op basis van deze curves kunnen worden getekend vertonen slechts geringe verschillen. Over het interval tussen rangnummer 100 en (ongeveer) 800 kan worden vastgesteld dat de woordtypen in de *juridische corpora* daar steeds een *hogere* frequentie bij gelijk rangnummer hebben. In het corpus algemeen Nederlands liggen de frequenties vanaf rangnummer 100 steeds *onder* die in de juridische corpora. Na rangnummer 800 kruist de curve van algemeen Nederlands die van de juridische corpora. Dat dit gebeurt is logisch, aangezien de corpora evenveel tokens bevatten. Lagere frequenties bij de veelvoorkomende typen moeten dan ook wel samengaan met hogere frequenties bij de minder vaak voorkomende typen.

Wanneer gebruik wordt gemaakt van Zipf-karakteristieken blijkt het wel mogelijk te zijn bepaalde verschillen in het frequentieverloop bij de drie corpora aan te tonen. Met name tekent zich in die karakteristieken voor ieder corpus een vrij duidelijke *kop* af, welke bij de twee juridische corpora *korter* is als bij het corpus algemeen Nederlands. Volgens De Mulder en Oskamp zou dit het gedeelte van de frequentielijst zijn waar de 'zeer vaak voorkomende' woordtypen te vinden zijn[9]. Ook bij de door hen onderzochte corpora bleek dit aantal 'zeer vaak voorkomende' typen in juridische corpora lager dan in een corpus met algemeen Nederlandse teksten. Aangezien bij de betekenis van Zipf-karakteristieken enige vraagtekens geplaatst kunnen worden (onder andere is het rangnummer - mede - bepalend voor de waarde van de gegevens op beide assen) is gezocht naar een manier om de 'echte kop' met (volgens De Mulder en Oskamp althans) de zeer vaak voorkomende typen ook op een andere manier zichtbaar te maken. Dit blijkt mogelijk te zijn door de *residuals* van de regressielijn van log frequentie op log rangnummer grafisch weer te geven. Deze grafieken vertonen dan een kop welke voor ieder corpus sterk lijkt op die in de Zipf-karakteristieken. Dat betekent dus dat aan de hand van de genoemde regressielijn in de corpora zowel een korte 'echte' kop kan worden gevonden van 32 tot 70 woordtypen die 'zeer vaak' zouden voorkomen, als een 'lange kop' van 120 tot 155 woordtypen waarvan we zouden kunnen zeggen dat ze 'zeer vaak tot vaak' voorkomen[10]. De omvang van de 'lange kop' laat zich wat eenvoudiger vaststellen dan die van de 'korte' (de plaats waar de datapunten niet langer de regressielijn volgen is nauwkeuriger aan te geven dan de plaats waar de piek in de residuals-grafiek ophoudt, zoals uit afbeeldingen 7-6a t/m c blijkt). Een voordeel van de 'korte' kop is dat deze bij twee van de drie onderzochte corpora minder dan de helft van het aantal woordtypen uit de 'lange' kop bevat, maar toch een bijna even hoog percentage van de woordtokens (de uitzondering vormt hier het corpus jurisprudentieteksten); de korte kop bevat dus de kern van de woordtypen met de hoogste frequenties, we kunnen deze woordtypen daarom met recht 'zeer vaak voorkomend' noemen. Daar komt bij dat er aanwijzingen zijn dat de omvang van de 'korte' kop niet of minder sterk afhankelijk is van de steekproefgrootte (het aantal woordtokens in het corpus). Op basis van de omvang van de 'korte' kop kan geconcludeerd worden dat de twee juridische corpora uit dit onderzoek evenals het corpus strafrechtelijke wetten uit het onderzoek van De Mulder en Oskamp een geringer aantal woordtypen tellen die kunnen worden aangemerkt als 'zeer vaak voorkomend'.

7.4.3 De spreiding van woordtypen

Met behulp van de gereduceerde juridische corpora kan ook de spreiding van woordtypen over de documenten waar de corpora uit bestaan nauwkeuriger worden vergeleken dan in de voorgaande hoofdstukken. Hierbij kan weer gebruik worden gemaakt van een grafiek, waarin de spreidingspercentages voor de drie corpora zichtbaar worden gemaakt. In de voorgaande hoofdstukken werd de spreiding op twee verschillende manieren in beeld gebracht:

---

9 De Mulder (1984, p. 144).

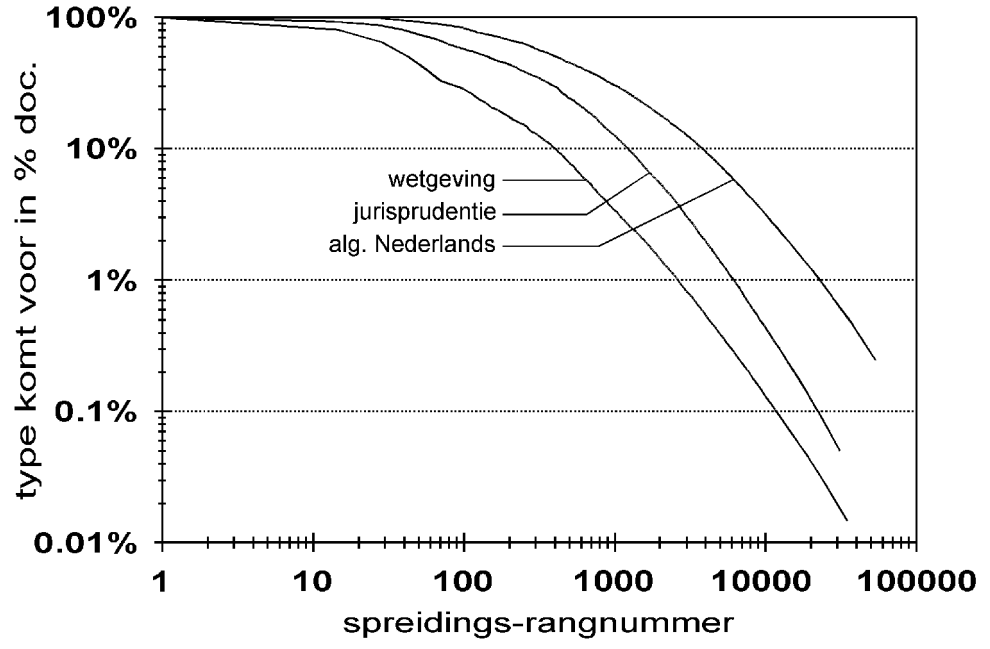
10 De Mulder en Oskamp (1979, p. 48-49) betitelen het laatste gedeelte van deze 'lange kop' als het 'middenstuk' van de Zipf-karakteristiek.

1. uitgaande van de volgorde van de frequentielijst werden de percentages die aangaven in hoeveel van de documenten ieder type voorkwam afgezet tegen het rangnummer;

2. de typen werden eerst gesorteerd op de mate van spreiding, en pas daarna werden de spreidingspercentages afgezet tegen het zogenaamde spreidings-rangnummer (een rangnummer toegekend op basis van de positie in de nu op mate van spreiding gesorteerde lijst).

In beide gevallen ontstond het beste overzicht wanneer voor beide assen gebruik werd gemaakt van een logaritmische schaalverdeling. Bij de eerste methode was dan sprake van een neerwaarts gekromde, enigszins uitwaaierende puntenwolk (zie bijvoorbeeld afbeelding 4-5) doordat een (frequentie-)rangnummer vaak verschillende woordtypen omvat, die elk hun eigen spreidingspercentage kunnen hebben. Bij de tweede methode waren de datapunten echter gegroepeerd in de vorm van een licht gekromde lijn of curve, als gevolg van het sorteren op mate van spreiding. Deze lijn volgde min of meer de bovenste begrenzing van de puntenwolk die het resultaat was van methode 1.

Voor het direct vergelijken van de mate van spreiding van de woordtypen in elk van de corpora is de tweede methode beter geschikt, met name omdat deze een curve oplevert in plaats van een puntenwolk. De aan de hand van deze methode vervaardigde curves voor de twee gereduceerde juridische corpora en het corpus algemeen Nederlands zijn samen afgebeeld in grafiek 7-7. Daarbij is voor de duidelijkheid, net als bij grafiek 7-3, afgezien van het weergeven van afzonderlijke datapunten maar is volstaan met een lijn die deze datapunten met elkaar verbindt.



Afb. 7-7 Spreiding van woordtypen in de drie corpora, rangnummers (x-as) op basis van de mate van spreiding

Alvorens deze grafiek te bespreken is echter een relativerende opmerking op zijn plaats. Het is namelijk zo dat nog steeds een bepaalde factor verstorend zou kunnen werken,

namelijk het *verschil in gemiddelde document-afmetingen* bij de drie corpora. Deze afmetingen lopen uiteen van 681 tokens in het corpus wetgevingsteksten, via 2318 tokens bij de jurisprudentieteksten, tot 11370 in het corpus algemeen Nederlands. Het is zeker niet uitgesloten dat de spreiding van woordtypen door deze grote verschillen in afmeting enigszins wordt beïnvloed. Aan het opheffen van dit punt van verschil, bijvoorbeeld door het samenstellen met behulp van steekproeven van kunstmatige, even grote documenten kon helaas in het kader van dit onderzoek geen aandacht worden besteed.

Verder kan ook niet onvermeld blijven dat de documenten in de corpora op verschillende manieren tot stand zijn gekomen. Bij de wetteksten heeft daarbij het toeval ook een zekere rol gespeeld (met name bij het samenvoegen van te kleine basisdocumenten, zie hoofdstuk 3), terwijl bij het corpus algemeen Nederlands de lengte van documenten wellicht zelfs een rol heeft gespeeld bij de beslissing ze al dan niet in het corpus op te nemen. Bij de jurisprudentieteksten hebben de documenten in dit opzicht nog de meest 'natuurlijke' opbouw, hoewel de redactie van het tijdschrift Nederlandse Jurisprudentie wellicht ook beperkingen aanlegt ten aanzien van de minimale en maximale omvang van te publiceren uitspraken. Wanneer we deze overwegingen in het achterhoofd houden vallen in de grafiek toch de volgende zaken op:

- De curves van de twee gereduceerde corpora hebben op het oog min of meer dezelfde vorm als die van de volledige corpora (zie afbeelding 4-6 en 5-4). Een uitzondering hierop vormt het eerste gedeelte van de curve voor de wetgevingsteksten (tot rangnummer 100). Hier is bij het gereduceerde corpus een kleine 'dip' zichtbaar die bij het volledige corpus niet of nauwelijks valt waar te nemen. De lichte 'knik' die bij het volledige corpus optrad in de buurt van de 10%-grens is juist bij het gereduceerde corpus niet meer zichtbaar. Van groot belang lijken deze verschillen echter niet.

- De drie curves lopen niet even ver door. Dit is een gevolg van het feit dat de corpora, die nu even veel woordtokens bevatten, uit een verschillend aantal documenten bestaan. Meer (en dus gemiddeld kleinere) documenten betekent dat ook lagere spreidingspercentages voorkomen.

De grafiek lijkt aan te geven dat de mate van spreiding van woordtypen duidelijk verschillend is in de drie corpora. Een voorbeeld: in het corpus wetgevingsteksten komen slechts zo'n 400 typen in 10% of meer van alle documenten voor, bij de jurisprudentieteksten zijn dat er al ruim 1200, en bij het corpus algemeen Nederlands bijna 4000, 10 keer zoveel dus. Deze verhouding blijft voor de corpora wetgevingsteksten en algemeen Nederlands daarna min of meer gehandhaafd (de curves lopen ongeveer parallel), bij het corpus jurisprudentieteksten daalt het spreidingspercentage iets sneller. Een voorzichtige conclusie die uit deze gegevens kan worden getrokken is dat *alle* woordtypen in de juridische corpora, en vooral in het corpus wetgevingsteksten, gemiddeld in een kleiner gedeelte van het corpus geconcentreerd lijken te zijn (dat wil zeggen, in minder documenten voorkomen). Dat zou kunnen betekenen dat veel typen in deze corpora minder 'universeel' zijn, hun gebruik beperkt zich slechts tot bepaalde (groepen van) documenten. Bij deze conclusie moet echter een slag om de arm worden gehouden, gezien de hierboven genoemde beperkingen ten aanzien van de documentgroottes en de totstandkoming van documenten.

7.4.4 Frequentieverdelingen

Van de gereduceerde juridische corpora zijn evenals van de volledige corpora frequentieverdelingen gemaakt, die een overzicht geven van de frequenties waarmee woordtypen in de corpora voorkomen.

Klasse	Grenzen	Klasse-midden	Corpus wetgevingsteksten:			Corpus jurispr. teksten:			Corpus alg. Nederlands:		
			Aantal freq.	Aantal typen	Aantal tokens	Aantal freq.	Aantal typen	Aantal tokens	Aantal freq.	Aantal typen	Aantal tokens
1	$10^{0} - 10^{0.1}$	1,13	1	34479	34479	1	32082	32082	1	70713	70713
2	$10^{0.1} - 10^{0.2}$	1,42	0	0	0	0	0	0	0	0	0
3	$10^{0.2} - 10^{0.3}$	1,79	0	0	0	0	0	0	0	0	0
4	$10^{0.3} - 10^{0.4}$	2,25	1	11136	22272	1	10090	20180	1	20315	40630
5	$10^{0.4} - 10^{0.5}$	2,84	1	5808	17424	1	5370	16110	1	10302	30906
6	$10^{0.5} - 10^{0.6}$	3,57	0	0	0	0	0	0	0	0	0
7	$10^{0.6} - 10^{0.7}$	4,50	2	6497	28637	2	6067	26767	2	10817	47644
8	$10^{0.7} - 10^{0.8}$	5,66	1	1986	11916	1	1870	11220	1	3313	19878
9	$10^{0.8} - 10^{0.9}$	7,13	1	1509	10563	1	1500	10500	1	2496	17472
10	$10^{0.9} - 10^{1.0}$	8,97	3	3137	27800	3	3108	27683	3	5306	47013
11	$10^{1.0} - 10^{1.1}$	11,29	2	1422	16324	2	1411	16201	2	2220	25402
12	$10^{1.1} - 10^{1.2}$	14,22	3	1583	21978	3	1594	22205	3	2444	34010
13	$10^{1.2} - 10^{1.3}$	17,90	4	1496	26000	4	1496	26042	4	2185	37948
14	$10^{1.3} - 10^{1.4}$	22,54	6	1405	31237	6	1508	33657	6	2193	48763
15	$10^{1.4} - 10^{1.5}$	28,37	6	1057	30018	6	1120	31781	6	1431	40417
16	$10^{1.5} - 10^{1.6}$	35,72	8	1011	35752	8	1004	35478	8	1344	47448
17	$10^{1.6} - 10^{1.7}$	44,96	11	944	42098	11	1012	45382	11	1348	60029
18	$10^{1.7} - 10^{1.8}$	56,61	13	748	42245	13	812	46001	13	1026	57901
19	$10^{1.8} - 10^{1.9}$	71,26	16	656	46464	16	692	49002	16	859	61162
20	$10^{1.9} - 10^{2.0}$	89,72	21	557	49604	21	618	55074	21	780	69626
21	$10^{2.0} - 10^{2.1}$	112,95	25	419	46937	25	500	56193	25	606	68253
22	$10^{2.1} - 10^{2.2}$	142,19	33	397	56182	33	465	65459	33	534	75188
23	$10^{2.2} - 10^{2.3}$	179,01	41	368	65233	41	389	68717	41	498	88921
24	$10^{2.3} - 10^{2.4}$	225,36	52	317	70761	52	350	77953	52	402	89341
25	$10^{2.4} - 10^{2.5}$	283,71	64	261	73505	63	306	86248	65	347	97701
26	$10^{2.5} - 10^{2.6}$	357,17	76	235	83539	77	234	83556	80	273	97123
27	$10^{2.6} - 10^{2.7}$	449,65	87	183	81228	87	197	86751	88	210	93522
28	$10^{2.7} - 10^{2.8}$	566,07	90	166	91935	90	166	94150	97	192	107265
29	$10^{2.8} - 10^{2.9}$	712,64	96	142	100745	96	152	107132	98	161	114101
30	$10^{2.9} - 10^{3.0}$	897,16	91	116	103994	96	132	118254	95	122	107770
31	$10^{3.0} - 10^{3.1}$	1129,46	79	94	106112	93	107	120780	76	87	97821
32	$10^{3.1} - 10^{3.2}$	1421,91	74	84	118495	99	114	162669	75	82	115205
33	$10^{3.2} - 10^{3.3}$	1790,08	80	88	156113	79	90	160667	86	73	129522
34	$10^{3.3} - 10^{3.4}$	2253,57	69	71	158424	62	69	153725	52	54	120961
35	$10^{3.4} - 10^{3.5}$	2837,08	44	46	130109	47	50	141119	25	25	72116
36	$10^{3.5} - 10^{3.6}$	3571,67	39	43	150963	40	41	143742	19	19	68372
37	$10^{3.6} - 10^{3.7}$	4496,47	24	26	116118	22	22	96674	27	27	120949
38	$10^{3.7} - 10^{3.8}$	5660,72	14	14	78821	16	16	93126	12	12	65535
39	$10^{3.8} - 10^{3.9}$	7126,43	12	12	84611	17	17	118600	5	5	35343
40	$10^{3.9} - 10^{4.0}$	8971,64	12	12	106812	8	8	68180	10	10	90802
41	$10^{4.0} - 10^{4.1}$	11294,63	4	4	45440	12	12	133355	8	8	93127
42	$10^{4.1} - 10^{4.2}$	14219,09	7	7	95255	6	6	79637	7	7	94882
43	$10^{4.2} - 10^{4.3}$	17900,78	9	9	158900	5	5	89668	6	6	106945
44	$10^{4.3} - 10^{4.4}$	22535,74	3	3	63289	3	3	66001	8	8	183384
45	$10^{4.4} - 10^{4.5}$	28370,82	8	8	235267	5	5	135120	3	3	81420
46	$10^{4.5} - 10^{4.6}$	35716,75	7	7	246361	5	5	169728	4	4	146790
47	$10^{4.6} - 10^{4.7}$	44964,72	3	3	129924	2	2	92337	3	3	142616
48	$10^{4.7} - 10^{4.8}$	56607,23	2	2	107569	2	2	112809	3	3	179837
49	$10^{4.8} - 10^{4.9}$	71254,28	0	0	0	2	2	142732	0	0	0
50	$10^{4.9} - 10^{5.0}$	89716,41	1	1	86673	1	1	81150	0	0	0
51	$10^{5.0} - 10^{5.1}$	112946,27	2	2	217514	2	2	218579	3	3	339747
52	$10^{5.1} - 10^{5.2}$	142190,93	1	1	157496	0	0	0	1	1	151236
53	$10^{5.2} - 10^{5.3}$	179007,78	0	0	0	1	1	160372	1	1	170425
54	$10^{5.3} - 10^{5.4}$	225357,44	0	0	0	1	1	226323	0	0	0
55	$10^{5.4} - 10^{5.5}$	283708,20	1	1	274895	1	1	300805	1	1	313273
56	$10^{5.5} - 10^{5.6}$	357167,47	1	1	325110						

Tabel 7-5 Frequentieverdeling van de drie corpora (juridische corpora gereduceerd)

Deze frequentieverdelingen kunnen anders dan die van de volledige corpora goed met elkaar, en met die van het corpus algemeen Nederlands worden vergeleken.
Bij dit vergelijken zal met name aandacht worden besteed aan de *aantallen woordtypen en -tokens* die in de diverse frequentieklassen voorkomen. De frequentieverdeling voor de gereduceerde corpora is op precies dezelfde manier samengesteld als bij de volledige corpora. Er is dus weer gebruik gemaakt van frequentie*klassen* die *logaritmisch in omvang toenemen*. Om het geheel zo overzichtelijk mogelijk te houden zijn de gegevens van de corpora die worden vergeleken (de gereduceerde juridische corpora en het corpus algemeen Nederlands) in één tabel samengebracht. Voor elk corpus wordt vermeld het aantal verschillende frequenties, het aantal typen en het aantal tokens per frequentieklasse (zie tabel 7-5). Ten aanzien van de tabel kunnen de volgende opmerkingen worden gemaakt:

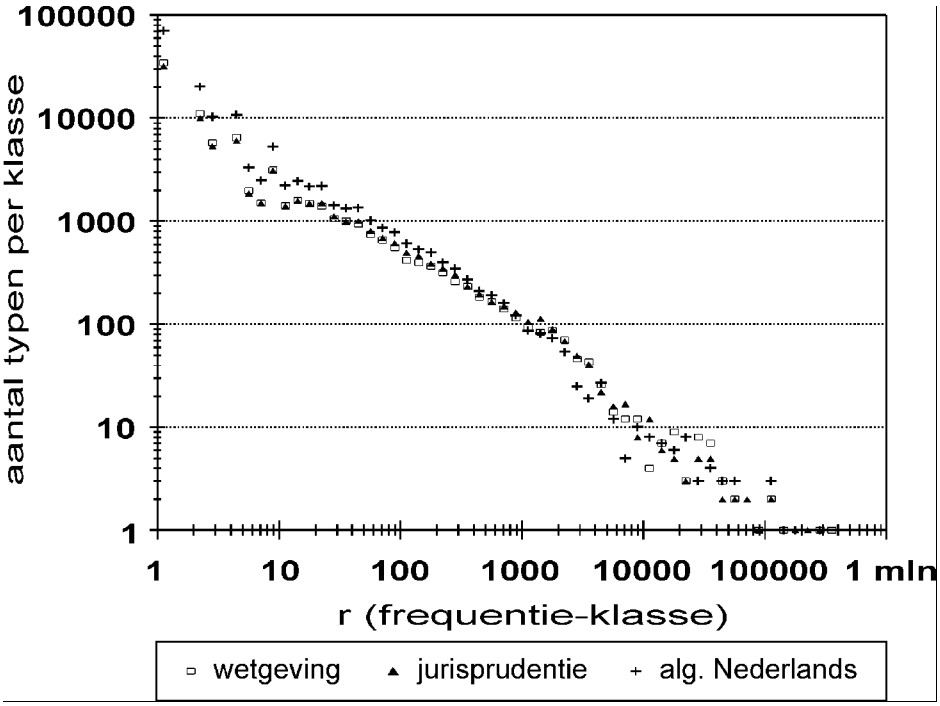
- Op het eerste gezicht is het opvallend dat dezelfde drie klassen (2, 3 en 6) bij alle drie corpora leeg zijn. Bij nadere bestudering blijkt dat echter geen toeval te zijn, deze klassen *kunnen* eenvoudig geen typen bevatten aangezien woordfrequenties altijd gehele getallen zijn. Typen die 1 keer voorkomen vallen in klasse 1, typen die 2 keer voorkomen in klasse 4, typen die 3 keer voorkomen in 5 en typen die 4 of 5 keer voorkomen in 7. Het gaat hier dus om een onvolkomenheid van de klasse-indeling, die zou verdwijnen wanneer we bijvoorbeeld het de klassebreedte zouden verhogen (waardoor overigens de nauwkeurigheid van de indeling zou verminderen).

- Het corpus wetgevingsteksten telt een frequentieklasse meer, die overigens slechts een enkel type bevat. Het blijkt te gaan om het type DE, dat 325110 keer in het corpus voorkomt, een aantal juist boven de klasse-grens van klasse 55 ($10^{5,5}$ = 316227).

- Het corpus jurisprudentieteksten bevat, afgezien van de klassen 2, 3 en 6, slechts 1 lege klasse, de andere corpora ieder drie.

- Bij de wetgevingsteksten en het corpus algemeen Nederlands is de frequentieverdeling min of meer symmetrisch in die zin dat een klasse ongeveer in het midden (in beide gevallen is dit klasse 29) het hoogste aantal verschillende frequenties bevat, en de klassen die daaraan voorafgaan en die erop volgen een vrij gelijkmatig toenemend c.q. afnemend aantal. Bij de jurisprudentieteksten is de klasse met de meeste verschillende frequenties klasse 32, maar de aantallen dalen op dat punt (na klasse 29) al weer zodat geen sprake is van een symmetrische opbouw. Afgezien hiervan vertonen de frequentieverdelingen van de gereduceerde corpora veel overeenkomst met die van de volledige corpora (zie de tabellen 4-2 en 5-2).

De volgende stap is het vergelijken van de wijze waarop woordtypen en -tokens zijn verdeeld over de frequentieklassen. Hiervoor zal gebruik worden gemaakt van een tweetal grafieken. Als eerste komt daarbij de verdeling van de aantallen woordtypen aan bod. Deze aantallen kunnen, afgezet tegen de frequentieklasse-middens, samen in een enkele grafiek worden afgebeeld. Aangezien de aantallen vrij dicht bij elkaar liggen levert dit een op het eerste gezicht niet zo duidelijk beeld op. Het trekken van lijnen tussen de datapunten biedt in dit geval geen oplossing, aangezien de lijnen elkaar bij de hoogste en laagste klassen veelvuldig zouden kruisen. Vandaar dat er in dit geval voogekozen is gebruik te maken van verschillende 'markers' voor ieder corpus; zie de legende onder de grafiek.



Afb. 7-8 Aantal woordtypen per frequentieklasse voor de drie corpora

Wanneer we de grafiek bekijken valt het volgende op:

- Er lijkt veel overeenstemming te bestaan tussen het patroon dat de datapunten van de gereduceerde juridische corpora in deze grafiek vertonen, en dat bij de volledige corpora in hoofdstuk 4 en 5. Wel is het zo dat het min of meer rechtlijnig verlopende middenstuk van de grafiek nu wat korter lijkt te zijn, en dat de onregelmatigheden aan het begin en einde groter zijn. Het geringere aantal typen per klasse is hier waarschijnlijk debet aan.

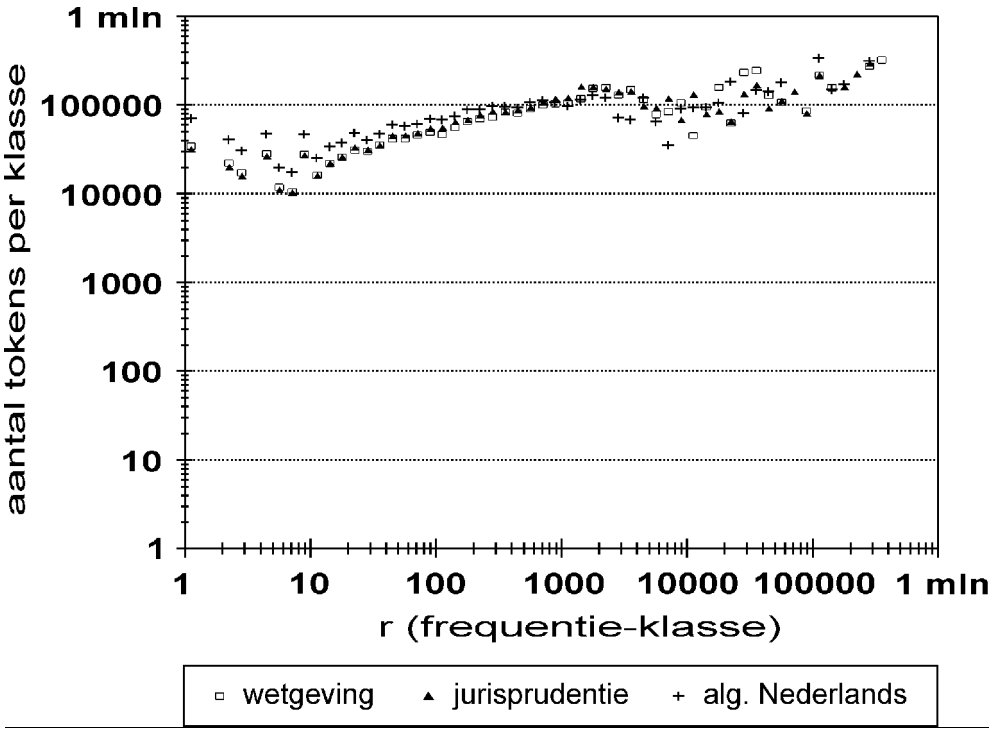
- De aantallen typen per klasse liggen voor de twee juridische corpora bijna steeds dicht bij elkaar, behalve bij de hoogste frequentieklassen (ongeveer vanaf klasse 37, die een klassemidden van 4496 heeft). Dit is overigens tevens het punt waar de *kop* van de frequentielijsten ongeveer begint.

- Ongeveer tot klasse 30 (klassemidden: 897) is het aantal typen per klasse in het corpus algemeen Nederlands steeds ruimschoots hoger (in klasse 1 zelfs ruim dubbel zo hoog) dan in de juridische corpora. Het verschil neemt echter tot deze klasse langzaam af, en het aantal typen is in de zes klassen die daarop volgen zelfs consequent lager dan bij de juridische corpora. Daarna wordt de situatie minder overzichtelijk, aangezien de aantallen voor alle corpora sterker fluctueren, maar in ieder geval is slechts in 3 van de overige 20 klassen het aantal typen in het corpus algemeen Nederlands het hoogst. Dat betekent dat veel van de 'extra' woordtypen in het corpus algemeen Nederlands (het corpus bevat er immers bijna twee keer zo veel als de twee juridische corpora) te vinden zijn in de eerste 30 frequentieklassen, wat dus wil zeggen dat deze typen voor het merendeel een lage frequentie hebben.

- Bij de eerste 10 frequentieklassen is weliswaar sprake van een wat onregelmatig verloop, maar de verhoudingen tussen de corpora blijven vrijwel gelijk. Het onregelmatige verloop wordt bijna zeker veroorzaakt door de gekozen indeling van de frequentieklassen. Doordat de klassen hier nog smal zijn en de klassegrenzen niet precies samenvallen met de weinige verschillende frequenties die er hier zijn kan het voorkomen dat een klasse juist twee frequenties omvat, en de klasse daarnaast geen enkele. Een remedie hiertegen zou zijn het werken met bredere klassen (wat ten koste van de precisie van de frequentieverdeling gaat), of eventueel het gebruik maken van een niet-logaritmische frequentieverdeling (wat helaas leidt tot een scheve verdeling van de verschillende frequenties over de klassen).

- Bij de hoogste frequentieklassen (ongeveer vanaf klasse 37), overeenkomend met de kop van de frequentielijsten, lopen de gegevens van de corpora enigszins door elkaar. Al in een van de voorgaande hoofdstukken is erop gewezen dat dit waarschijnlijk veroorzaakt wordt door het *geringe aantal typen per klasse*. Deze klassen bevatten namelijk nog uitsluitend *unieke frequenties*, wat betekent dat het aantal verschillende frequenties in een klasse gelijk is aan het aantal woordtypen. Een consequentie hiervan is echter dat het lastig is aan de hand van *deze* frequentieverdeling conclusies te trekken over verschillen tussen de aantallen woordtypen met de hoogste frequenties in elk corpus.

Ook de aantallen woordtokens per frequentieklasse kunnen in een enkele grafiek worden afgebeeld (zie afbeelding 7-9). Zaken die daarbij opvallen zijn:



Afb. 7-9 Aantal woordtokens per frequentieklasse voor de drie corpora

- Voor de overeenkomst tussen de gereduceerde en de volledige juridische corpora geldt hetzelfde als hierboven bij de woordtypen: een korter gebied waarin sprake is van een min of meer rechtlijnig verloop, en meer onregelmatigheden aan kop en staart. Ook hier is de meest voor de hand liggende verklaring daarvoor het in totaal geringere aantal tokens, waardoor de 'wet van de grote getallen' in iedere frequentieklasse minder sterk werkt.

- Het meeste dat voor de aantallen typen geldt, is ook hier van toepassing. Zo bevat ook hier het corpus algemeen Nederlands consequent meer tokens in de laagste 30 frequentieklassen, en in de meeste gevallen minder tokens in de klassen die daarna komen.

- Wat minder duidelijk zichtbaar is dat de juridische corpora 1 respectievelijk 2 niet-lege frequentieklassen méér bevatten dan het corpus algemeen Nederlands, hetgeen eveneens bijdraagt tot een hoger aantal tokens aan de kop van de frequentielijst.

Al met al komt uit vergelijking van de frequentieverdelingen ongeveer hetzelfde beeld naar voren als uit bijvoorbeeld de cumulatieve frequenties: in de juridische corpora omvatten de typen aan de kop van de frequentielijst een groter aantal woordtokens dan het corpus algemeen Nederlands. Bij dat corpus bevatten juist de laagste frequentieklassen een duidelijk hoger aantal tokens. Daaraan kunnen nu echter nog twee elementen worden toegevoegd. Allereerst kan nu nauwkeuriger het rangnummer c.q. de frequentieklasse - worden bepaald waar bovengenoemde zaken zich voordoen. Nog belangrijker is echter dat nu tevens vaststaat dat wat geldt voor de woord*tokens*, ook opgaat voor de woord*typen*. Ook daarvan komen er meer voor in de hoogste frequentieklassen, de klassen 31 en hoger bevatten bij de wetteksten 539 typen, bij de jurisprudentieteksten 582 typen en bij het corpus algemeen Nederlands 442 typen. Dit terwijl het corpus algemeen Nederlands in totaal ongeveer 2 keer zoveel typen bevat dan de juridische corpora; de 'extra' typen bevinden zich allemaal in de laagste frequentieklassen. Meer typen (zeker relatief gezien) aan de kop van de frequentielijst van de juridische corpora dus, die samen ook een groter aantal woordtokens (en dus een groter deel van het hele corpus) vertegenwoordigen. Opmerkelijk in dit verband is dat het aantal *verschillende frequenties* in het corpus algemeen Nederlands toch niet hoger is dan in de twee juridische corpora.

7.4.5 De verhouding tussen aantallen tokens en typen

Token-type ratio

Een duidelijk punt van verschil tussen de volledige corpora is de (overall) token/type ratio. Deze loopt uiteen van 151,18 bij de jurisprudentieteksten, via 89,47 bij de wetgevingsteksten tot 32,31 bij het corpus algemeen Nederlands. In de voorgaande hoofdstukken werd echter al vastgesteld dat de corpusgrootte op dit gegeven een sterke invloed uitoefent. Bij de gereduceerde corpora bedraagt de token/type ratio 58,78 (wetgeving) respectievelijk 61,71 (jurisprudentie). Dat betekent dat van de drie nu even grote corpora de twee juridische een token/type ratio hebben die dicht bij elkaar ligt. Deze is echter bijna twee maal zo hoog als in het corpus algemeen Nederlands, dat immers ook bijna twee maal zoveel typen bevat. In de voorgaande subparagraaf bleek dat de meeste van deze

'extra' typen te vinden zijn in de laagste frequentieklassen. Interessant is hier met name dat de token/type ratio in de juridische corpora zo dicht bij elkaar ligt.

Wanneer we de token/type ratio van de gereduceerde corpora vergelijken met de waarden voor de steekproeven die in de voorgaande hoofdstukken uit de corpora werden genomen blijkt in beide gevallen sprake te zijn van een hoge mate van overeenkomst. Het gereduceerde corpus wetgevingsteksten ligt qua omvang het dichtst bij de steekproef van 7000 documenten (zie tabel 4-3); de token/type ratio bedraagt daar 58,85, een verschil van slechts 0,07. Bij de jurisprudentieteksten komt de omvang van de steekproef van 2000 documenten het dichtst bij die van het gereduceerde corpus met een token/type ratio van 61,49, een verschil van 0,22. Ook op dit punt vertonen de in dit hoofdstuk samengestelde steekproeven dus karakteristieken die overeenkomen met de bevindingen in de voorgaande hoofdstukken.

De karakteristiek K

In de voorgaande hoofdstukken werd in dezelfde paragraaf als de token/type ratio ook steeds behandeld de karakteristiek *K* van Yule/Herdan, vanwege de mogelijkheden die deze karakteristiek (met name volgens Yule) zou hebben om de woordenschat (het aantal typen) in een corpus te berekenen. Zoals al eerder werd geconcludeerd is haar nut op dat punt echter beperkt. Wanneer we bijvoorbeeld de gereduceerde juridische corpora bekijken, blijkt dat deze een vrijwel gelijke token/type ratio hebben terwijl toch de waarde van *K* aanzienlijk verschilt, van 0,0128 bij de wetgevingsteksten tot 0,0111 bij de jurisprudentieteksten. Een waardevolle eigenschap van deze karakteristiek is echter wel dat deze praktisch onafhankelijk is van de afmetingen van een corpus, wanneer het aantal tokens daarin tenminste een zeker minimum bedraagt. Onder dit minimum kan de waarde van *K* nog steeds worden bepaald, maar deze fluctueert dan wat meer zodat de nauwkeurigheid geringer is. Bij de drie onderzochte corpora bleek dat bij steekproeven van zo'n 3 miljoen tokens en meer nog maar zeer weinig fluctuatie optrad. Wanneer dit ook bij andere corpora het geval is (en de kans daarop lijkt groot, gezien het uiteenlopende karakter van de hier onderzochte corpora) kan de genoemde grootte daarvoor eveneens als streefgetal gelden.

Het is dus zo dat voor de drie hier onderzochte corpora, zowel in volledige als in gereduceerde vorm, waarden voor *K* gelden die duidelijk van elkaar verschillen. Bij de wetgevingsteksten ligt de waarde rond de 0,0128 (berekend volgens de methode van Herdan), bij de jurisprudentieteksten is dat 0,0111 en bij de algemeen Nederlandse teksten 0,0106 (zie de voorgaande drie hoofdstukken). De vraag is nu welke conclusies daaruit getrokken kunnen worden ten aanzien van het woordgebruik. In de voorgaande hoofdstukken werd al geconstateerd dat het verschil in waarde tussen de juridische corpora niet samengaat met een andere token/type ratio, hetgeen in dit hoofdstuk nog eens is bevestigd bij het vergelijken van de gereduceerde corpora. In de hoofdstukken 2 en 4 werd daarnaast beschreven wat de karakteristiek volgens Herdan inhoudt, namelijk dat deze aangeeft in welke mate frequenties in een corpus 'van het gemiddelde afwijken'. Een hogere waarde staat daarbij voor méér of sterker van het gemiddelde afwijkende frequenties. Laten we dit eens in detail bekijken. De karakteristiek *K* is blijkens formule 4-5 gelijk aan

$$K = \frac{S_2}{S_1^{\,2}} = \frac{v_r^{\,2}}{N} \qquad (7\text{-}13)$$

waarbij v_r staat voor de variatie-coëfficiënt van r en N voor het aantal typen in het corpus. De variatie-coëfficiënt is gelijk aan $\sigma_r / \bar{r}$, het quotiënt van standaarddeviatie en gewogen gemiddelde van de frequenties. Wanneer we de gegevens waarmee K kan worden berekend voor de drie corpora op een rij zetten ziet dat er als volgt uit:

Corpus	Aantal typen (N)	σ_r	$\bar{r}$	v_r	v_r^2	K
Wetgeving	78574	1870,14	58,78	31,81	1011,88	0,0128
Jurisprudentie	74827	1783,84	61,71	28,90	835,21	0,0111
Algemeen Nederlands	142879	1257,10	32,31	38,91	1513,99	0,0106

Tabel 7-6 Gegevens die een rol spelen bij de berekening van K

Voor de twee juridische corpora, die ongeveer hetzelfde aantal woordtypen bevatten, is vergelijken vrij eenvoudig. Aangezien ook de gemiddelde frequenties dicht bij elkaar liggen moet het verschil tussen de variatie-coëfficiënten vrijwel geheel worden veroorzaakt door de *standaarddeviatie van de frequenties*. De hogere waarde van deze standaarddeviatie bij het corpus wetgeving betekent daarbij dat in dat corpus wat grotere verschillen bestaan tussen de aantallen woordtypen die dezelfde frequentie hebben. Wanneer we nog even terugkijken naar afbeelding 7-8 is daarvan ook in de geclassificeerde frequentieverdeling wel iets terug te vinden, de aantallen typen lijken daar bij de wetgevingsteksten vooral in de hoogste klassen (ongeveer vanaf r = 10000) meer te fluctueren dan bij de jurisprudentieteksten. Bij het corpus algemeen Nederlands ligt de beoordeling van de gegevens wat lastiger, aangezien *alle* variabelen verschillen vertonen. Wanneer we deze gegevens echter vergelijken met die van de wetteksten kan de redenering als volgt zijn: om eenzelfde waarde voor K te verkrijgen als bij de wetteksten zou, gegeven dat het aantal typen 1,82 maal hoger is dan bij de wetteksten, ook v_r^2 twee maal zo hoog moeten zijn. v_r zou dan $\sqrt{1,82}$ = 1,35 maal zo hoog moeten zijn. Daarvoor zou σ_r moeten stijgen tot 1695, of $\bar{r}$ moeten dalen tot 23,7. Aangezien de waarde van $\bar{r}$ nu al erg laag is (in vergelijking met de andere corpora) lijkt een waarde die nog ⅓ lager ligt geen erg realistische optie. Er lijkt daarom iets voor te zeggen dat de verschillende waarden voor K wederom vooral kunnen worden verklaard door de *lagere standaarddeviatie van de frequenties* bij het corpus algemeen Nederlands, wat dus betekent: minder verschillen tussen de aantallen woordtypen met dezelfde frequentie. Dit zou wel eens kunnen samenhangen met het feit dat de woordtypen in dit corpus minder geconcentreerd zijn aan de 'kop' van de frequentielijst, maar juist gelijkmatig zijn gespreid over de 'staart' (zie opnieuw afbeelding 7-8).

De constanten R en C

In paragraaf 7.3 werd bij de beoordeling van de twee gereduceerde corpora vastgesteld dat deze voor wat betreft de aantallen tokens en typen voldoen aan dezelfde voorwaarden als de volledige corpora. Het verband tussen aantal tokens en typen kan, met een geringe

foutmarge, worden weergegeven door dezelfde vergelijking van het type $V = R \,.\, N^c$ als bij de volledige corpora. Het is bij de hier onderzochte corpora dus mogelijk gebleken de waarde van de constanten *R* en *C* zodanig vast te stellen dat deze voor het gehele corpus en voor steekproeven van verschillende omvang daaruit gelden. Nu dat het geval is, is het zinvol om nog even stil te staan bij de *verschillende waarden* die deze constanten bij de drie corpora bleken te hebben, en na te gaan wat daarvan de betekenis is.

Gereduceerd Corpus	*R*	*C*
Wetgeving	13,2979	0,5675
Jurisprudentie	9,8383	0,5827
Algemeen Nederlands	6,9118	0,6484

Tabel 7-7 Waarde van de constanten *R* en *C* voor de drie corpora

De waarden, in tabel 7-7 nogmaals opgesomd, zijn afgeleid van de vergelijking van de lineaire regressielijn van log *V* op log *N* (*V* staat daarbij voor aantal typen, *N* voor aantal tokens). Deze vergelijkingen luiden:

$$\log V = 0{,}5675 \log N + 1{,}1237 \qquad (7\text{-}14)$$

voor het corpus wetgevingsteksten,

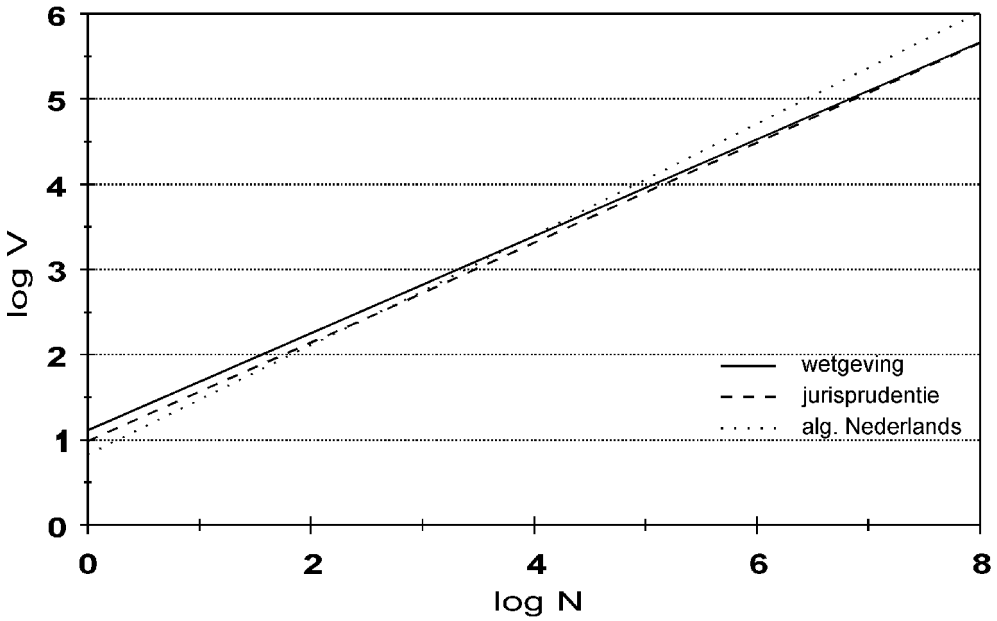
$$\log V = 0{,}5827 \log N + 0{,}9929 \qquad (7\text{-}15)$$

voor het corpus jurisprudentieteksten en

$$\log V = 0{,}6484 \log N + 0{,}8396 \qquad (7\text{-}16)$$

voor het corpus algemeen Nederlands (zie hoofdstuk 4, 5 en 6). Wanneer de lijnen samen in één grafiek worden getekend lijken de verschillen op het eerste gezicht niet eens zo groot (zie afbeelding 7-10). Het meest opvallende kenmerk is nog dat de lijn van het corpus algemeen Nederlands duidelijk sterker stijgt dan die van de juridische corpora; daarnaast stijgt de lijn van het corpus jurisprudentieteksten iets sterker dan die van de wetgevingsteksten. Daaruit kan worden afgeleid, zoals in de hoofdstukken 5 en 6 al werd aangegeven, dat het aantal woordtypen bij kleine steekproeven het hoogst zal zijn in het corpus wetgevingsteksten. De aantallen typen in de andere corpora liggen aanvankelijk lager, maar stijgen sneller bij een toename van het aantal tokens (dat wil zeggen, bij grotere steekproeven). Door substitutie kunnen de snijpunten van de drie regressielijnen worden vastgesteld. Daarbij blijkt dat de lijn van de jurisprudentieteksten die van de wetgevingsteksten pas snijdt bij een aantal van bijna 384 miljoen tokens. Alleen bij even grote corpora met meer dan dat aantal tokens zou het aantal typen in het corpus jurisprudentieteksten hoger zijn. Dat aantal tokens wordt hier nooit bereikt, waardoor bij steekproeven van een gelijk aantal tokens uit deze corpora, het corpus wetgevingsteksten steeds de meeste woordtypen zal bevatten. Anders ligt de zaak bij het corpus algemeen

Nederlands. Dat begint met het laagste aantal typen per token (dus met de hoogste token/type ratio), maar dit aantal stijgt sneller dan bij de juridische corpora, als gevolg van de hogere waarde van de exponent *C*. De regressielijn van dit corpus snijdt die van de jurisprudentieteksten al bij 216 tokens, en die van de wetgevingsteksten bij 3245 tokens. Steekproeven van meer dan dat aantal tokens uit het corpus algemeen Nederlands zullen dus steeds een hoger aantal woordtypen bevatten dan even grote steekproeven uit de andere twee corpora. Alle tot nu toe onderzochte steekproeven uit dit corpus (bijvoorbeeld in het vorige hoofdstuk) liggen ruimschoots boven deze grens, en vertonen dus ook allemaal een lagere token/type ratio dan steekproeven van overeenkomstige afmetingen uit de andere corpora.



Afb. 7-10 De regressielijn van log *V* op log *N* voor de drie corpora

Resumerend kan worden gesteld dat door het analyseren van het verband tussen aantallen woordtokens en -typen aan de hand van de formule van Erikstad $V = R \, . \, N^{c}$ nu ook veel meer duidelijkheid bestaat over de wijze waarop de token/type ratio samenhangt met de afmetingen van de corpora, of van steekproeven daaruit.

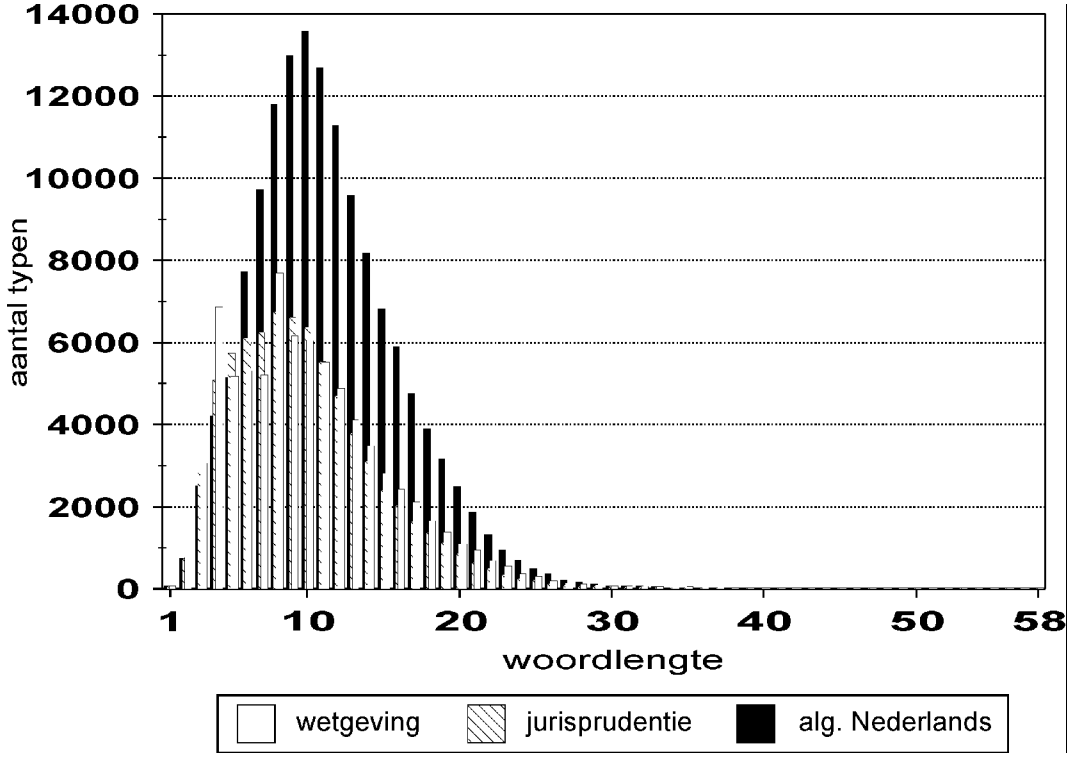
7.4.6 Woordlengtes

Bij het vergelijken van de woordlengtes in de corpora zal, net als in de voorgaande hoofdstukken, weer aan twee aspecten aandacht worden besteed. In de eerste plaats zullen de woordlengte*verdelingen* van de gereduceerde juridische corpora en van het corpus algemeen Nederlands worden vergeleken, en in de tweede plaats zal voor de drie corpora het verband tussen rangnummer en woordlengte worden geanalyseerd.

Woordlengte-verdelingen

In de voorgaande hoofdstukken kwamen al enkele verschillen tussen de woordlengte-verdelingen van de corpora aan het licht. Deze verschillen bleken het grootst te zijn bij de verdeling van de woord*typen*, waar de juridische corpora opvallend veel meer typen van 4, 5 of 6 karakters bleken te bevatten. Een mogelijke verklaring daarvoor werd in hoofdstuk 6 gegeven: de genoemde lengte-klassen blijken bij de juridische corpora een veel hoger percentage *getallen* te bevatten. Aan de hand van de gereduceerde corpora kan nu geverifieerd worden of dit verschijnsel zich ook voordoet wanneer de corpora qua omvang zoveel mogelijk aan elkaar gelijk zijn gemaakt, en wanneer dat het geval is, of dezelfde verklaring van toepassing is.

Verschillen tussen woordlengte-verdelingen kunnen het duidelijkst zichtbaar gemaakt wordt met behulp van een grafiek. In de voorgaande hoofdstukken werd daarbij steeds gebruik gemaakt van staafdiagrammen, waarbij elke lengte-klasse werd voorgesteld door een afzonderlijk balkje. Wanneer we voor de gereduceerde corpora net zulke diagrammen samenstellen, en deze samen met de gegevens van het corpus algemeen Nederlands combineren in een enkele grafiek ontstaat echter geen duidelijk beeld (zie afbeelding 7-11). De woordlengte-verdelingen zijn aan de hand van dit staafdiagram slechts met moeite te vergelijken. Eén reden daarvoor is dat de balkjes die bij de afzonderlijke corpora horen nauwelijks meer van elkaar te onderscheiden zijn; een staafdiagram is geen geschikte grafiekvorm wanneer het aantal datapunten (balkjes) te groot wordt. Daarnaast is echter met name bij deze verdeling van woord*typen* een probleem dat de aantallen woordtypen in het corpus algemeen Nederlands zo veel hoger zijn dan in de juridische corpora. Daardoor is nog steeds niet goed te zien welke woordlengtes in een corpus vaker, even vaak of juist minder vaak voorkomen.

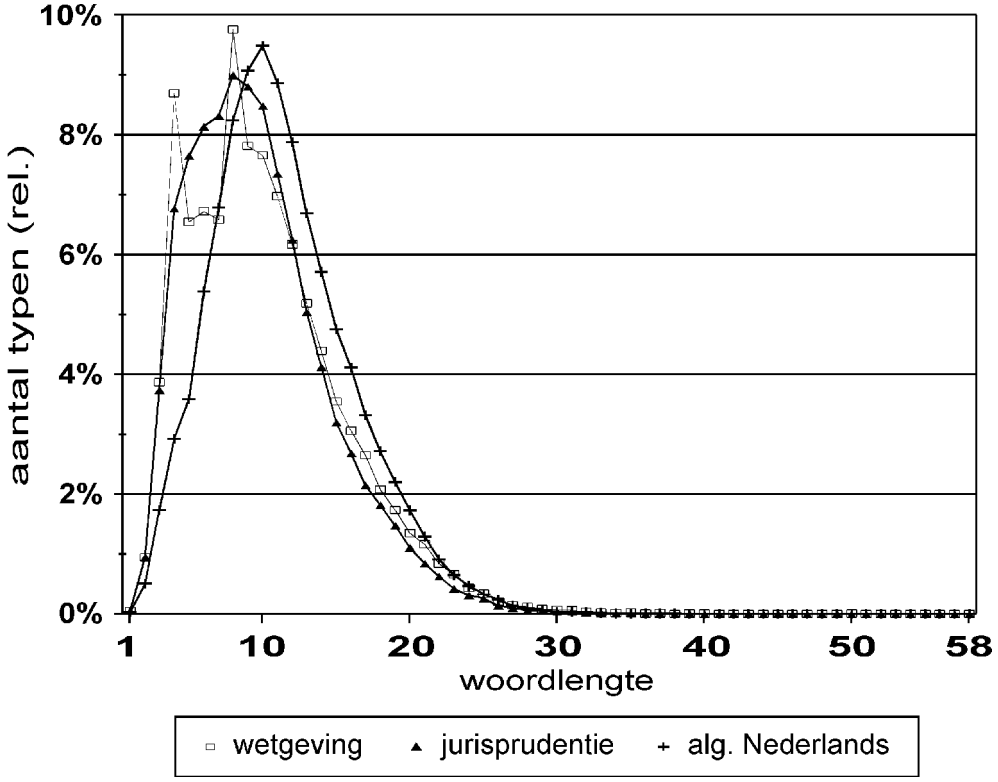


Afb. 7-11 De staafdiagrammen van de lengteverdeling van woordtypen gecombineerd

Vanwege deze twee problemen zijn de grafieken van de woordlengte-verdelingen in dit hoofdstuk anders samengesteld. In de eerste plaats is gekozen voor een ander grafiek-type, het lijn-diagram. Hierin zijn de aantallen typen en tokens met een bepaalde lengte aangegeven door middel van een enkel datapunt. Voor de duidelijkheid zijn alle datapunten van een corpus daarbij verbonden door een lijn.

Een tweede verandering is van fundamenteler aard. Om te compenseren voor de verschillende aantallen typen en tokens in de corpora (hoewel deze verschillen voor wat betreft de tokens slechts gering zijn) zijn in dit geval woordlengte-verdelingen samengesteld gebaseerd op *relatieve* aantallen typen en tokens. De aantallen typen en tokens per lengte-klasse zijn daarbij gedeeld door het totaal aantal typen c.q. tokens in het corpus. Wanneer gebruik wordt gemaakt van een normale schaalverdeling ontstaan dan grafieken met behulp waarvan de woordlengte-verdelingen wel redelijk goed kunnen worden vergeleken (zie afbeelding 7-12a en b). Het gebruik maken van relatieve aantallen heeft wel wat nadelen (zie ook paragraaf 7.2.1), maar die wegen in dit geval wat minder zwaar dan bij de volledige corpora, vooral omdat de verschillen tussen de absolute aantallen veel minder groot zijn (minder dan een factor 2). Voor de woordtokens is het bezwaar zelfs nog minder groot, aangezien de drie corpora op dit punt juist zoveel mogelijk aan elkaar gelijk gemaakt zijn.

Wanneer we de lengte-verdeling van de *woordtypen* bekijken blijkt dat de gereduceerde juridische corpora evenals de volledige een duidelijk ander patroon vertonen dan het corpus algemeen Nederlands.



Afb. 7-12a Woordlengte-verdeling (typen)

Dit geldt vooral voor de lengteklassen tot en met 10 karakters, hoewel ook de klassen daarna bij het corpus algemeen Nederlands afwijken van de andere corpora in die zin dat ze een aanmerkelijk hoger percentage woordtypen bevatten. Bij vergelijking van de gereduceerde en volledige juridische corpora vallen nog de volgende punten op:

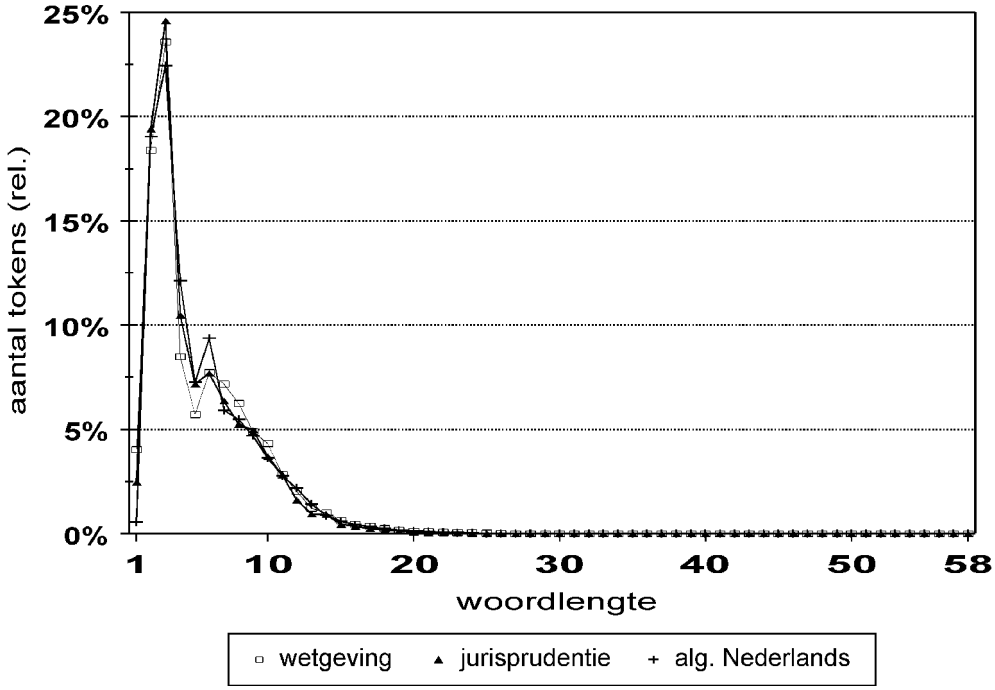
- Bij het corpus wetgevingsteksten telt lengteklasse 4 wat meer typen, terwijl de klassen 5 en 6 er juist wat minder bevatten. Voor het overige lijken de verdelingen sterk op elkaar, nog steeds is er een zeer duidelijke piek bij lengteklasse 8.

- De verdeling van de jurisprudentieteksten wijkt meer af van die van het volledige corpus. Vooral de lengteklassen 5 en 6, die bij het volledige corpus een duidelijke piek vertoonden, zien er nu anders uit. Toch bevatten deze klassen samen met klasse 4 nog steeds een bijna even groot percentage van de typen: 22,5%, tegen 23,5% bij het volledige corpus. Vooral klasse 4, die ruim een procent van de typen meer bevat, maakt hier een deel van het verlies in klasse 5 goed. Voor het overige heeft de verdeling weer min of meer dezelfde vorm als bij het volledige corpus.

De lengteklassen 4, 5 en 6 bevatten ook bij de gereduceerde juridische corpora een hoger percentage van alle typen dan bij het corpus algemeen Nederlands: 22,0% en 22,6% tegen 11,9%. Net als bij de volledige corpora vormen in deze klassen de *getallen* een duidelijk punt van verschil; ze zijn sterk oververtegenwoordigd in deze klassen. Bij de wetgevingsteksten zijn er 17260 typen van 4, 5 of 6 karakters, waarvan 9599 getallen zijn, een percentage van 55,6%. In totaal zijn in dit corpus 19,8% van alle typen getallen. Bij de jurisprudentieteksten zijn 6021 van de 16876 typen van 4, 5 of 6 karakters getallen, wat neerkomt op 35,7%, terwijl dat percentage voor alle klassen samen op 11,06% ligt. De waarden voor het corpus algemeen Nederlands zijn natuurlijk ongewijzigd: 9,1% getallen in de lengteklassen 4, 5 en 6, tegen 2,1% in het hele corpus. Het percentage getallen in het gereduceerde corpus jurisprudentieteksten is duidelijk gedaald ten opzichte van het volledige corpus, waar het 49,9% bedroeg; dat in het gereduceerde corpus wetgevingsteksten is ook iets gedaald, maar veel minder (het was 59,7%). Het dalen van beide percentages zou toeval kunnen zijn, wellicht is het zo dat de getallen wat ongelijk over de corpora zijn gespreid. In ieder geval is het zo dat het percentage getallen, die bij alle corpora toch al oververtegenwoordigd zijn in de lengteklassen 4, 5 en 6, bij de juridische corpora (volledig zowel als gereduceerd) veel hoger is dan in het corpus algemeen Nederlands.

Een ander duidelijk verschil tussen de lengteverdelingen is dat het corpus algemeen Nederlands een hoger percentage typen bevat in de klassen 9 tot en met 21. Daarnaast is de klasse met de meeste typen (klasse 10) bij dit corpus een hogere dan bij de juridische corpora. Het lijkt daardoor alsof de hele lengteverdeling van de juridische corpora naar links is verschoven ten opzichte van die van het corpus algemeen Nederlands. Wanneer typen van de meest voorkomende lengte (10 karakters) worden bekeken, valt daarin eigenlijk niets speciaals op, het is niet zo dat een bepaald soort typen erg vaak voorkomt (zoals het geval was met de getallen). Wel blijkt dat de typen van deze lengte in het corpus algemeen Nederlands gemiddeld een veel lagere frequentie hebben, namelijk 12,4 tegen 27,0 bij de jurisprudentieteksten en 33,2 bij de wetteksten.

Een lengteverdeling kan ook worden opgesteld voor alle woord*tokens* die in elk van de corpora voorkomen (zie afbeelding 7-12b). Met behulp daarvan kan om te beginnen worden vastgesteld dat lengteklasse 10 niet het hoogste percentage tokens bevat bij het corpus algemeen Nederlands maar bij de wetgevingsteksten.



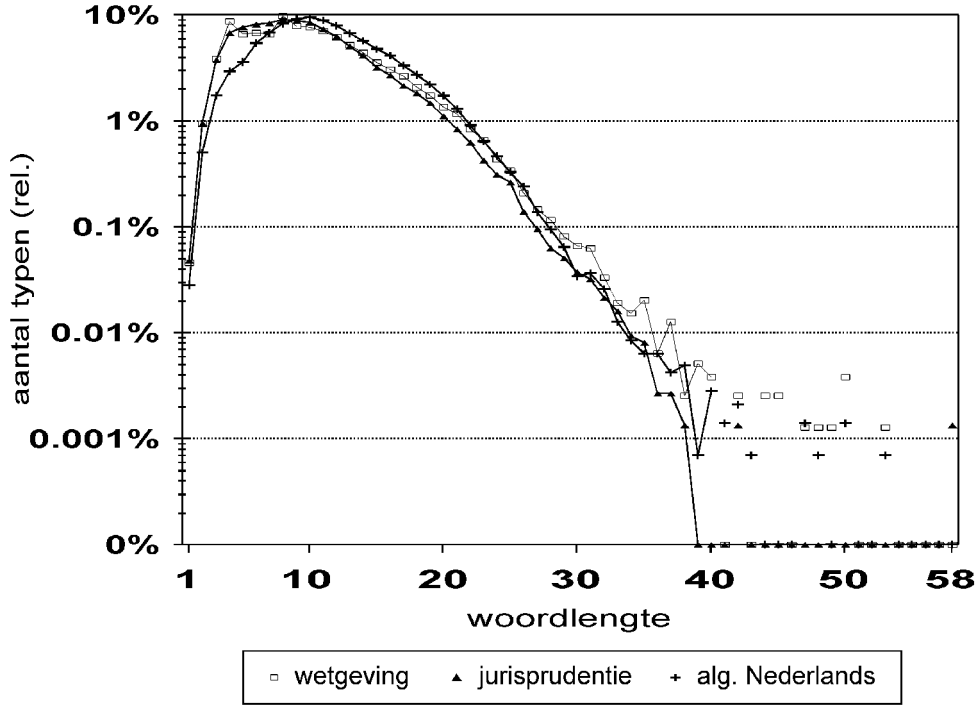
Afb. 7-12b Woordlengte-verdeling (tokens)

De lengteverdelingen van de woordtokens lijken voor de drie corpora overigens veel op elkaar, veel meer dan die van de woordtypen. De percentages woordtokens per klasse vertonen wel verschillen, maar deze vormen geen duidelijk patroon. De grootste verschillen treffen we aan in de lengteklassen 1 en 4. Het hogere percentage tokens met lengte 1 in de juridische corpora kan voor een deel weer worden toegeschreven aan de aanwezigheid van meer getallen (deze maken 59,6% respectievelijk 44,1% van de tokens met deze lengte uit bij de juridische corpora, tegen 37,1% bij het corpus algemeen Nederlands), maar ook veel (los voorkomende) letters, met name die aan het begin van het alfabet, hebben een hogere frequentie. Dit zou het gevolg kunnen zijn van het voorkomen van meer 'opsommingen' in de juridische teksten, waarbij elk element wordt aangeduid met een letter. Een laatste opvallend punt in deze lengteklasse is nog dat de letter U in het corpus algemeen Nederlands een erg hoge frequentie heeft (7188), maar in de juridische corpora niet (524 respectievelijk 483). Net als bij de volledige corpora bevat lengteklasse 5 weer de minste tokens, waarvoor in de voorgaande hoofdstukken als verklaring werd gegeven dat 5 karakters precies de grens is tussen één- en tweelettergreepswoorden.

Onder andere om ook de verdeling van de langste woordtypen en -tokens te kunnen beoordelen zijn hieronder beide lengteverdelingen nogmaals afgedrukt, met een logaritmische verdeling op de verticale as. Aan de rechterzijde van de verdeling, waar het percentage typen in elk corpus fluctueert en ook nogal wat lege klassen voorkomen, vormen lijnen

tussen de datapunten eerder een belemmering dan een hulpmiddel bij het

volgen van de verdeling van een bepaald corpus. Daarom zijn deze lijnen vanaf klasse 40 weggelaten.

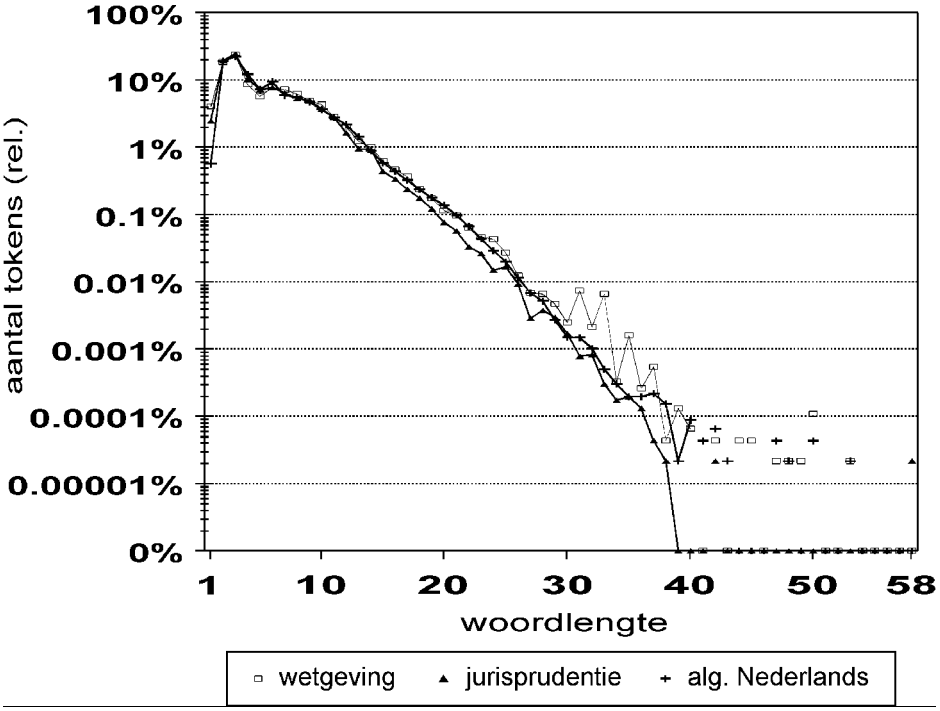


Afb. 7-13a Woordlengte-verdeling (typen, logaritmisch)

In de lengteverdeling van woordtypen is nog iets beter dan in de niet-logaritmische versie zichtbaar dat de juridische corpora voor wat betreft de eerste 10 lengteklassen sterke overeenkomsten vertonen, en afwijken van het corpus algemeen Nederlands in de zin dat ze een aanmerkelijk hoger percentage typen bevatten. Verder blijkt nu dat de lengteverdeling van het corpus wetgevingsteksten ongeveer vanaf klasse 14 steeds verder afwijkt van die van de jurisprudentieteksten, en rond klasse 25 zelfs die van het corpus algemeen Nederlands kruist. Vanaf die klasse en in de meeste die daarop volgen bevatten de wetgevingsteksten het hoogste percentage typen (totaal 1,16%), hoewel het verschil met het corpus algemeen Nederlands (1,02%) niet erg groot is. Het corpus jurisprudentieteksten bevat hier duidelijk minder typen (0,74%), de verdeling van dit corpus telt ook de meeste lege klassen (18 stuks). Daarnaast bevat het echter ook het langste type van alle corpora samen. Het blijkt te gaan om het type DRIEHUNDERTFUNFUNDNEUNZIGTAUSENDVIERHUNDERTNEUNUNDDREISSIG (58 karakters, duidelijk met een typefout). Bij de klassen waarin maar enkele typen aanwezig zijn is toch weer een nadeel van het gebruik van relatieve waarden zichtbaar: het lijkt af en toe, bijvoorbeeld in klasse 53, of het corpus algemeen Nederlands minder typen in een klasse bevat; het absolute aantal typen is echter in beide gevallen 1, het relatieve aantal is lager door het grotere aantal typen in het gehele corpus. Uit de verhoudingen tussen de relatieve aantallen typen kunnen dan ook, vooral bij de klassen met de laagste waarden, niet zonder meer conclusies met betrekking tot de verhoudingen tussen absolute aantallen worden getrokken.

Bij de logaritmische verdeling van de woordtokens valt in de eerste plaats het hogere percentage tokens in het corpus wetgevingsteksten op, in de lengteklassen 28 - 37. De

klassen 31, 33, 35 en 37 vormen daarbij duidelijke pieken. Deze blijken respectievelijk vooral voor rekening te zijn van de typen ARBEIDSONGESCHIKTHEIDSUITKERING (frequentie: 223), ARBEIDSONGESCHIKTHEIDSVERZEKERING (253), ACCOUNTANTSADMINISTRATIECONSULENTEN (44) en RADIOTELEGRAFIEVEILIGHEIDSCERTIFICAAT (10). De eerste drie typen worden al in hoofdstuk 4 genoemd, ze veroorzaakten ook pieken in de woordlengte-verdeling van het volledige corpus. Een tweede opvallend punt is dat het corpus jurisprudentieteksten ongeveer vanaf klasse 14 vrijwel steeds een lager percentage tokens bevat dan de andere corpora.



Afb. 7-13b Woordlengte-verdeling (tokens, logaritmisch)

Naar aanleiding van de woordlengte-verdelingen kunnen de volgende conclusies worden getrokken:

- De gereduceerde juridische corpora vertonen in grote lijnen ongeveer dezelfde karakteristieken als de volledige. Een uitzondering hierop vormen de lengteklassen 5 en 6 bij de jurisprudentieteksten. Deze klassen vormen niet zoals bij het volledige corpus een afzonderlijke piek. Toch is het aantal typen in de klassen 4, 5 en 6 samen nog ongeveer even hoog.

- De juridische corpora bevatten een aanmerkelijk hoger percentage verschillende getallen. Deze zijn in de lengteklassen 4, 5 en 6 oververtegenwoordigd, hetgeen waarschijnlijk het hogere percentage typen in deze klassen in de juridische corpora voor een belangrijk deel verklaart.

- Met betrekking tot de woordtypen zijn verder de meest opvallende verschillen het hogere percentage typen bij het corpus algemeen Nederlands in de klassen 9 - 24, en bij de wetgevingsteksten in de klassen 26 en hoger.

- Met betrekking tot de woordtokens zijn verschillen aanwezig in klasse 1 (de juridische corpora bevatten een hoger percentage). Daarnaast bevat het corpus wetgevingsteksten een hoger percentage tokens in de klassen 28 - 37, en het corpus jurisprudentieteksten vanaf klasse 14 vrijwel voortdurend een lager percentage dan de andere corpora. Bij alle drie corpora (en ook bij de ongereduceerde juridische corpora) vertoont de lengteverdeling van de woordtokens een inzinking bij lengteklasse 5, waarschijnlijk te verklaren doordat dit precies de grens is tussen één en tweelettergrepige woorden.

- Samenvattend kan gesteld worden dat het corpus wetgevingsteksten een wat hoger percentage lange woorden (30 karakters en meer) bevat. Het gaat daarbij zowel om woordtypen als woordtokens. Het corpus jurisprudentieteksten bevat juist wat een wat lager percentage woordtokens van 14 karakters en meer dan de andere corpora. Van de twee juridische corpora kan daarom alleen van de wetgevingsteksten worden gezegd dat deze een wat hoger percentage lange woorden bevatten; het gaat daarbij dan echter alleen om woorden van 30 karakters en meer.

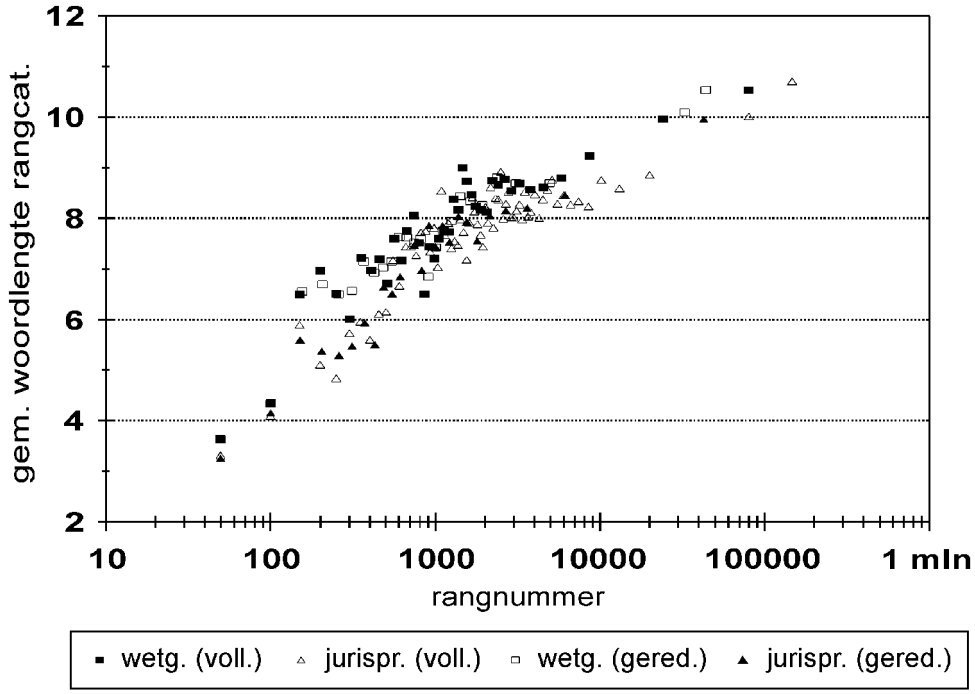
Het verband tussen het rangnummer en de lengte van woordtypen
In de voorgaande hoofdstukken bleek steeds een duidelijk verband te bestaan tussen het rangnummer en de lengte van woordtypen. De kop van de frequentielijst bevat veel korte woorden waardoor de gemiddelde lengte van de typen daar laag is, terwijl deze gemiddelde woordlengte bij de laagste rangnummers een stuk hoger blijkt te zijn. Om te zien of dit verschijnsel zich bij de drie corpora in gelijke mate voordoet, en of er wellicht verschillen zijn in het tempo waarmee de gemiddelde woordlengte stijgt zullen nu voor de gereduceerde juridische corpora en voor het corpus algemeen Nederlands de volgende gegevens worden vergeleken:

- de gemiddelde woordlengte gemeten over een bepaald interval van de frequentielijst, namelijk over 50 verschillende rangnummers, en
- het voortschrijdend gemiddelde van de woordlengtes, gemeten over telkens 200 woordtypen.

Bij de tweede methode zullen net als in de voorgaande hoofdstukken om praktische redenen en omdat de gemiddelde woordlengte in dat deel van de frequentielijst de sterkste ontwikkeling doormaakt alleen de eerste 5000 woordtypen van de frequentielijst van ieder corpus in aanmerking worden genomen.

Om te beginnen de interval-methode, waarbij steeds de gemiddelde lengte wordt berekend over een cluster van 50 opeenvolgende rangnummers. De gemiddelde lengte van elk cluster wordt afgezet tegen de hoogste frequentie die het cluster bevat. Om na te gaan hoe de gereduceerde corpora zich voor wat betreft dit kenmerk verhouden tot de volledige corpora, is allereerst een grafiek vervaardigd waarin de datapunten uit de grafieken 4-19 en 5-14 (die betrekking hebben op de volledige corpora) zijn opgenomen, met daarbij de datapunten die horen bij de twee gereduceerde corpora (zie afbeelding 7-14). Doordat de

punten op sommige plaatsen dicht bij elkaar liggen laat de overzichtelijkheid van vooral het middengedeelte van deze vergelijkingsgrafiek wat te wensen over. Toch zijn enkele overeenkomsten en verschillen tussen volledige en gereduceerde corpora wel zichtbaar.

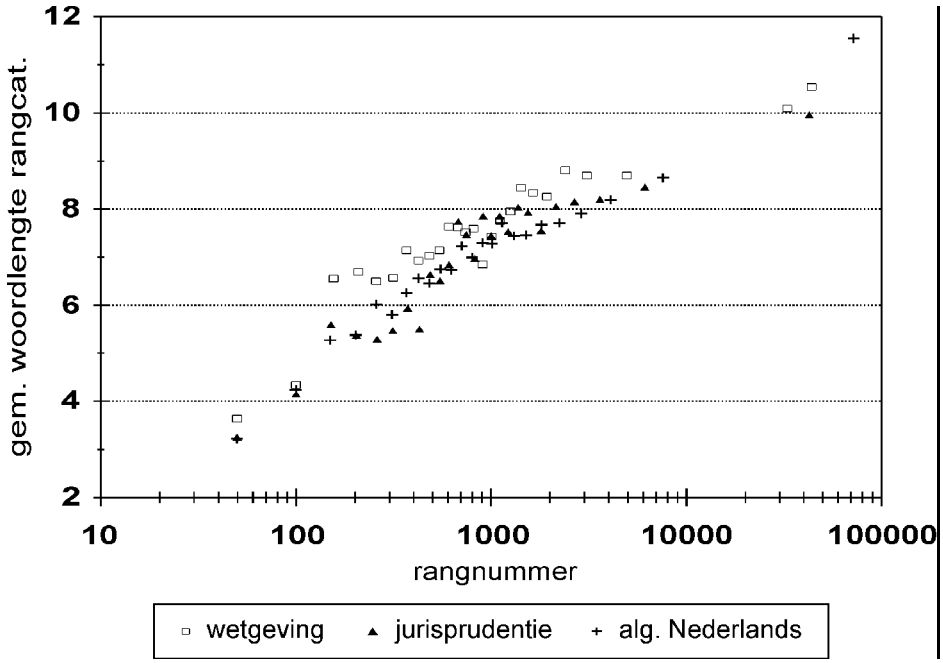


Afb. 7-14 Gemiddelde woordlengte per categorie van 50 rangnummers (volledige en gereduceerde corpora vergeleken)

Een verschil is bijvoorbeeld het geringere aantal datapunten voor beiden gereduceerde corpora, een gevolg van het eveneens geringere aantal verschillende rangnummers in deze corpora. Ondanks dat zijn er echter, vooral in het gedeelte tot rangnummer 1000, enkele vrij duidelijke overeenkomsten tussen de volledige en de overeenkomstige gereduceerde corpora zichtbaar, zoals bijvoorbeeld de praktisch gelijke gemiddelde lengte bij de twee laagste clusters (de datapunten vallen daar over elkaar heen) en de in alle corpora te herkennen sprong in de gemiddelde woordlengte tussen rangnummer 100 en 150. Ook is zowel bij de volledige als bij de gereduceerde corpora de gemiddelde woordlengte bij de wetgevingsteksten globaal steeds wat hoger dan bij de jurisprudentieteksten (een enkele uitbijter daargelaten). Een voorzichtige conclusie uit dit alles kan zijn dat de gereduceerde corpora voor wat betreft de gemiddelde woordlengtes globaal genomen een vrij sterke gelijkenis met de volledige corpora lijken te vertonen, hoewel bepaalde factoren, zoals het veel geringere aantal rangnummers en dus datapunten, vergelijken enigszins lastig maken.

Wanneer we de gegevens voor de gereduceerde corpora in één grafiek combineren met die van het corpus algemeen Nederlands ontstaat de grafiek uit afbeelding 7-15. Wanneer we de drie corpora vergelijken blijken vooral de gemiddelde lengtes in het corpus wetgevingsteksten boven die in de andere corpora te liggen. Het duidelijkst is dit het geval in het gebied ongeveer tot rangnummer 600 en bij rangnummers hoger dan 1000. Bij het corpus jurisprudentieteksten valt op dat de gemiddelde woordlengte aanvankelijk (bij de laagste rangnummers) erg laag is, maar rond rangnummer 600 snel stijgt en dan over een bepaald interval zelfs wat hoger is dan bij het corpus wetgevingsteksten. Bij het corpus

algemeen Nederlands stijgt de gemiddelde woordlengte het regelmatigst, maar deze ligt in de meeste gevallen onder die bij de wetgevingsteksten, en bij de hogere rangnummers ook onder die bij de jurisprudentieteksten. Daarop is één uitzondering: de gemiddelde woordlengte in het laatste cluster van dit corpus ligt hoger dan die in de hoogste clusters van de andere corpora. Wellicht zien we hier een gevolg van de wat afwijkende woordlengte-*verdeling* van dit corpus, met een groter aantal typen in de lengteklassen 9 - 24. Typen in die lengte-klassen zullen in de meeste gevallen geen hoge frequentie hebben (de lengte-verdeling van de woordtokens geeft dan ook geen hoger aantal tokens in deze klassen aan) zodat de kans groot is dat ze inderdaad in het laatste cluster uit grafiek 7-15 zijn vertegenwoordigd.

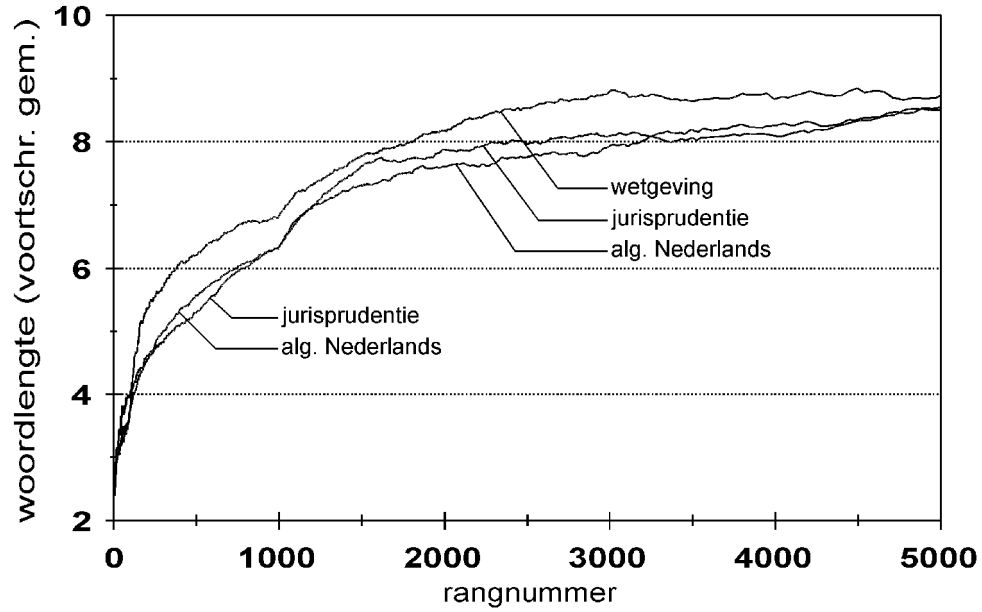


Afb. 7-15 Gemiddelde woordlengte voor de drie corpora, per categorie van 50 rangnummers

Afgezien van dit laatste punt verhouden de corpora zich ongeveer tot elkaar zoals men intuïtief zou verwachten: bij de juridische corpora, en vooral bij het corpus wetgevingsteksten, stijgt de gemiddelde woordlengte sneller bij toenemend rangnummer. Er komen in deze corpora dus meer typen voor die een grotere lengte paren aan een wat hogere frequentie. De verschillen zijn echter niet erg groot, en gemeten over de volledige corpora heeft het corpus algemeen Nederlands zelfs de grootste gemiddelde woordlengte.

Naast de gemiddelde woordlengte in clusters van rangnummers kan evenals in de voorgaande hoofdstukken het voortschrijdend gemiddelde van de lengte van woordtypen worden bestudeerd en voor de drie corpora worden vergeleken. In de voorgaande hoofdstukken werd dit voortschrijdend gemiddelde bepaald over telkens 200 opeenvolgende typen, dat wil zeggen, ieder datapunt vertegenwoordigde het gemiddelde van de lengte van een bepaald type en van de 199 daaraan voorafgaande typen. Het resultaat was een grafiek aan de hand waarvan nog vrij gedetailleerd de ontwikkeling van de woordlengte

bij toenemend rangnummer te volgen was. Nu zijn ook voor de gereduceerde corpora soortgelijke curves vervaardigd, en met de curve van het corpus algemeen Nederlands in één grafiek samengebracht. Daarbij is echter gebleken dat om de curves goed te kunnen vergelijken een wijziging noodzakelijk is, in de vorm van het verhogen van het aantal typen waarover het voortschrijdend gemiddelde wordt berekend. Het aantal van 200 leidt namelijk tot curves die, vooral bij de hogere rangnummers, enigszins grillig verlopen. Worden deze in één grafiek getekend, zelfs wanneer geen gebruik wordt gemaakt van datapunten maar alleen van lijnen, dan kruisen deze elkaar voortdurend, waardoor het volgen en vergelijken van de afzonderlijke curves niet meer mogelijk is. Om dit probleem te verhelpen is uiteindelijk, na enig proberen, gekozen voor een voortschrijdend gemiddelde over 1000 woordtypen. Dit geeft nog redelijk wat details, terwijl de curves voldoende gelijkmatig verlopen om ze met elkaar te kunnen vergelijken. Net als in de voorgaande hoofdstukken bevat de grafiek, onder andere omdat anders het aantal datapunten te groot zou worden, alleen de gegevens van de eerste 5000 rangnummers. De betreffende grafiek is te vinden in afbeelding 7-16.



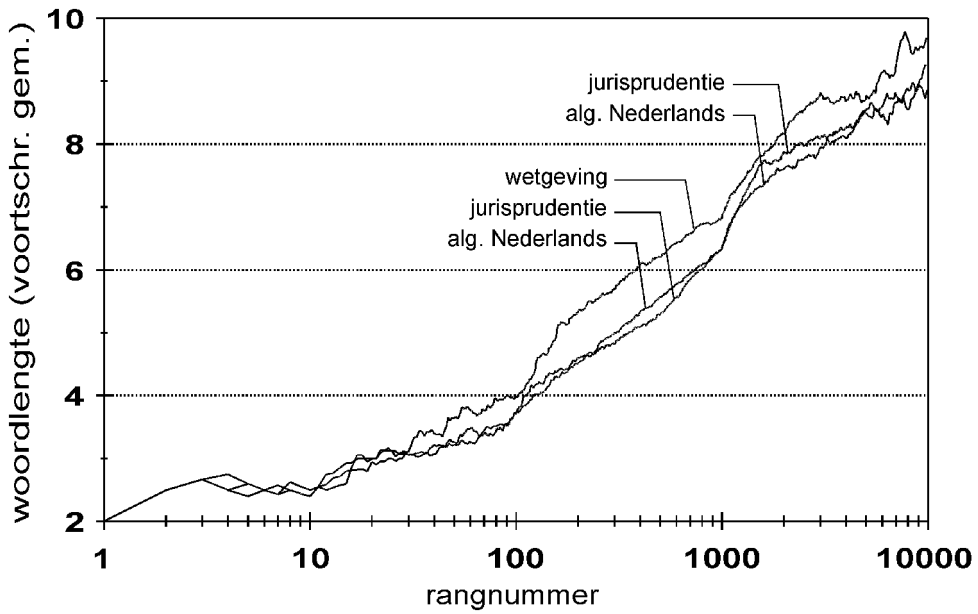
Afb. 7-16 Voortschrijdend gemiddelde van de lengte van woordtypen

Wanneer we deze afbeelding bekijken valt in de eerste plaats op dat alle drie de curves bij rangnummer 1000 een 'knik' vertonen: de gemiddelde woordlengte neemt daar over zo'n 100 à 200 rangnummers plotseling sneller toe. Dit verschijnsel is echter terug te voeren op de gebruikte techniek om het voortschrijdend gemiddelde te bepalen over een type en de 999 daaraan voorafgaande typen. Daarbij wordt het 'first in, first out' principe gevolgd: uit de reeks van 1000 wordt telkens het type met het laagste rangnummer verwijderd alvorens een nieuw type toe te voegen. Bij rangnummer 1001 en volgende worden dus de rangnummers 1, 2 etc. een voor een verwijderd. Deze typen hebben bijna allemaal een geringe lengte (2 of 3 karakters), de gemiddelde woordlengte van de eerste 16 typen ligt nog onder de 3. In plaats hiervan komen dan typen met rangnummers vanaf 1001, met een gemiddelde lengte van 7 à 8 karakters, waardoor het voortschrijdend gemiddelde plotseling sneller stijgt. Deze verklaring wordt bevestigd wanneer het voortschrijdend gemiddelde bij wijze van proef wordt berekend over bijvoorbeeld 700 typen. De knik in de curves

verplaatst zich dan naar dat punt.

Wanneer we de 'dip' die de curves rond rangnummer 1000 vertonen even negeren, blijkt verder uit de grafiek dat de curve voor het (ongewijzigde) corpus algemeen Nederlands nu een stuk vlakker verloopt (minder snel stijgt) dan in het vorige hoofdstuk (zie afbeelding 6-15). Dit verschijnsel houdt verband met het grotere aantal typen waarover de gemiddelde lengte wordt berekend. Ook de curves van de gereduceerde juridische corpora verlopen vlakker dan die uit hoofdstuk 4 en 5. Afgezien daarvan vertonen ze echter duidelijke overeenkomsten met die van de volledige corpora. De curve van de wetgevingsteksten stijgt net als die van het volledige corpus het snelst, en de curves gaan uiteindelijk min of meer horizontaal verlopen bij ongeveer dezelfde gemiddelde woordlengte (8 bij de jurisprudentieteksten, 8,5 tot 9 bij de wetteksten). De curve van het corpus wetgevingsteksten ligt daarbij steeds ruimschoots boven die van de andere corpora. De gemiddelde woordlengte neemt in dit corpus aanmerkelijk sneller toe, met name vanaf rangnummer 100 (zoals ook al bleek bij het vergelijken van de gemiddelde woordlengte per cluster van rangnummers), en ligt bij rangnummer 500 al bijna een karakter boven die in de andere corpora. Bij de andere corpora ligt de gemiddelde woordlengte steeds lager dan bij de wetteksten. Aanvankelijk blijft daarbij het corpus jurisprudentieteksten wat achter bij het corpus algemeen Nederlands, maar rond rangnummer 1000 kruisen de curves elkaar waarna tot ongeveer rangnummer 4500 het corpus algemeen Nederland van de drie de kleinste gemiddelde woordlengte heeft.

Om het gedeelte tot rangnummer 1000 wat nauwkeuriger te kunnen bestuderen is evenals in de voorgaande hoofdstukken van deze grafiek ook een versie gemaakt met een logaritmische schaalverdeling voor het rangnummer (zie afbeelding 7-14).



Afb. 7-17 Voortschrijdend gemiddelde van de lengte van woordtypen, logaritmische schaalverdeling voor rangnummers

Veel nieuwe inzichten geeft dit echter niet, onder andere doordat de curves tot rangnummer 100 dicht bij elkaar blijven en elkaar soms kruisen. De curve van het corpus

wetgevingsteksten blijft daarbij in de meeste gevallen de bovenste, maar bij de andere corpora is het lastiger om vast te stellen welk van de twee bij welk rangnummer de

hoogste gemiddelde woordlengte heeft. Na rangnummer 100 is de situatie zoals hierboven beschreven, aanvankelijk heeft het corpus algemeen Nederlands een grotere gemiddelde lengte en na rangnummer 1000 het corpus jurisprudentieteksten.

De overeenkomsten tussen de hier opgenomen curves van de gereduceerde juridische corpora en die van de volledige corpora zijn net als bij de niet-logaritmische versie groot. Bij de wetgevingsteksten is net als bij het volledige corpus sprake van een knik, alleen bevindt deze zich nu niet bij rangnummer 300 maar even voor 200. Ten aanzien van het verband tussen rangnummer en gemiddelde woordlengte kunnen de volgende conclusies worden getrokken:

- Bij het corpus wetgevingsteksten stijgt de gemiddelde woordlengte het snelst. Het corpus bevat kennelijk een hoger percentage langere typen aan de kop van de frequentielijst. Het verschil met de andere corpora neemt vooral in de buurt van rangnummer 100 toe.

- De gemiddelde woordlengte stijgt bij het corpus jurisprudentieteksten in grote lijnen ongeveer even snel als bij het corpus algemeen Nederlands. Een uitzondering hierop wordt gevormd door een bepaald gedeelte van de frequentielijst (ongeveer van rangnummer 600 tot 3 à 4000), waar de gemiddelde woordlengte sterker stijgt en uitkomt boven die van het corpus algemeen Nederlands. Daarna naderen de corpora elkaar wat dit punt betreft weer.

- Bij het corpus algemeen Nederlands stijgt de gemiddelde woordlengte het gelijkmatigst. Zij ligt vrijwel op ieder punt in de frequentielijst lager dan bij het corpus wetgevingsteksten, en in een bepaald interval ook lager dan bij het corpus jurisprudentieteksten. De gemiddelde woordlengte in het laatste cluster van rangnummers (dat wil zeggen bij de laagste woordfrequenties) is bij dit corpus echter hoger dan bij de andere corpora.

7.5 De kop van de frequentielijsten

7.5.1 Inleiding

Nadat in de voorgaande paragraaf diverse aspecten die samenhangen met de structuur van het woordgebruik werden vergeleken, is het in deze paragraaf de beurt aan de woordtypen zelf. Geprobeerd zal onder andere worden om vast te stellen of er woordtypen zijn die kunnen gelden als kenmerkend voor wetgevings- of jurisprudentieteksten, of voor beide. Gezien het grote aantal woordtypen is het daarbij noodzakelijk om beperkingen te hanteren. Deze beperkingen bestaan voornamelijk hieruit dat we slechts de typen zullen vergelijken die voorkomen in de *kop van de frequentielijsten van de corpora*. De kop van een frequentielijst is daarbij niet een arbitrair gegeven, deze volgt uit het verloop van de frequenties in relatie tot het rangnummer (zie paragraaf 7.4.2, hiervoor). Zoals in dit

hoofdstuk steeds het geval was bij het maken van directe vergelijkingen tussen de corpora zal weer gebruik worden gemaakt van de *gereduceerde corpora*, waarbij eerst zal worden bezien in hoeverre deze afwijken van de volledige. In eerste instantie zal daarbij worden uitgegaan van de zogenaamde 'lange kop' met de 'vaak tot zeer vaak' voorkomende woordtypen. Deze wordt aan de rechterkant begrensd door het punt waar de frequenties niet langer de regressielijn van log frequentie op log rangnummer volgen, en omvat bij het gereduceerde corpus wetgevingsteksten 120 rangnummers, bij het gereduceerde corpus jurisprudentieteksten 133 rangnummers en bij het corpus algemeen Nederlands 155 rangnummers.

7.5.2 Vergelijking van de 'lange' kop van de frequentielijsten

Het blijkt helaas zo te zijn dat het gedeelte van de frequentielijst dat aan de hand van het frequentieverloop als kop kan worden aangemerkt enigermate afhankelijk is van de corpusgrootte. De kop is bij de gereduceerde juridische corpora dan ook wat korter dan bij de volledige. Het blijkt dat daarbij uitsluitend typen zijn *verdwenen*, het is niet zo dat de kop van de gereduceerde corpora typen bevatten die niet in de kop van de volledige corpora voorkomen. Wanneer we de (lange) kop van de gereduceerde corpora nog wat nauwkeuriger vergelijken met de kop van het overeenkomstige volledige corpus blijkt dat de verschillen gering zijn:

- Bij het gereduceerde corpus wetgevingsteksten is de kop 16 woordtypen korter (120 in plaats van 136). Verdwenen (vergeleken met het volledige corpus) zijn daarbij de typen met rangnummer 121 en hoger, op drie typen na die nu een wat hoger rangnummer hebben, en de typen met rangnummer 105 (GRENS), 116 (SECTIE) en 120 (VERZOEK). Wanneer voor de overige typen de rangnummers worden vergeleken, blijkt dat daarbij voor de 19 hoogste rangnummers geen verschillen zijn, en dat daarna slechts geringe verschuivingen in rangnummer optreden (gemiddeld verschil: 1,1 rangnummer, standaarddeviatie 1,74 rangnummer).

- De kop van het gereduceerde corpus jurisprudentieteksten is 35 typen korter (133 in plaats van 168). Daarbij zijn opnieuw alleen typen verdwenen, en wel de typen met rangnummer 134 en hoger, op vier typen na die nu iets hoger in de lijst staan. Verder zijn verdwenen de typen met rangnummer 119 (GEMEENTE), 128 (S), 132 (WAT) en 133 (TWEEDE). In dit geval vertonen de eerste 26 rangnummers geen verschillen, daarboven is sprake van een gemiddelde verschuiving van 1,57 rangnummer (standaarddeviatie 2,85 rangnummer), iets meer dan bij de wetteksten dus.

Gezien deze geringe verschillen en gezien het doel, namelijk het maken van zo goed mogelijke vergelijkingen, lijkt het alleszins verantwoord gebruik te maken van de frequentielijsten van de gereduceerde corpora, naast die van het corpus algemeen Nederlands. Van de typen uit de kop van de drie frequentielijsten zal hier een vergelijkingslijst worden samengesteld. De typen zijn in deze lijst gesorteerd op alfabetische volgorde. Daarnaast bevat de lijst een drietal kolommen, één voor elk corpus. In die kolommen staat vermeld het rangnummer en de frequentie van een type in het betreffende corpus. Komt een type in een corpus niet voor, dan staan in deze kolommen streepjes (---). In plaats van de absolute frequentie van woordtypen wordt in dit geval de *relatieve frequentie* vermeld (absolute

frequentie gedeeld door aantal tokens in het corpus). Dit heeft als belangrijkste voordeel dat direct een indruk kan worden verkregen van het 'belang' van een bepaald type (welk percentage van alle tokens in het corpus wordt door het type 'beslagen') Voor het gebruik van deze relatieve frequenties gelden in dit geval nauwelijks nadelen, aangezien de drie corpora praktisch even groot zijn. De vergelijkingslijst, die in totaal 252 typen bevat, is achterin dit proefschrift opgenomen als bijlage 4. Bij wijze van voorbeeld is hieronder het eerste gedeelte van de lijst afgedrukt in tabel 7-8.

Een opvallend punt dat direct al uit tabel 7-8 kan worden afgelezen is dat onder andere geen van de cijfers 1 t/m 6, die als afzonderlijk type in beide juridische corpora elk een hoge frequentie hebben, bij het corpus algemeen Nederlands in de kop van de frequentielijst voorkomt. Voor cijfers hoeft dit op zich nog niet veel bevreemding te wekken, aangezien al eerder werd vastgesteld dat deze in de juridische corpora veel vaker voorkomen.
Precies hetzelfde geldt echter voor een op het oog heel gewoon en dikwijls voorkomend type als BEDOELD.

Woordtype	Wetgevingsteksten		Jurisprudentieteksten		Alg. Nederlands	
	Rangnr.	Rel. freq.	Rangnr.	Rel. freq.	Rangnr	Rel. freq.
1	19	0,6808%	38	0,2628%	---	---
2	26	0,5603%	49	0,1965%	---	---
3	36	0,3526%	66	0,1452%	---	---
4	48	0,2337%	87	0,1158%	---	---
5	62	0,1717%	100	0,0965%	---	---
6	75	0,1327%	113	0,0858%	---	---
7	95	0,1003%	---	---	---	---
8	109	0,0892%	---	---	---	---
10	107	0,0905%	---	---	---	---
A	40	0,3007%	67	0,1452%	---	---
AAN	14	0,7738%	13	0,7641%	17	0,6329%
AANTAL	---	---		---	106	0,0825%
AANZIEN	96	0,1002%	---	---	---	---
AF	---	---	---	---	129	0,0652%
AL	---	---	122	0,0808%	57	0,1851%
ALGEMENE	78	0,1252%	---	---	---	---
ALLE	106	0,0925%	---	---	79	0,1022%
ALLEEN	---	---	---	---	66	0,1268%
ALS	29	0,4407%	20	0,5841%	16	0,7107%
ALSMEDE	112	0,0865%	---	---	---	---
ALTHANS	---	---	124	0,0789%	---	---
ALTIJD	---	---	---	---	133	0,0628%
ANDERE	66	0,1537%	95	0,1019%	54	0,1910%
ARREST	---	---	75	0,1337%	---	---
ART	13	0,7969%	23	0,4909%	---	---
ARTIKEL	16	0,7259%	---	---	---	---
ARTIKELEN	79	0,1244%	---	---	---	---
B	44	0,2765%	90	0,1062%	---	---
BAL	---	---	---	---	150	0,0530%
BEDOELD	34	0,3703%	129	0,0760%	---	---
BEDOELDE	55	0,1954%	---	---	---	---
BEDRAG	88	0,1074%	---	---	---	---

Tabel 7-8 Vergelijkingslijst woordtypen uit de kop van de frequentielijsten (gedeelte van de eerste pagina)

Al met al kan een aantal verschillen tussen de corpora worden blootgelegd door aan de hand van de vergelijkingslijst die woordtypen op te sommen die in een bepaald corpus zeer vaak tot vaak voorkomen, maar in een ander corpus juist niet. 'Zeer vaak tot vaak voorkomen' wil in dit verband dan zeggen: *voorkomen in de 'lange' kop van de frequentielijst*. Op die manier is een achttal lijstjes van woordtypen samengesteld, die elk staan voor een bepaalde vergelijking van de corpora De lijstjes worden hieronder afgedrukt en kort besproken. We zullen beginnen met het vergelijken van de twee juridische corpora

met het corpus algemeen Nederlands.

1. Typen die zeer vaak tot vaak voorkomen in wetgeving, maar niet in algemeen Nederlands

Deze woordtypen zouden we kunnen aanduiden als specifiek voor wetgevingsteksten. Het gaat om:

ART	C	LEDEN
LID	HAND	7
ARTIKEL	5	AANZIEN
1	RAAD	NR
2	I	D
WET	BETREKKING	BESTUUR
MINISTER	II	INGEVOLGE
BEDOELD	KRACHTENS	VERSTAAN
3	6	10
INDIEN	ALGEMENE	GEMEENTEN
A	ARTIKELEN	8
B	ZAKEN	DAARVAN
GEMEENTE	F	BEPALINGEN
BESLUIT	DAG	ALSMEDE
TOEPASSING	DERDE	GROND
4	ENZ	JANUARI
STB	BEDRAG	WAARVAN
BEDOELDE	WAAROP	COMMISSIE
DEN	DATUM	STELLEN
TER	ZOVER	

De woordtypen zijn gesorteerd op hun frequentie in het corpus wetgevingsteksten. 'ART', als afkorting voor 'artikel' dikwijls gebruikt in teksten van wetgeving, is dus het woordtype met de hoogste frequentie uit het corpus wetgevingsteksten (uit de vergelijkingslijst blijkt dat het rangnummer 13 heeft en een relatieve frequentie van bijna 0,8%, dat wil zeggen dat het ongeveer 1 keer voorkomt op elk 125 tokens) dat niet voorkomt in de kop van de frequentielijst van het corpus algemeen Nederlands. Veel van de typen uit de lijst, zoals LID, ARTIKEL, WET, MINISTER, etc. lijken in zekere zin verwant met het onderwerp wetgeving, maar andere lijken op zich zeer algemeen, zoals BEDOELD, INDIEN, BEDOELDE, WAAROP, WAARVAN, etc. Verder valt het aantal getallen en letters (A, B, C, D, F) op. De letters zijn waarschijnlijk in de meeste gevallen gebruikt als aanduidingen bij een opsomming. STB is de afkorting voor Staatsblad, HAND is waarschijnlijk ook in de meeste gevallen gebruikt als afkorting, bijvoorbeeld van 'Handelingen Tweede Kamer'. DEN is een voorbeeld van archaïsch taalgebruik.

2. Typen die zeer vaak tot vaak voorkomen in jurisprudentie, maar niet in algemeen Nederlands

Deze woorden zouden we specifiek voor rechterlijke uitspraken kunnen noemen.

ART	MIDDEL	ARREST
HOF	F	WET
RB	VONNIS	PP
1	RECHT	VORDERING
BEROEP	O	NJ
2	3	OORDEEL
TER	A	MR
LID	GROND	4
CASSATIe	p	BESCHIKKING

GESTELD	ONDERHAVIGE	RO
B	HOGER	HETGEEN
HR	RAAD	ALTHANS
BESLISSING	6	REEDS
VERDACHTE	ONDERDEEL	GRIEF
5	WAARVAN	DOCH
GEGEVEN	ZAAK	BEDOELD
DERHALVE	OMSTANDIGHEDEN	RECHTER
INDIEN	CONCLUSIE	ZIN
BESTREDEN	NR	BW

Net als bij het eerste lijstje treffen we hier zowel typen die op het eerste gezicht aan rechtspraak gerelateerd lijken (ART, HOF, RB (afkorting van Rechtbank), CASSATIE, VONNIS etc.) als typen die op zich algemeen aandoen (INDIEN, OMSTANDIGHEDEN, WAARVAN, BEDOELD, ZIN, etc.). Opvallend is het ontbreken van het type ARTIKEL. Anders dan in het corpus wetgevingsteksten wordt dit woord door de redactie van het tijdschrift Nederlandse Jurisprudentie kennelijk consequent afgekort tot 'art'. De letter O wordt in arresten van de Hoge Raad dikwijls gebruikt als afkorting voor 'overwegende', en de letters P en PP als afkorting voor 'partij' respectievelijk 'partijen'. NJ staat voor Nederlandse Jurisprudentie, HR voor Hoge Raad, RO voor de Wet op de Rechterlijke Organisatie en BW voor Burgerlijk Wetboek. Al deze afkortingen worden in jurisprudentieteksten veelvuldig gebruikt. Verder vallen op de wat archaïsche c.q. plechtige termen DERHALVE, ONDERHAVIGE, HETGEEN en ALTHANS.

We kunnen de juridische corpora, naast met het corpus algemeen Nederlands, ook met elkaar vergelijken. Daaruit ontstaan de volgende lijstjes:

3. Typen die zeer vaak tot vaak voorkomen in wetgevingsteksten, maar niet in jurisprudentie en ook niet in algemeen Nederlands
Dit zijn typen (weer afgedrukt op volgorde van frequentie in jurisprudentieteksten) die als specifiek voor wetgeving kunnen gelden, waarbij tevens de meer algemene juridische termen (namelijk die welke ook in de jurisprudentieteksten een hoge frequentie hebben) zijn verwijderd.

ARTIKEL	ALGEMENE	BESTUUR
MINISTER	ARTIKELEN	INGEVOLGE
GEMEENTE	ZAKEN	VERSTAAN
BESLUIT	DAG	10
TOEPASSING	DERDE	GEMEENTEN
STB	ENZ	8
BEDOELDE	BEDRAG	DAARVAN
DEN	WAAROP	BEPALINGEN
C	DATUM	ALSMEDE
HAND	ZOVER	JANUARI
I	LEDEN	COMMISSIE
BETREKKING	7	STELLEN
II	AANZIEN	
KRACHTENS	D	

Enigszins merkwaardig doet de naam van de maand Januari aan; kennelijk is een meer dan gemiddeld aantal wetten en besluiten in die maand van kracht geworden, mogelijk verwijzen ook nogal wat regelingen naar deze maand bij het stellen van termijnen en dergelijke.

4. Typen die zeer vaak tot vaak voorkomen in jurisprudentie, maar niet in wetgeving en ook niet in algemeen Nederlands

Hiervoor geldt min of meer hetzelfde als bij het vorige rijtje: het gaat om die typen die specifiek zijn voor rechterlijke uitspraken, maar niet zo algemeen juridisch dat ze ook in wetgevingsteksten veel worden gebruikt

HOF	OORDEEL	ZAAK
RB	MR	OMSTANDIGHEDEN
BEROEP	BESCHIKKING	CONCLUSIE
CASSATIE	GESTELD	RO
MIDDEL	HR	HETGEEN
VONNIS	BESLISSING	ALTHANS
RECHT	VERDACHTE	REEDS
O	GEGEVEN	GRIEF
P	DERHALVE	DOCH
ARREST	BESTREDEN	RECHTER
PP	ONDERHAVIGE	ZIN
VORDERING	HOGER	BW
NJ	ONDERDEEL	

5. Typen die zeer vaak voorkomen in zowel wetgeving als jurisprudentie, maar niet in algemeen Nederlands

Dit zijn de woordtypen die de twee juridische corpora gemeenschappelijk hebben (en die dus uit het derde en vierde rijtje juist zijn weggelaten). Het zijn typen die kunnen gelden als kenmerkend voor juridische teksten in het algemeen.

ART	INDIEN	6
LID	A	F
1	B	NR
2	4	GROND
WET	TER	WAARVAN
BEDOELD	5	
3	RAAD	

Opvallend is dat de meeste van deze woorden niet direct een juridische connotatie hebben. Negen ervan zijn cijfers of letters (de letters zijn waarschijnlijk in de meeste gevallen gebruikt als aanduiding van elementen in een opsomming) en verder treffen we als 'niet-juridische' woorden aan BEDOELD, INDIEN, TER, NR (waarschijnlijk als afkorting van nummer) en WAARVAN. Ook twee andere typen kunnen naast een juridische ook heel goed een algemene betekenis hebben (RAAD en GROND).

De voorgaande vijf lijstjes woordtypen geven een goede indicatie van datgene wat de juridische corpora 'extra' hebben, dat wil zeggen welke woordtypen in deze corpora dikwijls voorkomen, die in het corpus algemeen Nederlands en eventueel ook in het andere juridische corpus niet of minder vaak voorkomen. We kunnen de zaak echter ook van de andere kant bekijken: zijn er woordtypen die in 'gewone' teksten (in het corpus algemeen Nederlands dus) veel voorkomen, maar in één of beide juridische corpora niet?

6. Typen die zeer vaak tot vaak voorkomen in algemeen Nederlands, maar niet in wetgevingsteksten

Dit zijn de typen die juist in wetgevingsteksten niet of minder vaak worden gebruikt.

	DUS	VERSCHILLENDE
OOK	TEGEN	LATEN
ER	GAAN	JAREN
MAAR	ONS	TOE
MEN	WEER	ZIEN
ZE	TOCH	GEVAL
IK	HIER	AF
WAS	NIEUWE	ZULLEN
WAT	GAAT	MINDER
NOG	WANNEER	ALTIJD
WE	TIJD	EENS
ZO	DAAR	MIJN
HAAR	VOORAL	MENS
JE	MOGELIJK	LAATSTE
AL	STAAT	TOEN
ZOU	DOEN	VRAAG
TUSSEN	TWEE	SLECHTS
U	ZELF	WIJZE
VEEL	KOMT	DUIDELIJK
NU	MENSEN	ZELFS
WERD	AANTAL	BELANG
ALLEENZOALS	STEEDS	WERK
EIGEN	WIL	ONDERZOEK
MAKEN	WAREN	BIJVOORBEELD
ECHTER	VAAK	BAL
WAAR	ELKAAR	UW
HAD	IETS	VORM
OMDAT	LEVEN	VERDER
KOMEN	GEVEN	ONTWIKKELING
HOE	ZONDER	WERDEN
GOED	ZEER	
GROTE	ÉÉN	

Hier valt bijvoorbeeld het type 'U' op, dat we ook al tegenkwamen bij de woordlengteverdelingen. Het heeft in de juridische teksten een veel lagere frequentie, omdat het niet of veel minder vaak als persoonlijk voornaamwoord wordt gebruikt; de juridische teksten bevatten onder andere weinig 'directe rede'. Verder valt nog op het type ÉÉN. Dit komt in de twee juridische corpora in het geheel niet voor, omdat deze corpora geen letters met accenten bevatten.

7. Typen die zeer vaak tot vaak voorkomen in algemeen Nederlands, maar niet in jurisprudentie

Hiervoor geldt hetzelfde als bij het voorgaande lijstje, maar dan ten aanzien van de jurisprudentieteksten

MEN	U	WAAR
ZE	VEEL	KOMEN
WAT	ALLEEN	ALLE
WE	EIGEN	HOE
ZO	MAKEN	GOED
JE	ECHTER	GROTE

DUS
WIJ
GAAN
ONS
WEER
TOCH
NIEUWE
GAAT
WANNEER
TIJD
DAAR
VOORAL
MOGELIJK
PLAATS
TWEE
ZELF
KOMT
MENSEN
AANTAL
BINNEN
STEEDS
WAARIN
WIL
WAREN
VAAK
ELKAAR
IETS
LEVEN
BEPAALDE
GEVEN
ZEER
ÉÉN
VERSCHILLENDE
LATEN
JAREN
TOE
ZIEN
JAAR
AF
ZULLEN
MINDER
ALTIJD
EENS
MIJN
MENS
LAATSTE
TOEN
WIJZE
DUIDELIJK
ONZE
ZELFS
WERK
ONDERZOEK
BIJVOORBEELD
TWEEDE
BAL
UW
VORM
VERDER
ONTWIKKELING
WERDEN

8. Typen die zeer vaak tot vaak voorkomen in algemeen Nederlands, maar niet in wetgevingsteksten en ook niet in jurisprudentie

Deze woordtypen blijken in dit geval niet of minder vaak te worden gebruikt in alle juridische teksten (zowel wetteksten als jurisprudentie).

MEN
ZE
WAT
WE
ZO
JE
U
VEEL
ALLEEN
EIGEN
MAKEN
ECHTER
WAAR
KOMEN
HOE
GOED
GROTE
DUS
GAAN
ONS
WEER
TOCH
NIEUWE
GAAT
WANNEER
TIJD
DAAR
VOORAL
MOGELIJK
TWEE
ZELF
KOMT
MENSEN
AANTAL
STEEDS
WIL
WAREN
VAAK
ELKAAR
IETS
LEVEN
GEVEN
ZEER
ÉÉN
VERSCHILLENDE
LATEN
JAREN
TOEZIEN
AF
ZULLEN
MINDER
ALTIJD
EENS
MIJN
MENS
LAATSTE
TOEN
WIJZE
DUIDELIJK
ZELFS
WERK
ONDERZOEK
BIJVOORBEELD
BAL
UW
VORM
VERDER
ONTWIKKELING
WERDEN

In dit rijtje vallen vooral de persoonlijke voornaamwoorden MEN, ZE, WE, JE en U op, die van deze typen de hoogste frequentie hebben maar in de juridische corpora niet of minder vaak voorkomen. Verder komen vrijwel alle woordsoorten (in taalkundige zin) voor: zelfstandige naamwoorden, werkwoordsvormen, bijvoeglijke naamwoorden, voornaamwoorden, voegwoorden, etc.

Conclusie ten aanzien van woordtypen in de 'lange' kop van de frequentielijsten
Het trekken van een conclusie aan het einde van dit gedeelte is moeilijk, vooral omdat het interpreteren van de gevonden verschillen tussen de corpora al snel neerkomt op het geven van interpretaties op semantisch niveau, terwijl het de bedoeling van dit onderzoek is dat we ons zoveel mogelijk beperken tot de vorm van de teksten. Twee aspecten, die respectievelijk volgen uit het vijfde en het achtste vergelijkingslijstje, zijn in dit verband het meest van belang:

- De juridische corpora uit dit onderzoek bevatten een duidelijk hoger aantal getallen. De meeste ééncijferige getallen komen zo vaak voor dat ze zijn te vinden in de kop van de frequentielijst, vooral bij het corpus wetgevingsteksten. Daarnaast hebben in de juridische corpora enkele 'gewoon' uitziende typen (zonder specifieke juridische connotatie) een aanmerkelijk hogere frequentie dan in algemeen Nederlands, zoals BEDOELD, INDIEN, TER en WAARVAN.

- In de juridische corpora hebben bepaalde persoonlijke voornaamwoorden een aanmerkelijk lagere frequentie. Voorbeelden daarvan zijn MEN, ZE, WE, JE en U, die in het corpus algemeen Nederlands een hoog rangnummer hebben (van 34 tot 61).

7.5.3 De 'korte' kop van de frequentielijsten

Tenslotte zal nu nog enige aandacht worden besteed aan het gedeelte van de frequentielijst dat hiervoor steeds is aangeduid als de 'korte' kop. De omvang van dit deel van de frequentielijst werd in paragraaf 7.4 bepaald aan de hand van de residuals van de regressielijn van log frequentie op log rangnummer. In dit deel van de frequentielijst treffen we de zeer vaak voorkomende woordtypen aan, die samen 35% tot 47% (afhankelijk van het corpus) van alle woordtokens omvatten. Voor het op dit punt vergelijken van de drie corpora zal een wat andere benadering worden gevolgd dan bij de lange kop (ook al omdat de hoeveelheid nieuwe informatie anders niet zo groot zou zijn). In plaats van een vergelijkingslijst van woordtypen zal nu een lijst van rangnummers worden opgesteld (gesorteerd van laag naar hoog) waarbij achter ieder rangnummer het bijbehorende woordtype uit elk van de drie corpora wordt vermeld, met de relatieve frequentie in dat corpus. Om een goede vergelijking mogelijk te maken zal net als bij de lange kop gebruik worden gemaakt van de gereduceerde juridische corpora en het corpus algemeen Nederlands. De vergelijkingslijst is te vinden in tabel 7-9, op de volgende bladzijde. De lijst lijkt veel op die welke aan het einde van hoofdstuk 3 werd afgedrukt voor de volledige corpora, alleen zijn nu niet voor elk corpus evenveel typen opgenomen (de 'korte' kop is immers niet in alle gevallen even lang) en worden ook de relatieve frequenties vermeld.

Wanneer we tabel 7-9 bekijken vallen verschillende zaken op, zoals:

- Het onbepaald lidwoord EEN heeft in de juridische corpora een 2 of 3 punten lager rangnummer en een duidelijk lagere relatieve frequentie dan in het corpus algemeen Nederlands. Daarentegen heeft het lidwoord DE bij de wetgevingsteksten een wat hogere relatieve frequentie dan in de andere corpora.

- Alleen de kop van het corpus wetgevingsteksten bevat een viertal getallen (1 - 4). Daarnaast hebben in dit corpus de letters A en B een hoge frequentie. Daar staat tegenover dat in het corpus algemeen Nederlands de letter U voorkomt, zij het pas bij rangnummer 61.

- De kop van het corpus wetgevingsteksten bevat vrij veel woorden waarvan direct duidelijk is dat ze samenhangen met de juridische inhoud van het corpus, zoals ART, LID, ARTIKEL, WET, MINISTER, GEMEENTE en STB (Staatsblad). De kop van het corpus jurisprudentieteksten bevat er daarvan slecht een drietal, namelijk ART, HOF en RB (Rechtbank). Opvallend is hier de afwezigheid van ARTIKEL.

- Sommige 'algemene' woordtypen zoals ALS, NIET en DIT komen in het corpus wetgevingsteksten duidelijk minder vaak voor dan in de andere corpora (en vooral het corpus algemeen Nederlands).

- De korte kop van het corpus algemeen Nederlands bevat net als de lange kop een aantal persoonlijke voornaamwoorden zoals HIJ, MEN, ZE en IK. Deze ontbreken in de kop van de juridische corpora.

Een vergelijking van de kop van de frequentielijsten met die uit het onderzoek van De Mulder en Oskamp

Aangezien de 'korte kop' van de frequentielijsten in dit onderzoek volgens dezelfde criteria is opgebouwd als de 'kop' in het onderzoek van De Mulder en Oskamp kunnen deze goed met elkaar worden vergeleken. Van de door hen samengestelde corpora komen daarvoor het meest in aanmerking het corpus 'Gemeenschappelijke wetten' (bestaande uit de tekst van de wetboeken van Strafrecht en van Strafvordering) en het corpus 'Nederlandse Schrijftaal' (een deel van het Eindhovens corpus). De kop van de frequentielijst van deze beide corpora is overgenomen uit het rapport 'Het woordgebruik van het Wetboek van Strafrecht en het Wetboek van Strafvordering' (De Mulder en Oskamp 1979), en weergegeven in tabel 7-10. De kolom van het corpus met strafwetgevingsteksten kan daarbij het beste worden vergeleken met de kolom van het corpus wetgevingsteksten uit tabel 7-9, en de kolom van het Eindhovens corpus met die van het corpus algemeen Nederlands. Zaken die daarbij opvallen zijn:

- Ook in het corpus strafwetgeving heeft EEN een wat lagere frequentie dan in het vergelijkingscorpus met algemeen Nederlandse teksten. DE heeft in dit geval echter *geen* hogere relatieve frequentie.

- ART heeft in het corpus strafwetgeving een wat hoger rangnummer en een relatieve frequentie die de helft hoger ligt dan in het corpus wetgevingsteksten. Hier speelt waarschijnlijk een rol dat beide strafrechtelijke wetboeken in artikelen zijn ingedeeld, terwijl het corpus wetgevingsteksten ook wetten bevat waarbij dat niet het geval is. Ook heeft ARTIKEL bij de strafwetgeving juist weer een lagere frequentie; het zou kunnen zijn dat dit woord in de strafwetboeken consequenter wordt afgekort tot ART.

Rangnr	Corpus wetgevingsteksten:		Corpus jurispr. teksten:		Corpus alg. Nederlands:	
	Woordtype	Rel. freq.	Woordtype	Rel. freq.	Woordtype	Rel. freq.
1	DE	7,0383%	DE	6,5142%	DE	6,7860%
2	VAN	5,9512%	VAN	4,9012%	VAN	3,6917%
3	HET	3,4096%	HET	3,4730%	HET	3,2760%
4	IN	2,4722%	IN	2,4443%	EEN	2,5696%
5	EN	2,2368%	DAT	2,2892%	EN	2,4894%
6	EEN	1,8764%	EN	1,7574%	IN	2,3005%
7	OF	1,2321%	EEN	1,6138%	IS	1,3066%
8	VOOR	1,0967%	TE	1,4772%	DAT	1,3024%
9	OP	1,0158%	IS	1,2616%	TE	1,2865%
10	IS	0,9086%	OP	1,1814%	OP	1,0566%
11	MET	0,8883%	NIET	1,0654%	ZIJN	1,0240%
12	TE	0,8381%	HEEFT	0,9342%	DIE	1,0088%
13	ART	0,7969%	AAN	0,7641%	VOOR	0,8557%
14	AAN	0,7738%	DOOR	0,7586%	MET	0,8084%
15	LID	0,7586%	ZIJN	0,7363%	NIET	0,8049%
16	ARTIKEL	0,7259%	MET	0,7119%	ALS	0,7107%
17	WORDEN	0,7221%	VOOR	0,7048%	AAN	0,6329%
18	DIE	0,7181%	DIE	0,6296%	OOK	0,5833%
19	1	0,6808%	TOT	0,6121%	OF	0,5475%
20	ZIJN	0,6788%	ALS	0,5841%	WORDEN	0,5407%
21	BIJ	0,6698%	BIJ	0,5561%	ER	0,5074%
22	WORDT	0,6678%	WORDEN	0,5443%	DOOR	0,5057%
23	DOOR	0,6397%	ART	0,4909%	DAN	0,4964%
24	DAT	0,6191%	DEZE	0,4829%	DEZE	0,4880%
25	TOT	0,5770%	OF	0,4555%	BIJ	0,4870%
26	2	0,5603%	HOF	0,4227%	OM	0,4801%
27	DEZE	0,4668%	RB	0,3719%	MAAR	0,4672%
28	NIET	0,4627%	DIT	0,3614%	TOT	0,4207%
29	ALS	0,4407%	TEN	0,3500%	WORDT	0,3889%
30	WET	0,4198%	OOK	0,3491%	HIJ	0,3888%
31	EERSTE	0,4161%	ZOU	0,3074%	KAN	0,3799%
32	MINISTER	0,4108%	KAN	0,2913%	DIT	0,3750%
33	TEN	0,4094%			ZICH	0,3633%
34	BEDOELD	0,3703%			MEN	0,3426%
35	DAN	0,3601%			NAAR	0,2981%
36	3	0,3526%			ZE	0,2979%
37	INDIEN	0,3518%			UIT	0,2904%
38	DER	0,3491%			KUNNEN	0,2770%
39	ONDER	0,3315%			OVER	0,2762%
40	A	0,3007%			HEEFT	0,2731%
41	DIT	0,2964%			IK	0,2680%
42	KAN	0,2963%			MEER	0,2675%
43	ONZE	0,2857%			WAS	0,2673%
44	B	0,2765%			WAT	0,2638%
45	GEMEENTE	0,2750%			NOG	0,2523%
46	BESLUIT	0,2713%			WE	0,2414%
47	TOEPASSING	0,2538%			HEBBEN	0,2330%
48	4	0,2337%			ZIJ	0,2240%
49	TWEEDE	0,2250%			ZO	0,2153%
50	STB	0,2120%			MOET	0,2110%
51					WEL	0,2101%
52					HUN	0,2057%
53					HAAR	0,1994%
54					ANDERE	0,1910%
55					GEEN	0,1875%
56					JE	0,1875%
57					AL	0,1851%
58					ZAL	0,1745%
59					ZOU	0,1583%
60					TUSSEN	0,1574%
61					U	0,1557%
62					VEEL	0,1523%
63					NU	0,1419%
64					WERD	0,1346%
65					MOETEN	0,1303%
66					ALLEEN	0,1268%
67					ZOALS	0,1252%
68					ONDER	0,1194%
69					EERSTE	0,1170%
70					EIGEN	0,1124%

Tabel 7-9 'Echte' kop van de gereduceerde juridische corpora en het corpus algemeen Nederlands

- Opvallend is dat bij de strafwetgeving de typen HIJ en HEM een hoog rangnummer hebben, terwijl deze typen in de kop van de wetgevingsteksten niet voorkomen. Dit heeft waarschijnlijk te maken met de manier waarop strafbepalingen vaak worden geformuleerd: "Hij die .... wordt gestraft met ....".

- Het corpus strafwetgeving bevat de typische strafrecht-typen VERDACHTE, GESTRAFT, GEVANGENISSTRAF, JUSTITIE en OFFICIER (meestal zal er Officier van Justitie hebben gestaan). Daarentegen ontbreken weer de typische wetgevingstypen WET, MINISTER, GEMEENTE en STB die het corpus wetgevingsteksten bevat.

Vergelijking van de kop van het Eindhovens corpus met die van het corpus algemeen Nederlands leert dat tussen deze corpora, die toch allebei bedoeld zijn als een goede 'dwarsdoorsnede' van Nederlandse schrijftaal, toch nog aanzienlijke verschillen bestaan. Zo is bijvoorbeeld opvallend dat VAN in het Eindhovens corpus niet rangnummer 2 maar 3 heeft, met een relatieve frequentie die bijna een procent lager is dan in het corpus algemeen Nederlands. Voor de dertig woordtypen met de hoogste frequentie uit elk van de corpora geldt dat ze ook in de kop van het andere corpus te vinden zijn, maar in veel gevallen met een heel ander rangnummer. Opvallend is bijvoorbeeld dat de persoonlijke voornaamwoorden HIJ en IK in het Eindhovens corpus een veel hoger rangnummer hebben dan in het corpus algemeen Nederlands. Redenen hiervoor zouden kunnen zijn dat de samenstelling van de corpora anders is (het hier gebruikte deel van het Eindhovens corpus bestaat bijvoorbeeld voor ongeveer de helft uit teksten van dag- opinie- en gezinsbladen, het corpus algemeen Nederlands bevat deze tekstsoorten bijna helemaal niet), dat ze verschillend van omvang zijn (het corpus algemeen Nederlands is bijna acht keer zo groot) of dat de teksten in het corpus algemeen Nederlands over het algemeen veel recenter zijn dan die in het Eindhovens corpus. In ieder geval kan worden vastgesteld dat zelfs tussen 'algemene' corpora, die niet uit één bepaald soort teksten zijn opgebouwd, nog aanzienlijke verschillen kunnen bestaan, zelfs waar het de zeer vaak voorkomende woordtypen betreft.

Conclusie ten aanzien van de vergelijking van de 'korte' kop van de diverse subcorpora
Vastgesteld kan worden dat tussen de 'korte' kop van het corpus wetgevingsteksten en het corpus strafwetgeving onmiskenbare overeenkomsten bestaan. Om te beginnen zijn beide koppen ongeveer even lang. Daarnaast bevatten ze beide een aantal getallen, en ligt de frequentie van bepaalde woordtypen in deze corpora lager dan in alle andere corpora.

Een verschil is dat de kop van het corpus strafwetgeving bepaalde woordtypen bevat die specifiek zijn voor dit type wetgeving, terwijl het corpus wetgevingsteksten (niet onverwacht) hier wat algemenere typen die iets met wetgeving te maken hebben bevat. Bij vergelijking van de kop van het corpus algemeen Nederlands met die van het Eindhovens corpus is het opvallend dat weliswaar veel typen in beide corpora voorkomen (de eerste 30 typen zelfs allemaal), maar dat frequentie en (als gevolg daarvan) rangnummer nogal eens behoorlijk verschillen. Een opvallende overeenkomst tussen deze corpora is weer dat de koppen precies even lang zijn.

Rangnr	**Corpus strafwetgeving:**		**Eindhovens corpus (alg. Ned.):**	
	Woordtype	Rel. freq.	Woordtype	Rel. freq.
1	DE	5,7131%	DE	6,0535%
2	VAN	5,4755%	HET	3,0475%
3	HET	3,1938%	VAN	2,8876%
4	OF	2,8767%	EEN	2,6951%
5	IN	1,9496%	EN	2,4024%
6	EEN	1,8953%	IN	2,0928%
7	IS	1,2779%	DAT	1,4812%
8	ART	1,2424%	IS	1,2096%
9	EN	1,2377%	TE	1,1972%
10	WORDT	1,2283%	ZIJN	1,0922%
11	DEN	1,1872%	DIE	1,0848%
12	TEN	1,1357%	NIET	0,9390%
13	MET	1,1048%	OP	0,9120%
14	DIE	1,0066%	MET	0,8898%
15	TE	0,9542%	IK	0,8316%
16	OP	0,9233%	HIJ	0,8200%
17	AAN	0,8120%	VOOR	0,7350%
18	1	0,8073%	MAAR	0,6703%
19	DOOR	0,7933%	ALS	0,6300%
20	2	0,7802%	ZE	0,5618%
21	HIJ	0,7652%	WAS	0,5441%
22	WORDEN	0,7297%	AAN	0,5174%
23	DAT	0,7241%	OOK	0,4779%
24	BIJ	0,7044%	DAN	0,4589%
25	HOOGSTE	0,7007%	DOOR	0,4160%
26	ZIJN	0,6717%	JE	0,4137%
27	INDIEN	0,6670%	ER	0,4096%
28	NIET	0,6436%	NOG	0,3941%
29	KAN	0,6174%	OM	0,3809%
30	TOT	0,6053%	HEEFT	0,3772%
31	ARTIKEL	0,5931%	BIJ	0,3658%
32	VOOR	0,5912%	NAAR	0,3597%
33	VERDACHTE	0,5903%	ZICH	0,3409%
34	DER	0,4556%	ZO	0,3371%
35	GESTRAFT	0,4547%	OF	0,3336%
36	GEVANGENISSTRAF	0,4500%	WAT	0,3312%
37	LID	0,4303%	TOT	0,3265%
38	HEEFT	0,4191%	HAAR	0,3254%
39	3	0,3882%	WORDEN	0,3112%
40	EENE	0,3882%	DEZE	0,3069%
41	RECHTER	0,3714%	DIT	0,2907%
42	JAREN	0,3564%	UIT	0,2691%
43	DAN	0,3452%	OVER	0,2658%
44	JUSTITIE	0,3368%	HEBBEN	0,2615%
45	ALS	0,3162%	WEL	0,2603%
46	FEIT	0,2956%	HAD	0,2590%
47	HEM	0,2872%	GEEN	0,2501%
48	TOEPASSING	0,2853%	MEER	0,2453%
49	OFFICIER	0,2825%	ZIJ	0,2389%
50	TER	0,2825%	AL	0,2387%
51	ONDERZOEK	0,2788%	WE	0,2359%
52			WORDT	0,2291%
53			HEM	0,2285%
54			KAN	0,2270%
55			ZOU	0,2250%
56			HUN	0,2197%
57			NU	0,2178%
58			MEN	0,2006%
59			WERD	0,1922%
60			MIJN	0,1745%
61			ZAL	0,1743%
62			KUNNEN	0,1702%
63			VEEL	0,1570%
64			U	0,1530%
65			WEER	0,1525%
66			MOET	0,1514%
67			TOEN	0,1501%
68			TOCH	0,1478%
69			JAAR	0,1431%
70			WIJ	0,1342%

Tabel 7-10 'Echte' kop van de corpora uit het onderzoek van De Mulder en Oskamp

7.6 Conclusies met betrekking tot overeenkomsten en verschillen in het woordgebruik

In deze paragraaf worden de voornaamste conclusies uit dit hoofdstuk herhaald, vanwege hun belang voor dit onderzoek als geheel.

- In dit hoofdstuk is onder andere gewerkt met steekproeven uit de juridische corpora. Voorzover op basis van enkele linguïstische constanten kan worden vastgesteld bestaan er sterke overeenkomsten tussen de structuur van het woordgebruik in deze steekproeven en in de volledige corpora. Bij het vergelijken van andere aspecten van het woordgebruik is dit nog enkele malen bevestigd.

- Naar aanleiding van de bevindingen ten aanzien van de steekproeven uit de juridische corpora kan worden vastgesteld dat zowel de karakteristiek *K* als de twee constanten *R* en *C* ook van toepassing zijn op willekeurige steekproeven uit het corpus waarvoor ze gelden, zelfs als die steekproeven maar een zeer beperkt deel van het corpus (zo'n 10% van de oorspronkelijke omvang) omvatten.

- Uit verschillen tussen de cumulatieve frequentie-grafieken kan worden geconcludeerd dat de aantallen woordtokens bij de twee juridische corpora anders verdeeld zijn over de rangnummers (lees: de woordtypen) dan bij het corpus algemeen Nederlands. De sterkere kromming van deze grafieken bij de twee juridische corpora kan worden verklaard uit het feit dat bij die corpora meer tokens te vinden zijn bij de lagere rangnummers. Dat leidt tot een relatief geringere toename van het aantal woordtokens bij de hoogste rangnummers. Dit is een eerste aanwijzing waaruit kan worden opgemaakt dat bij de juridische corpora de *meestgebruikte woordtypen* veelal ook *vaker worden gebruikt*.

- Uit analyse van het verband tussen rangnummer en frequentie blijkt dat aan de hand daarvan over de eerste 100 rangnummers eigenlijk voor geen van de corpora een structureel ander frequentieverloop kan worden aangetoond. Tussen rangnummer 100 en (ongeveer) 800 hebben woordtypen in de *juridische corpora* steeds een *hogere* frequentie bij gelijk rangnummer dan woordtypen in het corpus algemeen Nederlands. Na rangnummer 800 kruist de curve van algemeen Nederlands die van de juridische corpora. Dat dit gebeurt is verklaarbaar, gezien de gelijke aantallen tokens in ieder corpus. Lagere frequenties bij de veel-voorkomende typen worden hier gecompenseerd met hogere frequenties bij de minder vaak voorkomende typen. Onder gebruikmaking van de regressielijn van log frequentie op log rangnummer kan in de frequentiekarakteristiek een 'kop' worden herkend welke bij de verschillende corpora de eerste 120 tot 155 rangnummers omvat.

- Zowel met behulp van Zipf-karakteristieken als met een grafische weergave van de residuals van de regressielijn van log frequentie op log rangnummer blijkt het mogelijk te zijn een bepaalde structuur in de frequenties van de eerste 100 woordtypen te herkennen. Er tekent zich daarbij een tweede, kortere kop af van 32 tot 70 woordtypen, binnen de eerder vastgestelde 'lange' kop van 120 tot 150 typen. We zouden de typen in de korte kop de 'zeer vaak voorkomende' woordtypen kunnen noemen, en die in de langere kop de 'zeer vaak tot vaak voorkomende'. Op basis van de omvang van

de 'korte' kop en het aantal woordtokens dat deze omvat kan geconcludeerd worden dat de twee juridische corpora uit dit onderzoek evenals het corpus strafrechtelijke wetten uit het onderzoek van De Mulder en Oskamp een geringer aantal woordtypen tellen die kunnen worden aangemerkt als 'zeer vaak voorkomend', terwijl de meeste daarvan bovendien een hogere frequentie hebben.

- Met betrekking tot de spreiding van woordtypen over de documenten waaruit elk van de corpora bestaat kan slechts een voorzichtige conclusie worden getrokken, met name omdat niet precies is na te gaan wat de invloed is van verschillen in de gemiddelde documentgrootte per corpus. Voorzover uit de beschikbare gegevens kan worden afgeleid lijken alle woordtypen in de juridische corpora, en vooral in het corpus wetgevingsteksten, gemiddeld in een kleiner gedeelte van het corpus geconcentreerd te zijn. Dat zou kunnen betekenen dat veel typen in deze corpora minder 'universeel' zijn, hun gebruik beperkt zich slechts tot bepaalde (groepen van) documenten.

- Uit een vergelijking van de frequentieverdelingen van de corpora komt ongeveer hetzelfde beeld naar voren als uit de cumulatieve frequenties: in de juridische corpora omvatten de typen aan de kop van de frequentielijst een groter aantal woordtokens dan het corpus algemeen Nederlands. Bij dat corpus bevatten juist de laagste frequentieklassen een duidelijk hoger aantal tokens. Aan de hand van de frequentieverdelingen kan nauwkeuriger het rangnummer c.q. de frequentieklasse worden bepaald waar dit zich voordoet. Tevens kan daaruit worden afgeleid dat wat geldt voor de woord*tokens*, ook opgaat voor de woord*typen*: ook daarvan komen er meer voor in de hoogste frequentieklassen. Dit terwijl het corpus algemeen Nederlands in totaal ongeveer 2 keer zoveel typen bevat dan de juridische corpora; de 'extra' typen bevinden zich allemaal in de laagste frequentieklassen.

- De verschillende waarden die de constante *K* in de drie corpora heeft kan vooral worden verklaard uit de *lagere standaarddeviatie van de frequenties* bij het corpus algemeen Nederlands, wat dus betekent: minder verschillen tussen de aantallen woordtypen met dezelfde frequentie. Dit zou ermee kunnen samenhangen dat de woordtypen in dit corpus minder geconcentreerd zijn aan de 'kop' van de frequentielijst, maar juist gelijkmatig zijn gespreid over de 'staart'.

- Door het analyseren van het verband tussen aantallen woordtokens en -typen aan de hand van de formule van Erikstad $V = R \,.\, N^{c}$ is duidelijkheid verkregen over de wijze waarop de token/type ratio samenhangt met de afmetingen van de corpora, of van steekproeven daaruit. De formule legt de verhouding tussen aantallen tokens en typen voor een corpus effectief vast, waardoor het steeds mogelijk is een goede schatting te geven van het aantal woordtypen in een steekproef, wanneer het aantal woordtokens daarin bekend is. Dit kan relevant zijn voor verschillende vormen van kwantitatief taalkundig onderzoek, maar ook voor bijvoorbeeld het ontwikkelen van computerapplicaties waarin het verwerken van taal een rol speelt.

- De woordlengte-verdelingen van woordtypen voor de twee juridische corpora vertonen op bepaalde punten meer overeenkomst met elkaar dan met het corpus algemeen Nederlands. Ze bevatten beide een groter aantal woordtypen met lengtes van 4 tot 6 karakters. Een groot deel van de 'extra' typen in deze lengtecategorieën

blijken getallen te zijn. Het corpus algemeen Nederlands bevat daarentegen meer typen met lengtes van 9 tot 24 karakters. Alleen van het corpus wetgevingsteksten kan worden gezegd dat het een hoger percentage lange woorden bevat. Met name het percentage woordtokens van meer dan 25 karakters is bij dit corpus hoger, terwijl ook uit het verband tussen rangnummer en woordlengte blijkt dat in het bovenste deel van de frequentielijst gemiddeld langere woordtypen voorkomen. Het corpus jurisprudentieteksten wijkt op dat punt nauwelijks af van het corpus algemeen Nederlands.

- Uit vergelijking van de typen in de kop van de frequentielijsten blijkt dat in de juridische corpora bepaalde getallen een hoge frequentie hebben. De meeste ééncijferige getallen hebben bij deze corpora zo'n hoge frequentie dat ze inderdaad in de kop van de frequentielijst te vinden zijn. Daarnaast hebben in de juridische corpora enkele algemeen lijkende typen een aanmerkelijk hogere frequentie dan in algemeen Nederlands, zoals BEDOELD, INDIEN, TER en WAARVAN. In beide juridische corpora hebben bepaalde persoonlijke voornaamwoorden een aanmerkelijk *lagere* frequentie dan in algemeen Nederlands. Voorbeelden daarvan zijn MEN, ZE, WE, JE en U.

HOOFDSTUK 8 - DE SIMILARITEIT VAN DOCUMENTEN

8.1 Het begrip similariteit

8.1.1 Inleiding

In de voorafgaande hoofdstukken is uitgebreid aandacht geschonken aan de overeenkomsten en verschillen in het woordgebruik in de drie bestudeerde corpora. In dit hoofdstuk zal het onderzoek van de corpora worden afgerond met de bestudering van een aspect dat tot nu toe nog wat onderbelicht is gebleven. Het gaat hierbij om de afzonderlijke documenten waaruit elk corpus is opgebouwd, en met name om de *relaties* die er bestaan *tussen* deze documenten. Indirect is hieraan al wel aandacht geschonken in de vorm van de bestudering van de spreiding van woordtypen over (de documenten van) elk corpus. De gegevens die daarbij werden verzameld spelen in dit hoofdstuk opnieuw een rol. Om een volledig beeld te krijgen van de opbouw van een corpus en om de structuur die daarin mogelijk aanwezig is te kunnen onderkennen is bestudering van spreidingsgegevens echter niet voldoende. Met name vragen als "is het corpus, wat het woordgebruik betreft, homogeen van opbouw?", en nauw daaraan gerelateerd, "zijn er (groepen van) documenten waarin het woordgebruik duidelijk anders is?" kunnen alleen worden beantwoord wanneer de documenten onderling worden vergeleken. Het zou echter weinig zin hebben om daarbij bijvoorbeeld te werk te gaan zoals in het vorige hoofdstuk, en op dezelfde manier als waarop daar de volledige corpora werden vergeleken, nu te proberen de overeenkomsten en verschillen tussen meer dan 18000 documenten (in het geval van het corpus wetgevingsteksten) in kaart te brengen. Die methode zou een enorme overvloed aan gegevens opleveren, waarin samenhang maar moeilijk te ontdekken zou zijn. Daarom zal in dit hoofdstuk een andere weg worden gevolgd. Om enig inzicht te krijgen in de samenhang tussen documenten zal allereerst een maat worden geïntroduceerd waarmee kan worden uitgedrukt in hoeverre twee documenten met elkaar overeenkomen, of anders gezegd, hoeveel zij met elkaar gemeen hebben. De term die zal worden gebruikt voor deze 'overeenkomst-maat' is *similariteit.* Met behulp van deze maat, die dus betrekking heeft op telkens twee documenten, zal dan vervolgens worden nagegaan of ook samenhang kan worden aangetoond tussen omvangrijker groepen van documenten. Kennis die op die manier wordt verworven zal in de toekomst mogelijk kunnen leiden tot een beter inzicht in de opbouw en structuur van een corpus. Daarbij kan bijvoorbeeld worden gedacht aan het onderscheiden van *subcorpora* binnen een corpus, aan de hand van eigenschappen van de documenten die daarvan deel uitmaken. Bijna onvermijdelijk zal dan echter ook de semantiek een rol gaan spelen, aangezien te verwachten is dat met name die subcorpora interessant zullen blijken waarbij tussen de documenten naast overeenkomsten op syntactisch niveau, ook samenhang op semantisch

niveau kan worden aangetoond[1]. In dit hoofdstuk zal de nadruk zoveel mogelijk gelegd worden op syntactische aspecten, en daarvan met name op vorm-aspecten van overeenkomsten tussen documenten, waarbij het niveau van de semantiek waar mogelijk zal worden vermeden.

Zoals in het vervolg van dit hoofdstuk duidelijk zal worden, is similariteit een veelomvattend begrip. Een analyse van allerlei vormen van similariteit tussen documenten zou dan ook heel goed het onderwerp van een afzonderlijke, omvangrijke studie kunnen zijn. Een dergelijke uitputtende behandeling is hier niet aan de orde. In dit hoofdstuk wordt niet meer dan een eerste aanzet gegeven tot het analyseren van *sommige* vormen van similariteit tussen documenten in een omvangrijk tekstcorpus, om daarmee als het ware te demonstreren welke mogelijkheden op dit gebied bestaan. Gezien deze beperking is het onvermijdelijk dat niet alle vragen die aan de orde komen in dit hoofdstuk al beantwoord kunnen worden (hoewel daar in een aantal gevallen al wel een poging toe zal worden gedaan). Het verkennende karakter van dit onderzoek is tevens de reden voor een belangrijke beperking die gesteld is ten aanzien van het tekstmateriaal waarvan gebruik wordt gemaakt. Om de mogelijkheden van een similariteitsmaat te illustreren zal namelijk slechts gebruik worden gemaakt van één van de twee juridische corpora, te weten het corpus *wetgevingsteksten*. De belangrijkste redenen om voor dit corpus te kiezen zijn:

- het is een juridisch corpus, waardoor het onderzoek kan bijdragen aan het verkrijgen van kennis over overeenkomsten tussen bepaalde juridische documenten.
- de omvang van het corpus is weliswaar fors, maar het blijkt toch nog juist mogelijk de benodigde analyses uit te voeren met een personal computer (zoals ook het geval was in de voorgaande hoofdstukken). Bij het corpus jurisprudentieteksten zouden op dit punt problemen zijn ontstaan.
- De manier waarop het corpus in documenten is ingedeeld is weliswaar nauwkeurig gespecificeerd (zie hoofdstuk 3), maar draagt toch een enigszins arbitrair karakter. De grootste eenheden die in het corpus kunnen worden onderscheiden zijn de wetten. Deze werden echter in de meeste gevallen in een aantal afzonderlijke documenten gesplitst. Het is interessant om na te gaan of aan de hand van de similariteit van die documenten de grotere structuren van het corpus (wetten, of reeksen van wetten met bepaalde overeenkomsten) ook weer zichtbaar worden.

Belangrijk is dat het er in dit hoofdstuk in ieder geval *niet* om gaat nieuwe punten van overeenkomst of verschil tussen de corpora op te sporen. In plaats daarvan staat nu de vraag centraal of kwantitatieve gegevens over teksten, in dit geval vooral met betrekking tot de spreiding van woordtypen over die teksten, waarvan al is aangetoond dat zij een rol kunnen spelen bij het vergelijken van het taalgebruik in verschillende corpora, daarnaast ook een rol kunnen spelen bij het bestuderen van bepaalde aspecten van een op zichzelf staand corpus, en meer in het bijzonder van de interne structuur van dat corpus.

---

[1] Dergelijk onderzoek zou samenhangen met onderzoek dat binnen de taalkunde wordt verricht op het gebied van 'onderwerp-toekenning', het (eventueel automatisch) vaststellen waar een tekst over gaat door middel van syntactische en semantische analyse.

8.1.2 Vormen van similariteit

Het is van belang om zo nauwkeurig mogelijk te omschrijven wat in dit hoofdstuk precies wordt verstaan onder het begrip similariteit, gezien de centrale rol die dat begrip hier speelt. De term similariteit vindt zijn oorsprong in het Latijn (van similis = gelijk). Noch Van Dale (twaalfde druk, 1992), noch het Woordenboek der Nederlandse Taal vermeldt het woord echter. Wanneer we ons baseren op de betekenis van het Engelse 'similarity', kan onder similariteit worden verstaan:

1. vergelijkingspunt - gelijkenis, punt van overeenkomst
2. vergelijkbaarheid- gelijksoortigheid, soortgelijkheid, gelijkvormigheid, overeenkomst[2].

De tweede betekenis is hier het meest van toepassing. We kunnen dus aannemen dat similariteit tussen objecten of tussen begrippen bestaat, wanneer deze in een bepaald opzicht vergelijkbaar zijn, of in ieder geval zekere punten van overeenkomst hebben. De woorden 'vergelijkbaarheid' en vooral 'overeenkomst' komen we in Nederlandse teksten regelmatig tegen. Met het woord 'similariteit' is dat echter niet het geval, anders dan in het Engels waar 'similarity' een zeer algemeen en dikwijls gebezigd woord is. Zoals hierboven vermeld is het ook niet te vinden in de twee belangrijkste woordenboeken voor het Nederlands taalgebied. Toch maakt het deel uit van de vaktaal die in sommige takken van wetenschap wordt gebruikt, bijvoorbeeld in de wiskunde. Het proefschrift van Hendrickx (1994), waarover hieronder meer, is wat dat betreft illustratief. De meeste publikaties die gaan over het soort overeenkomsten tussen teksten die in dit hoofdstuk aan de orde komen zijn in de Engelse taal, en maken gebruik van de term 'similarity'. Als vertaling van deze term is veel te zeggen voor 'similariteit', waarmee beter dan met de termen 'overeenkomst' of 'gelijkenis' kan worden uitgedrukt dat tussen de teksten of documenten niet zozeer één bepaald punt van overeenstemming bestaat (bijvoorbeeld: ze zijn even lang of tellen evenveel woorden), maar dat die teksten in een bepaalde *relatie* tot elkaar staan, die beheerst wordt door *een aantal* punten van overeenkomst, en misschien ook wel van verschil. 'Similariteit' is dus een ruim begrip, het kan duiden op een *stelsel* van overeenkomsten en verschillen. Allerlei vormen van similariteit zijn denkbaar. In het kader van het hier beschreven onderzoek kunnen daarvan worden genoemd:

1. Similariteit als *wiskundige relatie*.

Hendrickx (1994) bespreekt de relatie wiskunde-werkelijkheid, en introduceert de *similariteitsrelatie*, welke zijn oorsprong hierin vindt dat de nauwkeurigheid waarmee we objecten uit de werkelijkheid kunnen observeren eindig is[3]. Tussen twee objecten die 'dichter bij elkaar liggen' dan de kleinste 'afstand' die we kunnen waarnemen bestaat volgens zijn definitie een similariteitsrelatie. De inhoud die hiermee aan het begrip similariteit wordt gegeven is aan de ene kant zeer ruim: similariteit kan bestaan tussen elk soort objecten. Aan de andere kant wordt similariteit hier eng geïnterpreteerd: er is alleen sprake van een similariteitsrelatie als tussen objecten een hoge mate van overeen
komst bestaat (het verschil tussen de objecten moet kleiner zijn dan we kunnen waarnemen).

---

[2] Bron: Van Dale Groot woordenboek Engels-Nederlands, tweede druk 1989, onder redactie van W. Martin en G.A.J. Tops. Van Dale Lexicografie, Utrecht/Antwerpen 1989.

[3] Hendrickx (1994, p. 15).

2. Similariteit tussen bepaalde *taalelementen van beperkte omvang*, zoals woorden of combinaties van woorden.

Van deze vorm van similariteit is in de literatuur een groot aantal voorbeelden te vinden. Het gaat daarbij, net als trouwens bij de categorie die hierna nog aan de orde komt, meestal om onderzoek dat in het teken staat van data-retrieval (terugzoeken van gegevens) in omvangrijke dataverzamelingen. Zo is door Stiles in 1961 een 'association factor' (een soort similariteitsmaat) tussen de aan een document toegekende *index-woorden* gedefinieerd[4]. Index-woorden die vaak samen voorkomen geeft hij een hoge association factor. Wanneer de gebruiker vervolgens op één van de index-woorden zoekt, worden ook de hiermee geassocieerde woorden in het zoekproces betrokken. Het resultaat is een soort automatische thesaurus- of synoniemenlijst-functie. Het onderzoek van Lesk (1969) vertoont hiermee veel overeenkomsten[5]. Ook hij gaat op zoek naar woorden die samen voorkomen in een document, met het idee om zo de terughaalfactor (recall) bij zoekoperaties in een database te vergroten. Zijn conclusie luidt echter dat de op deze wijze bereikte verbeteringen minder groot zijn dan wanneer gebruik wordt gemaakt van een (niet-automatisch samengestelde) thesaurus. Recenter onderzoek, van de hand van Cross and Kuo (1988), richt zich niet zozeer op associaties tussen woorden die samen voorkomen in één en hetzelfde document, maar op similariteit tussen woorden of combinaties van woorden in twee of meer *afzonderlijke* documenten. Zij beschrijven onder andere een efficiënt algoritme om 'largest common substrings' op te sporen, en passen dat toe in een databasesysteem met gegevens over handelsmerken[6]. Hoewel in elk van de hier genoemde publikaties interessante punten te vinden zijn, gaat het steeds om beperkte vormen van similariteit. De toegepaste algoritmen lenen zich niet voor het bepalen van de similariteit tussen documenten.

3. Similariteit tussen volledige *documenten*.

Het onderzoek in dit hoofdstuk gaat over deze vorm van similariteit. Ook hier is in het verleden al vrij veel onderzoek verricht, opnieuw meestal tegen de achtergrond van de opbouw en het gebruik van database systemen. Genoemd kan bijvoorbeeld worden het onderzoek van Salton in het kader van het SMART-project (een project waarin onderzoek werd gedaan naar diverse nieuwe technieken op het gebied van database management). In het overzichtsrapport 'The SMART System' (Salton 1971a) beschrijft hij een methode om in een database clusters te vormen van documenten die bepaalde overeenkomsten vertonen[7]. Daartoe gaat hij eerst na welke documenten als *centraal* voor een bepaald onderwerp kunnen gelden, en bepaalt de mate van similariteit van de overige documenten met deze centrale documenten. De keuze van centrale documenten kan eventueel tijdens dit proces worden bijgesteld. Uiteindelijk ontstaan groepen documenten, elk gerepresenteerd door hun centrale document. Als belangrijkste voordeel van het vormen van documentclusters noemt hij de mogelijkheid om het zoeken in de database te beperken tot bepaalde deelverzamelingen van documenten. Een nadeel van Salton's aanpak, evenals trouwens van de technieken die hierna zullen worden beschreven, is dat bij het vormen van clusters slechts gebruik wordt gemakt van an het document toege

---

4 Stiles (1961, p. 271).
5 Lesk (1969, p. 27-38).
6 Cross and Kuo (1988, p. 1-20).
7 Salton (1971b, p. 223 e.v.).

kende"idintifiers" (indexwoorden), wat een volledige automatische classificatie van documenten in de weg staat.

Een methode om uit een database sets van documenten te selecteren die onderling de hoogste mate van similariteit vertonen ("nearest neighbours" zijn) wordt beschreven door Smeaton and Van Rijsbergen (1981). Zij introduceren een algoritme om het zoeken naar documenten met een voldoende hoge mate van similariteit te bespoedigen door telkens een basis-document te kiezen, en daarmee alleen die documenten te vergelijken waarvan aan de hand van een inverted file kan worden vastgesteld dat ze *in theorie* een hogere mate van similariteit zouden kunnen vertonen dan de tot dat moment gevonden 'nearest neighbour'[8]. Een soortgelijke methodiek wordt beschreven door Willett (1981). Hij beschrijft een manier om, wederom met behulp van een inverted file, snel vast te stellen tussen welke documenten een zeer lage similariteit c.q. helemaal geen similariteit zal bestaan. Alleen de andere documenten, die bepaalde woordtypen gemeenschappelijk hebben en dus wel een bepaalde minimale similariteit vertonen, worden in meer detail vergeleken, waarbij het einddoel is - net als bij het onderzoek dat in dit hoofdstuk wordt beschreven - om documenten met hoge similariteit samen te voegen in een cluster[9]. Helaas geldt voor beide algoritmen, zowel voor dat van Smeaton and Van Rijsbergen als voor dat van Willett, dat ze alleen bruikbaar zijn wanneer de similariteit tussen documenten wordt berekend aan de hand van *aantallen* gemeenschappelijke woordtypen, en niet wanneer (zoals in dit hoofdstuk het geval is) wordt gewerkt met woordtypen waaraan een verschillend *gewicht* wordt toegekend.
Een recent artikel waarin een aantal similariteitsmaten met elkaar wordt vergeleken is "Comparing Similarity Measures" (Batagelj and Bren 1993). De auteurs definiëren naast het begrip similariteit ook *dissimilariteit* (afstand) tussen objecten, en gebruiken deze maat om de relaties tussen similariteitsmaten die in de literatuur worden voorgesteld uit te drukken[10]. Op enkele van deze similariteitsmaten zal verderop in dit hoofdstuk nog worden terugkomen.

Dit onderzoek beperkt zich tot een bepaalde vorm van de onder 3 genoemde similariteit tussen *documenten*. Om welke vorm het precies gaat wordt beschreven in de volgende subparagraaf.

8.1.3 Similariteit tussen documenten

Tussen twee objecten, in dit geval stukken tekst ofwel documenten, kunnen natuurlijk allerlei vormen van similariteit bestaan. Deze zouden bij documenten bijvoorbeeld betrekking kunnen hebben op de volgende punten:

---

8 Smeaton and Van Rijsbergen (1981, p. 83-87).
9 Willett (1981, p. 53-60).
10 Batagelj and Bren (1993, p. 8-16).

1. het lettertype dat is gebruikt bij het afdrukken van de documenten;
2. de algehele layout (pagina-opmaak, uiterlijk van kopjes, etc.);
3. het aantal woordtokens en -typen waaruit elk document bestaat;
4. de gebruikte woordtypen zelf, en hun frequentie;
5. de gemiddelde/maximale/minimale woord-, zins-, alinea- of paginalengte;
6. de taal waarin de documenten zijn gesteld;
7. de doelgroep waarop ze zich richten;
8. het onderwerp waarover ze gaan;
9. het effect dat ervan uitgaat (bijvoorbeeld wanneer ze worden gepubliceerd of gelezen);

etc.

De volgorde van dit rijtje mogelijke overeenkomsten is niet toevallig, maar gaat van zuivere vorm-overeenkomsten (1, 2, 3, 4, 5) via overeenkomsten waaraan verschillende aspecten zitten (6, 7) tot overeenkomsten die de betekenis (semantiek) betreffen (8) ofwel die vooral pragmatisch van aard zijn (9). In het kader van dit onderzoeksproject ligt het het meest voor de hand om aandacht te schenken aan de overeenkomsten in vorm. In feite zullen we ons bij het bepalen van de mate van similariteit tussen documenten zelfs nog beperken tot een enkel soort vormovereenkomst, namelijk die genoemd onder 4 (gebruikte woordtypen, eventueel met hun frequentie), hoewel ook overeenkomst 3 (aantal tokens/typen) in zekere mate een rol zal spelen. Een belangrijk voordeel van deze beperking is dat de similariteit automatisch (dat wil zeggen zonder menselijke tussenkomst) kan worden bepaald. Het gaat dus om similariteit met betrekking tot het *woordgebruik*. Het begrip similariteit, zoals dat in het vervolg van dit hoofdstuk zal worden gebruikt, kan daarom als volgt worden gedefinieerd:

> *Onder similariteit wordt verstaan de mate waarin twee documenten voor wat betreft hun vorm en in het bijzonder voor wat betreft de woordtypen waaruit zij zijn opgebouwd met elkaar overeenstemmen.*

Nu zou men zich natuurlijk kunnen afvragen of het onderzoeken van dit soort vorm-overeenkomsten tussen documenten eigenlijk wel interessante informatie kan opleveren. Worden we daardoor ook maar iets wijzer over de structuur die het corpus heeft, of over de samenhang tussen de documenten? Bij die twee zaken speelt immers de betekenis van de documenten een rol? Het antwoord daarop kan voorlopig als volgt luiden: natuurlijk is het denkbaar dat we uiteindelijk tot de conclusie komen dat alle documenten die wat hun vorm (lees: woordgebruik) betreft op elkaar lijken inhoudelijk totaal geen samenhang vertonen. Het omgekeerde is echter ook zeer wel mogelijk. Het is immers zo dat een menselijke lezer ook vaak bepaalde overeenkomsten in de vorm van documenten kan waarnemen wanneer die over hetzelfde of een verwant onderwerp handelen. Zo zullen de meeste juristen al snel in de gaten hebben of een rechterlijke uitspraak waarvan ze een fragment ter lezing voor zich hebben gekregen een uitspraak in een civiel geding betreft of een strafrechtelijk vonnis, en zo zal vrijwel iedereen al met een enkele blik kunnen vaststellen of een stuk tekst een recept uit een kookboek dan wel een installatiehandleiding voor een computerprogramma betreft. Ook teksten afkomstig uit gelijksoortige bronnen vertonen vaak overeenkomsten in vorm. Zo beschikken vrijwel alle kranteartikelen over kenmerken als een 'kop', de aanduiding "van onze correspondent ..." of de naam

van een persbureau. Door bestudering van de vorm van documenten kunnen in sommige gevallen dus uitspraken worden gedaan over de bron waaruit ze afkomstig zijn, en soms ook over bepaalde betekenisaspecten ervan.

Nadat is geconstateerd dat tussen twee documenten bepaalde overeenkomsten in vorm bestaan, wordt natuurlijk de vraag actueel of daaraan ook gevolgen moeten worden verbonden. Om die vraag te kunnen beantwoorden lijkt het op het eerste gezicht noodzakelijk om de *inhoud* van de documenten te beoordelen. Een dergelijke beoordeling zou vrijwel onvermijdelijk in belangrijke mate gebaseerd zijn op *betekenisaspecten*, en zou om die reden waarschijnlijk niet volledig kunnen worden geautomatiseerd. Deze vorm van natuurlijke taal verwerking met behulp van automatische systemen levert namelijk ook in de huidige tijd nog steeds veel problemen op. Een illustratie daarvan vormt bijvoorbeeld het geringe succes dat tot nu toe geboekt is bij het *automatisch vertalen* van teksten, ondanks het feit dat ondertussen al grote bedragen zijn geïnvesteerd in onderzoek op dat terrein. Het lukt gewoonweg niet om teksten door een computer te laten vertalen wanneer die computer de betekenis ervan niet tot in alle finesses begrijpt. Daarvoor is onder andere een hoeveelheid 'kennis van de wereld' noodzakelijk waarover nog geen enkel computerprogramma beschikt[11]. Zoals echter al in hoofdstuk 2 (paragraaf 2.2) werd vermeld, wordt op het gebied van dit automatisch vertalen de laatste jaren steeds vaker gebruik gemaakt van nieuwe methoden. Pogingen om een steeds vollediger (maar nooit volledig genoeg) bestand met achtergrond-kennis op te bouwen worden gestaakt en in plaats daarvan wordt nu onder andere gebruik gemaakt van statistische analyse van eerdere (door menselijke vertalers opgestelde) vertalingen, om aan de hand daarvan de meest waarschijnlijke vertaling te bepalen. Geen vertaling aan de hand van kwalitatieve, maar van formele, kwantitatieve criteria dus.

Dit laatste sluit nauw aan bij de weg die in dit hoofdstuk zal worden bewandeld. Bij het bepalen van de similariteit van documenten zal uitsluitend gebruik worden gemaakt van *kwantitatieve methoden*. Soms zal, op basis van de mate van similariteit, de beslissing worden genomen om twee documenten te beschouwen als behorend tot één en hetzelfde cluster (groep van documenten). Ook deze beslissing zal dan steeds gebaseerd zijn op formele criteria. Het gehele proces, dat in het vervolg van dit hoofdstuk in detail zal worden beschreven, heeft in grote lijnen de volgende kenmerken:

- Bij het bepalen van de mate van similariteit wordt ieder document vergeleken met *alle* andere documenten in het corpus.
- *Alle* woordtypen in een document zijn betrokken bij het bepalen van de mate van similariteit met een ander document. Wel is het zo dat woordtypen een verschillend 'gewicht' kunnen hebben, afhankelijk van hun mate van spreiding over de documenten van het corpus.
- De mate van similariteit tussen documenten wordt geacht te stijgen wanneer een bepaald woordtype in beide documenten *voorkomt*, of juist in beide *ontbreekt*.
- Nadat alle documenten in het corpus met elkaar zijn vergeleken worden clusters gevormd bestaande uit die documenten die de hoogste mate van similariteit met elkaarblijken te vertonen.

---

[11] vgl. Brandt Corstius (1978, p. 20-24).

- Bij het vormen van clusters kunnen nog extra eisen worden gesteld aan een documentenpaar, bijvoorbeeld dat de hoge mate van similariteit wederzijds is, of dat de similariteit een bepaalde drempelwaarde te boven gaat.
- De vorming van clusters wordt gedocumenteerd, in die zin dat wordt bijgehouden welke clusters in welke volgorde ontstaan, en tevens welke *woordtypen* door hun aan- of afwezigheid het meeste hebben bijgedragen tot de vorming van elk van de clusters. Deze documentatie kan mogelijk een rol spelen bij het beoordelen van de kwaliteit van de gevormde clusters.

8.2 Basisbegrippen met betrekking tot similariteit

8.2.1 Enkele uitgangspunten

We beschouwen in dit hoofdstuk elk corpus zoals gezegd niet langer als één geheel, als een tekststroom waarin duizenden woordtypen één of meer keren voorkomen, maar we gaan uit van de indeling van het corpus in documenten. Het is misschien goed om nog even in herinnering te roepen welke documenten in het corpus wetgevingsteksten, dat in dit hoofdstuk centraal staat, zijn vertegenwoordigd. Zoals in hoofdstuk 3 werd beschreven kon bij dit corpus de indeling van het basismateriaal - losse wetsartikelen, elk voorzien van referentiegegevens omtrent de wet of regeling waaruit ze afkomstig waren - anders dan bij de andere twee corpora niet zonder meer worden overgenomen. Onder andere zouden daardoor namelijk zeer veel tekstgedeelten (bijvoorbeeld de titels van wetten) een aantal keren zijn herhaald in het corpus, terwijl ook het aantal documenten te hoog zou zijn geworden (meer dan 100000). Het eerste bezwaar was met name relevant voor het opstellen van frequentielijsten en dergelijke, in de voorgaande hoofdstukken, het tweede bezwaar heeft vooral betrekking op het onderzoek dat in dit hoofdstuk wordt beschreven. Uiteindelijk werd het corpus, volgens nauwkeurig omschreven criteria, ingedeeld in ruim 18000 nieuwe documenten. Basisprincipe daarbij was dat de teksten werden ingedeeld op één niveau boven dat van de wetsartikelen. Bij de meeste wetten hield dat in dat steeds een 'titel' of paragraaf als afzonderlijk document werd beschouwd. Elk van die documenten bestond dus uit één of meer van de oorspronkelijke (wetsartikel-) documenten, die echter waren ontdaan van alle 'dubbele' informatie. Om verschillen in omvang tussen de documenten te beperken werden tenslotte nog kleine, opeenvolgende documenten samengevoegd.

Om deze documenten gaat het in dit hoofdstuk. Voor ieder afzonderlijk document in het corpus zal de mate van similariteit worden bepaald met elk ander document uit dat corpus. Daarna zal worden nagegaan welke documentenparen (wederzijds) de grootste similariteit vertonen. Deze documentenparen zullen nader worden bestudeerd, waarbij onder andere zal worden onderzocht welk effect het heeft wanneer de twee documenten verder als één enkel document worden behandeld (zie de paragrafen 8.4 en 8.5).

Vooral om praktische redenen is het nodig enige beperkingen te hanteren. De belangrijkste is dat bij het bepalen van de similariteit tussen documenten slechts rekening zal worden gehouden met het al dan niet *voorkomen* van een bepaald woordtype in een document, zonder daarbij ook nog de woordfrequentie *binnen* dat document te betrekken. Nog maar één frequentie-gegeven blijft dan voor ieder woordtype in het corpus van belang: het *aantal documenten* waarin het type voorkomt. Dit gegeven, dat we zouden kunnen aanduiden met de term *documenten-frequentie*, komt overeen met de *spreiding* van woordtypen die in eerdere hoofdstukken onder andere in de diverse frequentielijsten stond vermeld, alleen is deze spreiding nu niet uitgedrukt in een percentage (van het aantal documenten), maar in een absoluut aantal. Mocht uit dit eerste onderzoek blijken dat het afzien van de woordfrequentie binnen een document er bijvoorbeeld toe leidt dat geen bruikbare similariteitsscore kan worden vastgesteld - dat zou misschien het geval kunnen zijn wanneer bepaalde woordtypen een zeer ongelijke spreiding zouden vertonen, en in sommige documenten zeer vaak zouden voorkomen en in andere zeer weinig - dan kan de voorgestelde methode
zonder al te veel moeite worden aangepast, zodanig dat ook woordfrequenties worden meegewogen bij het bepalen van de mate van similariteit.

8.2.2 'Hits' en 'misses'

Wanneer we de mate van similariteit van twee documenten willen vaststellen aan de hand van de woordtypen die in die documenten voorkomen, kunnen zich op het eerste gezicht de volgende twee situaties voordoen:

- een woordtype komt in beide documenten voor; aangezien de documenten hiermee een punt gemeenschappelijk hebben, zou dit de mate van similariteit moeten *verhogen*. Voor deze situatie wordt hier de term *'hit'* geïntroduceerd.
- een woordtype komt in het ene document wel voor, maar in het andere niet; de documenten wijken op dit punt van elkaar af, de mate van similariteit zou dan ook moeten *dalen*. Deze situatie kan worden aangeduid als een *'miss'*.

Met name bij de misses kan het nogal een verschil maken *vanuit welk document* we de zaak bekijken, wanneer de documenten tenminste niet precies even groot zijn (wat buitengewoon toevallig zou zijn). Stel, we vergelijken 2 documenten, het ene bevat 100 typen, het andere 200. Er blijken 50 hits te zijn (50 typen komen in beide documenten voor), zodat het aantal misses bezien vanuit het eerste document 50 bedraagt, en bezien vanuit het andere 150, drie keer zo veel dus. We kunnen hieruit direct al concluderen dat het vergelijken van documenten een relatief gebeuren is; het hang er vaak maar net van af welk document men als uitgangspunt neemt. Meestal kan dan ook niet worden volstaan met het vergelijken van document X met document Y, maar moet daarnaast document Y ook worden vergeleken met document X.

Wanneer het alleen gaat om de misses is echter ook nog een andere verbetering denkbaar: we gaan in plaats van met één, met twee typen misses werken. Wanneer we weer uitgaan van twee documenten, X en Y, en document X als basis nemen, is sprake van:

- een *'miss-type-1'* (kortweg aangeduid als 'miss1') wanneer een woordtype in document X wel voorkomt, maar niet in document Y; en van
- een *'miss-type-2'* (kortweg 'miss2') wanneer een woordtype in document Y voorkomt, maar ontbreekt in X.

Met deze drie gegevens, het aantal hits, het aantal miss1's en het aantal miss2's kan de verhouding tussen twee documenten op zich effectief worden vastgelegd. Wanneer het echter niet alleen gaat om twee op zichzelf staande documenten, maar wanneer deze deel uitmaken van een database c.q. corpus met nog veel meer documenten, is er nog een vierde gegeven. Er zal in deze database c.q. in dit corpus namelijk waarschijnlijk nog een flink aantal *andere* woordtypen voorkomen die noch in document X, noch in document Y te vinden zijn. Goed beschouwd is het zelfs zo dat het ontbreken van een woordtype in zowel document X als document Y kan worden opgevat als een punt van *overeenkomst* tussen de documenten, dat dus hun mate van similariteit *verhoogt*. Hier is daarom ook sprake van een soort 'hit', net als in het geval van typen die in beide documenten juist wel voorkomen. Naast twee soorten misses zijn in totaal dus ook twee soorten hits denkbaar:

- een *'hit-type-1'* (kortweg 'hit1'), wanneer een woordtype in beide documenten voorkomt; en
- een *'hit-type-2'* (kortweg 'hit2'), wanneer een woordtype in beide documenten *niet* voorkomt (maar elders in het corpus wel).

Aan de vier mogelijke situaties die hier worden onderscheiden (2 soorten hits en 2 soorten misses) wordt ook in de literatuur gerefereerd. Daarbij worden de aantallen van elk type hit of miss nogal eens aangeduid met de letters *a*, *b*, *c* en *d*. In dat geval staat *a* meestal voor het aantal hit1's, *b* en *c* voor de aantallen miss1's en miss2's en *d* voor het aantal hit2's[12].

8.2.3 Het 'gewicht' van een hit of miss

Het is een feit dat niet ieder woordtype even vaak voorkomt in een corpus. Dat hoeft echter op zich nog geen invloed te hebben op de mate van similariteit van documenten, aangezien zoals gezegd de frequentie van een woordtype *binnen een document* daarbij geen rol speelt. Wel van invloed op de similariteit kan echter zijn de mate van *spreiding* van de woordtypen over het corpus, uitgedrukt in het *aantal documenten* waarin een bepaald woordtype voorkomt (de hierboven al beschreven *documenten-frequentie*). Omdat 'documenten-frequentie' een wat omslachtige term is, zal in het vervolg van dit hoofdstuk alleen nog worden gesproken over de 'frequentie' van woordtypen, waarmee dus van nu af aan uitsluitend nog wordt bedoeld: het aantal documenten uit het corpus waarin een bepaald woordtype te vinden is. *Deze* frequentie van woordtypen kan dus wel een rol spelen bij de berekening van de mate van similariteit van twee documenten. De kans dat een woordtype met hoge frequentie in beide documenten wordt aangetroffen en dus een 'hit1' oplevert is immers veel hoger dan de kans dat een type met lage frequentie dat doet. Anderzijds zal de kans op een hit2 juist weer groter zijn bij typen met een lage frequentie. Voor de twee typen misses geldt mutatis mutandis hetzelfde. In tabelvorm:

[12] Zie bijvoorbeeld Batagelj and Bren (1993, p. 7).

Kans op:	Frequentie van woordtypen: Hoog	Laag
Hit1	groot	klein
Hit2	klein	groot
Miss1	klein	groot
Miss2	groot	klein

Tabel 8-1 Het verband tussen frequentie en kans op hits en misses

Overwogen moet hierbij worden dat we voor de kans op hits en misses steeds uitgaan van één document. Laten we dit document X noemen. De vraag is dan steeds: hoe groot is de kans dat een zeker woordtype, dat op zich in veel resp. in weinig documenten voorkomt, in een willekeurig document Y voorkomt resp. ontbreekt, gegeven dat het in document X voorkomt resp. ontbreekt. De frequentie van een woordtype is dus rechtstreeks van invloed op de kans dat dat woordtype voor twee documenten een hit1, hit2, miss1 of miss2 oplevert. En dat heeft natuurlijk weer gevolgen voor het *belang* dat aan iedere hit of miss gehecht kan worden. Wanneer op een bepaalde gebeurtenis immers slechts een kleine kans bestaat, en die gebeurtenis doet zich toch voor, dan levert dat een grotere hoeveelheid informatie op, wat samen zou moeten gaan met een hogere similariteitswaarde, dan wanneer de kans op de gebeurtenis heel groot was. Zo zegt het bijvoorbeeld heel veel wanneer in twee documenten uit het corpus wetgevingsteksten die we aan het vergelijken zijn het woord AANKOOPBEDRAG voorkomt. Dat woord komt namelijk ook maar in twee documenten voor, en dat zijn dan net de documenten die we nu aan het vergelijken zijn. Het zegt daarentegen heel weinig wanneer we in beide documenten het woord VAN tegenkomen, aangezien dat woord in praktisch alle documenten voorkomt (om precies te zijn: in 18576 van de 18803). Dat betekent dat, wanneer we de similariteit van twee documenten bepalen, we niet kunnen volstaan met het optellen van de aantallen hits en misses. Bij iedere hit en iedere miss moet namelijk nog worden gecorrigeerd voor de *kans* dat die hit of miss zich voordoet in het onderzochte corpus. In de praktijk werkt dit zo, dat voor ieder woordtype van tevoren wordt uitgerekend wat de kans is dat dat type 'betrokken is' bij een hit1, een hit2, een miss1 en een miss2, en dat later, bij het 'turven' van de aantallen hits en misses per documentenpaar, met dit kans wordt vermenigvuldigd. Om de kansen uit te rekenen hoeven we nog niet te weten welke documenten vergeleken zullen worden, want deze volgen direct uit de *frequentie* van het woordtype in kwestie, en uit het aantal documenten in het corpus. In feite gaat het om twee verschillende kansen:

1. de kans om het woord *aan te treffen* in een bepaald document; met die kans hebben we te maken in geval van een hit1 of een miss2, omdat in die gevallen het woordtype in het document waar we mee vergelijken *aanwezig* is.
2. de kans om het woord *te missen* in een bepaald document; om die kans gaat het in geval van een hit2 of een miss1, omdat daarbij het woordtype in het document waar we mee vergelijken *niet aanwezig* is.

Kans nummer 1 is gelijk aan de *frequentie van het woordtype gedeeld door het aantal documenten in het corpus*. Kans nummer 2 is gelijk aan het *verschil tussen het aantal*

documenten en de frequentie van het woordtype, gedeeld door het aantal documenten in het corpus. In formulevorm:

$$P(i_{HIT\,1,MISS\,2}) = \frac{F_i}{D} \qquad (8\text{-}1)$$

$$P(i_{HIT\,2,MISS\,1}) = \frac{D - F_i}{D} \qquad (8\text{-}2)$$

Waarbij F_i staat voor de frequentie van woordtype *i* (het aantal documenten waar woordtype *i* in voorkomt dus) en *D* voor het aantal documenten in het gehele corpus. Daarbij moet nog worden opgemerkt dat strikt genomen deze kansen alleen gelden wanneer we ervan uitgaan dat het basisdocument (document X, in de terminologie die hiervoor werd gebruikt) zelf *geen* deel uitmaakt van de database. Is dat wel het geval, dan zou eigenlijk *D* in de noemer van beide breuken, en in het geval van een hit1 of een hit2 ook het getal in de teller, met 1 moeten worden verminderd. Om verschillende redenen is daarvan hier echter afgezien. De voornaamste reden is dat door af te zien van de correctiefactor een *symmetrisch* model ontstaat, waarbij de *aan*wezigheidskans van een woordtype dat in *ieder* document voorkomt gelijk is aan de *af*wezigheidskans van een type dat in *geen enkel* document voorkomt (beide kansen zijn gelijk aan 1), terwijl ook de *aan*wezigheidskans van een type dat in *geen enkel* document voorkomt gelijk is aan de *af*wezigheidskans van een type dat in *ieder* document voorkomt (beide gelijk aan 0). Het is daardoor ook niet nodig bij de berekening van de kansen allerlei voorwaarden na te lopen, om te voorkomen dat een kans negatief wordt of groter dan 1. Verder is ook de eenvoud gediend met het weglaten van de correctiefactor (er ontstaat een identieke berekening voor zowel hits als misses) terwijl bovendien het effect ervan voor de meeste typen gering is wanneer het aantal documenten hoog is (zoals bij het onderzochte corpus).

De kans op bijvoorbeeld een hit1 is dus hoog bij die woordtypen die in veel documenten voorkomen en de kans op een hit2 is hoog bij die woordtypen die maar in weinig documenten voorkomen. Tegelijkertijd 'zegt' zo'n hit1 resp. hit2 dan het minste voor wat betreft de similariteit van de twee documenten die worden vergeleken. Wanneer we daarmee rekening willen houden, zullen we iedere hit of miss daarom moeten vermenigvuldigen met een 'gewichtsfactor' *W*, met een waarde tegengesteld aan de kans. De twee meest voor de hand liggende methoden om deze gewichtsfactor te berekenen zijn door gebruik te maken van het complement of van de reciproque van de kans:

1. $W(i) = 1 - P(i)$ (8-3)

2. $W(i) = \frac{1}{P(i)}$ (8-4)

De tweede methode heeft echter het ernstige nadeel dat voor de gewichtsfactor een range geldt van [1..∞], waardoor de similariteitswaarden die met behulp van dit gewicht worden berekend zeer hoog zouden kunnen worden, met name wanneer het corpus omvangrijk is en de kans op aanwezigheid van sommige (laagfrequente) woordtypen daardoor zeer laag is. Verder moet bij deze methode, omdat de kans de noemer van een breuk vormt, voortdurend worden gecontroleerd of een kans niet gelijk is aan 0. Om deze redenen is ervoor gekozen het gewicht van een woordtype te berekenen volgens de eerste methode (8-3), en dus gelijk te stellen aan het complement van de kans op dat woordtype. Zie

echter ook paragraaf 8.2.6, hierna.

Met behulp van de op deze wijze berekende gewichtsfactoren, die we voor ieder woordtype kunnen uitrekenen, kunnen we voor een documentenpaar het *gecorrigeerd* aantal hit1's, hit2's, miss1's en miss2's bepalen. Die aantallen zijn namelijk gelijk aan de som van de gewichten van de betrokken woordtypen. Het gecorrigeerd aantal hit1's is dus bijvoorbeeld gelijk aan:

$$Hit1_{GECORR.} = \sum_{i=1}^{n} W(i) \qquad (8\text{-}5)$$

waarbij *n* het aantal hit1's voorstelt, ofwel het aantal typen dat in beide documenten voorkomt. De aantallen hits en misses die op deze manier worden berekend voor een documentenpaar zijn bruikbaar voor het bepalen van de similariteit van deze documenten, aangezien nu geen rol meer speelt *welke woordtypen* bij een hit of een miss betrokken waren.

8.2.4 Hit en miss percentages

Door het toepassen van 'gewichten' hebben we dus gecorrigeerd voor het feit dat een hit of miss meer of minder 'waard' kan zijn naar gelang de frequentie van het woordtype. Een maat voor de similariteit van documenten moet echter nog aan een andere voorwaarde voldoen: zij moet vergelijkbaar zijn voor verschillende documentenparen. Aangezien de *aantallen woordtypen per document* sterk uiteen kunnen lopen, zijn de absolute *aantallen* hits en misses (al dan niet gecorrigeerd) daarom niet zonder meer bruikbaar voor het berekenen van de similariteit. De meest voor de hand liggende oplossing voor dit probleem is om naast absolute, ook relatieve aantallen hits en misses, ofwel hit- en miss*percentages* te berekenen. De aantallen zouden dan een percentage moeten vormen van het *maximaal mogelijke* aantal hit1's, hit2's, miss1's of miss2's in een documentenpaar. Deze maximaal mogelijke aantallen kunnen voor ieder document in het corpus worden berekend door de hit- c.q. miss*gewichten* van alle woordtypen in het document bij elkaar op te tellen, op de volgende manier:

1. We gaan uit van de lijst van alle woordtypen die in alle documenten samen voorkomen.
2. Voor ieder woordtype dat in het basisdocument (document X) *wel* voorkomt verhogen we het maximale aantal hit1's met het bij dat woordtype behorende hit1-gewicht $(1 - (F / D))$, en het maximale aantal miss1's met het miss1-gewicht van dat type $(1 - ((D - F) / D))$. Bij een hit1 en een miss1 gaat het er immers om dat een woordtype *wel* in het basisdocument voorkomt (en ook, resp. en niet in het andere document).
3. Voor ieder woordtype dat in het basisdocument *niet* voorkomt verhogen we het maximale aantal hit2's met het bij dat type behorende hit2-gewicht $(1 - ((D - F) / D))$, en het maximale aantal miss2's met het miss2-gewicht van dat type $(1 - (F / D))$. Bij een hit2 en een miss2 gaat het immers om woordtypen die *niet* in het basisdocument voorkomen (en ook niet, resp. en wel in het andere document).

We verkrijgen op die manier het gecorrigeerde, maximale aantal hit1's, hit2's, miss1's en miss2's. Door de gecorrigeerde hit- en missaantallen uit de vorige subparagraaf te delen door deze maximale aantallen (en te vermenigvuldigen met 100%) worden de gecorrigeer-

de hit- en miss*percentages* verkregen. Met behulp van deze percentages

berekende similariteitsgetallen voor verschillende documentenparen kunnen onderling worden vergeleken, ook wanneer de documenten een verschillende omvang hebben.

8.2.5 Typen die in ieder document voorkomen

Wanneer we, zoals in de voorgaande twee subparagrafen, aan het rekenen zijn en bijvoorbeeld de aantallen typen tellen die wel of niet in een document voorkomen, is het zaak om ook rekening te houden met een bijzondere situatie. Het kan namelijk zo zijn dat in een document een type blijkt voor te komen, dat ook in *alle andere* documenten voorkomt. In de praktijk blijken er meestal maar weinig van dat soort typen te zijn, als ze er zijn gaat het om DE, VAN, HET, en dergelijke. Zo'n type dat in ieder document voorkomt is bijzonder omdat het voor het meten van de similariteit geen waarde heeft. Het *aan*wezig zijn van het type in een document kan geen verhoging van het aantal hit1's of miss1's opleveren (aangezien $F = D$, waardoor $P(i) = 1$, en dus $W(i) = 0$), het *af*wezig zijn in een document is niet mogelijk, maar zou theoretisch een hit2- en miss2gewicht van 1 opleveren ($D - F = 0$, $P(i) = 0$, $W(i) = 1$).

Goed beschouwd betekent een type dat in ieder document voorkomt voor de similariteit dus hetzelfde als een type dat in geen enkel document voorkomt: het heeft op de similariteit geen invloed. Het meest correct is dan ook om bij het bepalen van het aantal typen in het corpus, en van het aantal typen in een document, de typen die in ieder document voorkomen al niet mee te tellen, net zo min als typen die nergens in het corpus voorkomen worden geteld als afwezig in alle documenten (het aantal afwezige typen zou dan trouwens oneindig groot zijn). Verder heeft een en ander met name gevolgen voor het berekenen van de *maximale* aantallen hits en misses, zoals beschreven in de vorige subparagraaf. De procedure die daar wordt beschreven moet zodanig worden aangepast, dat woordtypen die in ieder document voorkomen niet worden meegeteld bij met name het maximale aantal *miss1*'s. Het miss1gewicht van een dergelijk type is immers gelijk aan 1. Het maximale aantal hit1's behoeft echter geen speciale aandacht; het hit1gewicht van een type dat in ieder document voorkomt is immers 0.

Tot slot nog eenmaal de aantallen typen waar we nu mee rekenen:

- het aantal typen in het corpus: gelijk aan het aantal verschillende typen in alle documenten gezamenlijk, verminderd met het aantal typen dat in ieder document voorkomt.
- het aantal typen dat niet in het corpus voorkomt: in principe oneindig groot (gegeven de oneindigheid van taaluitingen), maar verwaarloosd, en gelijk gesteld aan 0: het corpus wordt als een gesloten geheel beschouwd.
- het aantal typen in een document: het aantal in het document aangetroffen typen, verminderd met het aantal typen dat in ieder document voorkomt.
- het aantal typen dat niet in een document voorkomt: het aantal typen in het corpus, verminderd met het aantal typen in het document (al dan niet meetellen van typen die in ieder document voorkomen maakt daarbij geen verschil).

Daaraan kunnen nog worden toegevoegd:

- het *gecorrigeerd* aantal typen dat maximaal in een document kan voorkomen. Dit is het (gecorrigeerd) aantal typen dat zou voorkomen in het 'universele document', een document dat alle typen uit het corpus zou bevatten. Dit aantal is gelijk aan de som van de 'aanwezigheids'-gewichten (hit1/miss2gewichten) van alle typen in het corpus, met uitzondering van de typen die in ieder document voorkomen.
- het *gecorrigeerd* aantal typen dat maximaal in een document kan ontbreken. Het betreft hier het (gecorrigeerd) aantal typen dat aanwezig zou zijn in het 'nul-document', een document dat geen enkel type bevat. Dit aantal is gelijk aan de som van de 'afwezigheids'-gewichten (hit2/miss1gewichten) van alle typen in het corpus, wederom met uitzondering van typen die in ieder document voorkomen.
- het *gecorrigeerd* aantal typen in een document. Dit is gelijk aan de som van de 'aanwezigheids'- (hit1-/miss2-) gewichten van alle typen in het document, met uitzondering van de typen in ieder document. Tevens dus gelijk aan het (gecorrigeerde) maximale aantal hit1's (zie de voorgaande subparagraaf). In het vervolg van dit hoofdstuk zal voor dit gegeven de term *Hit1Max* worden gebruikt.
- het *gecorrigeerd* aantal typen dat niet in een document voorkomt. Dit is gelijk aan de som van de 'afwezigheids'- (hit2-/miss1-) gewichten van alle typen die niet in het document voorkomen. Tevens dus gelijk aan het (gecorrigeerde) maximale aantal hit2's. Dit gegeven zal in het vervolg *Hit2Max* worden genoemd.

8.2.6 Het kwantificeren van de hoeveelheid informatie per hit of miss

Al eerder in dit hoofdstuk werd opgemerkt dat de hoeveelheid informatie die verkregen wordt door het optreden van een bepaalde gebeurtenis omgekeerd evenredig is met de kans op die gebeurtenis. Voor het bepalen van de similariteit betekent dat bijvoorbeeld dat de aanwezigheid van een zeldzaam woordtype in twee documenten meer invloed heeft op de mate van similariteit dan de aanwezigheid in beide documenten van een woordtype dat in bijna ieder document voorkomt. Bij deze constatering hoeven we het echter niet te laten. De informatietheorie verschaft namelijk een middel om de hoeveelheid informatie precies te meten, en uit te drukken in een getal, de hoeveelheid *bits*. 'Bit' is de afkorting van 'binary digit', ofwel een getal uit het tweetallig stelsel. Een bit kan slechts twee waarden vertegenwoordigen: 0 en 1. Als zodanig vormt 1 bit de kleinst mogelijke hoeveelheid informatie; het is de hoeveelheid informatie die we verkrijgen wanneer we het antwoord op een tweekeuzevraag ontvangen, wanneer de kans op elk van de antwoorden 50% bedraagt.

De hoeveelheid informatie, in bits, die we verkrijgen wanneer een bepaalde gebeurtenis plaatsvindt kan worden afgeleid uit de kans op die gebeurtenis. Daarvoor kan de volgende formule worden gebruikt[13]:

$$I(a) = {}^{2}\log\frac{1}{P(a)} \qquad (8\text{-}6)$$

[13] De Mulder (1984, p. 22).

I(*a*), de hoeveelheid informatie verkregen door gebeurtenis *a*, is gelijk aan de 2-logaritme van de reciproque van de kans op *a*. Met behulp van deze formule kunnen we dus de hoeveelheid informatie die we verkrijgen, iedere keer als we bij het vergelijken van twee documenten een hit of een miss vinden, uitrekenen. De kans op ieder type hit en miss is immers bekend. Opgemerkt kan nog worden dat formule 8-6 sterke overeenkomst vertoont met 8-4, waarin het gewicht van een hit of miss wordt gesteld op de reciproque van de kans. De toepassing van de logaritme in 8-6 zorgt er echter voor dat de nadelen van 8-4 (met name de hoge waarde bij typen met geringe frequentie) in dit geval niet of in veel geringere mate optreden. De hoeveelheid informatie in bits zal in grote lijnen parallel lopen met het overeenkomstige (gecorrigeerde) aantal hits of misses, maar maakt het vooral voor woordtypen met geringe frequentie mogelijk om de mate van overeenkomst c.q. verschil nauwkeuriger te berekenen. Een ander voordeel van het gegeven is verder dat het ook gebruikt kan worden om de hoeveelheid informatie in documenten of zelfs in hele corpora te vergelijken. Zo kan bijvoorbeeld worden vastgesteld of de hoeveelheid informatie in het corpus toe- of afneemt wanneer bepaalde documenten worden samengevoegd, of juist gesplitst.

Ook de hoeveelheid informatie kan weer worden uitgedrukt in een percentage. Daarvoor moet, op overeenkomstige wijze als bij de (gecorrigeerde) aantallen hits en misses, de maximale hit- (en eventueel miss-) informatie van ieder document worden bepaald. Het informatiepercentage is dan weer gelijk aan de hoeveelheid informatie, gedeeld door de maximale hoeveelheid informatie, vermenigvuldigd met 100%.

Tenslotte kunnen dan nog worden berekend:

- de maximale hoeveelheid informatie die een document theoretisch kan bevatten. Dit is de hoeveelheid informatie die aanwezig zou zijn in het 'universele document', een document dat alle typen uit het corpus zou omvatten. Dit aantal is gelijk aan de som van de informatie van de 'aanwezigheids'-gewichten (hit1/miss2 gewichten) van alle typen in het corpus, met uitzondering van de typen die in ieder document voorkomen[14]. Voor dit gegeven zal verder de term Hit1MaxInfo worden gebruikt.
- de maximale hoeveelheid informatie die theoretisch in een document kan ontbreken. Het betreft hier de hoeveelheid informatie die aanwezig zou zijn in het 'nul-document', een document dat geen enkel type bevat. Dit aantal is gelijk aan de som van de 'afwezigheids'-gewichten (hit2/miss1gewichten) van alle typen in het corpus, wederom met uitzondering van typen die in ieder document voorkomen. Dit gegeven zal worden aangeduid met de term Hit2MaxInfo.

[14] Merk op dat de hoeveelheid informatie van typen die in ieder document voorkomen gelijk is aan $^{2}\log 1 = 0$.

8.3 Het berekenen van een similariteits-score

Met behulp de in de voorgaande paragraaf beschreven variabelen kan voor ieder documentenpaar in een corpus een similariteits-score worden berekend. Bij deze berekening kan gebruik worden gemaakt van verschillende formules, waarvan er enkele in deze paragraaf de revue zullen passeren. Onderzocht zal worden welke formule in dit geval het beste kan worden gehanteerd, om tot een similariteitsscore te komen waarmee voor het gebruikte corpus effectief kan worden vastgesteld welke documenten de hoogste similariteit ten opzichte van elkaar vertonen.

8.3.1 Formules voor de berekening van de similariteitsscore

Nadat twee documenten zijn vergeleken, waarbij is vastgesteld hoeveel woordtypen ze gemeenschappelijk hebben c.q. een van beide of allebei missen, kunnen we beschikken over vier variabelen. Twee daarvan (de hit1's en hit2's) vormen een aanduiding van de mate van *overeenkomst* tussen de documenten, de andere twee (miss1's en miss2's) van de mate van *verschil*. De waarde van elk van deze variabelen kunnen worden gecompenseerd voor de *kans* op voorkomen c.q. ontbreken van de woordtypen waarop ze betrekking hebben[15]. Bovendien kunnen, zoals in de vorige paragraaf beschreven, voor ieder document worden berekend de *maximale* aantallen hit1's en hit2's (en indien gewenst ook de maximale aantallen miss1's en miss2's, door gebruik te maken van de gecorrigeerde aantallen typen die *ten hoogste* af- respectievelijk aanwezig kunnen zijn in een document). Door de aantallen hits (c.q. misses) te delen door de desbetreffende maximale waarde kan dus een hit- c.q. miss*percentage* worden verkregen. Het ligt voor de hand om een similariteitsscore te baseren op deze percentages aangezien op die manier in principe een score kan worden verkregen met een nauwkeurig bepaalde range, namelijk [0..1]. Verder is voor een similariteits-score essentieel dat daarin *hits* (een of beide typen) een rol spelen, aangezien een score die uitsluitend bepaald wordt door aantallen misses in feite een *dis*similariteitsmaat is (waarnaar we in dit geval niet op zoek zijn).

Wanneer we met het bovenstaande als uitgangspunt de literatuur raadplegen, treffen we een aantal mogelijke similariteitsmaten aan. Het overzichtsartikel van Batagelj and Bren (1993) vormt daarbij een waardevolle bron. Het vermeldt in totaal niet minder dan 22 verschillende similariteits- en dissimilariteitsmaten, samen weer afkomstig uit meer dan 15 verschillende andere bronnen. Zonder uitzondering gaat het in dit artikel echter om maten die bedoeld zijn voor het berekenen van de similariteit bij *eigenschappen met gelijk gewicht*. De consequentie daarvan is dat er bij elk van deze maten van wordt uitgegaan dat de som van het aantal Hit1's, Miss1's, Miss2's en Hit2's (door Batagelj and Bren steeds aangeduid met de letters *a*, *b*, *c* en *d*) gelijk is aan het totale aantal mogelijke eigenschappen (waarvoor zij de letter *m* gebruiken), in het geval van een tekstcorpus dus aan het totaal aantal typen in het corpus. Aangezien bij het vergelijken van teksten de woordtypen echter een *verschillend gewicht* hebben (zie paragraaf 8.2.3) is geen van de similariteits-

[15] Voorlopig zullen we ervan uitgaan dat deze compensatie bestaat uit het gebruikmaken van *gewichten*. Verderop in deze paragraaf zal ook worden ingegaan op het gebruik van de bijbehorende hoeveelheden *informatie*.

maten zonder meer bruikbaar. Wel is het zo dat sommige van deze maten zonder veel moeite kunnen worden omgezet in een zodanige vorm dat ze *wel* bruikbaar zijn bij eigenschappen van ongelijk gewicht. Het gaat daarbij maten voor de similariteit waarbij aantallen hits en (in een enkel geval) misses worden omgerekend tot een *percentage* van het totaal aantal eigenschappen[16]:

$$S_1 = \frac{a}{m} \qquad (8\text{-}7)$$

$$S_2 = \frac{a + d}{m} \qquad (8\text{-}8)$$

$$S_4 = \frac{a + d - b - c}{m} \qquad (8\text{-}9)$$

Doordat in deze drie gevallen de similariteit wordt berekend door het aantal hits en/of misses te delen door *m* (in ons geval: het aantal woordtypen), ontstaan een soort similariteits*percentages* met waarden in de range [0..1] c.q. [-1..1] (bij 8-9). Deze waarden kunnen worden vergeleken met die voor andere documentenparen. Aangezien een dergelijke omrekening naar percentages ook mogelijk is voor hits en misses die een verschillend gewicht hebben (door gebruik te maken van de betreffende maximale waarden, zie hierboven) kunnen deze formules, met enige wijzigingen, ook worden toegepast in het onderhavige geval. Om precies te zijn bestaan deze wijzigingen hieruit, dat in plaats van door *m* steeds gedeeld moet worden door de som van het gecorrigeerd aantal typen in het document en het gecorrigeerd aantal typen dat ontbreekt in het document (Hit1Max plus Hit2Max dus, zie paragraaf 8.2.5). Bij 8-7 ligt het echter het meest voor de hand - aangezien in de teller van de breuk slechts het aantal hit1's voorkomt - om te delen door Hit1Max. De similariteit zou anders alleen de maximale waarde (1) kunnen bereiken bij documenten die alle typen bevatten die in het corpus voorkomen, hetgeen ongewenst lijkt. De formules worden dan:

$$S_1 = \frac{Hit1}{Hit1Max} \qquad (8\text{-}10)$$

$$S_2 = \frac{Hit1 + Hit2}{Hit1Max + Hit2Max} \qquad (8\text{-}11)$$

$$S_4 = \frac{Hit1 + Hit2 - Miss1 - Miss2}{Hit1Max + Hit2Max} \qquad (8\text{-}12)$$

Daarbij gaat het steeds om de aantallen hits en misses, *gecorrigeerd* voor het gewicht van de woordtypen, zoals beschreven in paragraaf 8.2.3. Elk van deze formules zou in principe kunnen worden toegepast bij het bepalen van de similariteit van documenten. Kenmerken van de formules zijn:

- In formule 8-10 wordt uitsluitend gebruik gemaakt van het hit1-percentage. De formule geeft dus aan in hoeverre het basisdocument (X) een ander document (Y) overlapt, uitgedrukt in een percentage van het aantal typen waaruit het basisdocument

[16] Batagelj and Bren (1993, p. 8). Als plaats van herkomst voor de genoemde formules noemen zij, echter zonder volledige bronvermelding:
S_1: Russel and Rao (1940); S_2: Kendall, Sokal-Michener (1958); S_4: Hamann (1961).

bestaat (range: [0..1]). Een voordeel van deze simpele similariteitsmaat is dat deze vooral hoge waarden oplevert wanneer documenten voor wat betreft het woordgebruik voor een belangrijk deel samenvallen, onafhankelijk van hoeveel woordtypen ze samen *niet* bevatten. Dit kan vooral een voordeel zijn wanneer de documenten sterk in afmetingen verschillen. Het kleinste document zal dan waarschijnlijk de hoogste mate van similariteit vertonen. Een nadeel is dat totaal geen rol speelt welke typen in beide documenten *ontbreken*, waardoor een zeer belangrijk similariteitscriterium helemaal blijft liggen (er lijkt veel voor te zeggen dat het ontbreken in twee documenten van een type dat in vrijwel ieder ander document voorkomt de similariteit zou moeten verhogen). Om die reden is het ook niet erg waarschijnlijk dat bij gebruik van deze maat dikwijls sprake zal zijn van *wederzijdse* similariteitsverhoudingen (waarbij document X het meest lijkt op Y, en Y ook het meest op X), behalve wanneer de documenten ongeveer even groot zijn. Tenslotte is een mogelijk probleem van deze similariteitsmaat dat *bijna alle* documenten elkaar wel enigszins overlappen. Sommige woordtypen komen immers in praktisch elk document voor[17]. Deze typen zorgen er, ondanks hun geringe gewicht, voor dat een document met vrijwel alle andere documenten een zekere similariteit (berekend met behulp van 8-10) heeft. Om te voorkomen dat deze 'basis'similariteit een te grote rol gaat spelen zal het misschien nodig zijn te werken met een niet te lage *drempelwaarde*, bijvoorbeeld van 0,5 (50% van document X overlapt document Y). Alleen wanneer de similariteit deze drempel overschrijdt zou deze dan als significant moeten worden beschouwd.

- Formule 8-11 kent niet het nadeel dat de hit2's worden verwaarloosd, en maakt dus gebruik van alle overeenkomsten tussen de documenten (hit1's en hit2's). De range is hier eveneens [0..1]. Merk op dat de uitkomst van 8-11 niet overeenkomt met het gemiddelde van het hit1percentage en het hit2percentage, behalve wanneer Hit1Max en Hit2Max precies gelijk zijn aan elkaar, en dat is alleen het geval wanneer een document precies de helft van het gecorrigeerd aantal typen in het corpus bevat. Dit heeft tot gevolg dat de similariteit van een zeer klein of een zeer groot document waarschijnlijk in belangrijke mate gereduceerd zal worden, hetgeen verdedigbaar is gezien de ongelijke verhouding tussen aantallen hit1's en hit2's die waarschijnlijk in zo'n document zal bestaan. Het is te verwachten dat met deze similariteitsmaat vaker dan bij de voorgaande sprake zal zijn van een wederzijdse similariteitsverhouding, aangezien bijvoorbeeld verschil in afmeting tussen de documenten, vanwege de wisselwerking tussen hit1percentage en hit2percentage, een veel geringere rol speelt. Een nadeel van formule 8-11 zou misschien kunnen zijn dat bij kleine documenten (die maar weinig hit1's en zeer veel hit2's kunnen opleveren) het hit2percentage allesoverheersend kan worden (hit1's die er nog zijn tellen nauwelijks meer mee), terwijl bij grote documenten hetzelfde geldt voor het hit1percentage.

- In formule 8-12 wordt van alle beschikbare gegevens gebruik gemaakt, waarbij het totale miss-percentage in mindering wordt gebracht op het totale hit-percentage uit 8-11. Het is echter de vraag of dat in dit geval een juist resultaat oplevert, gegeven

[17] Woordtypen die in *elk* document voorkomen tellen bij de berekening van de similariteit niet mee (ze hebben een gewicht van 0).

dat gebruik wordt gemaakt van *gewichten*. Daarbij vindt namelijk al correctie plaats voor miss1's en miss2's, op de volgende manier. Voor een document X wordt berekend Hit1Max, de maximale hit1-score die het ooit zal kunnen behalen, en Hit2Max, de maximale hit2-score. De waarde van deze variabele wordt berekend door het optellen van de gewichten van alle typen die in het document voorkomen resp. die in het document ontbreken. Bij het bepalen van de similariteit met een ander document Y zal meestal slechts een deel van de typen in X een hit1 opleveren, de rest is miss1. Doordat zowel de typen die op dat moment een hit1 opleveren als die een miss1 opleveren zijn meegeteld in Hit1Max, zal het hit1-*percentage* lager worden naarmate het aantal miss1's hoger is. Hetzelfde geldt mutatis mutandis voor de miss2's, die zijn meegeteld in Hit2Max. Het aantal miss2's komt hier in mindering op het aantal hit2's, wat weer het hit2-*percentage* drukt. In formule 8-12 treedt deze 'automatische' correctie ook op, maar daarbovenop wordt dan het misspercentage nog eens in mindering gebracht op het hitpercentage. Dit leidt hoogstwaarschijnlijk tot een situatie waarin het belang van de misses dat van de hits overtreft, hetgeen ingaat tegen de essentie van de similariteitsmaat, die de mate van *overeenkomst* tussen documenten bedoelt aan te duiden. Een ander nadeel van 8-12 is dat de uitkomst ook negatief kan worden (in theorie is de range [-1..1]), wat enigszins onhandig kan zijn bij het maken van berekeningen. Al met al lijkt de toepassing van formule 8-12 in dit geval niet voor de hand te liggen.

Op basis van deze kenmerken kunnen voorlopig de volgende *conclusies* worden getrokken:

- De gepresenteerde formules vormen slechts een greep uit de verschillende mogelijkheden om de similariteit van documenten te berekenen. Een gemeenschappelijk kenmerk van alle drie is echter dat zij zodanig kunnen worden omgewerkt dat toepassing ook mogelijk is wanneer eigenschappen die bepalend zijn voor de similariteit een verschillend gewicht hebben. Dit geldt voor vrijwel geen van de andere formules die in de literatuur worden genoemd.
- Bij deze specifieke formules en in dit specifieke geval, waarin gewerkt wordt met *gewichten* van hits en misses en waarin deling plaatsvindt door een (van tevoren berekend) hoogst mogelijk totaalgewicht, leidt het apart in de berekening verwerken van aantallen *misses* tot een onjuist resultaat. De formule waarin dat gebeurt (8-12) zal dan ook verder niet meer worden gebruikt.
- Voor de formules die dan overblijven (8-10 en 8-11) geldt dat 8-11 er het meest geschikt uitziet, aangezien in deze formule ook de hit2's een rol spelen, die een belangrijk similariteitskenmerk vormen. Het is echter niet uit te sluiten dat ook 8-10 goede resultaten kan opleveren, vooral wanneer in een corpus documenten voorkomen die sterk in afmetingen verschillen. Aangezien dat bij het corpus wetgevingsteksten in zekere mate het geval is, zullen beide formules voor dat corpus worden beproefd, om zo te kunnen nagaan welke het beste resultaat oplevert.
- Het is mogelijk dat in sommige gevallen een *drempelwaarde* gehanteerd zal moeten worden. De mate van similariteit wordt in dat geval alleen als significant beschouwd wanneer deze boven de drempelwaarde ligt. Het hanteren van een zo'n drempel lijkt op theoretische gronden vooral nodig wanneer gebruik wordt gemaakt van formule 8-10, maar heeft wellicht ook voordelen bij toepassing van 8-11.

8.3.2 Het gebruik van hit-informatie in plaats van hit-aantallen

In het voorgaande gedeelte is bij de berekening van de similariteit steeds uitgegaan van de voor het type-gewicht gecorrigeerd hit- en miss*aantallen*. Zoals in paragraaf 8.2.6 werd beschreven kan echter, met behulp van de aldaar gepresenteerde formule (8-6), ook nauwkeurig worden berekend hoeveel bits *informatie* iedere hit of miss precies oplevert. Het belangrijkste verschil tussen het gebruik van deze informatie en het gebruik van gecorrigeerde hitaantallen is dat informatie een logaritmische maat is. Het gewicht van een hit of miss (waarop de gecorrigeerde hitaantallen zijn gebaseerd) is het complement van de *kans* op het betreffende type, de informatie die de hit of miss oplevert is gelijk aan de *2-logaritme van de reciproque van de kans* op dat type. Het verschil tussen de twee is daardoor het grootst wanneer de kans op een type *laag* is (hoe lager de kans, hoe groter het verschil). Voor hit1's is dat het geval bij typen die maar in weinig documenten voorkomen, en voor hit2's bij typen die in bijna alle documenten voorkomen.

De mate van similariteit kan ook worden berekend op basis van deze hitinformatie, in plaats van op basis van (voor gewicht gecorrigeerde) hitaantallen. Het effect van deze alternatieve wijze van berekenen is dat de typen met de kleinste kansen een veel sterkere invloed krijgen op de similariteits-score. Dat betekent dat vooral document-paren die enkele zeldzame typen gemeenschappelijk hebben, en/of die allebei enkele veelvoorkomende typen missen, een hogere mate van similariteit zullen hebben. De similariteit van documentenparen die alleen gemiddeld vaak voorkomende typen gemeenschappelijk hebben en/of missen zal daarentegen lager zijn dan wanneer voor de berekening gecorrigeerde hitaantallen zouden zijn gebruikt. Wanneer de similariteits-score dus wordt berekend op basis van hit-informatie heeft dit tot gevolg dat veel sterker de nadruk komt te liggen op eigenschappen die voor een document *het meest specifiek* zijn, die het document het meest onderscheiden van de rest. Het is niet onaannemelijk dat dit juist bij omvangrijke corpora zal leiden tot een betrouwbaarder similariteits-score. Vooral wanneer daarin grote aantallen documenten en woordtypen voorkomen wordt de kans op toevallige overeenkomsten tussen die documenten namelijk ook groter. Door het gebruik van hitinformatie komt de nadruk echter meer te liggen op de meest specifieke kenmerken en wordt de invloed op de similariteits-score van woordtypen die ook in veel andere documenten voorkomen teruggedrongen. De keerzijde hiervan is dat twee documenten die wel veel typen met een gemiddelde frequentie gemeenschappelijk hebben of juist missen, maar geen of heel weinig typen met een hoge informatiewaarde, een beduidend lagere similariteits-score zullen hebben dan wanneer gebruik gemaakt zou worden van gecorrigeerde aantallen.

Wanneer gebruik wordt gemaakt van hitinformatie in plaats van van gecorrigeerde aantallen, kan de similariteit als volgt worden berekend:

$$S_1 = \frac{Hit1Info}{Hit1MaxInfo} \qquad (8\text{-}13)$$

$$S_2 = \frac{Hit1Info + Hit2Info}{Hit1MaxInfo + Hit2MaxInfo} \qquad (8\text{-}14)$$

8.3.3 De similariteits-score bij tekstcorpora

De similariteits-score van documenten die deel uitmaken van een bepaald tekstcorpus kan worden berekend aan de hand van de formules uit de voorgaande twee subparagrafen. De formule waarin aantallen *misses* apart van de aantallen hits worden afgetrokken (8-12) zal daarbij niet worden gebruikt. Dit heeft als consequentie dat de aantallen misses hier *niet* behoeven te worden bijgehouden. Met betrekking tot de volgende punten werd nog geen keuze gemaakt:

- berekening van de similariteits-score met behulp van alleen hit1's, of van zowel hit1's als hit2's
- gebruik van voor gewicht gecorrigeerde hitaantallen, of van hitinformatie.

De verschillende mogelijkheden die hierdoor ontstaan kunnen worden ondergebracht in een dubbele dichotomie, waaruit tevens blijkt volgens welke formule de similariteits-score in dat geval moet worden berekend:

	Alleen hit1's	Hit1's *en* hit2's
Hitaantallen (voor gewicht gecorrigeerd)	Berekening vlg. 8-10	Berekening vlg. 8-11
Hitinformatie	Berekening vlg. 8-13	Berekening vlg. 8-14

Tabel 8-2 Vier methoden voor de berekening van de similariteits-score

Eén van de voornaamste doelstellingen van het onderzoek dat in dit hoofdstuk wordt beschreven is om deze vier methoden te beproeven voor het corpus wetgevingsteksten, en zo mogelijk conclusies te trekken ten aanzien van hun respectievelijke eigenschappen, en van hun geschiktheid voor de bepaling van de similariteit van documenten. Om dat mogelijk te maken zal met behulp van deze vier methoden afzonderlijk de mate van similariteit worden berekend tussen elk van de 18803 documenten uit het corpus en de 18802 andere documenten. De werkwijze kan daarbij globaal als volgt zijn:

1. Allereerst worden de documenten ingelezen, en wordt per document een lijst aangelegd van de woordtypen die erin voorkomen.
2. Met behulp van de typelijsten per document wordt een typelijst voor het hele corpus aangelegd. Daarin wordt tegelijkertijd vastgelegd in hoeveel documenten een bepaald type voorkomt (de frequentie van dat type, zoals dat in dit hoofdstuk genoemd wordt).
3. Nu kan worden berekend wat het gewicht c.q. de hoeveelheid informatie van een bepaald type zou zijn, zowel in het geval dat het type betrokken zou zijn bij een hit1 als bij een hit2. Deze cijfers worden opgeslagen in de lijst bij het desbetreffende type. Er wordt tevens voor gezorgd dat gewicht en informatie van een type dat in alle documenten voorkomt 0 is.
4. Met behulp van de gewichten c.q. de hoeveelheid informatie van de typen kan nu voor ieder document worden berekend het maximale (gecorrigeerde) aantal hit1's en hit2's c.q. de maximale hoeveelheid hit1-informatie en hit2-informatie die het zou kunnen bevatten (Hit1Max, Hit2Max, Hit1MaxInfo en Hit2MaxInfo). Deze cijfers worden opgeslagen in de lijst bij het desbetreffende document.

5. Nu volgt het meest tijdrovende deel van het proces: de typelijst van ieder document wordt vergeleken met die van elk ander document in het corpus. Het aantal vergelijkingen ligt daarbij in principe in de orde van N^2, het kwadraat van het aantal documenten (eigenlijk: $N * (N - 1)$, aangezien een document niet met zichzelf behoeft te worden vergeleken). Gelukkig is het zo dat documentvergelijkingen slechts 'eenzijdig' behoeven plaats te vinden: gecorrigeerde hitaantallen en hitinformatie zijn voor twee documenten die worden vergeleken steeds gelijk, alleen moeten deze waarden voor de berekening van de similariteits-score gedeeld worden door de *maximale* waarde(n) die horen bij het betreffende document. Het vergelijken van twee documenten X en Y levert dus zowel de similariteits-score van X ten opzichte van Y op, als de similariteits-score van Y ten opzichte van X. Dit vermindert het aantal document-vergelijkingen met de helft: $N * (N - 1) / 2$. Voor het corpus wetgevingsteksten houdt dat in dat voor het vaststellen van de similariteit tussen alle documenten 18803*18802/2 = 176.767.003 documentvergelijkingen moeten plaatsvinden, waarbij steeds alle woordtypen uit de twee documenten (vaak enkele honderden stuks) zijn betrokken. Dit aantal stelt hoge eisen aan de efficiëntie van het vergelijkingsalgoritme en aan de snelheid van de computer waarop het betreffende programma wordt uitgevoerd. Tijdens het vergelijkingsproces wordt voor ieder document bijgehouden met welk document het de hoogste similariteit vertoont, met daarbij de similariteits-score zelf en de gegevens op basis waarvan die werd berekend.
6. Nadat het vergelijkingsproces is afgerond, en alle documenten met elkaar zijn vergeleken, wordt een rapport gemaakt van de resultaten. Dit rapport heeft de vorm van een lijst, waarin ieder document staat vermeld met daarbij het document waarmee het de hoogste similariteit heeft, de similariteits-score zelf en de gegevens (aantallen hits e.d.) waarop die is gebaseerd. Deze lijst is gesorteerd op similariteits-score, de hoogste score bovernaan.

Aan het Centrum voor Informatica en Recht is een computerprogramma ontwikkeld dat deze similariteitsberekening uitvoert[18]. Het programma werd geschreven in de programmeertaal Pascal, met enkele tijds-intensieve gedeelten in machinetaal. Het grootste deel van de ontwikkeltijd van dit programma is besteed aan het optimaliseren van het vergelijkings-algoritme (het programmaonderdeel dat de typelijsten van twee documenten naast elkaar legt, waarna op basis daarvan de aantallen hits en de bijbehorende informatiehoeveelheid worden berekend). De eerste versies waren op dit punt nog veel te traag, de similariteitsberekening van het corpus wetgevingsteksten (waarvoor zoals gezegd ruim 176 miljoen document-vergelijkingen nodig zijn) zou met die versies ongeveer 700 uur rekentijd hebben gekost. Uiteindelijk werd een versie ontwikkeld die op een 90 Mhz Pentium PC ongeveer 2000 document-vergelijking per seconde kan uitvoeren waardoor de rekentijd nog 'slechts' zo'n 28 uur bedraagt.

Het programma stelt na een vergelijkings'ronde' een rapport op zoals hierboven bij punt 6 beschreven. Een gedeelte van zo'n rapport, dat vervaardigd werd na een proef-run van het programma met alleen de *eerste 300* documenten uit het corpus wetgevingsteksten, is

[18] Programmeurs: J.P. van der Landen en ondergetekende.

hieronder afgedrukt (tabel 8-3). Bovenaan dit rapport staan enkele gegevens met betrekking tot de instellingen van het programma en de onderzochte data. Uit die gegevens blijkt bijvoorbeeld dat de similariteit hier werd berekend op basis van hit-gewichten (gecorrigeerde hit-aantallen), en rekening houdend met zowel hit1's als hit2's (dat betekent dat gebruik is gemaakt van formule 8-11). Het aantal typen in de 300 documenten bedraagt 10837, waarvan er 2 eigenlijk niet meedoen omdat ze in alle documenten voorkomen. De gegevens 'gecorrigeerd aantal typen' en 'gecorrigeerd aantal afwezige typen', met bijbehorende hoeveelheid informatie, hebben betrekking op de som van hit1- resp. de hit2gewichten van alle typen in het corpus.

Score berekening gebaseerd op hit-GEWICHT (met Hit1's en Hit2's)
Ronde: 1
Totaal aantal typen: 10837; zonder "typen-in-allemaal": 10835
Gecorrigeerd aantal typen: 10595.5; Max. informatie: 75888.1
Max. afwezige typen: 10837; zonder "typen-in-allemaal": 10835
Gecorrigeerd aantal afw. typen: 239.8; Max. informatie: 429.2

Doc X	Doc Y	Score	TpX	Tp-X	TpY	Tp-Y	TpXc	Tp-Xc	TpYc	Tp-Yc	H1c	H2c	H1inf	H2inf
D00194	D00193	9648	44	10793	46	10791	23,1	220,9	25,5	221,3	18,6	216,8	64,9	359,1
D00032	D00046	9643	143	10694	172	10665	97,3	196,1	122,8	192,6	92,4	190,5	355,0	301,8
D00193	D00194	9538	46	10791	44	10793	25,5	221,3	23,1	220,9	18,6	216,8	64,9	359,1
D00026	D00039	9505	52	10785	66	10771	28,6	218,4	36,4	212,2	23,9	210,9	91,1	341,0
D00195	D00196	9500	133	10704	137	10700	87,4	196,3	90,2	195,1	77,0	192,5	314,8	304,2
D00197	D00198	9470	146	10691	141	10696	97,4	193,2	96,0	196,8	84,1	191,1	357,6	299,6
D00039	D00026	9446	66	10771	52	10785	36,4	212,2	28,6	218,4	23,9	210,9	91,1	341,0
D00196	D00195	9446	137	10700	133	10704	90,2	195,1	87,4	196,3	77,0	192,5	314,8	304,2
D00031	D00045	9434	128	10709	140	10697	87,8	201,6	98,4	200,2	77,0	196,1	307,8	310,8
D00198	D00197	9399	141	10696	146	10691	96,0	196,8	97,4	193,2	84,1	191,1	357,6	299,6
D00020	D00163	9339	28	10809	27	10810	16,7	230,5	17,3	232,2	3,5	227,3	9,4	394,1
D00138	D00139	9254	83	10754	88	10749	53,2	212,0	56,4	210,2	36,3	209,1	114,1	343,5
D00163	D00020	9252	27	10810	28	10809	17,3	232,2	16,7	230,5	3,5	227,3	9,4	394,1
D00192	D00191	9208	99	10738	155	10682	58,8	201,7	105,2	192,0	53,2	186,6	192,6	290,6
D00139	D00138	9204	88	10749	83	10754	56,4	210,2	53,2	212,0	36,3	209,1	114,1	343,5
D00149	D00148	9190	108	10729	120	10717	73,4	207,2	82,6	204,4	58,4	199,4	200,0	322,2
D00041	D00036	9150	57	10780	47	10790	29,5	214,3	25,8	220,6	12,6	210,4	49,8	342,0
D00045	D00031	9142	140	10697	128	10709	98,4	200,2	87,8	201,6	77,0	196,1	307,8	310,8
D00134	D00122	9111	151	10686	183	10654	104,9	195,7	137,2	196,0	85,3	188,7	396,2	294,2
D00036	D00023	9102	47	10790	65	10772	25,8	220,6	41,6	218,4	12,4	211,8	41,5	346,2

Tabel 8-3 Voorbeeld van een deel van een rapport van het similariteitsprogramma

In de kolommen treffen we achtereenvolgens aan: Naam basisdocument, naam document waarmee het de hoogste similariteit heeft, similariteits-score (voor goede leesbaarheid vermenigvuldigd met 10000), aantal typen in document X, aantal typen dat ontbreekt in document X, aantal typen in Y en aantal typen dat ontbreekt in Y. Daarna volgen vier kolommen met dezelfde opschriften, maar met een 'c' erachter. Daarin treffen we aan de som van de gewichten van alle typen die aanwezig resp. afwezig zijn in X en Y; het gaat daarbij dus in feite om Hit1Max en Hit2Max van document X en Y. In de laatste vier kolommen vinden we tenslotte de aantallen hit1's en hit2's (gecorrigeerd voor gewicht) en de bijbehorende informatie-cijfers. Deze gegevens gelden voor document X en Y gezamenlijk.

In principe komt ieder document uit het corpus 1 keer voor in de eerste kolom (Doc X). De overige kolommen vermelden dan met welk ander document het de hoogste similariteit vertoont. Uit de tabel kan bijvoorbeeld worden opgemaakt dat de documenten 193 en 194 *wederzijds* een zeer hoge similariteit vertonen. Document 194 heeft de hoogste similariteit met document 193 (Score: 9648), terwijl 193 ook de hoogste similariteit heeft met 194 (Score: 9538). Een dergelijk geval van wederzijdse similariteit treedt lang niet bij alle documenten op. Zo heeft document 41 de hoogste similariteit met document 36, maar 36 scoort het hoogst met 23. Zoals in paragraaf 8.3.1 al werd aangegeven is de kans op

wederkerige similariteitsverhoudingen in theorie het grootst bij het type similariteitsberekening zoals dat hier is toegepast, en waarbij zowel hit1's als hit2's een rol spelen. Er

lijkt dan ook veel voor te zeggen om wanneer voor de similariteits-score een berekening wordt toegepast waarin ook hit2's een rol spelen (formule 8-11 of 8-14 dus), de nadruk te leggen op de wederkerigheid van de similariteitsrelatie, en alleen als daarvan sprake is ook gevolgen aan de similariteit te verbinden. Wanneer bij de berekening alleen hit1's zijn betrokken ligt wederkerigheid juist niet voor de hand, aangezien daarvan meestal slechts sprake zal zijn wanneer documenten ongeveer even veel woordtypen bevatten.

Nadat voor een corpus de similariteitsscores zijn berekend, en is vastgesteld tussen welke documenten een similariteitsrelatie bestaat (eenzijdig of wederkerig, afhankelijk van de toegepaste berekening) kan aan de hand daarvan onder andere worden bepaald welke documenten in het corpus zoveel overeenkomsten vertonen (op het gebied van het woordgebruik) dat ze kunnen worden aangemerkt als één enkel 'cluster'. Hierop zal in de volgende paragraaf nader worden ingegaan.

8.4 Het vormen van document-clusters met behulp van de similariteits-score

8.4.1 Inleiding

De similariteits-score, beschreven in het voorafgaande gedeelte van dit hoofdstuk, kan op zichzelf al een nuttig hulpmiddel zijn bij het onderzoeken van corpora. Met behulp van de score kan worden nagegaan welke documenten voor wat betreft het woordgebruik het meeste op elkaar lijken. Dit geeft inzicht in de structuur van een corpus. Verder wordt bijvoorbeeld duidelijk of bepaalde documenten meer dan een keer voorkomen - kopieën van een document hebben een similariteits-score van 100% - wat bij een omvangrijk corpus wel eens onbedoeld het geval is. Nog meer mogelijkheden ontstaan echter wanneer we na het berekenen van de similariteits-scores nog een stap verder gaan, op de volgende wijze. We kunnen ertoe overgaan om twee documenten waartussen een hoge mate van similariteit bestaat *samen te voegen tot een document-cluster*. Door zo'n samenvoeging ontstaat dus één nieuw document in het corpus, en verdwijnen er twee. Vervolgens zouden we opnieuw de similariteits-scores voor alle documenten (inclusief het zojuist gevormde) kunnen uitrekenen, en nagaan welke documenten *nu* het meest op elkaar lijken. Op basis van die gegevens zouden opnieuw documenten kunnen worden samengevoegd, etc.

Het resultaat van deze werkwijze zou uiteindelijk kunnen zijn dat in een corpus documentclusters ontstaat van een aanzienlijke omvang. Deze clusters, gebaseerd op relaties die gelegd zijn tussen de oorspronkelijke documenten, zouden nog aanzienlijk meer informatie kunnen verschaffen over de structuur van een corpus dan de similariteits-scores alleen. Wel is het zo dat de *manier waarop de similariteits-score wordt berekend* op dit proces een aanzienlijke invloed zou kunnen uitoefenen. Clustering vindt immers uitsluitend op basis daarvan plaats. Daarnaast kunnen verschillende cluster-algoritmen worden gebruikt, en is het mogelijk (en vaak ook verstandig) om aan de similariteit bepaalde

specifieke eisen te stellen (zoals een zekere minimumwaarde). Op deze punten zal in de volgende subparagraaf worden ingegaan.

8.4.2 Methode

Het eerste punt dat in overweging moet worden genomen bij het vaststellen van de meest geschikte methode waarop documenten kunnen worden samengevoegd in clusters betreft de *berekening van de similariteits-score*. Het is immers de bedoeling om op die score het gehele clusterings-proces te baseren. In de voorgaande paragraaf kwamen voor berekening uiteindelijk vier formules als in principe geschikt naar voren (8-10, 8-11, 8-13 en 8-14). De vraag is echter of deze ook alle vier even bruikbaar zijn als basis voor het vormen van document-clusters. De punten van verschil tussen de formules zijn:

- gebruik van (voor gewicht gecorrigeerde) hit*aantallen* of van hit*informatie*
- rekening houden met *hit2's* (dat wil zeggen, met woordtypen die in beide documenten *af*wezig zijn) of niet.

Met betrekking tot het eerste punt is het moeilijk om alle consequenties van een keuze voor een van de alternatieven precies te overzien. Gebruik van informatie zorgt voor meer nadruk op 'zeldzame eigenschappen': documenten die enkele woordtypen met lage frequentie gemeenschappelijk hebben, of enkele woordtypen met hoge frequentie missen hebben daardoor samen al snel een hogere mate van similariteit dan documenten die vooral typen met gemiddelde frequentie gemeenschappelijk hebben of missen. Dit lijkt op het eerste gezicht een gewenst patroon, maar er zijn situaties denkbaar waarin het juist nadelig zou kunnen uitpakken. Te denken valt dan aan documenten waartussen een bepaalde relatie bestaat, maar die vrijwel uitsluitend woordtypen van gemiddelde en hoge frequentie bevatten, en niet of nauwelijks typen met hoge frequentie missen. Dergelijke documenten zouden wellicht bij gebruik van gecorrigeerde hitaantallen wel, maar bij gebruik van hitinformatie niet in een cluster worden samengevoegd (vanwege een te lage similariteit). Toch zal, gezien het grote aantal documenten waar het in dit geval om gaat, gebruik van hit-informatie misschien onvermijdelijk blijken om de invloed van toevallige overeenkomsten - die in zo'n omvangrijke dataverzameling vaker zullen optreden dan in een klein corpus - zo klein mogelijk te houden. De conclusie moet dan ook luiden dat voor beide benaderingen iets te zeggen is, en dat ze het beste ook allebei kunnen worden uitgeprobeerd.

Voor het tweede punt geldt tot op zekere hoogte hetzelfde, de exacte gevolgen van de keuze om al dan niet gebruik te maken van hit2's bij het bepalen van de similariteit en dus bij het vormen van documentclusters zijn eigenlijk niet te voorspellen. Wel kan in theorie bij het gebruik van alleen hit1's een ongewenst effect ontstaan, dat gevolgen heeft voor het cluster-proces. Dit effect treedt vooral op doordat similariteit op basis van hit1's vaak *enkelzijdig* zal zijn (behalve wanneer de documenten ongeveer even groot zijn, zie paragraaf 8.3.1), waarbij het kleinste van twee documenten de hoogste similariteit zal hebben. Het resultaat daarvan is dat een klein document door een groot wordt aangetrokken, *juist als gevolg van de asymmetrie in de similariteitsrelatie*. Bovendien is het denkbaar, wanneer op basis van deze similariteit wordt besloten om documenten samen te voegen, dat daarbij clusters zullen ontstaan die de neiging hebben steeds meer documenten

aan te trekken, net zolang tot er geen documenten met voldoende similariteit meer over zijn of het cluster-proces om een andere reden stopt. Het eindresultaat zou dan wel eens kunnen bestaan uit slechts een of twee zeer omvangrijke clusters, waaruit misschien weinig is af te leiden over de structuur van het corpus. Om na te gaan of dat probleem inderdaad reëel is moet de methode echter toch eerst in de praktijk worden uitgeprobeerd.

Aangezien dus niet één van de berekeningsmethoden al bij voorbaat terzijde kan worden gelegd is het nodig om, teneinde een volledig beeld van de voor- en nadelen van elke berekeningsmethode te krijgen, het gehele proces van similariteitsberekening en clustering voor het corpus wetgevingsteksten in totaal vier keer te doorlopen. De resultaten daarvan kunnen dan met elkaar worden vergeleken. Daarnaast moeten nog enkele andere voorwaarden worden besproken die op het clusteringsproces van toepassing zijn. Het betreft hier met name regels die gelden bij de vorming van de clusters.

- In de eerste plaats worden binnen een ronde nooit meer dan twee documenten tot één enkel cluster samengevoegd. Wanneer een document X een hoge mate van similariteit vertoont met een document Y dat in dezelfde ronde al met een ander document Z is samengevoegd dan wordt X niet nog eens toegevoegd aan het cluster van Y en Z. De ratio achter deze regel is dat door het samenvoegen van documenten een nieuwe situatie ontstaat, waarbij nog geen zekerheid bestaat over similariteiten tussen het nieuw gevormde document en de rest van het corpus. Daarom wordt het nieuwe cluster met rust gelaten tot de volgende ronde begint, en de similariteits-scores opnieuw zijn uitgerekend. N.B.: in één ronde kunnen natuurlijk wel verschillende clusters ontstaan, maar het gaat daarbij steeds slechts om de samenvoeging van twee documenten.
- Bij het zoeken naar mogelijke clusters wordt altijd gewerkt op volgorde van similariteit, het documenten met de hoogste similariteits-score wordt eerst bekeken, dan het document met de op één na hoogste score, etc.
- Wanneer de similariteit op basis van alleen hit1's wordt berekend (volgens formule 8-10 of 8-13) wordt uitgegaan van een asymmetrisch similariteitscriterium, dat wil zeggen dat wanneer een document X de hoogste similariteit vertoont met document Y het er niet toe doet of Y ook de hoogste similariteit vertoont met X. De volgorde waarin documenten worden bekeken (hoogste similariteitsscore eerst) is hier dus van essentieel belang. De reden waarom het het meest voor de hand ligt om te werken met een enkelzijdig criterium is eerder in dit hoofdstuk al enkele malen besproken (zie onder andere paragraaf 8.3.1). De hoofdgedachte is dat de similariteit op basis van alleen hit1's sterk afhankelijk is van de afmetingen van documenten, of om precies te zijn van Hit1Max in elk document. Dat betekent dat voor een documentenpaar slechts vergelijkbaar hoge similariteitsscores zijn te verwachten wanneer die documenten een ongeveer even hoge Hit1Max hebben, wat met name het geval zal zijn wanneer ze ongeveer even groot zijn.
- Wanneer de similariteit op basis van hit1's *en* hit2's wordt berekend (volgens formule 8-11 of 8-14) geldt een tweezijdig similariteitscriterium: vorming van een cluster vindt alleen plaats wanneer documenten wederzijds de hoogste similariteit met elkaar vertonen.
- Zowel wanneer alleen hit1's worden gebruikt als wanneer de similariteit gebaseerd is op hit1's *en* hit2's wordt gewerkt met een significantie-drempel. Alleen wanneer de similariteit deze drempel overschrijdt kan een document (mits ook aan de overige voorwaarden is voldaan) worden samengevoegd met een ander. Op basis van enig

vooronderzoek met kleine hoeveelheden documenten is besloten deze drempels als volgt vast te stellen:

- Bij gebruik van alleen hit1's: 50%, of 5000 wanneer de similariteits-score met 10000 is vermenigvuldigd. Dit relatief hoge percentage vormt een zekere compensatie voor het feit dat het hier om een eenzijdig similariteitscriterium gaat. Alleen een document dat door een ander document met tenminste 50% wordt overlapt[19] kan eventueel aan dat document worden toegevoegd, en aldus een cluster vormen.
- Bij gebruik van zowel hit1's als hit2's: 20% of 2000. Dit veel lagere percentage is eigenlijk alleen toegepast om te voorkomen dat documenten worden samengevoegd op basis van een zeer geringe, of zelfs een nul-similariteit. Het laatste blijkt vooral tegen het einde van een cluster-proces nog wel eens op te treden, wanneer bijvoorbeeld uiteindelijk twee document-clusters zijn ontstaan zonder hit1's, die samen alle typen van het corpus bevatten. N.B.: het is zeer waarschijnlijk dat in die situatie een groot aantal typen in beide documenten voorkomen, maar deze tellen niet mee omdat ze daarmee tevens in *alle* documenten voorkomen[20]. Overigens is het clusteren van twee documenten uit een totaal van twee ook niet zinvol; het resultaat daarvan zou immers weer het volledige corpus zijn.

Vooral de drempel van 50% kan het clusterings-proces vrij sterk beïnvloeden. Om het aantal variabelen niet te groot te maken is er hier voor gekozen deze drempel (en trouwens ook die van 20%) als een vast gegeven te beschouwen. Het is echter niet uitgesloten dat het proces verder geoptimaliseerd kan worden door het nauwkeurig afstemmen van deze drempels (en met name van die van 50%).

- Het aantal 'ronden' waarin getracht wordt documenten samen te voegen is om praktische redenen beperkt tot 300. Dit 'beperkt' de afmetingen van de te vormen clusters tot een theoretisch maximum van 2^{300} documenten (mochten er in een corpus zoveel aanwezig zijn). Het hoogste aantal ronden dat wordt doorlopen blijkt in de praktijk minder te zijn, namelijk 253, zodat het maximale aantal hier feitelijk geen betekenis heeft.

8.5 Document-clusters in het corpus wetgevingsteksten

8.5.1 Inleiding

In deze paragraaf worden de resultaten beschreven van de pogingen om documenten uit het corpus wetgevingsteksten samen te voegen tot clusters. Het betreffende proces werd in totaal vier keer doorlopen, waarbij steeds gebruik werd gemaakt van verschillende berekeningsmethoden voor de similariteits-score, zoals in de voorgaande paragraaf beschreven. In totaal vergde dit ongeveer een maand rekentijd, merendeels op 90 Mhz Pentium PC's. Elke berekeningsmethode bleek haar specifieke kenmerken te hebben, waarbij een aantal verwachtingen uit de voorgaande paragraaf juist bleken te zijn. Het totale proces leverde een grote hoeveelheid gegevens op, waarvan de belangrijkste hier

---

[19] Het gaat hier natuurlijk om een *similariteits*-score van 50%, waarin ook de frequenties van woordtypen (in de vorm van gewichten of van de bijbehorende hoeveelheid informatie) een rol spelen.

[20] Typen die in ieder document voorkomen hebben immers een gewicht en een informatiewaarde van 0.

worden beschreven en geanalyseerd. Bij het uitvoeren van deze analyses was het nodig enkele keuzes te maken, waarover in de volgende subparagrafen meer.

8.5.2 Gegevens met betrekking tot de gevormde clusters

Allereerst is het van belang om enig inzicht te krijgen in het verloop van het cluster-proces en in het aantal en de omvang van de gevormde clusters. Het blijkt daarbij nogal wat verschil te maken welke van de vier verschillende methoden gebruikt wordt om de similariteitsscore te berekenen en vast te stellen welke documenten worden samengevoegd. Om daarin enig inzicht te geven is in tabel 8-4 een aantal statistische gegevens met betrekking tot de resultaten van elke clustermethode opgenomen (zie de volgende bladzijde). Om te beginnen treffen we daarin een aantal gegevens over de clustervorming in het algemeen aan: gedurende hoeveel ronden zijn er clusters gevormd, hoeveel clusters waren dat in totaal, hoeveel van de oorspronkelijke documenten zijn in een van die clusters vertegenwoordigd en hoeveel documenten (clusters en overgebleven 'solitaire' documenten) telt het corpus nadat het clusterproces is afgelopen? Bij deze gegevens wordt al direct het grote verschil zichtbaar tussen de eenzijdige, asymmetrische clustermethoden en de tweezijdige, symmetrische. Het clusterproces gaat bij de eerste veel langer door, het aantal gevormde clusters is dan ook veel hoger (drie tot vier keer zo hoog als bij de symmetrische methoden). Het aantal documenten na clustervorming is bij de asymmetrische methoden laag (101 c.q. 45), de oorspronkelijke documenten zijn dus effectief in een hanteerbaar aantal groepen verdeeld. Dat het *aantal* hanteerbaar is betekent daarbij echter nog niet dat ook alle clusters even zinvolle combinaties van documenten bevatten; op dat punt zal hierna nog worden teruggekomen.

Hoewel het aantal ronden (en daarmee ook de benodigde rekentijd) bij de twee asymmetrische methoden veel hoger is, worden per ronde gemiddeld wat minder clusters gevormd. Bij raadpleging van de cluster-tabellen die door het programma zijn opgesteld blijkt dat in beide gevallen gedurende de laatste ronden van het cluster-proces nog maar mutaties zijn opgetreden in een enkel cluster, waaraan in iedere ronde een document (of een eerder gevormd cluster van documenten) is toegevoegd. Dit lage aantal van 1 cluster per ronde drukt natuurlijk het gemiddelde. Het hoogste aantal clusters werd bij alle vier methoden gevormd in de eerste ronde. Dit aantal varieert ook bij de vier methoden, maar niet zoveel als misschien verwacht kon worden op basis van de grote verschillen in aantallen clusters die elke methode uiteindelijk oplevert. Opmerkelijk is verder nog dat bij de asymmetrische methoden de similariteitsberekening op basis van hit-*informatie* een hoger maximum aantal clusters (in de eerste ronde) oplevert dan de berekening op basis van hit-*gewicht*, terwijl dit bij de symmetrische methoden precies andersom is.

De grootste verschillen tussen de vier methoden worden echter zichtbaar wanneer we kijken naar de afmetingen van de clusters die uiteindelijk ontstaan. In de tabel zijn deze afmetingen op drie manieren uitgedrukt: in aantallen oorspronkelijke documenten (documenten van vóór het cluster-proces), in aantallen woordtokens en in aantallen bytes (waarbij een byte gelijk staat aan een karakter). Verder worden bij ieder gegeven steeds twee getallen vermeld. Het eerste c.q. het bovenste heeft uitsluitend betrekking op *clusters* (van een of meer oorspronkelijke documenten), het tweede c.q. het onderste

	Similariteit op basis van:			
	Informatie, hit1's	Gewichten, hit1's	Informatie, hit1's en hit2's	Gewichten, hit1's en hit2's
Clustervorming algemeen:				
Aantal ronden	253	126	11	17
Aantal gevormde clusters	18702	18758	4150	5674
Aantal betrokken documenten	18729	18772	6796	8831
Aantal doc. na clustervorming	101	45	14653	13129
Per ronde:				
Hoogste aantal gevormde clusters	4322	4105	2807	3325
Laagste aantal gevormde clusters	1	1	2	1
Gemiddeld aantal	73,9	148,9	377,3	333,8
Standaarddeviatie	323,7	447,9	802,4	799,4
Afmetingen eindclusters **excl / incl** enkelvoudige documenten:				
Aantal documenten per cluster:				
Grootste cluster	10897 / 10897	18745 / 18745	23 / 23	56 / 56
Kleinste cluster	2 / 1	2 / 1	2 / 1	2 / 1
Gemiddelde omvang	693,7 / 186,2	1340,9 / 417,8	2,6 / 1,3	2,8 / 1,4
Standaarddeviatie	2264,2 / 1210,1	4827,0 / 2762,9	1,4 / 0,8	2,0 / 1,3
Aantal tokens per cluster:				
Grootste cluster	7068255 / 7068255	12776726 / 12776726	52839 / 52839	115632 / 115632
Kleinste cluster	81 / 7	81 / 10	11 / 6	11 / 6
Gemiddelde omvang	472132,5 / 126808,5	913316,9 / 284614,6	1750,7 / 874,1	1659,6 / 975,5
Standaarddeviatie	1493839,4 / 800040,9	3290317,8 / 1883257,0	2920,8 / 1523,4	3183,9 / 1794,2
Aantal bytes per cluster:				
Grootste cluster	48313572 / 48313572	87570070 / 87570070	354709 / 354709	768263 / 768263
Kleinste cluster	1053 / 61	1053 / 360	78 / 46	78 / 46

Tabel 8-4 Statistische gegevens clustervorming

heeft betrekking op *alle documenten* waaruit het corpus in de eindsituatie bestaat, inclusief enkelvoudige (oorspronkelijke) documenten die niet met andere documenten zijn samengevoegd (en waar dus tijdens het cluster-proces niets mee is gebeurd). Nu blijkt dat vooral de tweede methode (similariteit op basis van hit1-gewichten) te ver is doorgeschoten: deze methode heeft een cluster opgeleverd bestaande uit 18745 van de oorspronkelijke 18803 documenten, en met bijna alle 12,8 miljoen woordtokens uit het corpus (op 30933 na, om precies te zijn). Het praktisch nut van dit omvangrijke cluster is zeer twijfelachtig, hoewel bestudering ervan mogelijk informatie kan opleveren over de reden waarom zoveel documenten in één cluster terecht zijn gekomen. Verder is niet uit te sluiten dat de methode nog andere (kleine) clusters heeft opgeleverd die interessant zijn. Ook bij de eerste methode (similariteit op basis van hit1-informatie) is tenminste 1 zeer omvangrijk cluster ontstaan (10897 documenten, ruim 7 miljoen tokens), het aantal overige clusters en enkelvoudige documenten is daar echter wat groter (100 stuks). De grootste clusters die gevormd zijn bij de twee symmetrische methoden zijn van een heel andere grootte-orde (23 respectievelijk 56 documenten, 52839 respectievelijk 115632 tokens).

Een conclusie die aan de hand van deze gegevens kan worden getrokken is de volgende: hoewel nog niet vaststaat of met behulp van de ene methode betere of zinvoller clusters kunnen worden gevormd dan met de andere, functioneert geen van de methoden optimaal met betrekking tot de *aantallen* clusters die uiteindelijk overblijven, en met betrekking tot de *afmetingen* van die clusters. Bij de twee asymmetrische methoden wordt een veel groter aantal ronden doorlopen (waarbij in de laatste ronden overigens nog maar mutaties in een enkel cluster optreden), maar deze methoden leveren elk tenminste 1 onhanteerbaar groot cluster op. De twee symmetrische methoden hebben dat nadeel niet, maar daarbij stopt het clusterproces juist weer relatief snel, waardoor de gemiddelde omvang van de gevormde clusters erg laag is (2 tot 3 documenten) en het aantal overblijvende documenten hoog.

8.5.3 Beoordeling van de gevormde clusters

Na wat hierboven gezegd is over het cluster*proces*, zal hier aandacht besteed worden aan de gevormde clusters zelf. Daartoe zal in eerste instantie vooral het begin van het clusterproces onder de loep worden genomen, waarbij zal worden getracht na te gaan welke verschillen er zijn tussen de bij elke methode gevormde clusters. Aangevangen zal daarbij worden met een opmerkelijk verschijnsel: het blijkt zo te zijn dat sommige documenten een similariteit van 100% vertonen, en dus in principe identiek zijn. Verder zal worden gekeken naar clusters die wat later (om precies te zijn in ronde 10) worden gevormd, en naar de clusters en documenten die overblijven nadat het clusterproces is gestopt. Geprobeerd zal worden om met name aandacht te besteden aan clusters die in bepaalde opzichten als interessant kunnen worden aangemerkt. Dat kan bijvoorbeeld het geval zijn wanneer

- een cluster direct aan het begin van het clusterproces ontstaat op basis van zeer hoge similariteitswaarden
- een cluster duidelijk sneller groeit dan andere clusters
- een cluster vanaf een bepaald moment niet meer verandert, terwijl het clusterproces nog wel verder doorgaat
- een cluster uitsluitend of bijna uitsluitend bestaat uit een reeks documenten met opeenvolgende nummers (het zou kunnen gaan om een serie documenten uit een en dezelfde wet)
- een cluster juist uitsluitend bestaat uit documenten waarvan de nummers *niet* opeenvolgend zijn (het zou kunnen gaan om documenten over een verwant onderwerp uit verschillende wetten).

Documenten met een similariteit van 100%

Door het cluster-programma wordt in iedere ronde een lijst aangelegd van similariteitswaarden. In deze lijst wordt, nadat de daartoe noodzakelijke berekeningen zijn gemaakt, voor ieder document vermeld met welk ander document het in die ronde de hoogste similariteit heeft. De hoogste 500 waarden voor iedere ronde worden daarbij telkens bewaard in een bestand (meer is niet praktisch, het uitvoerbestand zou anders onoverzichtelijk groot worden). Uit de lijst van de eerste ronde kan worden afgelezen dat bij elk van de

gehanteerde clustermethoden een aantal documenten een similariteit van 100% met een ander document vertonen. Bij de twee symmetrische methoden moet het daarbij gaan om volledig identieke documenten, aangezien daar zowel hit1's als hit2's een rol spelen bij de berekening van de similariteit. Alle typen die aanwezig zijn in document X moeten ook aanwezig zijn in document Y, en alle typen die ontbreken in X moeten ontbreken in Y, anders kan de similariteit geen 100% zijn. Bij de asymmetrische methoden behoeft alleen het aantal hit1's gelijk te zijn aan 100%. Dat betekent dat een klein document, waarvan alle typen ook voorkomen in een ander, veel groter document, met dat grotere document toch een similariteit van 100% heeft. Vooral de geheel identieke documenten roepen daarbij vragen op, zoals: gaat het hier om dubbel opgenomen tekstgedeelten (in wezen een fout in het corpus) of om gedeeltelijke overeenkomsten tussen voor het overige verschillende wetten of regelingen? Wanneer een lijst van in het corpus aanwezige documenten wordt geraadpleegd blijkt dat beide categorieën vertegenwoordigd zijn. Onderscheiden kunnen worden:

- Kennelijke doublures in het corpus. Een voorbeeld hiervan vormen de documenten 834 en 1342, wederzijds met een similariteit van 100%. Beide documenten bevatten dezelfde wettelijke regeling, de "Wet houdende machtiging tot deelneming door Nederland in de Tweede Aanvulling van het Aziatisch Ontwikkelingsfonds" Deze wet is kennelijk tweemaal in het corpus terechtgekomen, een nog niet eerder opgemerkte fout. Ook het "Veiligheidsbesluit Binnenvaart 1934" blijkt voor een groot deel dubbel opgenomen, enkele tientallen documenten uit beide versies hebben een similariteit van 100%. Opmerkelijk bij deze laatste wet is echter dat ook verschillende documenten uit beide exemplaren *niet* identiek zijn, diverse woorden blijken onder andere anders gespeld. Het moet hier wel gaan om los van elkaar ingevoerde versies, er is geen sprake van kopieën.

- Interessanter dan bovenstaande fouten zijn identieke documenten in twee verschillende wetten. Zo blijken bijvoorbeeld de "Algemeen Burgerlijke Pensioenwet 1966" en de "Spoorwegpensioenwet 1967" in totaal zes identieke documenten te bevatten, elk bestaande uit een paragraaf. Ook het "Slagerij Hinderwet Besluit" en het "Besluit Brood- of banketbakkerijen Hinderwet" hebben in totaal 5 documenten gemeenschappelijk, en de "Pensioenwet voor de Landmacht" en de "Pensioenwet voor de Zeemacht" in totaal 7. Een vijftal "Wetten voor Gemeentelijke Herindeling" bevat hetzelfde document met "Algemene bepalingen".

- Een bijzonder geval vormen twee paren identieke documenten uit één en dezelfde wet. Het blijkt daarbij te gaan om het "Onderzoeksregulatief 1957", waarin voorschriften worden gegeven voor allerlei vormen wetenschappelijk onderzoek. Sommige voorschriften zijn daarbij voor twee verschillende onderzoeksvormen gelijk.

In totaal worden bij de symmetrische methoden in de eerste ronde 81 paren, 6 drietallen en 1 vijftal identieke documenten gevonden. Deze documenten met een similariteit van 100% worden als eerste tot een cluster samengevoegd. Bij de 81 paren is één clusterronde daarvoor voldoende, voor de drietallen zijn twee en voor de vijftallen vier ronden nodig. Bij de asymmetrische methoden worden daarnaast nog 62 documenten gevonden die éénzijdig een similariteit van 100% met een ander document vertonen, wat betekent dat alle woordtypen uit deze documenten ook te vinden zijn in dat andere document, maar

andersom niet (meestal is het andere document groter). Ook deze documenten worden, samen met de documentenparen, direct in de eerste ronde tot clusters samengevoegd. Het samenvoegen van deze documenten, die ofwel identiek zijn ofwel geheel door een ander document overlapt worden is bij beide methoden zeker goed te verdedigen, aangezien daardoor in de eerste plaats fouten (in de vorm van dubbel opgenomen documenten) aan het licht worden gebracht en geneutraliseerd, en in de tweede plaats relaties tussen wetten en regelingen gelegd kunnen worden die in alle bestudeerde gevallen logisch en zinvol blijken te zijn.

Clusters, gevormd in de eerste ronde

In de eerste ronde is bij iedere methode naast documenten met een similariteit van 100% ook een flink aantal documenten met een lagere similariteit in clusters samengevoegd. Zelfs bij de minst 'produktieve' methode (similariteit op basis van hit1- en hit2informatie) zijn dat er nog altijd ruim 2700. Wanneer we net als het cluster-programma de lijst met similariteitsgegevens van boven naar beneden doorlopen blijkt dat de eerste clusters (afgezien van die waar sprake was van 100% similariteit) gevormd zijn op basis van een similariteit van ruim 99%, terwijl bij de laatste clusters de similariteit (bij de symmetrische methoden: een van beide similariteiten) juist boven de drempelwaarde van 50 c.q. 20% ligt[21]. In tabel 8-5 worden de eerste vijf clusters met een similariteit van minder dan 100% en de laatste vijf clusters van de eerste ronde opgesomd die met elk van de vier methoden worden gevormd.

Informatie, hit1's		Gewichten, hit1's		Informatie, hit1's en hit2's		Gewichten, hit1's en hit2's	
nr	documenten	nr	documenten	nr	documenten	nr	documenten
161	12288 + 12298	161	9225 + 9319	100	2006 + 14944	99	7046 + 7052
162	9280 + 9299	162	16561 + 17384	101	9273 + 9292	100	7284 + 7346
163	562 + 584	163	3552 + 4228	102	6313 + 6329	101	2006 + 14944
164	16106 + 16850	164	16106 + 16850	103	5986 + 6106	102	9273 + 9292
165	9225 + 9319	165	16127 + 16871	104	7726 + 7728	103	6313 + 6329
4318	1601 + 1989	4101	4702 + 6071	2803	6899 + 6901	3321	18256 + 18267
4319	15583 + 15594	4102	2029 + 2038	2804	15907 + 15998	3322	10737 + 10739
4320	7110 + 7292	4103	18475 + 18478	2805	2266 + 2267	3323	18092 + 18093
4321	15335 + 15341	4104	8584 + 8585	2806	14100 + 14109	3324	4589 + 4801
4322	15739 + 16146	4105	10949 + 11064	2807	4271 + 4272	3325	14100 + 14109

Tabel 8-5 Eerste en laatste clusters gevormd in de eerste ronde (exclusief clusters met 100% similariteit)

Het blijkt dat de twee asymmetrische methoden twee clusters uit het begin-rijtje gemeenschappelijk hebben, en de twee symmetrische methoden drie. Alleen de symmetrische methoden hebben een gemeenschappelijk cluster in de eind-rijtjes. Geen enkel cluster blijkt voor te komen in de begin- of eindrijtjes van een van de asymmetrische *en* een van de symmetrische methoden. Voor wat betreft de *eind*-rijtjes is dat niet zo vreemd, aangezien de similariteit daar niet op hetzelfde niveau zal liggen (voor de symmetrische en asymmetrisch methoden gelden immers andere drempelwaarden). Dat de *begin*-rijtjes op geen enkel punt overeenkomen komt niet doordat bij de eerste documentenparen met een similariteit net onder de 100% geen sprake is van *wederzijdse* similariteit (in feite blijkt het

[21] Zie voor uitleg met betrekking tot deze drempelwaarden de voorgaande paragraaf.

begin-rijtje uit kolom 1 al drie documentenparen te bevatten waarvan de similariteit wederzijds is), het is een gevolg van de andere uitkomsten van de similariteitsberekening wanneer hit2's daarin niet of juist wel een rol spelen. Zelfs bij de documenten met de hoogste similariteitswaarden (98 àl 99%) is dat merkbaar, vooral wanneer de omvang van de documenten verschillend is. Zo blijkt uit de lijsten met similariteitswaarden dat document 12283 (61 tokens) een similariteit heeft van 99,51% met document 12289 (78 tokens) wanneer de berekening wordt uitgevoerd op basis van hit1-informatie, en van 98,49% wanneer deze is gebaseerde op hit1- en hit2informatie. Dit lijkt een gering verschil, maar het is goed voor een tientallen plaatsen lagere positie in de gesorteerde similariteitslijst. Bovendien zal het verschil in de meeste gevallen toenemen wanneer de similariteit tussen documenten lager is dan in dit voorbeeld.

Van belang is vooral de vraag of de gevormde clusters ook als zinvol kunnen gelden. Bij de clusters die aan het begin van ronde 1 gevormd zijn en die bestaan uit documenten die ofwel identiek zijn, of een zeer hoge similariteit vertonen zal dat meestal wel het geval zijn. Anders ligt dat mogelijk met de clusters die in ronde 1 als laatste zijn gevormd[22] en die een veel lagere similariteit hebben. Om enig inzicht te verkrijgen in hoeverre ook de documenten in deze clusters gemeenschappelijke kenmerken hebben (afgezien van overeenkomsten in woordgebruik) is van enkele ervan, te weten van de documenten in de laatste twee clusters in elke kolom, nagegaan wat hun inhoud is:

Doc. 15335: de paragraaf "Behandeling in raadkamer" uit de Tweede afdeling ("Behandeling van klachten") van het "Reglement medisch tuchtrecht en oplossing van geschillen".
Doc. 15341: de paragraaf "Bij het Gerechtshof" uit het hoofdstuk "Hoger beroep" van hetzelfde reglement.

Doc. 15739: de paragraaf "Commissies van de Arboraad" uit de "Arbeidsomstandighedenwet".
Doc. 16146: de paragraaf "Bedrijfsgeneeskundige diensten" uit de "Veiligheidswet 1934".

Doc. 8584: de afdeling "Inpassing bestaande scholen" uit de "Wet op het hoger beroepsonderwijs".
Doc. 8585: de afdeling "Tijdelijke handhaving opleidingen en examens oude stijl" uit dezelfde wet.

Doc. 10949: de complete "Beschikking geldelijke steun warmte-isolatie bestaande woningen".
Doc. 11064: de paragraaf "Brandwering, brandvoortplanting en vliegvuurbestendigheid" uit de "Regeling geldelijke steun huurwoonwagens".

Doc. 14100: bijlage IV (over plantenziekten) uit de "Beschikking in- en uitvoereisen planten 1979".
Doc. 14109: de complete "Regeling aanwijzing schadelijke organismen 1984".

---

[22] Bij het zoeken naar mogelijke clusters wordt altijd gewerkt op volgorde van similariteit. Zie ook de voorgaande paragraaf.

Doc. 4271: bijlagen P en Q uit de "Aanpassingsregeling pensioenen 1982-I".
Doc. 4272: bijlage R uit dezelfde regeling.

Doc. 4589: de tekst van het "Besluit houdende aanwijzing van de rijkseigendommen die in eigendom, beheer en onderhoud overgaan op de gemeente Lelystad en bepaling van de voorwaarden en tijdstippen van overgang". Dit besluit bevat onder andere een lange lijst getallen (nummers van percelen, met hun oppervlakte).
Doc. 4801: de tekst van het "Besluit houdende aanwijzing van de rijkseigendommen die in eigendom, beheer en onderhoud overgaan op het openbaar lichaam 'Zuidelijke IJsselmeerpolders' en bepaling van de voorwaarden en tijdstippen van overgang". Ook dit besluit bevat een lijst getallen, soortgelijk aan die in het voorgaande document.

Doc. 14100 en 14109: zie hierboven.

Het is niet eenvoudig om uit de gegevens in deze lijst conclusies te trekken, aangezien het om zeer verschillende documenten gaat en uit het opschrift niet altijd goed is op te maken wat de inhoud is. Het vermoeden rijst wel dat de similariteit in sommige gevallen (bijvoorbeeld bij de documenten over warmte-isolatie en brandwering) gebaseerd is op slechts een of enkele elementen die beide documenten gemeenschappelijk hebben (zoals in dit geval mogelijk de elementen brand en brandwering), waardoor het enigszins twijfelachtig is of het betreffende cluster direct zinvol is. Het lijkt erop dat deze 'similariteit gebaseerd op een beperkt aantal elementen' meer voorkomt in de clusters gevormd door beide asymmetrische methoden dan in die gevormd door de symmetrische methoden; er kan daarbij echter sprake zijn van toeval. In ieder geval blijkt in alle bestudeerde gevallen sprake te zijn van een zekere inhoudelijke verwantschap tussen de documenten in bovenstaande clusters.

Naar aanleiding van dit alles kan de voorzichtige conclusie worden getrokken dat in de eerste ronde bij elk van de vier toegepaste methoden een aantal zinvolle clusters worden gevormd. Zoals verwacht kon worden zijn de overeenkomsten, voor wat betreft de clusters die gevormd worden, aanzienlijk groter tussen beide asymmetrische methoden en beide symmetrische methoden dan tussen bijvoorbeeld de asymmetrische en de symmetrische methode gebaseerd op hit-informatie.

Clusters, gevormd na 10 ronden
Bij alle vier clustermethoden worden tenminste 10 ronden voltooid. Het aantal clusters dat in die tien ronden wordt gevormd verschilt echter aanzienlijk, van 11389 respectievelijk 12259 bij de asymmetrische methoden tot 4148 respectievelijk 5651 bij de symmetrische methoden. Het aantal clusters dat per ronde wordt gevormd is dus bij de asymmetrische methoden gemiddeld meer dan twee keer zo hoog. Iedere volgende ronde levert steeds minder clusters op dan de ronde daarvoor, maar het is opvallend dat de daling bij beide asymmetrische methoden veel geringer is. Deze bedraagt bijvoorbeeld in de vierde ronde al minder dan 30%, tegen 60% respectievelijk 46% bij de twee symmetrische methoden. De daling van het aantal clusters neemt bij de asymmetrische methoden daarna steeds verder af, maar stabiliseert zich bij de symmetrische methoden op 40 àl 50%. Dit resulteert erin dat in ronde tien bij de asymmetrische methoden 342 respectievelijk 471 clusters

worden gevormd, terwijl dat er bij de symmetrische methoden maar 3 respectievelijk 11 zijn.

Het is interessant om na te gaan of zich bij de verschillende methoden, naast een hoger of lager aantal, ook andersoortige clusters hebben gevormd. Om praktische redenen zullen daarbij alleen de clusters die zijn gevormd in ronde 10 in beschouwing worden genomen, clusters waar in die ronde niets mee gebeurt blijven in dit geval buiten beschouwing. Bij iedere methode worden enkele clusters geselecteerd en vervolgens geanalyseerd. Het gaat hierbij niet om een representatieve steekproef uit alle clusters, maar om clusters die om bepaalde redenen (bijvoorbeeld door hun omvang, door de opeenvolging van document-nummers etc.) opvallen. De clusters zullen worden geanalyseerd door na te gaan wat de inhoud is van de documenten waar ze uit bestaan, en verder door het bestuderen van het 'top-20' lijstje van hit1- en hit2-woordtypen dat bij de vorming van een cluster door het clusterprogramma is opgeslagen (zie de vorige paragraaf). Om te beginnen een drietal clusters die het resultaat zijn van *similariteitsberekening op basis van hit1-informatie*:

11154: &10-107 (57 documenten) =	5010+5011+5068+5070+5092+5267+5270+5286+5288+5289+5290+ 5297+5298+5306+5307+5319+5322+5330+5333+5403+5448+5464+ 5467+5520+5521+5535+5538+5544+5547+5630+5631+5632+5663+ 5772+5773+5774+5783+5784+5789+5790+6497+6532+6533+6534+ 6540+6541+6542+6545+6546+6547+6548+6549+6573+6574+6575+ 6576+6577
11185: &10-138 (23 documenten) =	7373+7413+7418+7419+7420+7421+7422+7423+7424+7425+7426+ 7427+7428+7429+7430+7431+7432+7433+7434+7435+7436+7437+ 7438
11277: &10-230 (33 documenten) =	14320+14324+14329+14333+14337+14338+14341+14342+14345+ 14346+14349+14350+14356+14362+14363+14367+14368+14371+ 14372+14375+14376+14377+14380+14381+14384+14385+14391+ 14397+14398+14402+14403+14407+14408

Het eerste getal is het volgnummer van het cluster, daarna volgt de cluster-naam bestaande uit het teken '&', het ronde-nummer (10) en het volgnummer binnen deze ronde. Vervolgens wordt vermeld hoeveel documenten het cluster telt, en worden de nummers van deze documenten opgesomd (in de volgorde waarin ze in het corpus voorkomen). Het eerste cluster is in dit geval geselecteerd omdat het het grootste is in ronde 10, en omdat het steeds reeksen van twee of drie documenten met opvolgende nummers bevat. Het tweede cluster bevat bijna alleen documentnummers die elkaar direct opvolgen, en het derde heeft weer ongeveer dezelfde structuur als het eerste. De clusters blijken het volgende te bevatten:

11154: paragrafen uit wetten die gaan over gemeentelijke her-indelingen (de paragraaf "Verkiezing vertegenwoordigende lichamen" komt bijvoorbeeld een aantal keren voor). Voorzover valt na te gaan gaat het in de meeste gevallen om vrij recente herindelingswetten (van na 1970). Woordtypen die door hun *aan*wezigheid een rol hebben gespeeld bij de vorming van dit cluster (hit1-woordtypen) zijn vrijwel uitsluitend de namen van gemeenten: GIETHOORN, BUURMALSEN, ZOELEN, DEIL, OLDEMARKT. Aangezien bij deze clustermethode alleen hit1's van belang zijn, zijn er geen woordtypen waarvan de *af*wezigheid een rol heeft gespeeld.

11185: een aantal paragrafen uit wetten en regelingen over het onderwijs, zoals de "Lager-onderwijswet 1920" en het "Besluit buitengewoon onderwijs 1967". Hit1-woordtypen zijn onder andere: HOOFDLEIDSTER, BEZITTERS, ONDERWIJZER, GETUIGSCHRIFTEN en ONDERWIJSRAAD.

11277: een aantal paragrafen uit regelingen die betrekking hebben op zuivelprodukten, zoals de "Landbouwkwaliteitsbeschikking kaasprodukten" en de "Landbouwkwaliteitsbeschikking boterprodukten". Hit1-woordtypen zijn: CONTROLETEKEN, RIJKSKAASMERK, KAASSOORT, KAASSOORTEN, AFGEBEELDE en RIJKSTOEZICHT

Elk van deze drie clusters (en ook een groot aantal andere die werden onderzocht alvorens deze drie geselecteerd werden) kan als zinvol worden aangemerkt, hoewel ze geen van alle 'volledig' zijn. Ze bevatten steeds wel een aantal paragrafen uit verschillende wetten of regelingen, maar nooit een gehele regeling, of bijvoorbeeld een heel hoofdstuk uit een regeling.

Enkele clusters die gevormd werden bij de *similariteitsberekening op basis van hit1-gewichten* zijn:

11947: &10-159 (28 documenten) = 8371+8373+8379+8381+8382+8384+8385+8386+8387+8388+8391+
8871+8873+8877+8878+8879+8880+8881+8882+8884+8885+8886+
8887+8888+8894+8895+8896+8897

11961: &10-173 (16 documenten) = 11108+11111+11112+11113+11114+11115+11141+11197+11199+
11200+11201+11202+11203+11223+11224+11225

12217: &10-429 (74 documenten) = 5003+5013+5018+5034+5041+5045+5061+5073+5074+5080+5096+
5097+5104+5111+5123+5180+5192+5193+5281+5300+5335+5388+
5394+5406+5413+5414+5416+5417+5420+5424+5430+5442+5450+
5451+5461+5476+5523+5527+5531+5549+5552+5569+5570+5572+
5586+5602+5612+5613+5614+5625+5654+5695+5696+5760+5763+
5764+5766+5775+5776+5777+5793+6412+6413+6414+6418+6420+
6421+6441+6442+6443+6445+6446+6447+6579

De eerste twee clusters vallen met name op vanwege de opeenvolgende documentnummers, het laatste cluster werd geselecteerd omdat het veel weg heeft van het eerste cluster van de vorige serie, bij de clustermethode op basis van hit1-informatie. De clusters bevatten het volgende:

11947: een groot deel van het "Bekostigingsbesluit WBO" (betreffende bekostiging van basisscholen) en het "Bekostigingsbesluit ISOVSO" (betreffende bekostiging van speciaal en voortgezet speciaal onderwijs). Hit1-woordtypen zijn hierbij: SCHOOLBADEN, SPEELLOKAAL, SPEELLOKALEN, HUISVESTINGSVOORZIENINGEN, GYMNASTIEKLOKALEN, MOTORISCHE en THERAPIE.

11961: delen van regelingen die te maken hebben met subsidiëring van huurwoningen, te weten enkele paragrafen uit de "Regeling geldelijke steun huurwoningen normkostensysteem" en de "Regeling geldelijke steun voorzieningen aan huurwoningen". Hit1-woorden zijn: STREEFSUBSIDIE, STREEFSUBSIDIES, FLEXIBILITEIT, STREEFAANTAL, STREEFAANTALLEN, BUDGETTEN en DEELBUDGETTEN.

12217: een aantal wetten (meestal volledig, een enkele keer gedeeltelijk) die betrekking hebben op gemeentelijke herindelingen, zoals de "Wet tot vereeniging der gemeenten Alkemade en Vrije en Lage Boekhorst" en de "Wet tot vereeniging der gemeenten Almkerk en Emmikhoven". Anders dan bij het cluster 11154 uit de vorige serie blijkt hier in bijna alle gevallen al uit de oude spelling in de titel van de wet dat het om herindelingen van wat langer geleden gaat, uit de vorige eeuw of het begin van deze eeuw. Het is dan ook niet verwonderlijk dat onder de hit1-typen, naast namen van betrokken gemeenten, ook woorden in oude spelling en enkele jaartallen uit die periode zijn: VOOREN, REENEN, LINDT, ACTEN, STEMOPNEMING, VEREENIGDE, DINGSDAG, NADEELE, JUNIJ, COLLEGIEN, BURGERLIJKEN, 1855 en 1857. Woorden in oude spelling komen weinig voor in het corpus, en hebben daardoor een vrij hoog hit1-gewicht (en hoge hit1-informatie).

Bij de twee *symmetrische methoden* van similariteitsberekening werden in ronde 10 veel minder clusters gevormd dan bij de asymmetrisch methoden. Bij de *methode op basis van hit1- en hit2informatie* waren dit er in totaal slechts 3:

4146: &10-1 (22 documenten) =	5003+5073+5080+5097+5335+5420+5430+5442+5451+5527+5531+5613+5614+5763+5764+5766+5776+5793+6420+6442+6443+6579
4147: &10-2 (13 documenten) =	4942+5006+5054+5116+5265+5533+5768+5779+6482+6503+6528+6551+6571
4148: &10-3 (23 documenten) =	5034+5045+5074+5111+5123+5180+5281+5300+5388+5394+5406+5413+5414+5417+5424+5461+5476+5572+5625+5760+5777+6421+6441

In dit geval zijn dus alle clusters uit ronde 10 opgenomen, het maken van een keuze is hier niet aan de orde. De inhoud ervan is als volgt:

4146: een subset van de wetten over gemeentelijke herindeling uit cluster 12217 van de vorige serie. In alle gevallen gaat het om oude wetgeving (onder andere uit 1855), wat soms al blijkt uit de opschriften ("Wet van de 11den Julij 1855, tot vereeniging der gemeenten ...."). Hit1-typen met een grote hoeveelheid informatie zijn onder andere: CURTIUS, REENEN, ACTEN, REGTSGEBIED, 11DEN, DINGSDAG, VEREENIGDE, JULIJ, BURGELIJKEN, 185455 en 1855. De *af*wezigheid van woordtypen heeft in dit geval ook een rol gespeeld bij het berekenen van de similariteit. Afwezige typen (hit2-typen) met veel informatie zijn onder andere: EEN, ARTIKEL, LID, ZIJN, INDIEN, BEDOELD en BEDOELDE[23].

4147: een cluster dat sterk doet denken aan cluster 11050 uit de eerste serie (zie hierboven). De documenten waar het uit bestaat zijn hoofdstukken "Algemene bepalingen" uit een aantal gemeentelijke herindelingswetten. Typen met veel hit1-informatie zijn onder andere: GESPROKEN, HERINDELING, KALENDERMAAND, TOEGEVOEGD, DAARONDER en INWERKINGTREDING. Ontbrekende typen met veel hit2-informatie zijn: EN, MET, VOOR, WORDEN, ARTIKEL,

[23] Merk op dat van de meeste van deze hit2-typen in het vorige hoofdstuk werd vastgesteld dat ze juist *kenmerkend* zijn voor wetgevingsteksten (zie paragraaf 7.5). De documenten in dit cluster komen op dit punt dus niet overeen met 'typische' wetgevingsdocumenten.

ZIJN, LID, BEDOELD en INDIEN.

4148: opnieuw een subset uit hetzelfde cluster 12217 uit de vorige serie. Net als in cluster 4146 gaat het steeds om oude wetgeving over gemeentelijke herindelingen. Hit1-typen zijn: DINGSDAG, KOCK, ACTEN, RAPPARD, STEMOPNEMING, 1857, 1859, 1860, VEREENIGDE, JUNIJ en BURGERLIJKEN. Hit2-typen komen voor het grootste deel overeen met die bij cluster 4146: EEN, ARTIKEL, ZIJN, NIET, KAN, BEDOELD, INDIEN, BEDOELDE en BEPAALDE.

Op het eerste gezicht is niet duidelijk waarom de documenten in de clusters 4146 en 4148 niet net als bij de vorige serie één groot cluster vormen. Uit de lijstjes met hit1- en hit2-typen blijkt eigenlijk (afgezien van de namen van de gemeenten waar de wetten over gaan) maar één verschil: de wetten in cluster 4148 lijken één of enkele jaren recenter te zijn (ze bevatten de jaartallen 1857, 1859 en 1860) dan die in cluster 4146 (waarin de jaartallen 1854/55 en 1855 voorkomen). Het zou kunnen zijn dat sprake is van een tweetal *series* herindelingswetten, enkele jaren na elkaar uitgevaardigd en met een iets afwijkend woordgebruik (zie ook de iets afwijkende hit2-lijstjes). Als deze verklaring inderdaad juist is zou dat betekenen dat deze symmetrische cluster-methode wat gevoeliger is voor geringe variaties in het woordgebruik dan de asymmetrische methode (op basis van hit1-gewichten) waarvan cluster 12217 het resultaat was.

Bij de methode op basis van hit1- en hit2gewichten werden in ronde 10 onder andere de volgende clusters gevormd:

5646: &10-6 (30 documenten) =	16091+16093+16095+17864+17867+17870+17872+17878+17879+ 17881+17895+17897+17908+17913+17957+17958+17959+17960+ 17961+17962+17963+18020+18024+18063+18076+18099+18138+ 18150+18158+18166
5648: &10-8 (21 documenten) =	18500+18509+18517+18527+18546+18566+18574+18582+18590+ 18600+18609+18624+18646+18655+18684+18697+18706+18719+ 18731+18744+18757
5649: &10-9 (18 documenten) =	6648+6649+6650+6738+6741+6742+6744+6745+6746+6748+6750+ 6752+6753+6754+6764+6765+6768+6769

5646: een aantal besluiten tot aanwijzing van een psychiatrische afdeling van een bepaald ziekenhuis of verpleeghuis als een "inrichting, welke niet als een gesticht voor krankzinnigen wordt beschouwd". Hit1-typen zijn hierbij: PSYCHIATRISCHE, VERPLEEGD, AMERSFOORT, PATIENTEN, KRANKZINNIGEN, CENTRUM, ZIEKENHUIS en GESTICHT. Hit2-typen zijn: DIE, EERSTE, A, B, INDIEN, TOEPASSING, TER, HEEFT, UIT, ANDERE, KRACHTENS en DAARVAN.

5648: 21 paragrafen uit het omvangrijke "Stoffenbesluit Cosmetica" die allemaal het opschrift "Herhaalbaarheid" hebben. Hit1-typen zijn: HERHAALBAARHEID1, MEETUITKOMSTEN, ANALIST, IDENTIEKE, ZELFDE, GEHALTE en MONSTER. Hit2-typen zijn: ART, TE, DIE, WORDT, ARTIKEL, LID, BEDOELD, A en B.

5649: een aantal paragrafen die voornamelijk salarisschalen voor politieambtenaren bevatten uit het "Bezoldigingsreglement politie 1958" en het "Besluit herstructurering politiesalarissen". Hit1-typen zijn in dit geval: een aantal getallen, zoals 2952 en 3326 (mogelijk salarissen in guldens) en de typen HOOFDAGENT en SAL (waarschijnlijk als afkorting voor salaris). Hit2-typen zijn: ART, WORDEN, TE,

DAT, DOOR, LID, NIET, TEN, BEDOELD, INDIEN, EERSTE, TWEEDE en BEPAALDE.

Een conclusie met betrekking tot clusters, gevormd in ronde 10 kan zijn dat alle vier clustermethoden kennelijk in principe geschikt zijn voor de vorming van zinvolle clusters. De twee asymmetrische clustermethoden produceren een groter aantal clusters, die gemiddeld ook omvangrijker zijn. Daarentegen lijkt het erop dat de symmetrische methoden wat nauwkeuriger te werk gaan, en in staat zijn op basis van zeer geringe verschillen in woordgebruik documenten in aparte clusters in te delen, zoals blijkt uit het voorbeeld van de gemeentelijke herindelingswetten. Door die precisie vallen echter ook sommige documenten uit de boot: de twee afzonderlijke clusters uit het voorbeeld tellen samen minder documenten dan het allesomvattende cluster dat gevormd werd door de asymmetrische methode (op basis van hit1-gewichten). Opvallend is daarnaast nog, dat in geen van de onderzochte clusters - volgens welke methode ook gevormd - noch in tientallen andere clusters die werden onderzocht maar hier niet konden worden besproken, documenten werden aangetroffen die duidelijk *niet* in het cluster thuishoorden. In alle gevallen was dus sprake van een zekere relatie, ook inhoudelijk, tussen de documenten in een cluster. Dit is op zich al een opvallend resultaat van de clustervorming op basis van similariteit. Een keerzijde is dat van de meeste clusters in deze ronde kan worden gezegd dat ze niet 'volledig' zijn in de zin dat er nog bepaalde documenten aan ontbreken (zoals bij een cluster dat ongeveer de helft van de paragrafen uit een bepaalde regeling omvat). In het hierna volgende gedeelte zal worden bekeken of dit nog steeds het geval is wanneer het clusterproces is geëindigd.

Clusters aan het einde van het cluster-proces

Bij de twee symmetrische clustermethoden stopt het clusterproces vrij kort na de zojuist besproken ronde 10, namelijk na in totaal 11 respectievelijk 17 ronden. Het aantal clusters dat in deze laatste ronden is gevormd is gering: niet meer dan 2 stuks in ronde 11 bij de methode op basis van hit1- en hit2-informatie, en 23 stuks in de ronden 11 t/m 17 bij de methode op basis van hit1- en hit2-gewichten. Zoals verwacht kon worden is het aantal mutaties dat nog optreedt in de clusters uit ronde 10, waarvan enkele hierboven werden besproken, dan ook gering. Bij de methode op basis van hit1- en hit2-informatie worden nog slechts twee clusters gevormd, een uit twee oorspronkelijke documenten, en een door samenvoeging van twee eerder gevormde clusters uit de ronden 1 en 5. In alle gevallen gaat het weer om wetten voor gemeentelijke herindelingen, waarvan ook al enkele (andere) clusters uit ronde 10 werden besproken. Cluster 4148 uit ronde 10 blijft met 23 documenten het grootste cluster dat met deze methode wordt gevormd.

Bij de methode op basis van hit1- en hit2-gewichten worden zoals gezegd nog 23 clusters gevormd. Het analyseren van al deze clusters gaat wat ver, maar wel kan worden nagegaan of er nog mutaties optreden in de drie hierboven besproken clusters uit ronde 10. Dat blijkt alleen het geval te zijn met cluster 5646 (het cluster met aanwijzingsbesluiten van psychiatrische afdelingen). In vier stappen worden hieraan nog 10 (al eerder tot kleinere clusters samengevoegde) documenten toegevoegd, in alle gevallen hetzelfde type aanwijzingsbesluiten als al in het cluster aanwezig waren. In de andere twee clusters uit ronde 10 die hierboven werden besproken treden in de laatste 7 ronden geen mutaties meer op, het zijn dus in feite 'eind-clusters'. In het grootste cluster dat met deze methode

wordt gevormd (56 documenten) blijkt voor het laatst een wijziging op te treden in ronde 14. Het gaat daarbij opnieuw om een cluster met gemeentelijke herindelingswetten, een subset van de wetten in cluster 12217 dat gevormd werd bij de asymmetrische methode op basis van hit1-gewichten. Al met al kan gezegd worden dat het laatste gedeelte van het cluster-proces bij de symmetrische methoden weinig verrassingen te zien geeft. Er ontstaan nog enkele nieuwe clusters en de clusters die al in eerdere ronden werden gevormd groeien nog wat, waarna uiteindelijk het clusterproces stopt. Dit stoppen is het gevolg van het niet langer optreden van *wederzijdse* similariteit tussen de documenten, aangezien in de laatste ronde nog een groot aantal documenten een hoge similariteit, zelfs tot 90%, vertonen met een ander document.

Heel anders ligt de situatie bij de asymmetrische clustermethoden. Na ronde 10 volgt bij deze methoden nog een groot aantal volgende ronden (243 respectievelijk 116), waarin nog een groot aantal mutaties optreedt in de tot en met ronde 10 gevormde clusters. Uiteindelijk worden 18702 respectievelijk 18758 clusters gevormd, waardoor er in totaal maar 101 (18803 - 18702) respectievelijk 45 (18803 - 18758) documenten overblijven. Een deel van deze documenten zijn clusters (waaronder 1 zeer omvangrijk), maar er zijn ook 74 respectievelijk 31 'oorspronkelijke' documenten bij, waarmee dus gedurende het clusterproces niets is gebeurd. In detail is de opbouw in documenten aan het eind van het clusterproces als volgt:

- clustermethode op basis van hit1-informatie:
  - 74 enkelvoudige (oorspronkelijke) documenten
  - 17 clusters van 2 documenten
  - 4 clusters van 3 documenten
  - 2 clusters van 4 documenten
  - 1 cluster van 12 documenten
  - 1 cluster van 2763 documenten
  - 1 cluster van 5003 documenten
  - 1 cluster van 10897 documenten

- clustermethode op basis van hit1-gewichten:
  - 31 enkelvoudige (oorspronkelijke) documenten
  - 12 clusters van 2 documenten
  - 1 cluster van 3 documenten
  - 1 cluster van 18745 documenten

Het is duidelijk dat bij beide methoden (maar vooral bij de methode van de hit1-gewichten) één cluster wel erg groot is geworden. Het blijkt zo te zijn dat de opbouw van dit grootste cluster in beide gevallen is terug te volgen tot aan de allereerste ronde, en dat er dus in iedere ronde een document of een cluster van documenten aan is toegevoegd. Het is interessant om te zien of de eindclusters en -documenten zich net als in ronde 10 laten benoemen. In het geval van de enkelvoudige documenten en de clusters van 2 en 3 documenten wordt daarbij echter volstaan met een steekproef van enkele stuks. Als eerste komen aan bod de eindclusters van de methode op basis van hit1-informatie:

- Bij de enkelvoudige documenten gaat het meestal om bijlagen bij regelingen als het "In- en uitvoerbesluit bedreigde en uitheemse dier- en plantensoorten" of het "Regle

ment geur- en smaakstoffen", waarin termen voorkomen (bijvoorbeeld namen van dieren of van stoffen) die nergens anders in het corpus te vinden zijn. Daarnaast treffen we een aantal documenten aan die in het Frans gesteld blijken te zijn (paragrafen uit een mijnwet uit de Franse tijd).

- Bij de clusters van 2 documenten, die overigens allemaal in de eerste ronde gevormd blijken te zijn, gaat het eigenlijk om hetzelfde (documenten met zeer weinig gebruikte termen), alleen is er in dit geval nog precies 1 ander document dat dezelfde zeldzame term bevat. Een voorbeeld zijn hier de documenten 1431 en 1576, die allebei over zeldzame vissen gaan. Het ene document komt uit bovengenoemd in- en uitvoerbesluit, het andere uit de "In- en uitvoer*beschikking* bedreigde en uitheemse dier- en plantensoorten".
- Clusters bestaande uit 3 documenten vallen eigenlijk in dezelfde categorie als die van twee documenten. De documenten waar ze uit bestaan bevatten een term, bijvoorbeeld een plaatsnaam, die verder (praktisch) niet voorkomt. Bovendien treffen we hier een cluster bestaande uit drie Franstalige documenten uit bovengenoemde mijnwet.
- Eén van de clusters bestaande uit 4 documenten bevat enkele paragrafen uit de "Aanpassingsregeling pensioenen 1980" en de "Aanpassingsregeling pensioenen 1981". De documenten hebben voornamelijk een aantal getallen (waarschijnlijk geldbedragen) gemeenschappelijk, en termen als PENSIOENUITKERING, INVALIDITEITSTOESLAG en WEZENONDERHOUD. Verder treffen we een cluster aan met documenten over plantenziekten en schadelijke organismen.
- Het cluster van 12 documenten is een merkwaardig samenstelsel van documenten over allerlei onderwerpen, van de schadelijkheid van straling tot de samenstelling van cosmetica en de afkoop van pensioenrechten. Hit1-woordtypen blijken in dit geval vooral *getallen* te zijn, zoals 259, 261, 265, etc. Het is heel goed mogelijk dat deze bij de samenstellende documenten geheel verschillende betekenissen hebben, zoals bijvoorbeeld nummers van wettelijke bepalingen, atoomnummers en geldbedragen.
- Bij het cluster van 2763 documenten is het al bijzonder moeilijk om een samenhang tussen de documenten te vinden. Het cluster omvat bijvoorbeeld een groot aantal van de al eerder genoemde gemeentelijke herindelingswetten, maar ook delen van de "Comptabiliteitswet", over de opstelling van de rijksbegroting. Het bestuderen van lijstjes met hit1-woorden zegt bij dergelijke omvangrijke clusters weinig meer, aangezien het lijstje dat bij het cluster hoort slechts betrekking heeft op de laatste mutatie die heeft plaatsgevonden. Wanneer een cluster erg groot wordt is het mogelijk dat de typen met de meeste hit1-informatie slechts betrekking hebben op een gedeelte van de documenten waar het cluster uit bestaat.
- Nog sterker geldt dit voor de clusters van 5003 en van 10897 documenten. Deze clusters omvatten allebei wetten, regelingen en delen daarvan zonder dat een en ander zich precies laat benoemen. Verschillende wetten, zoals wederom de "Comptabiliteitswet" zijn verdeeld over beide clusters, zonder dat daarvoor direct een reden is aan te geven.

Bij de clustermethode op basis van hit1-gewichten gaat het om de volgende eindclusters:

- Bij de enkelvoudige documenten treffen we hetzelfde soort documenten aan als bij de methode op basis van hit1-informatie: ze bevatten woordtypen die nergens anders in het corpus voorkomen. De Franse documenten zijn er in dit geval niet bij, die zijn terecht gekomen in het grootste cluster van 18745 documenten.

- ook de clusters van 2 documenten lijken sterk op die uit het rijtje hierboven (het gaat voor een deel weer om dezelfde documenten.
- Er is maar 1 cluster dat uit 3 documenten bestaat. De drie documenten blijken vooral namen van geneesmiddelen te bevatten, het zijn bijlagen uit regelingen op dat gebied.
- Het cluster van 18745 documenten omvat alle andere documenten in het corpus. Het is vanzelfsprekend dan ook niet mogelijk dit cluster nader te benoemen.

Op basis van deze analyse blijkt dat vooral de grootste clusters bij de twee asymmetrische methoden zich niet of nauwelijks meer laten benoemen. Er is bij deze clusters geen voor de hand liggende reden meer te vinden waarom documenten er wel of geen deel van uitmaken. Waarschijnlijk hebben we hier te maken met een ongewenst neveneffect van asymmetrische clustermethoden dat al eerder in dit hoofdstuk werd genoemd, namelijk dat clusters van documenten waarin een bepaald aantal woordtypen met een hoog hit1-gewicht (c.q. hoge hit1-informatie) voorkomen de neiging hebben om alle documenten waarin die woordtypen ook voorkomen 'naar zich toe te trekken'. Zo'n wat omvangrijker cluster overlapt kleinere documenten vaak voor een aanzienlijk deel (terwijl dat kleinere document het cluster slechts voor een klein deel overlapt), waardoor het kleine document een hoge similariteit met het cluster heeft, en op basis daarvan met dat cluster zal worden samengevoegd. Het cluster heeft intussen slechts een geringe similariteit met het kleine document, maar dat is bij een asymmetrische clustermethode niet van belang. Dat dit 'magneet-effect' van grote clusters in de praktijk inderdaad optreedt kan ook worden opgemaakt uit de manier waarop de clustervorming in de laatste ronden verloopt. Bij beide asymmetrische methoden wordt dan per ronde steeds maar een enkel cluster gevormd, bestaande uit het omvangrijke cluster van de ronde daarvoor en een enkelvoudig document of klein cluster dat van alle documenten die nog over zijn 'toevallig' de hoogste similariteit met het omvangrijke cluster vertoont, en daardoor aan dat cluster 'vastplakt'. Dit proces herhaalt zich in de ronden daarna, totdat het clusterproces stopt bij gebrek aan documenten met een voldoende hoge similariteit (>50%). Dit 'samenklonteren' aan het eind van het clusterproces kan als een belangrijk nadeel van asymmetrische clustermethoden worden gezien.

8.5.4 Conclusie

Het onderzoek dat in deze laatste paragraaf werd beschreven laat nog vele vragen onbeantwoord. De voorgestelde clustermethoden werken alle vier wel, maar hebben elk ook nadelen. Een belangrijk nadeel is dat het clusterproces bij de twee symmetrische methoden te vroeg stopt, waardoor veel losse documenten overblijven, terwijl het bij de twee asymmetrische methoden juist te lang doorgaat, waardoor één of enkele clusters ontstaan die een te groot deel van het corpus omvatten en waarin geen gemeenschappelijke kenmerken van de documenten meer herkenbaar zijn. Toch kunnen wel enkele conclusies worden getrokken met betrekking tot het vormen van clusters op basis van similariteit:

- De similariteit van documenten wordt uitsluitend berekend op basis van de woordtypen die in die documenten voorkomen c.q. ontbreken. Wanneer echter documenten die een hoge similariteit ten opzichte van elkaar vertonen worden bekeken blijkt in een aantal hier bestudeerde gevallen ook sprake te zijn van inhoudelijke overeenkomsten tussen deze documenten, dat wil zeggen van overeenkomsten op semantisch niveau.

- Het berekenen van de similariteit tussen alle 18803 documenten van het corpus wetgevingsteksten, welke gemiddeld rond de 211 woordtypen bevatten, blijkt met behulp van een op dit moment gangbaar type personal computer mogelijk te zijn in ongeveer 24 uur.

- De similariteit van documenten kan worden gebruikt als basis voor het vormen van *clusters* van documenten. Daarbij kan gewerkt worden in een aantal *rondes*. Bij onderzoek van clusters van documenten uit het corpus wetgevingsteksten die zijn gevormd na in totaal 10 ronden blijkt in alle gevallen sprake te zijn van clusters die als relevant kunnen worden aangemerkt, in de zin dat tussen de documenten in een cluster ook een inhoudelijke relatie bestaat.

- In dit hoofdstuk zijn twee methoden van similariteitsberekening behandeld, die respectievelijk zijn gebaseerd op de (voor *gewicht* gecorrigeerde) *aantallen* woordtypen die documenten gemeenschappelijk hebben c.q. allebei missen, en op de met die aantallen woordtypen samenhangende *hoeveelheid informatie*. De methode welke gebruik maakt van de hoeveelheid informatie legt meer nadruk op de aanwezigheid van zeldzame woordtypen en de afwezigheid van veel voorkomende woordtypen. Wanneer deze laatste vorm van similariteitsberekening wordt toegepast voor het vormen van clusters van documenten uit het corpus wetgevingsteksten, blijkt dat in totaal minder clusters worden gevormd dan wanneer de methode op basis van voor gewicht gecorrigeerde aantallen wordt toegepast. Er zijn echter wel aanwijzingen dat met de methode op basis van hoeveelheid informatie een nauwkeuriger onderscheid kan worden gemaakt tussen documenten die slechts geringe verschillen vertonen.

- De eis dat twee documenten *wederzijds* de hoogste similariteit met elkaar moeten hebben om ze te kunnen samenvoegen in een cluster (in dit hoofdstuk ook wel aangeduid als de *symmetrische* clustermethode) leidt er bij het onderzochte corpus toe dat het cluster-proces al vrij snel stopt, terwijl nog niet alle documenten die voor bepaalde clusters relevant zouden kunnen zijn daarin zijn opgenomen.

- Wanneer *eenzijdige* hoge similariteit als voldoende voorwaarde voor het vormen van een cluster wordt beschouwd, blijkt een ongewenst neveneffect te zijn dat vooral in de laatste fase van het clusterproces omvangrijke clusters de neiging hebben voortdurend nog meer documenten aan te trekken, en zo steeds meer in omvang toe te nemen. Dit leidt uiteindelijk tot clusters die een groot deel van het corpus omvatten, en waar niet langer een bepaalde betekenis aan kan worden toegekend op basis van de documenten waaruit ze bestaan.

Het is duidelijk dat het onderzoek dat in dit hoofdstuk werd beschreven niet meer dan een eerste verkenning vormt van de mogelijkheden die de similariteit van documenten biedt. Voor de hand liggende opties voor verder onderzoek zijn bijvoorbeeld het beproeven van de effecten van andere clustermethoden (zoals bijvoorbeeld symmetrische methoden gebaseerd op alleen hit1's, en asymmetrische methoden gebaseerd op hit1's en hit2's) en het gebruikmaken van andere corpora. Verder zou een belangrijke verbetering zijn wanneer aan de vorming van clusters een bepaalde *richting* kon worden gegeven, bijvoorbeeld door het stellen van aanvullende (inhouds)eisen aan nieuw te vormen clusters of door het geven van *voorbeelden* van documenten die in het cluster zouden moeten

voorkomen. Een probleem van de hier toegepaste clustermethoden is namelijk dat het toeval nog een te sterke invloed heeft wanneer het er om gaat of er clusters worden gevormd van documenten die over een bepaald onderwerp gaan. Daarnaast is de vraag interessant of de hier besproken similariteits-scores (op welke manier ook berekend) kunnen worden toegepast voor andere doeleinden. Gedacht kan daarbij bijvoorbeeld worden aan systemen voor data-retrieval, waar de similariteitsrelatie van documenten deel zou kunnen uitmaken van de te stellen zoekvragen, of waarin automatisch gezocht zou kunnen worden naar documenten die met aanvankelijk geselecteerde documenten uit de database een hoge similariteit hebben. Uiteraard zijn ook daarbij bepaalde problemen te verwachten die eerst moeten worden opgelost; op een aantal punten is waarschijnlijk lering te trekken uit de ervaringen met het vormen van document-clusters, zoals in dit hoofdstuk beschreven. Dergelijk vervolgonderzoek over similariteit en de vorming van clusters valt echter buiten het kader van dit boek over juridisch woordgebruik. Vast staat echter intussen wel dat kennis over dat woordgebruik ook praktisch inzetbaar is en wellicht in de toekomst kan leiden tot nieuwe manieren om juridische tekstbestanden in te richten en daaruit gegevens die relevant zijn voor wetenschap en praktijk van het recht te verzamelen.

HOOFDSTUK 9 - SAMENVATTING EN CONCLUSIES

Het onderzoek dat in dit proefschrift wordt beschreven richt zich op kwantitatieve eigenschappen van juridisch taalgebruik. Het zo nauwkeurig mogelijk beschrijven van deze eigenschappen vormt onder andere een noodzakelijke voorwaarde voor de ontwikkeling, in een later stadium, van een nieuwe generatie juridische informatiesystemen, zoals systemen die het mogelijk zullen maken om juridische tekstbestanden te raadplegen aan de hand van juridische 'concepten'. Om inzicht te krijgen in specifieke eigenschappen van het woordgebruik in bepaalde juridische teksten worden deze teksten vergeleken met algemeen Nederlandse tekstmateriaal. De juridische teksten waar het om gaat zijn teksten van wet- en regelgeving en jurisprudentieteksten. Met behulp daarvan zijn voor dit onderzoek twee afzonderlijke juridische corpora opgebouwd. Daarnaast is gebruik gemaakt van een vergelijkingscorpus met algemeen Nederlandse teksten.

Alvorens het woordgebruik in de corpora in kaart te brengen wordt eerst enige aandacht besteed aan onderzoek op het gebied van 'recht en taal' dat in het verleden werd verricht. Daarbij blijkt dat kwantitatieve aspecten van juridisch taalgebruik nog maar zelden voorwerp van onderzoek zijn geweest. Onder andere omdat de (kwantitatieve) taalkunde voor de meeste juristen onbekend terrein is, wordt voorts een hoofdstuk aan deze tak van wetenschap gewijd. Daarin wordt onder andere de plaats aangegeven die de kwantitatieve taalkunde inneemt binnen de taalkunde als geheel. Verder wordt een overzicht gegeven van kwantitatieve methoden en technieken die in dit onderzoek een belangrijke rol spelen, en wordt aandacht besteed aan de voor dit onderzoek relevante begrippen 'subtaal' en 'vaktaal'.

Na dit inleidende gedeelte worden de drie corpora beschreven die in deze studie centraal staan. Het betreft achtereenvolgens:

- een corpus dat de teksten bevat van alle wet- en regelgeving die op het moment van samenstellen in Nederland in werking was. Dit materiaal is afkomstig uit de elektronische databank NLEX, waarbij alleen de oorspronkelijke wetteksten zijn verwerkt, zonder toevoegingen als trefwoorden of annotaties.

- een corpus dat een ruime selectie bevat van teksten van rechterlijke uitspraken. Het gaat hier om alle uitspraken die zijn gepubliceerd in het tijdschrift 'Nederlandse Jurisprudentie' in de jaren 1965 - 1989. Van deze uitspraken is opnieuw alleen de tekst van de oorspronkelijke uitspraak opgenomen, zonder samenvattingen of annotaties.

- een corpus met Nederlands tekstmateriaal van algemene aard, bedoeld als vergelijkingsmateriaal. Dit corpus bevat onder andere teksten van romans, korte verhalen, studieboeken en artikelen, zowel wetenschappelijk als niet-wetenschappelijke, over uiteenlopende onderwerpen.

De corpora worden gesplitst in afzonderlijke woorden ('woorden' betekent hier: rijtjes lettertekens; een andere term ervoor is *woordtokens*). De omvang van de drie corpora loopt nogal uiteen. Het corpus jurisprudentieteksten is het grootst (ruim 40 miljoen woordtokens), het corpus wetgevingsteksten telt ruim 12 miljoen woordtokens, en het corpus algemeen Nederlands is met 4,5 miljoen woordtokens het kleinst. Voor elk corpus wordt een lijst aangelegd van alle verschillende woorden (alle *woordtypen*) die erin voorkomen, met daarbij de *frequentie* van die woorden in het corpus, en het *aantal documenten* waarin ze voorkomen. Deze lijst wordt gesorteerd op woordfrequentie, het meest voorkomende woord bovenaan, waarna aan ieder woordtype een *rangnummer* kan worden toegekend: 1 voor het meest voorkomende woord, 2 voor het op een na meest voorkomende, etc. Deze *frequentielijst* vormt de basis voor een groot deel van de verdere analyses.

Nadat op deze manier is beschreven waaruit het te onderzoeken tekstmateriaal bestaat, worden vervolgens de drie corpora eerst afzonderlijk onderzocht. Het doel daarbij is om systematisch het woordgebruik in deze corpora in kaart te brengen, onder gebruikmaking van kwantitatieve methoden. Aspecten die daarbij aan de orde komen betreffen met name de woordfrequenties, de frequentieverdeling, de verhouding tussen de aantallen woordtokens en woordtypen, en de woordlengtes. Bij het analyseren van deze aspecten wordt veel gebruik gemaakt van grafieken. De nadruk ligt daarbij steeds op het blootleggen van kenmerken die als specifiek voor een bepaald corpus zouden kunnen gelden, waardoor ze een rol zouden kunnen spelen bij het trekken van *vergelijkingen* tussen de corpora. Op elk van de genoemde aspecten zal hier kort worden ingegaan.

Woordfrequenties
Bij het analyseren van de in een corpus voorkomende woordfrequenties is vooral aandacht besteed aan het verband tussen het rangnummer van woordtypen en hun frequentie. Daartoe zijn zowel de woordfrequenties zelf als de cumulatieve frequenties (de som van de frequenties van alle woordtypen tot en met een bepaald rangnummer) bestudeerd. Onder andere aan de hand van een grafische weergave van deze beide gegevens werd vastgesteld dat logaritmische verhoudingen hier een belangrijke rol spelen. De logaritme van de frequentie, afgezet tegen de logaritme van het rangnummer, levert een dalende curve op die over een bepaald gebied (ongeveer tot rangnummer 100) een rechte lijn benadert, en voor het overige een lichte neerwaartse kromming heeft. Bij bestudering van deze curve is de positie waar de ene vorm in de andere overgaat vrij nauwkeurig vast te stellen. De woordtypen vóór dit punt worden in het vervolg aangeduid als de 'kop' van de frequentielijst. Bij de cumulatieve frequenties is sprake van een vrijwel rechte stijgende lijn tot ongeveer rangnummer 1000, daarna vertoont de curve ook een neerwaartse kromming.

Frequentieverdeling
Een frequentieverdeling is een lijst van alle in een corpus voorkomende woordfrequenties, met bij iedere frequentie het aantal woordtypen dat met die frequentie in het corpus voorkomt. Deze lijst wordt gewoonlijk op frequentie gesorteerd, de meest voorkomende frequentie (meestal is dat 1) bovenaan. Wanneer de frequentiegegevens op deze manier worden gerangschikt kan een indruk worden verkregen van de *structuur* van het woordgebruik. Bij een omvangrijk corpus is echter een probleem dat de frequentieverde

ling ook buitengewoon omvangrijk en daardoor onoverzichtelijk wordt. In het corpus wetgevingsteksten komen bijvoorbeeld bijna 2000 verschillende frequenties voor. Daarom is in dit geval gebruik gemaakt van *geclassificeerde* frequentieverdelingen, waarbij voor elk van de corpora de verschillende frequenties zijn ondergebracht in rond de 60 frequentieklassen. Omdat anders de verdeling zeer scheef zou worden (de lage frequenties liggen dicht bij elkaar, tussen de hoge bestaat een aanzienlijke ruimte) is gebruik gemaakt van frequentieklassen die *logaritmisch* (dat wil zeggen, door toename van een macht bij een constant grondtal) in omvang ('breedte') toenemen. Op die manier ontstaat een verdeling waarbij het aantal verschillende frequenties in de middelste klassen het hoogst is, terwijl de klassen daarvoor en daarna een gelijkmatig toenemend c.q. afnemend aantal bevatten. Voor vergelijkingsdoeleinden zijn met name interessant de aantallen woordtypen en -tokens per frequentieklasse, aangezien deze aantallen informatie geven over het patroon volgens welk de woorden in het corpus worden gebruikt. Zo blijken bij de juridische corpora zowel meer woordtokens als woordtypen voor te komen in de *hoogste* frequentieklassen, en blijken bij het corpus algemeen Nederlands de aantallen tokens en typen juist hoger te zijn in de *laagste* klassen.

Token/type ratio's
Wanneer de verhouding tussen aantallen woordtokens en typen in een tekstcorpus wordt bestudeerd, blijkt al snel dat deze verhouding in belangrijke mate samenhangt met de omvang van een corpus. Wanneer uit een corpus namelijk steekproeven worden genomen van steeds toenemende omvang blijkt de token/type ratio in die steekproeven ook voortdurend te blijven stijgen. Deze ratio vormt daarom op zich niet zo'n geschikt middel om corpora te vergelijken, tenzij die corpora exact even groot zijn. Dat ligt anders voor een verhouding die aan de hand van de frequentieverdeling kan worden berekend, de zogenaamde 'karakteristiek *K*' van Yule en Herdan. Deze karakteristiek, die een indicatie is voor de gemiddelde herhalingsfrequentie van woordtypen en daarmee, althans volgens een der opstellers, ook voor de grootte van de woordenschat in een corpus, blijkt wel in voldoende mate stabiel te zijn bij steekproeven van verschillende omvang uit een corpus. Datzelfde geldt eveneens voor een ander verband tussen aantallen woordtokens en woordtypen (gedefinieerd door Erikstad), waarbij het aantal woordtypen in een corpus gelijk wordt gesteld aan een macht *C* van het aantal woordtokens, vermenigvuldigd met een constante *R*. Door gebruik te maken van enkele bekende waarden voor de aantallen tokens en typen in verschillende steekproeven kan met behulp van regressie-analyse de waarde van *C* en *R* worden vastgesteld.

Woordlengtes
De lengte van woordtypen en -tokens in een corpus is een ander kwantitatief gegeven aan de hand waarvan mogelijk vergelijking met andere corpora kan plaatsvinden. Daarbij kan gedacht worden aan de *verdeling* van woordlengtes (hoeveel typen en tokens hebben een bepaalde lengte), maar ook aan het verband tussen het *rangnummer* van woordtypen en hun lengte (veelgebruikte typen blijken in het algemeen korter te zijn dan zeldzame). Zoals blijkt kunnen woordlengte-verdelingen voor de meest voorkomende lengtes (ongeveer tot 15 karakters) het beste worden weergegeven met een normale schaalverdeling. Voor de minder vaak voorkomende lengtes (de corpora bevatten typen tot 58 karakters lang) kan het beste gebruik worden gemaakt van een logaritmische schaalverdeling voor de aantallen typen. Het verband tussen rangnummer en lengte kan ook op een aantal verschillende manieren worden weergegeven. Genoemd kunnen daarvan worden het

indelen van de woordtypen in klassen bestaande uit een gelijk aantal rangnummers waarna voor iedere klasse de gemiddelde woordlengte wordt berekend, en het berekenen van een voortschrijdend gemiddelde over alle typen, op volgorde van rangnummer. De laatste methode blijkt, in ieder geval voor de eerste 5000 rangnummers in een corpus, het duidelijkste resultaat op te leveren.

Het vergelijken van de corpora
Aan de hand van de hierboven genoemde aspecten kunnen al globale vergelijkingen tussen de corpora worden gemaakt, waardoor al een indruk van de overeenkomsten en verschillen ontstaat. In een aantal gevallen blijkt het echter zo te zijn dat het verschil in omvang tussen de corpora van invloed is op het resultaat van deze vergelijkingen. Daarom is uit de twee grootste corpora (dat zijn de beide juridische) een steekproef getrokken ter grootte van het kleinste corpus met algemeen Nederlandse teksten. Bij onderzoek van deze steekproeven blijkt dat zij in grote lijnen dezelfde kenmerken hebben als de volledige corpora. Bij het in detail vergelijken van de verschillende aspecten van het woordgebruik is van deze gereduceerde corpora gebruik gemaakt, die dus allemaal ongeveer hetzelfde aantal woord*tokens* bevatten. De twee juridische corpora bevatten daarnaast ook ongeveer hetzelfde aantal woord*typen*, het corpus algemeen Nederlands bevat er echter bijna twee maal zo veel.

De belangrijkste overeenkomsten en verschillen tussen de corpora zijn als volgt samen te vatten:

- Uit *cumulatieve frequentiegegevens* blijkt dat bij de twee juridische corpora een groter aantal woordtokens vertegenwoordigd is in het bovenste gedeelte van de frequentielijst. De woordtypen met de hoogste frequenties hebben in deze corpora dus gemiddeld een *hogere* frequentie dan in het corpus algemeen Nederlands.

- Uit analyse van de (normale, niet cumulatieve) *frequenties* blijkt dat aan de hand daarvan in de frequentielijst een tweeledige 'kop' van wat genoemd kunnen worden de 'zeer vaak' en de 'vaak' voorkomende woordtypen kan worden vastgesteld. Het gedeelte met de *zeer vaak* voorkomende woordtypen lijkt daarbij korter te zijn bij de twee juridische corpora dan bij het corpus algemeen Nederlands. Dat zou betekenen dat in het onderzochte juridische tekstmateriaal minder typen zeer vaak voorkomen. Blijkens de cumulatieve frequentieverdeling zouden de meeste van deze typen daarnaast ook nog eens een hogere frequentie hebben.

- Uit de *frequentieverdeling* kan onder andere worden opgemaakt dat de 'extra' woordtypen die het corpus algemeen Nederlands bevat vooral te vinden zijn aan de 'staart' van de frequentielijst; ze hebben dus voor het merendeel een lage frequentie.

- Geconstateerd wordt dat bij elk van de corpora de zogenaamde karakteristiek *K* een andere waarde heeft. Dit is waarschijnlijk het gevolg van verschillen op het niveau van de *frequentieverdeling*. Er blijkt namelijk een verband te bestaan tussen de waarde van *K* en de standaarddeviatie van de verschillende frequenties die in het corpus voorkomen.

- Aan de hand van de door Erikstad gedefinieerde verhouding tussen de aantallen tokens en typen in een corpus kan nauwkeurig worden vastgesteld hoe de verschillen tussen de token/type ratio's van de corpora precies tot stand komen, en wat het verband tussen deze ratio's en de corpusgrootte is. De constanten *R* en *C* kunnen worden gebruikt om dit aspect van de structuur van het woordgebruik voor een corpus vast te leggen.

- Met betrekking tot de woordlengte-verdelingen van de corpora kan worden geconstateerd dat de juridische corpora een aanmerkelijk groter aantal woordtypen telt in de lagere lengteklassen (met name in de klassen van 4 - 6 karakters). Dit blijkt in belangrijke mate toe te schrijven te zijn aan het grotere aantal *getallen* dat in de juridische corpora voorkomt.

- Met betrekking tot de spreiding van de woordtypen over de documenten waaruit de corpora zijn opgebouwd wordt slechts met enige reserve een conclusie getrokken, aangezien niet precies duidelijk is of de gemiddelde documentgrootte (deze loopt nogal uiteen in de drie corpora) op dit gegeven invloed heeft. Het lijkt er echter op dat woordtypen vooral in het corpus algemeen Nederlands gemiddeld een sterkere spreiding vertonen (dat wil zeggen, gemiddeld in meer documenten voorkomen).

In het laatste hoofdstuk van dit proefschrift wordt tenslotte nog een onderwerp besproken dat ook nauw verband houdt met de mate van spreiding van woordtypen in een corpus. Beproefd wordt een model met behulp waarvan de *similariteit* (mate van overeenkomst) tussen paren documenten kan worden vastgesteld. Met behulp van dit gegeven wordt vervolgens getracht in een van de corpora (het corpus wetgevingsteksten) *clusters van documenten* te vormen aan de hand van overeenkomsten en verschillen in het woordgebruik. Hoewel geen van de toegepaste clustermethoden een optimaal resultaat blijkt op te leveren, wordt wel een groot aantal relevante clusters gevormd, die elk benoemd kunnen worden aan de hand van de documenten waaruit ze zijn opgebouwd.

Conclusies

De structuur van het woordgebruik blijkt in de juridische corpora op bepaalde punten af te wijken van die in het corpus algemeen Nederlands. Onder andere blijkt het zo te zijn dat in de juridische corpora *minder woordtypen zeer vaak worden gebruikt*, waarbij elk van die woordtypen tevens gemiddeld een *hogere frequentie* heeft. De juridische corpora vertonen ook *onderling* verschillen voor wat betreft de structuur van het woordgebruik. De verschillen tussen de corpora laten zich met behulp van bepaalde linguïstische constanten kwantificeren. Met name de karakteristiek *K* van Yule en Herdan en de verhouding $V = R \, . \, N^{C}$ van Erikstad blijken daarbij praktisch ongevoelig te zijn voor de omvang van een corpus.

Voor wat betreft de woordlengtes kan worden geconcludeerd dat alleen het corpus wetgevingsteksten een hoger percentage lange woorden bevat dan de andere corpora. Het

corpus jurisprudentieteksten wijkt op dat punt nauwelijks af van het corpus algemeen Nederlands.

Wanneer de meest voorkomende woordtypen uit de corpora worden vergeleken vinden we ook een aantal verschillen. Zo hebben in de juridische corpora bepaalde getallen een hoge frequentie. Daarnaast vallen in de juridische corpora enkele woorden op die op het eerste gezicht niet specifiek juridisch zijn, maar wel een aanmerkelijk hogere frequentie hebben dan in het corpus algemeen Nederlands, zoals de woorden BEDOELD, INDIEN, TER en WAARVAN. Bepaalde persoonlijke voornaamwoorden komen in de juridische corpora juist minder vaak voor.

Het onderzoek dat in het laatste hoofdstuk is beschreven, naar de similariteit van documenten in een juridisch corpus, heeft een exploratief karakter. Een van de conclusies hieruit is dat het aan de hand van het woordgebruik mogelijk is clusters van documenten uit een corpus te vormen, waarbij de documenten niet alleen overeenkomsten vertonen voor wat betreft hun vorm, maar ook op semantisch niveau.

SUMMARY

The research project described in this thesis is concerned with the quantitative aspects of legal language. Among other things, the accurate description of these aspects is a necessary precondition for the future development of a new generation of legal information systems, such as systems which will make it possible to retrieve data from legal databanks using legal 'concepts'. In order to gain insight into the specific properties of the word use in certain legal texts, these texts have been compared to general text material in the Dutch language. The legal texts concerned are statute law and case law texts. Using these texts, two separate legal corpora have been built for this research project. Alongside these, a corpus with general Dutch texts has been developed to carry out comparison.

Before the word use in the corpora is outlined, attention is paid to research in the field of 'law and language' which has been carried out in the past. From this review, it appears that the quantitative aspects of legal language have rarely been the subject of research. As (quantitative) linguistics is not a familiar field for most lawyers, a chapter is devoted to this subject. In this chapter, among other matters, the position of quantitative linguistics in the field of linguistics as a whole is specified. Furthermore, a survey is given of the quantitative methods and techniques which play an important role in this research project and attention is paid to the terms 'sub-language' and 'language for specific purposes' which are relevant to this research project.

After this preliminary part, the three corpora which are central to this study are described. These corpora are:

- a corpus which contains the texts of every piece of statute law which was in force at the moment the corpus was created. This material comes from the electronic legal databank NLEX. Only the original statute law texts have been used, without additions such as index words or annotations.

- a corpus which contains a broad selection of case law texts. These are all cases which were published in the magazine 'Nederlandse Jurisprudentie' (Dutch Case Law) between 1965 and 1989. Of these cases again only the text of the original decision has been used, without any summaries or annotations.

- a corpus with Dutch text material of a general nature, intended for comparative purposes. This corpus contains, inter alia, texts of novels, short stories, text books and articles, scientific as well as non-scientific, on various subjects.

The corpora are split into separate words ('words' in this case means: rows of characters; another term for these is *word tokens*). The size of the three corpora rather varies. The corpus containing case law texts is the largest (more than 40 million word tokens), the corpus containing statute law texts consists of more than 12 million word tokens and the corpus with

general Dutch texts is the smallest with 4.5 million word tokens. A list is compiled for each corpus containing all the different words (all *word types*) which are present in it, with the *frequency* of those words in the corpus, and the *number of documents* in which they are found. This list is sorted by word frequency, the most frequently used word being at the top, after which a *rank number* can be given to every word: 1 for the most common word, 2 for the next most common word, etc. This *frequency list* is basic to most of the analyses that follow.

After the text material involved has been described, each of the three corpora are first examined individually. The goal here is to characterise the word use in these corpora systematically, using quantitative methods. Aspects which come up for discussion primarily are word frequencies, the frequency distribution, the ratio between the number of word tokens and the number of word types, and word lengths. To analyze these aspects, extensive use is made of graphs. The accent is on distinguishing characteristics which could be specific to a certain corpus, and therefore could play a role when *comparing* the corpora. Each of the aspects mentioned will be described briefly below.

Word frequencies
While analyzing the word frequencies which are present in a corpus, most attention has been focused on the relationship between the rank number of word types and their frequency. For that purpose the word frequencies themselves as well as the cumulative frequencies (the sum of the frequencies of all word types up to a certain rank number) have been studied. By studying, among other things, a graphical representation of both these data, it was found that logarithmic ratios play an important role here. The logarithm of the word frequency plotted against the logarithm of the rank number yields a downward curve which approaches a straight line in a certain area (about up to rank number 100) and beyond that point has a light downward slope. By studying this curve, the point where the sloping part starts can be determined quite accurately. All word types before this point from now on will be called the 'head' of the frequency list. It has been found that cumulative frequencies form a more or less straight line up to about rank number 1000, thereafter this curve shows a downward slope as well.

Frequency distribution
A frequency distribution is a list of all word frequencies which are present in a corpus, with each frequency being coupled with the number of word types that have that frequency in the corpus. This list is usually sorted by frequency, the most common frequency (which most of the times is 1) on top. When the frequency data are laid out in this way, an impression can be obtained of the *structure* of the word use. However, for a large corpus a problem is that the frequency distribution can be very bulky, and therefore difficult to survey. In the statute law corpus, for example, almost 2000 different frequencies are present. Therefore, in this case, *classified* frequency distributions are used for each corpus, in which the different frequencies are put into around 60 frequency classes. As otherwise the distribution would be very skewed (low frequencies are close together, but between high frequencies there are large intervals) frequency classes are used which grow larger ('wider') *logarithmically* (i.e. by the increase of the power to which a constant is raised). This way, a distribution is formed in which the number of different frequencies is the highest in the middle classes, while the preceding and succeeding classes contain a steadily increasing and decreasing number respectively. In

comparing corpora, the numbers of word types and word tokens in every frequency class are of especial interest, because these numbers provide information about the pattern according to which the words are used in the corpus. It is found, for instance, that in the legal corpora the number of word tokens as well as word types is higher in the *upper* frequency classes, whereas in the general Dutch corpus the number of tokens and types is higher in the *lower* classes.

Token/type ratios
When the relationship between the number of word tokens and word types in a text corpus is studied, it quickly becomes apparent that this relation is connected to the size of the corpus to a large extent. When random samples of increasing size are taken from a corpus, the token/type ratio proves to be increasing continuously as well. Therefore, this ratio would not appear to be a very suitable means to compare corpora, unless these corpora are of equal size. This is not the case for a relationship which can be calculated using the frequency distribution, the so-called 'characteristic *K*' of Yule and Herdan. This characteristic is an indication of the average frequency of the repetition of wordtypes. According to at least Yule this characteristic is therefore also an indication of the size of the lexicon in a corpus. It appears to be sufficiently stable in samples of different size taken from a corpus. The same goes for another relationship between the number of word tokens and word types (defined by Erikstad), in which the number of word types in a corpus is considered equal to a power *C* of the number of word tokens, multiplied by a constant *R*. By using some known values for the numbers of tokens and types in different samples, the value of *C* and *R* can be calculated by means of regression analysis.

Word lengths
The length of word types and tokens in a corpus is another quantitative measure by means of which corpora could be compared to each other. For this, the *distribution* of word lengths can be considered (how many types and tokens share a certain length), but also the relationship between the *rank number* of word types and their length (types which are used frequently in general prove to be shorter than rare types). It has been found that word length distributions for the most common lengths (up to about 15 characters) are best depicted in a graph with normal scale. For less common lengths (the corpora contain types of up to 58 characters long) a logarithmic scale for the number of types is preferable. The relationship between rank and length can also be represented in a number of ways. Examples are the distribution of word types among a number of classes with an equal number of rank numbers each, after which the average length is calculated for every class, and the calculation of the moving average of the length of all types, in order of their rank number. This last method proves to yield the most comprehensible results, at least for the first 5000 rank numbers in every corpus.

Comparing corpora
Evaluating the aspects mentioned above, global comparisons between the corpora can already be made, which makes it possible to give an impression of similarities and dissimilarities. In some cases, however, the differences in size between the corpora prove to be of influence on the results of these comparisons. Therefore, samples have been taken from the two largest corpora (these are the two legal corpora), of the same size as the smallest corpus with general Dutch texts. When these samples are examined, they are found to have, in general, the same characteristics as the unabridged corpora. For detailed comparisons of the different aspects of word use these reduced corpora have been used, which all contain about the same number of

word *tokens*. Apart from that, the two legal corpora both contain about the same number of word *types* as well, whereas the general Dutch corpus contains about twice as many types.

The most important similarities and dissimilarities between the corpora can be summarized as follows:

- From *cumulative frequency data* it has been found that in the two legal corpora a larger number of word tokens is represented in the upper part of the frequency list. This means that the word types with the highest frequencies have a higher frequency on average in these corpora than in the general Dutch corpus.

- From analysis of (normal, non cumulative) *frequencies* it follows that by means of these frequencies a two part 'head' of the frequency list can be recognized, containing what could be called 'very common' and 'common' word types. The part which contains the *very common* word types seems to be shorter in the two legal corpora than in the general Dutch corpus. This would indicate that in the legal texts which were examined, less word types are very common. From the cumulative frequency distribution it follows that most of these types have a higher frequency as well.

- From the *frequency distribution* it can be seen, among other things, that the 'extra' word types the general Dutch corpus contains can be found mainly at the 'tail' of the frequency list. Therefore, most of these types have low frequencies.

- The so-called characteristic *K* can be noticed to have a different value in each of the corpora. This probably is caused by differences at the level of the *frequency distribution*. This is because there is a connection between the value of *K* and the standard deviation of the different frequencies in a corpus.

- By means of the ratio between the number of tokens and types in a corpus as defined by Erikstad, the causes of differences between token/type ratios in the three corpora and the connection between these ratios and the size of the corpora can be assessed accurately. The constants *R* and *C* can be used to record this aspect of the structure of word use in a corpus.

- With respect to the word length distributions of the corpora, it can be seen that the legal corpora contain a considerably larger number of word types in the lower length classes (especially in the classes from 4 to 6 characters). It is found that to a large extend this is caused by the larger number of *numbers* which is present in the legal corpora.

- With respect to the distribution of word types over the documents in each of the corpora, conclusions can only be drawn with some reservations. This is because it is not quite clear if the average size of documents (which varies quite a lot over the corpora) is of any influence on this measurement. It seems, however, that especially in the general Dutch corpus word types are spread out more (on average these types are present in a larger number of documents).

Finally, in the last chapter of this thesis, a subject is discussed which also has a strong

connection with the distribution of word types in a corpus. A model is examined which can serve to determine the *similarity* (the measure of resemblance) of a pair of documents. By means of this measure an attempts is made to form *clusters of documents* in one of the corpora (the statute law corpus), using similarities and dissimilarities in word use. Although none of the clustering methods which have been applied has been found to yield optimal results, a large number of relevant clusters have been formed by each of these methods, all of which can be designated by means of the documents they contain.

Conclusions

The structure of word use in the legal corpora proves to differ at certain points from that in the general Dutch corpus. Among other things it was found that in the legal corpora *fewer word types are very common*, whereas each of these word types has on average a *higher frequency* as well. The legal corpora show some mutual differences as far as the structure of word use is concerned. The differences between the corpora can be verified by means of certain linguistic constants. Of these, especially the characteristic K of Yule and Herdan and the ratio $V = R \,.\, N^{C}$ of Erikstad seem to be more or less insensitive to the size of a corpus.

With respect to the word lengths, it can be concluded that only the statute law corpus contains a higher percentage of long words than the other corpora. The case law corpus in this respect is virtually the same as the general Dutch corpus.

When the most common word types in the corpora are compared, differences are found as well. Certain numbers have a much higher frequency in the legal corpora. Furthermore, in the legal corpora some words are encountered which, at least at first sight, are not specifically legal, but do have a considerably higher frequency than in the general Dutch corpus, such as the words BEDOELD (intended), INDIEN (in case), TER (to) and WAARVAN (from which). However, certain personal pronouns are found to be less common in the legal corpora.

The research described in the last chapter, on the similarity of documents in a legal corpus, is of an exploratory nature. One of the conclusions which can be drawn from it is that clusters of documents can be formed from a corpus based on word use. Documents in these clusters are found not only to have similarities in form, but also at the semantic level.

LITERATUURLIJST

Aarts en Meijs 1988 - J. Aarts en W. Meijs, "Corpustaalkunde". In: *Spektator*, jaargang 18 (1988-1989), nr. 1, p. 6-23. Foris publications, Dordrecht 1988.

Austin 1970 - W.B. Austin, "The authorship of certain Renaissance English pamphlets: an informal account of work in progress". In: F.M. Burelbach Jr. (ed.), *Proceedings: Computer Applications to Problems in the Humanities*. SUNY College at Brockport, Brockport, New York 1970.

Baayen 1989 - R.H. Baayen, *A corpus-based approach to morphological productivity*. Proefschrift, Centrum voor Wiskunde en Informatica (Vrije Universiteit), Amsterdam 1989.

Bailey 1969 - R.W. Bailey, "Statistics and Style: a historical survey". In: Doležel and Bailey (1969, p. 217-236).

Batagelj and Bren 1993 - V. Batagelj and M. Bren, *Comparing Similarity Measures*. University of Ljubljana, Ljubljana 1993.

Battus 1983 - H. Battus, *Rekenen op taal*. Querido, Amsterdam 1983.

Bax 1986 - C.J. Bax, *De geslotenheid van het recht*. Acco, Leuven 1986.

Bennett 1969 - P.E. Bennett, "The Statistical Measurement of a Stylistic Trait in 'Julius Caesar' and 'As You Like It'". In: Doležel and Bailey (1969, p. 29-41).

Van Berckel *et al.* 1963 - J.A.Th.M. van Berckel, H. Brandt Corstius, R.J. Mokken en A. van Wijngaarden, *Formal Properties of Newspaper Dutch*. Amsterdam 1963.

Van den Bergh en Broekman 1979 - G.C.J.J. van den Bergh en J.M. Broekman, *Recht en Taal*. Preadvies voor de jaarlijkse algemene vergadering van de Koninklijke Notariële Broederschap. Kluwer, Deventer 1979.

Biber and Finegan 1986 - D. Biber and E. Finegan, "An Initial Typology of English Text Types". In: J. Aarts en W. Meijs, *Corpus Linguistics II - New Studies in the Analysis and Exploitation of Computer Corpora*, p. 19-46. Rodopi, Amsterdam 1986.

Uit den Boogaart 1975 - P.C. Uit den Boogaart (red.), *Woordfrequenties in geschreven en gesproken Nederlands*. Oosthoek, Scheltema en Holkema, Utrecht 1975.

Brandt Corstius 1974 - H. Brandt Corstius, *Algebraïsche taalkunde*. Oosthoek, Utrecht 1974.

Brandt Corstius 1978 - H. Brandt Corstius, *Computer-taalkunde*. Coutinho, Muidenberg 1978.

Butler 1985 - C.S. Butler, *Computers in Linguistics*. Basil Blackwell, Oxford 1985.

Carroll 1938 - J.B. Carroll, "Diversity of Vocabulary and the Harmonic Series Law of Word-Frequency Distribution". In: *The Psychological Record*, II (1938), p. 379-386.

Carroll 1967 - J.B. Carroll, "On sampling from a lognormal model of wordfrequency distribution". In: Kučera and Francis (1967, p. 406-424).

Carroll 1970 - J.B. Carroll, "An alternative to Juilland's Usage Coefficient for Lexical Frequencies, and a Proposal for a Standard Frequency Index (SFI)". In: *Computer Studies in the Humanities and Verbal Behaviour*, 3, 1970, p. 61-65.

Cherubini 1975 - W. Cherubini, "The 'Goldenness' of Sidney's 'Astrophel and Stella': test of a quantitative-stylistics routine". In: *Language and Style*, vol. 8, p. 47-59.

Cluett 1976 - R. Cluett, *Prose Style and Critical Reading*. Columbia University Teachers' College Press, New York 1976.

Condon 1928 - E.U. Condon, "Statistics of Vocabulary". In: *Science*, LXVII, no. 1733 (1928), p. 300.

De la Court 1937 - J.F.H.A. de la Court, *De meest voorkomende woorden en woordcombinaties in het Nederlandsch*. Batavia 1937

Cross and Kuo 1988 - G.R. Cross and S. Kuo, *Algorithms for Word Similarity in a Trademark Retrieval System*. Washington State University, Washington 1988.

Davis 1986 - D. Davis, "Semantic Analysis in Legal Text Information Retrieval". In: Martino, A.A. (ed.), *Automated Analysis of Legal Text: Logic, Informatics, Law*, p. 473-481. North Holland, Amsterdam 1986.

Deville 1989 - G. Deville, *Modelization of task-oriented utterances in a man-machine dialogue system*. Universitaire Instelling, Antwerpen 1989.

Dik en Kooij 1988 - S.C. Dik en J.G. Kooij, *Algemene Taalwetenschap*. Vijfde druk, Het Spectrum (Aula), Utrecht 1988.

Doležel and Bailey 1969 - L. Doležel and R.W. Bailey (eds.), *Statistics and Style*. Elsevier, New York 1969.

Van Eemeren *et al.* 1991 - F.H. van Eemeren, E.T. Feteris, R. Grootendorst, T. van Haaften, W. den Harder, H. Kloosterhuis, T. Kruiger en J. Plug, *Argumenteren voor juristen*. Wolters-Noordhoff, Groningen 1991.

Ehrenberg 1982 - A.S.C. Ehrenberg, *A Primer in Data Reduction*. John Wiley & Sons, Chichester 1982.

Ellegård 1962a - A. Ellegård, *Who was Junius?* Almqvist and Wiksell, Stockholm 1962.

Ellegård 1962b - A. Ellegård, *A Statistical Method for Determining Authorship: The Junius Letters 1769-1772*. Gothenburg Studies of English no. 13, University of Gothenburg Press, Gothenburg 1962.

Erikstad 1980 - O.M. Erikstad, "Appropriate Document Units for Text Retrieval Systems". In: J. Bing and K.S. Selmer, *A Decade of Computers and Law*, p. 220-238. Universitetsforlaget, Oslo 1980.

Estoup 1916 - J.B. Estoup, *Gammes stenographiques*. 4th edition, Paris 1916.

Guiraud 1954 - P. Guiraud, *Les Caractères Statistiques du Vocabulaire*. Presses Universitaires de France, Paris 1954.

Guiraud 1959 - P. Guiraud, *Problèmes et Méthodes de la Statistique Linguistique*. Reidel publishing company, Dordrecht 1959.

Guiter and Arapov 1982 - H. Guiter and M.V. Arapov (eds.), *Studies on Zipf's Law*. Brockmeyer, Bochum 1982.

Hempel 1967 - C.G. Hempel, "Fundamentals of Concept Formation in Empirical Science". In: *International Encyclopedia of Unified Science*, vol. II, 1967, no. 7, p. 1-62.

Hendrickx 1994 - R.L.L. Hendrickx, *Over wiskunde en werkelijkheid*. Proefschrift Katholieke Universiteit Brabant, Tilburg 1994.

Herdan 1960 - G. Herdan, *Type-Token Mathematics*. Mouton, 's Gravenhage 1960.

Herdan 1964 - G. Herdan, *Quantitative linguistics*. Buttersworths, London 1964.

Herdan 1966 - G. Herdan, *The advanced theory of language as choice and chance*. Springer-Verlag, Berlin 1966.

Hofhuis 1988 - H.F.M. Hofhuis, "De taal van de rechter". In: *Meesterlijke taal*, bundel uitgegeven ter gelegenheid van het achtste lustrum van de Jonge Balie Utrecht, p. 51-64. Tjeenk Willink, Zwolle 1988.

Juilland and Chang-Rodriguez 1964 - A.G. Juilland and E. Chang-Rodriguez, *Frequency dictionary of Spanish words*. The Hague 1964.

Juilland *et al.* 1965 - A.G. Juilland, P.M.H. Edwards and I. Juilland, *Frequency dictionary of Rumanian words*. The Hague 1965.

Juilland *et al.* 1970 - A.G. Juilland, D. Brodin and C. Davidowitch, *Frequency dictionary of French words*. The Hague 1970.

Juilland and Traversa 1973 - A.G. Juilland and V.P. Traversa, *Frequency dictionary of Italian words*, The Hague 1973.

Kenny 1977 - A. Kenny, "A stylometric study of Aristotle's Ethics". In: S. Lusignan and J.S. North (eds.), *Computing in the Humanities*. University of Waterloo Press, Waterloo, Ontario 1977.

Kenny 1978 - A. Kenny, *The Aristotelian Ethics*. Oxford University Press, Oxford 1978.

Kerkmeester 1989 - H.O. Kerkmeester, *Recht en Speltheorie*. Vermande, Lelystad 1989.

Knottenbelt en Torringa 1986 - B. Knottenbelt en R.A. Torringa, *Inleiding in het Nederlandse Recht*, vierde druk, bewerkt door J.W.P. Verheugt. Gouda Quint, Arnhem 1986.

Kučera and Francis 1967 - H. Kučera and N.W. Francis, *Computational analysis of present-day American English*. Brown University Press, Providence 1967.

Lesk 1969 - M.E. Lesk, "Word-Word Associations in Document Retrieval Systems". In: *American Documentation*, Volume 20 (1969), p. 27-38. American Society for Information Science, Baltimore 1969.

Loevinger 1971 - L. Loevinger, "Jurimetrics, the next step forward". In: *Jurimetrics Journal*, september 1971, p. 3-41. American Bar Foundation, Chicago 1971.

Loth 1991 - M.A. Loth, *Recht en taal. Een kleine methodologie*. Gouda Quint, Arnhem 1991.

Linschoten 1963 - J. Linschoten, *De la Court's Frequentie-telling van Nederlandse woorden*. Utrecht 1963.

Maes 1991 - A.A. Maes, *Nominal Anaphors and the Coherence of Discourse*. Tilburg 1991.

Mandelbrot 1953 - B. Mandelbrot, "An Information Theory of the Statistical Structure of Language". In: *Proceedings of the Symposium on Applications of Communication Theory*, London, Sept. 1952. Butterworth, London 1953.

Mandelbrot 1962 - B. Mandelbrot, "On the Theory of Word Frequencies and on Related Markovian Models of Discourse". In: R. Jacobson (ed.), *Structure of Language and its Mathematical Aspects. Proceedings of Symposia in Applied Mathematics*, Vol. XII, p. 190-219. American Mathematical Society, Providence, Rhode Island 1962.

Martin 1968 - W. Martin, *De inhoud van krant en roman*. Antwerpen 1968.

Martin 1970 - W. Martin, *Analyse van een vocabularium met behulp van een computer*. AIMAV, Brussel 1970.

Martin 1973 - W. Martin, "Some quantitative aspects of a Dutch poem". In: A.J. Aitken, R.W. Bailey and N. Hamilton-Smith (eds.), *The Computer and Literary Studies*, p. 61-68. Edinburgh University Press, Edinburgh 1973.

Martin 1976 - W. Martin, "On the evolution of Word-length in Dutch". In: A. Jones and R.F. Churchhouse, *The Computer in Literary and Linguistic Studies*, p. 271 - 284. The University of Wales Press, Cardiff 1976.

Martin 1981 - W. Martin, "Kwantitatieve taalkunde". In: W. de Geest *et al.* (eds.), *Twintig facetten van de taalwetenschap*, p. 273-291. Acco, Leuven 1981.

Martin 1988 - W. Martin, "Corpora voor woordenboeken". In: *Corpusgebaseerde woordanalyse, jaarboek 1988*, p. 91-99. Vakgroep taalkunde, Vrije Universiteit, Amsterdam 1988.

Martin 1990 - W. Martin, "The Frequency Dictionary". In: F.J. Hausmann *et al.* (eds.), *Wörterbücher, Dictionaries, Dictionnaires*, vol. 2, p. 1314-1322. De Gruyter, Berlin/ New York 1990.

Martin en Ten Pas 1991 - W. Martin en E. ten Pas, "Subtaal en Lexicon". In: *Spektator*, jaargang 20 (1991), nr. 3/4, p. 361-375. Foris publications, Dordrecht 1991.

Meier 1967 - H. Meier, *Deutsche Sprachstatistik*. Hildesheim 1967.

Meijs 1991 - W. Meijs, "De empirische dimensie". In: *Spektator*, jaargang 20 (1991), nr. 3/4, p. 313-336. Foris publications, Dordrecht 1991.

Mendenhall 1887 - T.C. Mendenhall, "The Characteristic Curves of Composition". In: *Science*, IX, no. 214 (March, 1887), p. 237-249.

Milic 1967 - L.T. Milic, *A Quantitative Approach to the Style of Jonathan Swift*. Mouton, The Hague 1967.

Mosteller and Wallace 1964 - R. Mosteller and D.L. Wallace, *Inference and Disputed Authorship: 'The Federalist'*. Addison-Wesley, Reading, Mass. 1964.

De Mulder en Oskamp 1979 - R.V. De Mulder en A. Oskamp, *Het woordgebruik van het wetboek van Strafrecht en het wetboek van Strafvordering*. Rapport Erasmus Universiteit, Rotterdam 1979.

De Mulder 1984 - R.V. De Mulder, *Een model voor juridische informatica*. Vermande, Lelystad 1984.

De Mulder *et al.* 1993 - R.V. De Mulder, M.J. van den Hoven and C. Wildemast, "The concept of concept in 'conceptual legal information retrieval'", in: *8th BILETA Conference*

Pre-proceedings, p. 79-92. CTI Law Technology Centre, University of Warwick, Warwick 1993.

Muller 1969 - Ch. Muller, "Lexical Distribution reconsidered: the Waring-Herdan Formula". In: Doležel and Bailey (1969, p. 42-56).

Muller 1977 - Ch. Muller, *Principes et méthodes de statistique lexicale*. Hachette, Paris 1977.

Muller 1979 - Ch. Muller, "Du nouveau sur les distributions lexicales: la formule de Waring-Herdan". In: Ch. Muller, *Langue Française et Linguistique Quantitative*, p. 177-195. Slatkine, Genève 1979.

Van Naelten 1991 - M.A.M.G. Van Naelten, *Requiem voor een 'simplistisch' model; de 'Rank-size-wet van Zipf' als statistisch fata morgana*. Afscheidscollege, Katholieke Universiteit Nijmegen, Nijmegen 1991.

Neijt en Bakker 1988 - A. Neijt en D. Bakker, "Ter introductie", bij themanummer Computerlinguistiek. In: *Spektator*, jaargang 18 (1988-1989), nr. 1, p. 3-5. Foris publications, Dordrecht 1988.

Orlov 1982 - Ju.K. Orlov, "Ein model der Häufigkeitsstruktur des Vokabulars". In: Guiter and Arapov (1982, p. 154-233).

Van der Pot - Donner 1989 - Van der Pot - Donner, *Handboek van het Nederlandse staatsrecht*, bewerkt door Prof. Mr. L. Prakke. Twaalfde druk, Tjeenk Willink, Zwolle 1989.

Rapoport 1982 - A. Rapoport, "Zipf's Law Re-visited". In: Guiter and Arapov (1982, p. 1-28).

Van Reenen en Siertsema 1972 - P. Th. van Reenen en B. Siertsema, "De-stigmatisering". In: *Delikt en Delinkwent*, nr. 2 (1972), p. 399-411. Brill, Leiden 1972.

Van Reenen 1990 - P. Th. van Reenen, "De Hollandse Expansie, gebruiksfrequenties en het gebruik van drempelwaarden". In: *Corpusgebaseerde woordanalyse, jaarboek 1990*, p. 77-116. Vakgroep taalkunde, Vrije Universiteit, Amsterdam 1990.

Reinsma en Reinsma 1976 - M. Reinsma en R. Reinsma, "'De vrouw in wier lichaam zich eerstbedoeld leven ontwikkelt', of Zestig jaar Nederlandse rechtstaal". In: *Nederlands Juristenblad*, 1976, p. 857-872. Tjeenk Willink, Zwolle 1976

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

Rescher 1969 - N. Rescher, *Introduction to logic*. New York 1969.

Salton 1971a - G. Salton (ed), *The SMART retrieval system, experiments in automatic document processing*. Prentice-Hall, Englewood Cliffs N.J. 1971.

Salton 1971b - G. Salton, "Cluster Search Strategies and the Optimization of Retrieval Effectiveness". In: Salton (1971a, p. 223-242).

Sciarone 1977 - A.G. Sciarone, *Vocabolario fondamentale della lingua Italiana*. Bergamo 1977.

Smeaton and Van Rijsbergen 1981 - A.F. Smeaton and C.J. van Rijsbergen, "The Nearest Neighbour Problem in Information Retrieval". In: *Proceedings of the Fourth International Conference on Information Storage and Retrieval*, p. 83-87. ACM, New York 1981.

Štejnfel'dt 1963 - E.A. Štejnfel'dt, *Častotnyj slovar' sovremennogo russkogo literaturnogo jazyka* (Vertaling: *Frequentiewoordenboek van moderne Russische schrijftaal*). NII Pedagogiki ESSR, Tallin 1963.

Van Sterkenburg 1989 - P.G.J. van Sterkenburg, *Taal van het journaal: momentopname van hedendaags Nederlands*. SDU, 's Gravenhage 1989.

Stiles 1961 - H.E. Stiles, "The Association Factor in Information Retrieval". In: *Journal of the Association for Computing Machinery (ACM)*, Volume 8 (1961), nr 2, p. 271-279. ACM, New York 1961.

Thomson and Thompson 1915 - G.H. Thomson and J.R. Thompson, "Outlines of a Method for the Quantitative Analyis of Writing Vocabularies". In: *British Journal of Psychology*, VIII (1915), p. 52-69.

Werkgroep 1980 - Werkgroep, ingesteld bij besluit van de Minister van Justitie d.d. 28 maart 1979, *Een nationaal geautomatiseerd systeem van wetgevings- en rechtspraakinformatie*. Interim-advies, 's Gravenhage, april 1980.

Wildemast en De Mulder 1992 - R.V. De Mulder en C.A.M. Wildemast, "Some design considerations for a conceptual legal information retrieval system". In: C.A.F.M. Grütters *et al., Legal knowledge based systems, Information Technology and Law, proceedings of the fifth conference of the Foundation for Legal Knowledge Systems (JURIX)*, p. 81-92. Vermande, Lelystad 1992.

Willett 1981 - P. Willett, "A Fast Procedure for the Calculation of Similarity Coefficients in Automatic Classification". In: *Information Processing and Management*, Volume 17, p. 53-60. Pergamon Press, Oxford 1981.

Witteveen 1988 - W.J. Witteveen, "De taal van de wetgever". In: *Meesterlijke taal*, bundel uitgegeven ter gelegenheid van het achtste lustrum van de Jonge Balie Utrecht, p. 3-18. Tjeenk Willink, Zwolle 1988.

Yule 1944 - G.U. Yule, *The Statistical Study of Literary Vocabulary*. Cambridge University Press, Cambridge 1944.

Zipf 1935 - G.K. Zipf, *The Psycho-Biology of Language*. Houghton Mifflin, Boston 1935.

Zipf 1949 - G.K. Zipf, *Human Behavior and the Principle of the Least Effort. An Introduction to Human Ecology*. Hafner, New York 1949.

LIJST VAN BIJLAGEN

1. Lijst van titels en auteurs van de documenten die samen het corpus algemeen Nederlands vormen 286

2. Lijst van alle woorden van meer dan 30 lettertekens, per corpus 296

3. Eerste deel van de frequentielijst (eerste 300 rangnummers) van elk van de drie corpora 303

4. Vergelijkingslijst van alle woordtypen uit de 'kop' van de frequentielijsten van de drie corpora 310

BIJLAGE 1 - INHOUD CORPUS ALGEMEEN NEDERLANDS

Doc nr.	Aantal tokens	Auteur(s)	Titel(s) Indien cursief: titel hoofdwerk, ingedeeld in afzonderlijke documenten
	-	Anoniem	*Sociobiologie ter discussie. Evolutionaire wortels van menselijk gedrag?*
1	12745	Hooff, J. van	deeltitels: 1 Inleiding in de evolutionaire verklaring van sociaal gedrag 2 Repliek 3 De kiem van censuur 4 Een hardnekkig schijnprobleem
2	9101	Huis in 't Veld, G.	deeltitels: 1 Het sociobiologiedebat is nog niet gesloten 2 Sociobiologie: een maatschappelijk gevaar of academische borrelpraat? 3 Repliek 4 Rechts laten liggen
3	7082	Wind, J.	deeltitels: 1 De genen met egoïsme in degenen zonder egoïsme? 2 Repliek 3 Problemen met de sociobiologie? Maak er gewoon een karikatuur van 4 Overlevingswaarden en morele waarden
4	10050	Vroon, P.A.	deeltitels: 1 Extreem reductionisme 2 Problemen? Voor alles is gewoon een gen 3 Repliek
5	8475	Hartog, H. den	deeltitels: 1 Sociobiologie: Een oude hypothese verpakt als nieuwe synthese 2 Repliek 3 De mens. Een uitzonderlijk deel van de natuur
6	8750	Musschenga, A.W.	deeltitels: 1 Is altruïsme een verdwijnende eigenschap? 2 De bijdrage van de sociobiologie aan moraalwetenschap en ethiek 3 Repliek
7	9368	Falger, V.S.E.	deeltitels: 1 Wetenschap en politiek 2 Ongeloof in het neo-darwinisme als ideologie 3 Biopoliticologie: Sociobiologie in het perspectief van de wetenschap der politiek 4 Repliek
8	11265	Waal, F. de	deeltitels: 1 Inleiding 2 Het is niet òf cultuur òf erfelijkheid 3 Het menselijk voetstuk; gedragsovereenkomsten tussen de mens en andere primaten 4 Repliek 5 Waarom zouden we de harmonisatie uitstellen? 6 Nawoord
9	5152	Reinders, L.	deeltitels: 1 Drie problemen met de reconstructie van moraal 2 Overeenkomsten in argumentatiestructuur 3 Zijn menswetenschappen en biologie aan een gelukkig huwelijk toe? 4 Repliek
10	1588	Rijswijk, Maria J. van	Over vrouwen en mannen
11	1739	Kalma, A.	Evolutie en de menswetenschappen: zin en onzin
12	1848	Schuyt, C.J.M.	Sociobiologie en sociale wetenschappen: een kritische bijdrage
13	1371	Rhijn, J. van	De speltheorie is wel degelijk toetsbaar
	-	Anoniem	*Mens: man - vrouw - wezen. Een bezinning op d androgyne aard van de mens.*
14	13623	Douven, K.	deeltitels: 1 Kosmische oorsprong 2 Mystieke eenheidsbeleving
15	15687	Buuren, Hanneke van	Man-vrouw in de kanteling der tijden
16	15524	Elbers, Jita	Verankering in het hogere zelf
17	7631	Heijboer, Magnolia	Androgynie in het fysieke lichaam
18	8856	Kuijer, K. de	Wording van de nieuwe mens
	-	Anoniem	*Wetenschap en rekenschap.* *1880-1980: een eeuw wetenschapsbeoefening en wetenschapsbeschouwing aan de Vrije universiteit.*
19	15662	Wieringa, W.J.	De Vrije Universiteit als bijzondere instelling
20	23316	Veenhof, J.	Honderd jaar theologie aan de Vrije Universiteit
21	23578	Diepenhorst, I.A.	De juridische faculteit (1880-1980)
22	12305	Kuilman, M.	Geschiedenis van de psychiatrie aan de Vrije Universiteit

Doc nr.	Aantal tokens	Auteur(s)	Titel(s) Indien cursief: titel hoofdwerk, ingedeeld in afzonderlijke documenten
23	17503	Langevoort, H.L.	Faculteit der geneeskunde
24	12324	Jonker, C.C.	Natuurkunde en scheikunde
25	13037	Lever, J. / Vlijm, L.	Biologie
26	5744	Fliert, J.R. van de	De aardwetenschappen aan de Vrije Universiteit
27	15021	Deursen, A.Th. van	De Vrije Universiteit en de geschiedwetenschappen
28	16821	Knol, J.G.	Economie op weg
29	16689	Kuiper Hzn., G.	Sociologie. Nederland en de Vrije Universiteit.
30	18360	Sanders, C. / Eisenga, L.K.A.	De psychologie aan de Vrije Universiteit
31	25630	Klapwijk, J.	Honderd jaar filosofie aan de Vrije Universiteit
	-	Anoniem	*Mensen bij gijzelingen.*
32	26505	Bastiaans, J.	Gijzelingen
33	42439	Mulder, D.	Terrorisme
34	12674	Dijk, W.F. van	Gijzelaars en hun familie. Een voorbeeld.
	-	Anoniem	*Spiegel op de toekomst. Laatste advies van de Innovatie Commissie Middenschool.*
35	5508	Geerars, C.M.	Geïntegreerd onderwijs
36	5765	Stuart, M.J.W.	Achtergronden van de Bielefeldse schoolexperimenten
37	4773	Schoonenberg, H.H.H.T.M.	Effecten van innovatieplannen op scholen
38	3757	Bloo, G.	De inspectie in het middenschoolproces
39	4704	Schüssler, E.J.	Waar verandering eindigt en vernieuwing begint
40	4720	Jacobs, H.J.	Balans van een innovatie
41	3863	Sap-Jongkees, Els J.	Onderwijsvernieuwing en maatschappelijk draagvlak
42	5280	Karstanje, P.N.	Onderzoek als indicator van ambivalente onderwijspolitiek?
43	3253	Lagerweij, N.A.J.	Ondersteuning van veranderend onderwijs
44	1759	Meertens, J.	Van bestaande naar nieuwe regelgeving
45	5580	Veenstra, E.J.	Brede basisvorming
46	4181	Pedroli, J.A.J.	Alternatieve scholen of een alternatief voor school
47	4805	Doornbos, K.	Verticale groepen in het funderend onderwijs
48	6501	Jong, H.G. de	Zonder transplantatie geen systeemdoorbraak
49	4039	Vries, J. de	Een middenschool-conceptie in de volwassenen-educatie
	-	Anoniem	*De smalle marge van de onafhankelijkheid. Opstellen over Afrika. (Contact Tijdsdocumenten)*
50	20756	Geschiere, P.	Politiek-economische patronen in de geschiedenis van Afrika
51	7485	Roon, G. van	Imperialisme na de dekolonisatie
52	7989	Cohen, B.	De Europese Gemeenschap en Afrika
53	11739	Jonge, K. de	deeltitels: 1 Frankrijk in Afrika 2 Senegal: een voorbeeld van een neo-kolonie
54	3232	Davidson, B.	De Cubanen in Afrika
55	3078	Knol, Aly	Eritrea: imperialisme gaat ook in rood gekleed
56	3489	Stoof, J.	Zaïre: speeltuin van het Westen
57	6780	Ojik, B. van	Afrika in (onder)ontwikkeling
58	4724	Fortuin, Johanna	'Mama, wat eten we vandaag?'; vrouwen en hun dagelijkse strijd om het bestaan in zwart Afrika
59	5286	Dijk, M.P. van	Boven-Volta: overstroomd met hulp, maar nog steeds stagnerend en afhankelijk
60	6282	Gerla, A.	Tanzania: de onopgemerkte dood van het ujamaa-socialisme
61	6301	Veenstra, W.	Kenia: succes en zijn schaduwzijden
62	5571	Schellenberger, B.	Nigeria: groei of ontwikkeling
63	4292	Overbeek, H.	De Nederlandse economische betrekkingen met tropisch Afrika
64	7382	Geuns, R. van	Nederlandse ontwikkelingshulp aan Afrika
	-	Anoniem	*Koningin Wilhelmina.*
65	2031	Tamse, C.A.	Inleiding
66	1496	Bernhard, prins der Nederlanden	Herinneringen aan mijn schoonmoeder
67	6959	Beelaerts van Blokland, G.	1880-1890. Het beeld van Koningin Wilhelmina
68	4581	Beerman, J.C.	Twintig jaar in persoonlijke dienst van Koningin Wilhelmina
69	1739	Bischoff van Heemskerk, W.F.K.	Wilhelmina en haar Staldepartement
70	3579	Burger, J.A.W.	Koningin Wilhelmina toen!
71	2592	Drees sr., W.	Koningin Wilhelmina
72	3518	Gaay Fortman, W.F. de	Koningin Wilhelmina en de ontwikkeling van het constitutionele koningschap
73	2598	Geldens, Jeanette C.M.M.	Koningin Wilhelmina: toeverlaat en vluchthaven
74	1962	Gelissen, H.C.J.H.	Dankbare herinnering van een vooroorlogs minister van Economische Zaken
75	5101	Hulleman, A.	Koningin Wilhelmina, Het Loo en de mensen
76	1008	Jong, L. de	Gesprekken met Koningin Wilhelmina
77	3451	Kohnstamm, M.	Herinneringen aan mijn eerste werkgeefster
78	1764	Kolfschoten, H.A.M.T.	Koningin Wilhelmina en de eenheid rond de troon

Doc nr.	Aantal tokens	Auteur(s)	Titel(s) Indien cursief: titel hoofdwerk, ingedeeld in afzonderlijke documenten
79	3487	Lawick van Pabst-Doude van Troostwijk, J.H.M.A.	Een hoofdstuk apart
80	2983	Lieftinck, P.	Openingswoord herdenkingscolloquium 30 augustus 1980
81	3276	Loois, J.F.	Herinneringen van een oud-chauffeur
82	3271	Manning, A.F.	Vereisten voor een Wilhelmina-biografie
83	870	Mansholt, S.L.	Persoonlijke herinnering aan Koningin Wilhelmina
84	3809	Meijer, P.H.	Wilhelmina's humor, menselijkheid en trouw
85	1477	Oosterlee, Anna E.	Koningin Wilhelmina als moeder
86	4152	Puchinger, G.	Koningin Wilhelmina in de omgang met haar ministers
87	2568	Ranitz-de Brauw, A.M. / Schimmelpenninck-de Brauw, L.C.E.	Enige grepen uit de oude doos
88	925	Ranitz, C.J.A. de	Impressies van een oud-Leienaar
89	3123	Reinders, E.	Het Loo: het huis van Koningin Wilhelmina
90	1264	Roovers, P.	De vorstin en een beeldhouwer
91	3915	Sesink, C.	Impressies van dertig jaar veiligheidsdienst voor het Koninklijk Huis
92	3251	Touw, H.P. van der	Koningin Wilhelmina als expert op het gebied van paarden en tuigage
93	3614	Tromp, Th.P.	Herinneringen aan een markante en grote vrouw
94	6645	Veld, J. in 't	Oranje en rood in fel bewogen tijden
95	3928	Zinnicq Bergmann, R.J.E.M. van	Herinneringen aan drie jaar arbeid in dienst van Koningin Wilhelmina
	-	Anoniem	*Jaarboek van de Stichting Instituut voor Nederlandse Lexicologie. Overzicht van de jaren 1981 en 1982.*
96	3701	Geeraerts, D.	De ontwikkelingen in de lexicografische theorievorming
97	6603	Wortel, Th.P.F.	Dr. J.A.N. Knuttel, peuterige kamergeleerde tussen Marx en Kluyver
	-	Anoniem	*Methoden in de literatuurwetenschap.*
98	3430	Laan, N.	Over wetenschap
99	5155	Rees, C.J. van / Verdaasdonk, H.	Literatuurwetenschap en literatuuropvattingen
100	6933	Hoek, L.H.	Verhaalstrategieën. Aanzet tot een semiotisch georiënteerde narratologie
101	3231	Rutten, F.	De verhaallogica van Bremond
102	4691	Nef, F.	Inleiding in de narratieve semiotiek van A.J. Greimas
103	4811	Zoest, A.J.A. van	Peirciaanse semiotiek
104	6276	Alonso Hernàndez, J.L.	Linguistiek, literatuur en stylistiek
105	3344	Lintvelt, J.	Discursieve verteltypen
106	5321	Wierenga, L.	Retorische tekstbeschrijving
107	5489	Gumbrecht, H.U.	Receptie esthetika en handelingstheoretische literatuurwetenschap
108	6212	Zima, P.	Literatuursociologie als tekstsociologie
109	5542	Briosi, S.	Psychoanalyse en literatuurwetenschap
110	6200	Alexandrescu, S.	Logica en literatuurwetenschap
111	7349	Adriaens, M.	Modellen van linguistische tekstgrammatica's
	-	Anoniem	*Dezer jaren. Buitenlands beleid & internationale werkelijkheid. Beschouwingen aangeboden aan J.L. Heldring bij diens afscheid als directeur van het Nederlands Genootschap voor Internationale Zaken. (Anthos)*
112	3883	Beugel, E.H. van der	Een bezem door ons denken. Jéröme Louis Heldring.
113	8537	Bos, B.R.A. van den / Meerts, P.W.	Commentaar op de commentator. Een beeld van Heldrings werk aan de hand van vraaggesprekken.
114	8613	Leurdijk, J.H.	J.L. Heldring. Het denken van een atlanticus.
115	6964	Brands, M.C.	Patricische en plebejische elementen in de Nederlandse buitenlandse politiek. Heldrings paradoxen.
116	8527	Kossmann, E.H.	Progressiviteit en conservatisme in de Westeuropese staat
117	7448	Dunk, H.W. von der	Macht en moraal. De vereniging van het onverenigbare.
118	2850	Baehr, P.R.	De zorg voor de mensenrechten in het Nederlandse buitenlandse beleid
119	4452	Neuman, H.J.	Spanning en ontspanning
120	6457	Wellenstein, E.P.	Europese integratie en Atlantische samenwerking. Wederzijdse stimulans of hinderpaal?
	-	Anoniem	*Seksualiteit bij gehandicapten: (on)gewoon? (Leuvense Cahiers voor Seksuologie nr. 7)*
121	5972	Nijs, P.	Seksualiteit en gehandicapten: recht op en nood aan seksualiteit
122	26651	Christiaens, M.	deeltitels: 1 Seksualiteit bij gehandicapten: (on)gewoon? 2 Seksuele ethiek en lichamelijke handicap 3 Ethiek en seksualiteit van mentaal gehandicapten
123	5889	Bertha-Biro, Judith	Alleen seksualiteit? Seksuele opvoeding van kinderen met lichamelijke handicaps.
124	2206	Hoppenbrouwers, L.	Seksualiteit bij verlamden na ruggemergletsel
125	3448	Beusen, L.	Seksualiteit bij mentaal gehandicapten

Doc nr.	Aantal tokens	Auteur(s)	Titel(s) Indien cursief: titel hoofdwerk, ingedeeld in afzonderlijke documenten
126	3215	Pauwels, J.	Juridische aspecten van contraceptie, seksualiteit en ouderschap bij mentaal gehandicapten
127	4537	Lambrechts, G.	De opvoeder en zijn seksualiteit
	-	Anoniem	*De audiovisuele revolutie. De rol van de audiovisuele massamedia in de komende 25 jaar. (Sociale en culturele reeks)*
128	4279	Kooyman, A.	De nieuwe televisiecultuur
129	8808	Hamelink, C.	Informatiesamenleving en communicatiebeleid
130	7475	Knulst, W. / Staay, A. van der	De nieuwe kijker
131	7174	Manschot, B. / Stegeren, Th. van	Hebben we nog wat te zeggen?
132	1180	Anoniem	Overzicht nieuwe technologische ontwikkelingen
	-	Anoniem	*Vrijwillig werken aan ons welzijn. (Sociale en culturele reeks)*
133	8238	Ruygers, H.	Wat rondneuzen in de werkwinkel
134	9379	Fransen, W.	Vrijwilligers in de samenleving
135	13004	Bots, K.	Van spontane dienstverleners naar gemobiliseerde medewerkers. En vice versa?
136	5466	Ruygers, J.	Het ondersteunen van niet cliënt-gerichte vrijwilligers-organisaties
137	696	Anoniem	Elf stellingen tot besluit
	-	Anoniem	*Het belang van politieke partijen.*
138	3406	Lipschits, I.	Functies van politieke partijen
139	5217	Blom, J.C.H.	Partijen en verzuiling rond de eeuwwisseling
140	6760	Dunk, H.W. von der	De partijen en de parlementaire democratie in het interbellum
141	3775	Bank, J.	Van maatschappijvisie naar beleidsnota
142	5115	Baehr, P.R.	Nederlandse politieke partijen en de buitenlandse politiek
143	8164	Koole, R.A. / Soetendorp, R.B.	De partij als stootblok tussen overheid en samenleving: politisering of democratisering van het Nederlandse buitenlandse beleid
144	4487	Keman, J.E.	Het belang van politieke partijen en het politieke belang van partijen
145	3461	Brink, H. van den	Politieke partijen en participatie
146	3749	Kortmann, C.A.J.M.	De vrijheid van interne organisatie van de politieke partij
147	6836	Elzinga, D.J.	Partijbinding en politieke moraliteit
	-	Anoniem	*Adieu Dries. Overpeinzingen bij het afscheid van een liefhebber in de politiek.*
148	660	Wiegel, H.	Dat was nog eens formeren!
149	923	Gaay Fortman, W.F. de	De schok der herkenning
150	2341	Mulder, A.	Van Agt als minister van justitie
151	1161	Grinten, W.C.L. van der	Ter herinnering
152	1188	Puchinger, G.	De erkende leider én de kop van Jut
153	1051	Plas, M. van der	A Dieu, Van Agt
154	855	Kremers, J.	Dries van Agt: Gouverneur van Limburg in buitengewone dienst
155	895	Veen, C. van	Adieu Dries, au revoir
156	660	Geertsema, W.J.	Beste Dries
157	886	Steenkamp, P.A.J.M.	Ontzettend bedankt
158	753	Terlouw, J.	Over en aan Dries van Agt
159	540	Meekren, J. van	Man van het jaar
160	853	Lubbers, R.F.M.	Dries
161	505	Agt, A.A.M. van	De melancholieke nevels van de herfst
	-	Anoniem	*Het gezin vandaag en morgen.*
162	6002	Boer, C. den	Het gezin in de Bijbel
163	7160	Exalto, K.	Het gezin in de Reformatie en de nadere Reformatie
164	6754	Velema, W.H.	Gevaren die het gezin bedreigen
165	7374	Vergunst, A.	De verhoudingen in het gezin
166	6473	Hoek, J. / Hoek-van Kooten, A.B.F.	Vragen rond huwelijk en seksualiteit
167	5845	Malkenhorst-Visser, S.T. van	Het gezin en de zorgen in het gezin
168	7976	Kole, I.A.	Het gezin en de (kinder)bijbel
169	6132	Westerink, J.	Het gezin en de recreatie
170	4868	Blenk, C.	Het gezin in een apocalyptische tijd
171	6699	Graaf, J. van der	Gezin - school - catechese
172	11049	Huizer, W.	Onze dagelijkse arbeid
	-	Anoniem	*Korte verhalen uit de jaren zeventig.*
173	3106	Alberts, A.	De onbekende maarschalk
174	4367	Biesheuvel, J.M.A.	De heer Mellenberg
175	1024	Boer, H.P. de	Het feest
176	1657	Deelder, J.A.	'The Eddie Daté Story'
177	3760	Hart, M. 't	Ouderlingenbezoek
178	2869	Heeresma, H.	Vals van hart maar hoog het lied...
179	4521	Hotz, F.B.	De tramrace
180	2722	Jong, O. de	Dekker, Koolen & Buis

Doc nr.	Aantal tokens	Auteur(s)	Titel(s) Indien cursief: titel hoofdwerk, ingedeeld in afzonderlijke documenten
181	3593	Keulen, Mensje van	Tigertits Rosie
182	2000	Kuik, H.	Trix en de onzekerheid
183	3954	Matsier, N.	Flatus vocis
184	1917	Peskens, R.J. (ps. van G.A. van Oorschot)	De wasmachine
185	1689	Plomp, H.	Zoals de waard is, vertrouwt hij zijn gasten
186	3836	Portnoy, Ethel	Melk
187	4387	Romijn Meijer, H.	De eerste foto van het album
188	2736	Santen, S.	Een geintje
189	4220	Uijl, B. den	Het morele verval
190	5520	Vervoort, H.	Heden mosselen, morgen gij
191	1994	Waasdorp, J.	Neef Kas, de behanger
	-	Anoniem	*Wellustig knallen de kampernoelies.*
192	1035	Rybel, R. de	IJsbergman
193	835	Schrauwen, Griet	Amsterdam
194	1984	Mommaerts, M.	Wellustig knallen de kampernoelies
195	1728	Winsemius, Dieuwke	Een mens alleen
196	3554	Timmermans, B.	deeltitels: 1 De gefronste jongen 2 Het offer
197	3200	Hoeck, A. van	Pro Deo
198	2891	Impe, M. van	Stoom
199	1256	Weiwerd, M. van	Maar morgen Rotterdam
200	841	Vanmarcke, D.	Het konijn
201	2043	Dikker, Marianne	deeltitels: 1 Maffia 2 De kranen van het bad
202	888	Smet, M. de	deeltitels: 1 Fluitjes snijden 2 Droomvakantie
203	2049	Auwera, F. van der	Nutteloze handelingen
204	1935	Molle, B. van	De hamer
205	5685	Vermeulen, J.	Bedorven vlees
206	1903	Damme, B. van	Het veenlijk
207	1008	Hanchart, Joyce	De steenbokskeerkring
208	6629	Misselyn, J.	Vluchtend
209	4019	Vanbossuyt, P.	Het ratton rapport
210	573	Thielen, J. van	Close encounters of the third kind
211	1080	Delafontaine, R.	De kleine acht
212	810	Clemens, E.	De kalender van Homeros
	-	Anoniem	*Levensgemeenschappen. (Natuurbeheer in Nederland dl. 1)*
213	3192	Hanekamp, G.	Grienden
214	4908	Higler, L.W.G.	deeltitels: 1 Beken en kleine rivieren 2 Sloten
215	4137	Houte de Lange, S.M. ten	Bossen
216	13757	Leentvaar, P.	deeltitels: 1 Water en wateren 2 Duinmeren 3 Laagveenplassen en -meren 4 Oude kreken 5 Oude rivierlopen 6 Wielen
217	5568	Leys, H.N.	deeltitels: 1 Jeneverbesvegetaties 2 Akkers 3 Stinsemilieus
218	33159	Londo, G.	deeltitels: 1 Duinen 2 Hagen, houtwallen, houtsingels en bomenrijen 3 Struwelen 4 Graslanden 5 Ruigtkruiden en zoomvegetaties 6 Bermen en taluds 7 Wegen en paden
219	12350	Reijnders, T.	deeltitels: 1 Droge heide 2 Natte heide 3 Stuifzand
220	4987	Schroevers, P.J.	Vennen
221	4480	Smit, J.J.	Eendenkooien
222	1506	Werf, S. van der	Heide en heiden
223	13619	Wirdum, G. van	deeltitels: 1 Veen, venen en moerassen 2 Laagveenmoerassen 3 Hoogvenen
224	3893	Wolff, W.J.	Wadden en estuariën
225	5469	Wijngaarden, A. van	deeltitels: 1 Gebouwen 2 Ondergrondse kalksteengroeven

Doc nr.	Aantal tokens	Auteur(s)	Titel(s) Indien cursief: titel hoofdwerk, ingedeeld in afzonderlijke documenten
	-	Anoniem	*Filosofen van de 20e eeuw.*
226	4332	Doorman, S.J.	Bertrand Russell
227	4418	Schoonbrood, C.A.J.A.	George Moore
228	5925	Hubbeling, H.G.	Ludwig Wittgenstein
229	4683	Nuchelmans, G.	Peter Strawson
230	4942	Boer, Th. de	Edmund Husserl
231	5346	IJsseling, S.	Martin Heidegger
232	7937	Fretz, L.C.	Jean-Paul Sartre
233	7149	Bakker, R.	Maurice Merleau-Ponty
234	6227	Plattel, M.G.	Ernst Bloch
235	6266	Korthals, M.	Herbert Marcuse
236	6918	Görtzen, R.	Jürgen Habermas
237	6906	Locher, G.W.	Claude Lévi-Strauss
238	4524	Bertels, C.P.	Michel Foucault
239	5129	Heilbron, J.	Gaston Bachelard
240	5159	Wesly, P.G.E.	Rudolf Carnap
241	4574	Boon, L.	Karl Popper
242	6222	Vries, G.H. de	Thomas Kuhn en Imre Lakatos
	-	Anoniem	*Betalingsverkeer.*
243	6226	Boll, J.M.	De afroep(bank)garantie
244	7445	Bannier, F.A.W.	Handelskoop
245	10731	Mijnssen, F.H.J.	deeltitels: 1 De rekening-courantverhouding 2 Documentair krediet
246	7688	Rooy, R.E. de	Gegarandeerde betaalmiddelen
247	4562	Rank, W.A.	Civielrechtelijke aspecten van girale betaling
	-	Anoniem	*Postmoderne aardrijkskunde. De sociografische traditie voortgezet.*
248	5326	Smidt, M. de	Amsterdams territorium. Tendensen in de beoefening van de Amsterdamse geografie
249	5223	Paassen, Chr. van	Heinrich Rickert's interesse in het individuele - een geografische ouverture?
250	3718	Nauta, A.P.N.	Sociaal beleidsonderzoek voor de overheid
251	5067	Jolles, H.M.	Burengesprek. Ontmoetingen tussen sociale geografie en sociologie
252	4557	Beenakker, J.J.J.M.	Een onderzoek naar de beheersing van het binnenwater; gevolgen voor landschap en bewoning
253	4985	Klerk, A.P. de	Om de toekomst van het verleden. De rol van de historische geografie in het toegepaste landschapsonderzoek
254	5185	Amersfoort, J.M.M. / Surie, B.	Immigratieland tegen wil en dank. Nederland 1970-1985
255	3981	Ostendorf, W.	Culturele differentiatie binnen Nederland: van territoriale binding naar sociaal-ruimtelijke uitsortering
256	4566	Knippenberg, H. / Kwaasteniet, M. de	De nieuwe schoolstrijd in geografisch perspectief: de spreiding van openbaar en bijzonder onderwijs 1973-1984
257	6037	Dijkink, G.J. / Wusten, H. van der	Geweld als politiek-geografisch onderzoeksthema
258	5531	Hart, H.W. ter	Amsterdam: het financiële hart van Nederland
259	5398	Lambooy, J.G.	Transformatie en cyclus; gedachten over regionale ontwikkelingstheorie
260	4129	Dietz, T.	Geografisch determinisme in vermomming? Over agrarische landevaluatie, bevolkingsdraagvlak en geïntegreerde rurale ontwikkelingsgeografie
261	4689	Bruijne, G.A. de	Over niet-westerse en koloniale geografie, sociale geografie van ontwikkelingslanden en Amsterdam
262	4423	Engelsdorp Gastelaars, R. van / Vijgen, J.	Bezigheden buitenshuis hebbende. De betekenis van woning en woonomgeving voor enkele huishoudtypen in opkomst
263	6419	Anoniem	Kleine gemeenten in het spanningsveld van stadsgewest en provincie; een institutionele benadering
264	6759	Amersfoort, J.J.M. van	Sociale geografie als nuttige bezigheid. Overpeinzingen aan de oever van de derde geldstroom.
265	6646	Wusten, H. van der / Heldring, O.	De materiële grondslagen van het samenleven. Een programmatische beschouwing over de ontwikkeling van de sociale geografie.
	-	Anoniem	*Euthanasie. Knelpunten in een discussie.*
266	12303	Wal, G.A. van der	Bestaat er een recht op sterven?
267	9645	Musschenga, A.W.	Kwaliteit van leven: grond voor levensbeëindiging?
268	8250	Jacobs, F.C.L.M.	De figuur van de veronderstelde wil
269	8945	Feber, H.R.G.	De wederwaardigheden van artikel 293 van het Wetboek van Strafrecht vanaf 1981 tot heden
270	5934	Cohen, H.S.	Genoeg is genoeg
271	7447	Beaufort, I.	Op weg naar het einde?
	-	Anoniem	*Vreemd volk, gemengde gevoelens. Etnische verhoudingen in een grote stad.*
272	33415	Wouters, H.	De wijk in het midden van de stad.

Doc nr.	Aantal tokens	Auteur(s)	Titel(s) Indien cursief: titel hoofdwerk, ingedeeld in afzonderlijke documenten
273	50962	Bruin, K.	Het Oostersch feest.
274	45324	Anoniem	deeltitels: 1 Inleiding 2 De stad Utrecht en haar buitenlandse immigranten 3 Vreemdelingen in de buurt 4 Conclusie
	-	Anoniem	*Kleinschaligheid in de psychiatrie. De inrichting een vergissing?*
275	2046	Bökenkamp, K.	Is er al tijd voor kleinschaligheid?
276	790	Donker, Marianne	Wat is kleinschaligheid?
277	1050	Bruning, H.J.J.	Gemeente Den Haag zet eerste stappen op terrein kleinschaligheid
278	1545	Euwema, M.	Een goede visie is het halve werk
279	1013	Lucassen-Stauttener, Anne-Marie	We gaan de goede kant op
280	1422	Toussaint, R.	Nota Geestelijke Volksgezondheid is slechts vage aanzet
281	1878	Boot, P.	Dichter bij is beter
282	1790	Kolkman-Nijsen, M.E.	Klein is mooi, groot is machtig, slechts een combinatie ervan levert succes op
283	1974	Joele, L.J.	Kleinschaligheid, hoezo?
284	1056	Poel, E. van der	Het cacaoblik van Droste en de warmte van kleinschaligheid
285	1510	Bolle, Anneke	Almacht en onmacht
286	1563	Wiegant, H.	Mensen moeten zelf kunnen kiezen
287	1893	Brits, H.	'Belangstelling belangrijker dan deskundigheid'
288	1994	Hekkelman, J.W.	Kleinschaligheid uitgangspunt van provinciaal beleid
289	1013	Ramlal, D.	We moeten anders aankijken tegen psychiatrische patiënten
290	735	Binnema, W.J.	Enkele vraagtekens bij de voordelen van kleinschaligheid
291	2979	Verzijde, G.	Discussie over kleinschalige psychiatrie begon al in 1923
292	1921	Koen, Nel	Psychiatrie moet meer oog hebben voor maatschappelijke misstanden
293	1512	Brink, H. ten	Kleinschaligheid en huisvesting
294	1294	Faber, Annelies	Patiënten hebben recht op sex
295	1184	Anoniem	deeltitels: 1 Belang van kleinschaligheid nog onvoldoende onderkend 2 Kleinschaligheid: voor een beter persoonlijk contact 3 Lijst van afkortingen
296	893	Spruit, A.	Psychosociale problemen vaak gevolg van grootschaligheid
297	561	Groot, Carolien de / Akkermans, Marina	Onderzoek naar de acceptatie en behoefte aan kleinschalige woonvoorzieningen
298	2910	Reijden, van der	Samenvatting van de Nota Geestelijke Volksgezondheid
	-	Reynaerts, W.H.J. / Nagelkerke, A.G.	*Arbeidsverhoudingen, theorie en praktijk. Deel I: Inleiding, schets, interpretatie.*
299	34718	Reynaerts, W.H.J.	deeltitels: 1 De leer der arbeidsverhoudingen, een inleidend overzicht 2 Arbeidsverhoudingen in Nederland, schets van het institutionele patroon
300	42782	Nagelkerke, A.G.	De Nederlandse arbeidsverhoudingen in het spanningsveld van centralisatie en decentralisatie
	-	Anoniem	*Lees je knetter*
301	3674	Dros, Imme	Wil en dank
302	3526	Kerkwijk, H. van	Jam
303	3831	Hofman, W.	Billy (een wildwestverhaal)
304	3697	Meinkema, Hannes (ps. van Hannemieke Stamperius)	Trein
305	3518	Zanger, J. de	De klap
306	2855	Bogaards, Carla	Je moet het lef maar hebben om een held te zijn
307	4162	Ede, B. van	Café De Oude Snik
308	1640	Jekkers, H.	Oorlog & vreten
309	4069	Bloem, Marion	Bittere limonade met wat ijs
310	4082	Kromhout, R.	Terroristen
311	2746	Meinderts, K.	De grote verdwijntruc
312	5113	Vos-Dahmen von Buchholz, Tonny	Triomf van de rede
	-	Anoniem	*Ziekenhuis en strategie. Samenspel tussen manager en medisch specialist.*
313	6544	Peters, J.H.	deeltitels: 1 Inleiding: verantwoording, opzet, samenvatting 2 Management en medische staf
314	5972	Pal, J.H.N. van der / Peters, J.H.	Strategische beleidsvorming
315	3129	Peters, J.H. / Veltkamp, J.J.	Beleidsplanning
316	8011	Custers, H.L.H.M. / Henneke, J.A.M.	Budgettering
317	10379	Bakker, A.R.	Automatisering
318	4625	Peters, J.H. / Schouten, L. / Zwaan, F.E.	Innovatie

Doc nr.	Aantal tokens	Auteur(s)	Titel(s) Indien cursief: titel hoofdwerk, ingedeeld in afzonderlijke documenten
319	3670	Dam, P. van / Peters, J.H.	Patiëntenbeleid en public relations
320	5310	Pluyter-Wenting, E.S.P. / Vries, J.J.	Management van de kliniek: patiëntenclassificatie
321	7983	Buren, D.J. van	Management van de polikliniek
322	7411	Vissers, J.M.H.	Samenwerking met de eerste lijn
323	3986	Schouten, L.	Samenwerking met andere ziekenhuizen
	-	Anoniem	*Cursusdeel 1. Staatsrecht centrale overheid. Blok 1+6.*
324	75013	Pot, C.W. van der / Donner, A.M.	Overige staatsrechtelijke onderwerpen
	-	Anoniem	*Wetenschap & taal.*
325	4981	Tervoort, B.T.	Wat is taalpathologie?
326	4647	Dik, S.C.	Wat is pragmatiek?
327	5118	Walraven, Th.L.M.	Wat is sociolinguistiek?
328	5082	Goossens, J.	Wat is dialectologie?
329	5517	Holmes, J.S.	Wat is vertaalwetenschap?
330	4458	Els, T.J.M. van	De studie der vreemdetaalverwerving: taalonderwijskundig
331	4977	Carpay, J.A.M.	De studie der vreemde taalverwerving: leerpsychologisch
332	5306	Geest, T. van der	De studie van de kindertaal
333	4007	Kempen, G.	Wat is psycholinguistiek?
334	5134	Luijten, T.	Het toetsen van taalvaardigheid
335	5204	Landsbergen, S.P.J.	Wat is computerlinguistiek?
336	3731	Brandt Corstius, H.	Wat is algebraïsche taalkunde?
337	3014	Kuipers, A.H.	Wat is descriptieve linguistiek?
	-	Anoniem	*Succesvol rechten studeren.*
338	16714	Klein, M.	De beroepen
339	3671	Smits, R.	De faculteit en de universiteit
340	5064	Aller, H.B. van / Klein, M.	Het recht
	-	Anoniem	*Vakbondswerk moet je leren.* *Een leerboek over industriebonden in Engeland, Frankrijk, Duitsland en Nederland, de kaderleden en hun scholing.*
341	65403	Visser, J.	Deel IV. Dertig jaar scholing en vorming in de Industriebond-NVV (1945-1975)
342	5008	Gevel, A.A.J.S. van de	Cursus De Gemeenteraad. Les 1: Plaats en taak van de gemeente.
343	8534	Gevel, A.A.J.S. van de	Cursus De Gemeenteraad. Les 2: Samenwerking in en van de gemeenten.
344	10120	Gevel, A.A.J.S. van de	Cursus De Gemeenteraad. Les 3: Politiek in de gemeente.
345	9742	Gevel, A.A.J.S. van de	Cursus De Gemeenteraad. Les 4: De relatie bestuur-bestuurden.
346	6865	Emster, F.W.H. van den	Cursus De Gemeenteraad. Les 5: Rechtsbescherming.
347	17226	Gevel, A.A.J.S. van de	Cursus De Gemeenteraad. Les 6: De gemeenteraad.
348	11755	Heijden, F.J. van der	Cursus De Gemeenteraad. Les 7: Het raadslid.
349	14074	Heijden, F.J. van der	Cursus De Gemeenteraad. Les 8: Het raadslid in en buiten de raad.
350	12561	Rappard, W.L.F.G. ridder van	Cursus De Gemeenteraad. Les 9: Burgemeester en wethouders.
351	4261	Bonnema, W.	Cursus De Gemeenteraad. Les 11: Het financiële beleid.
352	4378	Bonnema, W.	Cursus De Gemeenteraad. Les 12: Begroting en meerjarenramingen.
353	8420	Otterloo, G.F. van	Cursus De Gemeenteraad. Les 13: Ruimtelijke ordening.
354	13681	Doorninck, D.J. van / Reijnen, J.L. van	Cursus De Gemeenteraad. Les 14: Volkshuisvesting en stadsvernieuwing.
355	8437	Meijer Drees, F.J.	Cursus De Gemeenteraad. Les 15: Milieu.
356	6047	Kuijers, F.J.	Cursus De Gemeenteraad. Les 16: Verkeer en vervoer.
357	12206	Schouten, A.	Cursus De Gemeenteraad. Les 17: Economie en werkgelegenheid.
358	18467	Ginjaar-Maas, N(el).J.	Cursus De Gemeenteraad. Les 18: Onderwijs.
359	10926	Fijen, J.Chr.G.	Cursus De Gemeenteraad. Les 19: Welzijn.
360	14258	Eissens, W.J.G.M.	Cursus De Gemeenteraad. Les 20: Sociaal-cultureel werk.
361	11194	Dekkinga, T(ineke).	Cursus De Gemeenteraad. Les 21: Emancipatie.
362	4641	Gerrits, R.W.M.	Cursus De Gemeenteraad. Les 22: Maatschappelijke dienstverlening.
363	14168	Wiersma, A(nke). / Schröder, J.	Cursus De Gemeenteraad. Les 23: Sport en recreatie.
364	8075	Martini, C.V.	Cursus De Gemeenteraad. Les 24: Volksgezondheid.
365	9682	Pop, J.J.H.	Cursus De Gemeenteraad. Les 25: Openbare orde en hulpverlening.
366	99215	Heeresma, H.	Een hete ijssalon en andere verhalen.
367	61146	Zeeuw, G. van der	Helder weten. Ervaringen uit onzichtbare werelden.
368	119318	Crul, Heleen	Wij willen kinderen. Voorlichtingsboek voor toekomstige ouders over gezinsvorming, zwangerschap en de baby.
369	91519	Quick, R.	Duivensport totaal.
370	82374	Klis, T. van der / Mortier, J.W. du	Oppervlaktebehandelingen van constructiestaal.
371	33356	Anoniem	Bejaard, hoe en wat. (Recht & raad)
372	60035	Postma, P.	Marketing voor iedereen. Compleet handboek voor directe toepassing. (Mens en bedrijf)
373	46055	Overkleeft, D. / Groosman, L.E.	Het Dekker perspectief

Doc nr.	Aantal tokens	Auteur(s)	Titel(s) Indien cursief: titel hoofdwerk, ingedeeld in afzonderlijke documenten
374	36126	Haak, R.	Het speuren van honden. In theorie en praktijk.
375	29056	Anoniem	Onderhoudsbewust ontwerpen nu en in de toekomst. (Toekomstbeeld der Techniek)
376	71162	Rachels, J.	Leven en sterven. Over euthanasie en ethiek. (Aula)
377	32895	Peursen, C.A.	Filosofie van de wetenschappen.
378	38343	Reijnders, F.	Kunst - geschiedenis, verschijnen en verdwijnen.
379	48704	Poll, K.	De beklemde elite. Essay over het relativiteitsbeginsel in de cultuur.
380	43294	Droste, F.G.	Het taaldier mens. Een pamflet.
381	21518	Siccama, J.G.	Roepnaam: Airland Battle. Een nieuwe militaire aanpak voor de NAVO. (Clingendael-reeks)
382	96273	Teitler, G.	De wording van het professionele officierskorps. Een sociologisch-historische analyse.
383	55175	Walraven, Th.	Taalgebruik en taalwetenschap. Inleiding in het onderzoek van verbale kommunikatie.
384	65960	Fleischhacker, L.E.	Over de grenzen van de kwantiteit.
385	34370	Veenendaal, W. van	Autokennis. Het voorkomen van reparaties.
386	24055	Aalders, W.	Isoleren van muren, vloeren, daken en ramen.
387	125641	Nierop, M. van	Nieuwe woorden. Verklarend en verhalend woordenboek van modern taalgebruik. (De taal waarmee wij leven)
388	87509	Weeszenberg, J. van	Smalfilmhobbyboek.
389	56523	Blok, H.	Volleybal.
390	27396	Best, Kr. de / Burdorf, M.	Solliciteren met succes. HBO-ers en academici op weg naar de juiste baan.
391	21835	Roethof, O.	Karate. Een handboek voor trainer, coach en karateka.
392	70000	Louwerse, C.	Studeren kun je leren.
393	68295	Anoniem	Wegwijzer voor de werkende vrouw. Tweede, bijgewerkte druk.
394	24554	Sanders, L.F.J.	Huiswerk. Leidraad voor ouders en docenten.
395	22432	Eck, R. van / Kapteijns, J.H.M.	Wegwijzer wonen in groepsverband. Overwegingen bij het maken van programma's van eisen.
396	80470	Ban, A.W. van den	Inleiding tot de voorlichtingskunde.
397	78394	Rottier, H.	Stedelijke strukturen. Een inleiding tot de ontwikkeling van de Europese stad.
398	57488	Röling, B.V.A.	Inleiding tot de wetenschap van oorlog en vrede. (Polemologische Studiën VIII)
399	41594	Hoogendoorn, J.	Inleiding personeelplanning.
400	63410	Lint, P. van	Vormen van taalverkeer. Een elementaire inleiding in de leer van de verstandhoudingsmiddelen.
401	71610	Joosen, A.W.A.	Autoleasing en wagenparkbeheer, vormen van externe dienstverlening.
402	27299	Bauwens, R.	Snooker.
403	75310	Hay, A.	Handboek golf. (Sport)

BIJLAGE 2 - LIJST VAN WOORDEN VAN MEER DAN 30 KARAKTERS

Corpus Wetgevingsteksten

Nr.	Woordtype	Lengte	Abs. freq.	Rel. freq.	In % doc.
1	AANSPRAKELIJKHEIDSVERZEKERINGEN	31	4	0,000031%	0,0213%
2	ACCOUNTANTADMINISTRATIECONSULENT	32	1	0,000008%	0,0053%
3	ACCOUNTANTSADMINISTRATIECONSULENTEN	35	4	0,000031%	0,0213%
4	ACETALDEHYDEBENZYLMETHOXYETHYLACETAAT	37	1	0,000008%	0,0053%
5	ADEMHALINGSBESCHERMINGSAPPARAAT	31	1	0,000008%	0,0053%
6	ADEMHALINGSBESCHERMINGSAPPARATEN	32	2	0,000016%	0,0053%
7	ADEMHALINGSBESCHERMINGSMIDDELEN	31	3	0,000023%	0,0160%
8	ARBEIDSGESCHIKTHEIDSVERZEKERING	31	1	0,000008%	0,0053%
9	ARBEIDSONGECHIKTHEIDSVERZEKERING	32	1	0,000008%	0,0053%
10	ARBEIDSONGESCHIKHEIDSVERZEKERING	32	3	0,000023%	0,0160%
11	ARBEIDSONGESCHIKTGHEIDSVERZEKERING	34	1	0,000008%	0,0053%
12	ARBEIDSONGESCHIKTHEIDSCRITERIUM	31	3	0,000023%	0,0106%
13	ARBEIDSONGESCHIKTHEIDSPERCENTAGE	32	4	0,000031%	0,0160%
14	ARBEIDSONGESCHIKTHEIDSREGELINGEN	32	13	0,000102%	0,0106%
15	ARBEIDSONGESCHIKTHEIDSSUITKERING	32	1	0,000008%	0,0053%
16	ARBEIDSONGESCHIKTHEIDSUITKERING	31	828	0,006465%	0,4574%
17	ARBEIDSONGESCHIKTHEIDSUITKERINGEN	33	18	0,000141%	0,0479%
18	ARBEIDSONGESCHIKTHEIDSVERZEKERING	33	762	0,005950%	1,0264%
19	ARBEIDSONGESCHIKTHEIDSVERZEKERINGEN	35	2	0,000016%	0,0106%
20	ARBEIDSONGESCHIKTHEIDSVOORZIENING	33	46	0,000359%	0,1223%
21	ARBEIDSONGESHIKTHEIDSVERZEKERING	32	1	0,000008%	0,0053%
22	ARBEIDSONTSCHIKTHEIDSVERZEKERING	32	1	0,000008%	0,0053%
23	AUTOMATISERINGSINSTALLATIEBEDRIJF	33	1	0,000008%	0,0053%
24	BASISREMIGRATIESUBSIDIEREGELING	31	5	0,000039%	0,0160%
25	BEDRIJFSVERGUNNINGENBESCHIKKING	31	10	0,000078%	0,0053%
26	BEDRIJFSZELFBESCHERMINGSORGANISATIE	35	2	0,000016%	0,0053%
27	BEJAARDENZIEKENFONDSVERZEKERING	31	7	0,000055%	0,0266%
28	BEJAARDENZIEKENFONDSVERZEKERINGENVAN	36	1	0,000008%	0,0053%
29	BENZALDEHYDEPROPYLEENGLYCOLACETAAT	34	1	0,000008%	0,0053%
30	BEROEPSAANSPRAKELIJKHEIDSVERZEKERINGEN	38	1	0,000008%	0,0053%
31	BESTRIJDINGSMIDDELENBESCHIKKING	31	8	0,000062%	0,0213%
32	BIJZONDEREZIEKTEKOSTENVERZEKERING	33	1	0,000008%	0,0053%
33	BURGEMEESTERSVERTEGENWOORDIGING	31	1	0,000008%	0,0053%
34	CONTRACTVERZEKERINGSOVEREENKOMST	32	3	0,000023%	0,0160%
35	CONTRACTVERZEKERINGSOVEREENKOMSTEN	34	3	0,000023%	0,0160%
36	DEARBEIDSONGESCHIKTHEIDSVERZEKERING	35	1	0,000008%	0,0053%
37	DESOXYRIBONUCLEINEZUURMOLECULEN	31	1	0,000008%	0,0053%
38	DIISOPROPYLBENZEENHYDROPEROXYDE	31	3	0,000023%	0,0106%
39	DIMETHYLFENYLETHYLCARBINYLACETAAT	33	1	0,000008%	0,0053%
40	DIMETHYLFENYLETHYLCARBINYLISOBUTYRAAT	37	1	0,000008%	0,0053%
41	DINATRIUMWATERSTOFFOSFAATOPLOSSING	34	1	0,000008%	0,0053%
42	DISTRICTSCONSULTATIEBUREAUDIENST	32	1	0,000008%	0,0053%
43	EENENZESTIGDUIZENDACHTHONDERDENDRIEENTACHTIG	44	1	0,000008%	0,0053%
44	FURFURYLTRIMETHULAMMONIUMZOUTEN	31	1	0,000008%	0,0053%
45	GEHEIMHOUDINGSBESLUITKERNENERGIEWET	35	1	0,000008%	0,0053%
46	GEZONDHEIDSBESCHERMINGSONDERZOEK	32	1	0,000008%	0,0053%
47	HARTCATHETERISATIEVOORZIENINGEN	31	1	0,000008%	0,0053%
48	HERSTELFINANCIERINGSMAATSCHAPPIJ	32	1	0,000008%	0,0053%
49	HOEVEELHEIDSAANDUIDINGENBESLUIT	31	72	0,000562%	0,0904%
50	HOGEDRUKVLOEISTOFCHROMATOGRAFIE	31	2	0,000016%	0,0106%
51	HONDERDNEGENTIGDUIZENDZESHONDERDENACHT	38	1	0,000008%	0,0053%
52	HONDERDNEGENTIGDUIZENDZEVENHONDERDENDRIEEN-NEGENTIG	50	1	0,000008%	0,0053%
53	HONDERDTWEEENZESTIGDUIZENDACHTHONDERDVIJFEN-TWINTIG	50	1	0,000008%	0,0053%
54	HONDERDZESENNEGENTIGDUIZENDVIERHONDERDENZEVEN-TIEN	49	1	0,000008%	0,0053%
55	HOOGRENDEMENTSVERWARMINGSKETELS	31	4	0,000031%	0,0053%
56	HOOGSPANNINGSONTSTEKINGSSYSTEEM	31	1	0,000008%	0,0053%
57	HUURPRIJSUITVOERINGSBESCHIKKING	31	2	0,000016%	0,0053%
58	INSTRUMENTENWEERSOMSTANDIGHEDEN	31	1	0,000008%	0,0053%
59	INSTRUMENTVLIEGVERKEERSVOORSCHRIFTEN	36	2	0,000016%	0,0053%
60	KALIUMDIWATERSTOFORTHOFOSFAATOPLOSSING	38	1	0,000008%	0,0053%

Nr.	Woordtype	Lengte	Abs. freq.	Rel. freq.	In % doc.
61	KALIUMHEXACYANOFERRAATOPLOSSING	31	1	0,000008%	0,0053%
62	KALIUMHEXACYANOFERRAATTRIHYDRAAT	32	1	0,000008%	0,0053%
63	KETELVOEDINGWATERBEHANDELINGSAPPARATUUR	39	1	0,000008%	0,0053%
64	KREDIETVERZEKERINGSMAATSCHAPPIJ	31	6	0,000047%	0,0053%
65	KREDIETVERZEKERINGSMAATSCHAPPIJEN	33	7	0,000055%	0,0053%
66	LEVENSVERZEKERINGSONDERNEMINGEN	31	2	0,000016%	0,0053%
67	LUCHTVAARTTERREINVERKEERSGEBIED	31	2	0,000016%	0,0106%
68	MEDEVERANTWOORDELIJKHEIDSHEFFING	32	3	0,000023%	0,0106%
69	MEDEZEGGENSCHAPSAANGELEGENHEDEN	31	2	0,000016%	0,0106%
70	MOLESTBEDRIJFSSCHADEVERZEKERING	31	1	0,000008%	0,0053%
71	MOLESTVERZEKERINGSMAATSCHAPPIJEN	32	1	0,000008%	0,0053%
72	NADERINGSVERKEERSLEIDINGSDIENST	31	8	0,000062%	0,0106%
73	NADERINGSVERKEERSLEIDINGSDIENSTEN	33	1	0,000008%	0,0053%
74	NADERINGSVERKEERSLEIDINGSGEBEID	31	1	0,000008%	0,0053%
75	NADERINGSVERKEERSLEIDINGSGEBIED	31	3	0,000023%	0,0160%
76	NATRIUMDIWATERSTOFFOSFAATOPLOSSING	34	1	0,000008%	0,0053%
77	NATRIUMORTHOFENYLFENOLAATGEHALTE	32	1	0,000008%	0,0053%
78	NEGENHONDERDZEVENENNEGENTIGDUIZEND	34	1	0,000008%	0,0053%
79	NOODSTROOMVOORZIENINGSINSTALLATIES	34	1	0,000008%	0,0053%
80	ONDERWIJSVERZORGINGSINSTELLINGEN	32	3	0,000023%	0,0160%
81	ONDERWIJZERSSTAATSEXAMENBESLUIT	31	1	0,000008%	0,0053%
82	ONREGELMATIGSIGNAALDISCRIMINATOR	32	1	0,000008%	0,0053%
83	OPPERVLAKTEWATERKWALITEITSBEHEERDER	35	1	0,000008%	0,0053%
84	OVERGEVOELIGHEIDSVERSCHIJNSELEN	31	1	0,000008%	0,0053%
85	OVERWEGBEVEILIGINGSINRICHTINGEN	31	2	0,000016%	0,0053%
86	PARAMENTENVERVAARDIGINGSBEDRIJF	31	1	0,000008%	0,0053%
87	PASPOORTENBESTANDSADMINISTRATIE	31	4	0,000031%	0,0053%
88	PLANTENZIEKTENBESTRIJDINGSMIDDELEN	34	2	0,000016%	0,0106%
89	PROGRAMMAVOORBEREIDINGSCOMMISSIE	32	3	0,000023%	0,0053%
90	PROJEKTONTWIKKELINGSMAATSCHAPPIJ	32	1	0,000008%	0,0053%
91	RADIOFREQUENTVERMOGENSVERSTERKERS	33	3	0,000023%	0,0106%
92	RADIOTELEGRAFIEVEILIGHEIDSCERTIFICAAT	37	1	0,000008%	0,0053%
93	RECHTSPOSITIEBESLUITONDERWIJSPERSONEEL	38	1	0,000008%	0,0053%
94	RECLAMEVERLICHTINGSINSTALLATIES	31	1	0,000008%	0,0053%
95	RHEINSCHIFFAHRTPOLIZEIVERORDNUNG	32	1	0,000008%	0,0053%
96	RIOOLWATERZUIVERINGSINSTALLATIES	32	2	0,000016%	0,0106%
97	SCHEEPVAARTVERKEERSREGLEMENTERING	33	4	0,000031%	0,0053%
98	SCHOOLBEGELEIDINGSOVEREENKOMSTEN	32	1	0,000008%	0,0053%
99	STUURSTROOMBEVEILIGINGSTOESTELLEN	33	1	0,000008%	0,0053%
100	TENTOONSTELLINGSINSTALLATIEBEDRIJF	34	1	0,000008%	0,0053%
101	TRIACETYLDIHYDROXYPHENYLISATINUM	32	1	0,000008%	0,0053%
102	TWEEENZESTIGDUIZENDNEGENHONDERDACHTENTWINTIG	44	1	0,000008%	0,0053%
103	TWEEHONDERDDUIZENDNEGENHONDERDDRIEENDERTIG	42	1	0,000008%	0,0053%
104	TWEEHONDERDENNEGENENDERTIGDUIZEND	33	2	0,000016%	0,0106%
105	VEILIGHEIDSCONSTRUCTIECERTIFICAAT	33	9	0,000070%	0,0106%
106	VERBINDINGSINLICHTINGENPERSONEEL	32	1	0,000008%	0,0053%
107	VERKEERSONGEVALLENCONCENTRATIES	31	2	0,000016%	0,0053%
108	VIERENTWINTIGDUIZENDZESHONDERDENVEERTIG	39	1	0,000008%	0,0053%
109	VIERENTWINTIGDUIZENDZEVENHONDERDENEENENDERTIG	45	1	0,000008%	0,0053%
110	VIJFENTWINTIGDUIZENDVIJFHONDERDENNEGENENVIJFTIG	47	1	0,000008%	0,0053%
111	VISSERIJONDERZOEKINGSVAARTUIGEN	31	1	0,000008%	0,0053%
112	VLIEGTUIGTUIGONDERZEEBOOTBESTRIJDER	35	1	0,000008%	0,0053%
113	VLUCHTVOORBEREIDINGSFORMULIEREN	31	1	0,000008%	0,0053%
114	VOEDINGSWAARDEAANDUIDINGENBESLUIT	33	3	0,000023%	0,0053%
115	VOLKSVERZEKERINGSOVEREENKOMSTEN	31	2	0,000016%	0,0053%
116	VOORSCHOTVERSTREKKINGSPROCEDURE	31	1	0,000008%	0,0053%
117	WATERLEIDINGSCHADEVERZEKERINGEN	31	1	0,000008%	0,0053%
118	WATERLEIDINGSSCHADEVERZEKERINGEN	32	1	0,000008%	0,0053%
119	WONINGSTOFFEERDERSWERKZAAMHEDEN	31	1	0,000008%	0,0053%
120	ZESENTWINTIGDUIZENDHONDERDENEENENTACHTIG	40	1	0,000008%	0,0053%
121	ZEVENENZESTIGMILJOENVIJFHONDERDENZESTIGDUIZEND	46	1	0,000008%	0,0053%
122	ZEVENHONDERDTWEEENZESTIGDUIZEND	31	1	0,000008%	0,0053%
123	ZICHTVLIEGVERKEERSVOORSCHRIFTEN	31	1	0,000008%	0,0053%
124	ZIEKENHUISVERPLEGINGSVERENIGINGEN	33	1	0,000008%	0,0053%
125	ZIEKTEKOSTENVERZEKERINGSBEDRIJF	31	21	0,000164%	0,0372%
126	ZIEKTEKOSTENVERZEKERINGSMAATSCHAPPIJEN	38	2	0,000016%	0,0106%
127	ZWAVELZUURKOPERSULFAATOPLOSSING	31	1	0,000008%	0,0053%

Corpus Jurisprudentieteksten

Nr.	Woordtype	Lengte	Abs. freq.	Rel. freq.	In % doc.
1	AANDEELHOUDERSVERGADERINGSBESLUIT	33	2	0,000016%	0,0061%
2	AANSPRAKELIJKHEIDSUITSLUITINGEN	31	2	0,000016%	0,0122%
3	AANSPRAKELIJKHEIDSVERZEKERINGEN	31	12	0,000094%	0,0365%
4	AANSPRAKELIJKHEIDSVERZEKERINGSOVEREENKOMST	42	5	0,000039%	0,0061%
5	AANSPRAKELIJKHEIDSVERZEKERINGSPOLIS	35	4	0,000031%	0,0122%
6	AANSPRAKELIJKHEIDUITSLUITINGSGROND	34	1	0,000008%	0,0061%
7	ACCOUNTANTADMINISTRATIECONSULENT	32	1	0,000008%	0,0061%
8	ACHTENNEGENTIGDUIZENDVIERHONDERDVIJF	36	2	0,000016%	0,0061%
9	ADVERTENTIEKOSTENDOORBEREKENING	31	1	0,000008%	0,0061%
10	ALLEENVERTEGENWOORDIGINGSBEVOEGD	32	1	0,000008%	0,0061%
11	ALLEENVERTEGENWOORDIGINGSCONTRACT	33	1	0,000008%	0,0061%
12	ALLEENVERTEGENWOORDIGINGSOVEREENKOMST	37	6	0,000047%	0,0122%
13	ALLEENVERTEGENWOORDIGINGSRECHTEN	32	2	0,000016%	0,0061%
14	ALLEENVERTEGENWOORDIGINGSRELATIE	32	1	0,000008%	0,0061%
15	ARBEIDSGESCHILLENLANDSVERORDENING	33	1	0,000008%	0,0061%
16	ARBEIDSONGESCHIKHEIDSVERZEKERING	32	2	0,000016%	0,0122%
17	ARBEIDSONGESCHIKTHEIDSPERCENTAGE	32	6	0,000047%	0,0243%
18	ARBEIDSONGESCHIKTHEIDSPERCENTAGES	33	2	0,000016%	0,0061%
19	ARBEIDSONGESCHIKTHEIDSUITKERING	31	51	0,000398%	0,1522%
20	ARBEIDSONGESCHIKTHEIDSUITKERINGEN	33	15	0,000117%	0,0548%
21	ARBEIDSONGESCHIKTHEIDSVERKLARING	32	1	0,000008%	0,0061%
22	ARBEIDSONGESCHIKTHEIDSVERKLARINGEN	34	2	0,000016%	0,0122%
23	ARBEIDSONGESCHIKTHEIDSVERZEKERING	33	137	0,001070%	0,3043%
24	ARBEIDSONGESCHIKTHEIDSVERZEKERINGEN	35	1	0,000008%	0,0061%
25	ARBEIDSONGESCHIKTHEIDSVOORZIENING	33	11	0,000086%	0,0183%
26	ARBEIDSONGESCHIKTHEIDSVOORZIENINGEN	35	1	0,000008%	0,0061%
27	ARBEIDSONGESCHIKTHEIDSWETGEVING	31	1	0,000008%	0,0061%
28	ARBEIDSONGESCHIKTHEIDVERZEKERING	32	1	0,000008%	0,0061%
29	ARBEIDSOVEREENKOMSTENVERORDENING	32	2	0,000016%	0,0122%
30	ARBEIDSVOORWAARDENOVEREENKOMSTEN	32	2	0,000016%	0,0122%
31	BEDRIJFSAANSPRAKELIJKHEIDSVERZEKERING	37	1	0,000008%	0,0061%
32	BEDRIJFSAANSPRAKELIJKHEIDSVERZEKERINGEN	39	1	0,000008%	0,0061%
33	BEDRIJFSBEEINDIGINGSOVEREENKOMST	32	5	0,000039%	0,0122%
34	BEDRIJFSBEVEILIGINGSACTIVITEITEN	32	1	0,000008%	0,0061%
35	BEDRIJFSONDERBREKINGSVERZEKERING	32	6	0,000047%	0,0061%
36	BEDRIJFSONGEVALLENAANSPRAKELIJKHEID	35	1	0,000008%	0,0061%
37	BEDRIJFSVERGUNNINGENBESCHIKKING	31	2	0,000016%	0,0061%
38	BENZEENSULFONYLUREUMVERBINDINGEN	32	1	0,000008%	0,0061%
39	BEROEPSAANSPRAKELIJKHEIDSASSURADEUR	35	2	0,000016%	0,0061%
40	BEROEPSAANSPRAKELIJKHEIDSASSURADEUREN	37	1	0,000008%	0,0061%
41	BEROEPSAANSPRAKELIJKHEIDSVERZEKERAAR	36	4	0,000031%	0,0061%
42	BEROEPSAANSPRAKELIJKHEIDSVERZEKERAARS	37	2	0,000016%	0,0061%
43	BEROEPSAANSPRAKELIJKHEIDSVERZEKERING	36	27	0,000211%	0,0183%
44	BEROEPSAANSPRAKELIJKHEIDSVERZEKERINGEN	38	1	0,000008%	0,0061%
45	BEROEPSKEUZEVOORLICHTINGSRAPPORT	32	1	0,000008%	0,0061%
46	BESTRIJDINGSMIDDELENBEWAARPLAATS	32	1	0,000008%	0,0061%
47	BOERDERIJVERPLAATSINGSOVEREENKOMST	34	4	0,000031%	0,0061%
48	BOUWGRONDBELASTINGVERORDENINGEN	31	1	0,000008%	0,0061%
49	BRANDVERZEKERINGSOVEREENKOMSTEN	31	2	0,000016%	0,0122%
50	BRUTOBEROEPSINKOMENSVERMINDERING	32	1	0,000008%	0,0061%
51	BUNDESVERFASSUNGSGERICHTSGESETZ	31	2	0,000016%	0,0122%
52	CHARTERLUCHTVAARTMAATSCHAPPIJEN	31	1	0,000008%	0,0061%
53	COMPTABILITEITSLANDSVERORDENING	31	23	0,000180%	0,0061%
54	CONJUNCTUURAFBETALINGSBESCHIKKING	33	1	0,000008%	0,0061%
55	CREDIETVERZEKERINGSMAATSCHAPPIJ	31	1	0,000008%	0,0061%
56	DIAETHYLSTILBOESTROLDIPROPIONAAT	32	1	0,000008%	0,0061%
57	DIEFSTALVERZEKERINGSOVEREENKOMST	32	1	0,000008%	0,0061%
58	DISTRIBUTIEREGELINGSBESCHIKKING	31	1	0,000008%	0,0061%
59	DOELMATIGHEIDSHEIDSOVERWEGINGEN	31	1	0,000008%	0,0061%
60	DRIEENDERTIGDUIZENDDRIEHONDERDDRIEENDERTIG	42	1	0,000008%	0,0061%
61	DRIEHONDERDNEGENENTACHTIGDUIZEND	32	1	0,000008%	0,0061%
62	DRIEHUNDERTFUNFUNDNEUNZIGTAUSENDVIERHUNDERT-NEUNUNDDREISSIG	58	1	0,000008%	0,0061%
63	DRIEPARTIJENHUURKOOPTRANSACTIES	31	1	0,000008%	0,0061%
64	DUIZENDVIERHONDERDNEGENENZEVENTIGGULDEN	39	1	0,000008%	0,0061%
65	DWANGLICENTIEVERLENINGSPROCEDURE	32	1	0,000008%	0,0061%
66	EIGENDOMSOVERDRACHTSOVEREENKOMSTEN	34	1	0,000008%	0,0061%
67	ELECTRICITEITSPRODUCTIEBEDRIJVEN	32	1	0,000008%	0,0061%
68	FAMILIENRECHTSANDERUNGSGESETZES	31	1	0,000008%	0,0061%

Nr.	Woordtype	Lengte	Abs. freq.	Rel. freq.	In % doc.
69	GASVERWARMINGSINSTALLATIEBEDRIJF	32	1	0,000008%	0,0061%
70	GELDKREDIETVERLENINGSOVEREENKOMST	33	1	0,000008%	0,0061%
71	HANDELSVERTEGENWOORDIGINGSOVEREENKOMST	38	2	0,000016%	0,0061%
72	HOOGFREKWENTVERMOGENVERSTERKERS	31	1	0,000008%	0,0061%
73	HOOGSTONWAARSCHIJNLIJKHEIDSFORMULE	34	1	0,000008%	0,0061%
74	HUURPRIJSBEHEERSINGSVOORSCHRIFTEN	33	1	0,000008%	0,0061%
75	HUURPRIJSVASTSTELLINGSBEVOEGDHEID	33	1	0,000008%	0,0061%
76	HUURPRIJSVASTSTELLINGSPROCEDURE	31	2	0,000016%	0,0061%
77	HUURPRIJSVASTSTELLINGSPROCEDURES	32	2	0,000016%	0,0061%
78	HUWELIJKSGOEDERENGEMEENSCHAPPEN	31	1	0,000008%	0,0061%
79	HUWELIJKSVOORWAARDENOVEREENKOMST	32	1	0,000008%	0,0061%
80	INBESCHULDIGINGSTELLINGSFUNCTIE	31	1	0,000008%	0,0061%
81	INBOEDELVERZEKERINGSOVEREENKOMST	32	1	0,000008%	0,0061%
82	INDUSTRIEWATERVOORZIENINGSMAATSCHAPPIJ	38	2	0,000016%	0,0061%
83	INKOMSTENBELASTINGCONSEQUENTIES	31	1	0,000008%	0,0061%
84	INKOMSTENBELASTINGSCHADEVERGOEDING	34	1	0,000008%	0,0061%
85	INTERNATIONAALPRIVAATRECHTELIJK	31	5	0,000039%	0,0243%
86	INTERNATIONAALPRIVAATRECHTELIJKE	32	11	0,000086%	0,0548%
87	INVALIDITEITSPENSIOENBETALINGEN	31	1	0,000008%	0,0061%
88	INVALIDITEITSPENSIOENVOORZIENING	32	1	0,000008%	0,0061%
89	INVALIDITEITSVERZEKERINGSOVEREENKOMST	37	1	0,000008%	0,0061%
90	JEUGDMAATSCHAPPELIJKWERKDESKUNDIGEN	35	1	0,000008%	0,0061%
91	KAMPEERCENTRUMEXPLOITATIEVERGUNNING	35	24	0,000187%	0,0061%
92	KAMPEERCENTRUMEXPLOITATIEVERUNNING	34	1	0,000008%	0,0061%
93	KOSTWINNERSVERGOEDINGSBESCHIKKING	33	1	0,000008%	0,0061%
94	KREDIETVERGOEDINGSVOORSCHRIFTEN	31	1	0,000008%	0,0061%
95	KREDIETVERZEKERINGSAUTORITEITEN	31	1	0,000008%	0,0061%
96	KREISJUGENDWOHLFAHRTSAUSSCHUSSES	32	1	0,000008%	0,0061%
97	KWALIFICATIEGRENSOVERSCHRIJDING	31	1	0,000008%	0,0061%
98	LANDBOUWLUCHTVAARTONDERNEMINGEN	31	1	0,000008%	0,0061%
99	LEGITIMATIONSFESTSTELLUNGSBESCHLUSSES	37	1	0,000008%	0,0061%
100	LEVENSVERZEKERINGSMAATSCHAPPIJEN	32	4	0,000031%	0,0243%
101	LEVENSVERZEKERINGSOVEREENKOMSTEN	32	7	0,000055%	0,0122%
102	MEDEZEGGENSCHAPSAANGELEGENHEDEN	31	1	0,000008%	0,0061%
103	METAALOPPERVLAKTEVEREDELINGSBEDRIJF	35	1	0,000008%	0,0061%
104	MILCHFETTVERBILLIGUNGSVERORDNUNG	32	2	0,000016%	0,0122%
105	NEGENTIENHONDERDNEGENENZEVENTIG	31	1	0,000008%	0,0061%
106	OLIESTOOKVERWARMINGSINSTALLATIE	31	1	0,000008%	0,0061%
107	ONBEWOONBAARVERKLARINGSPROCEDURE	32	4	0,000031%	0,0061%
108	ONDERCURATELESTELLINGSPROCEDURE	31	1	0,000008%	0,0061%
109	ONGEVALLENVERZEKERINGSMAATSCHAPPIJ	34	3	0,000023%	0,0061%
110	ONGEVALLENVERZEKERINGSOVEREENKOMST	34	3	0,000023%	0,0122%
111	ONGEVALLENVERZEKERINGSOVEREENKOMSTEN	36	1	0,000008%	0,0061%
112	ONGEVALSVERZEKERINGSGENEESKUNDIGE	33	1	0,000008%	0,0061%
113	OPPERVLAKTEWATERVERONTREINIGING	31	1	0,000008%	0,0061%
114	OVERDRACHTSBELASTINGVERORDENING	31	8	0,000062%	0,0061%
115	PASSAGIERSACCOMMODATIEVERGOEDING	32	3	0,000023%	0,0061%
116	PENSIOENVERZEKERINGMAATSCHAPPIJ	31	3	0,000023%	0,0061%
117	PENSIOENVERZEKERINGSMAATSCHAPPIJ	32	3	0,000023%	0,0061%
118	PENSIOENVERZEKERINGSOVEREENKOMSTEN	34	4	0,000031%	0,0122%
119	PERSOONLIJKHEIDSVORMINGSMOGELIJKHEDEN	37	1	0,000008%	0,0061%
120	PLAATSVERVANGENDGROEPSCOMMANDANT	32	1	0,000008%	0,0061%
121	PRIORITEITSAANDEELHOUDERSVERGADERING	36	1	0,000008%	0,0061%
122	PRODUKTENAANSPRAKELIJKHEIDSSCHADE	33	1	0,000008%	0,0061%
123	PROJECTONTWIKKELINGSMAATSCHAPPIJ	32	4	0,000031%	0,0061%
124	PROJECTONTWIKKELINGSMAATSCHAPPIJEN	34	1	0,000008%	0,0061%
125	PROJECTVERGOEDINGSOVEREENKOMSTEN	32	18	0,000141%	0,0061%
126	PROJEKTONTWIKKELINGSMAATSCHAPPIJ	32	1	0,000008%	0,0061%
127	QUALIFICATIEGRENSOVERSCHRIJDING	31	2	0,000016%	0,0122%
128	RECHTSPERSOONLIJKHEIDSPROBLEMATIEK	34	1	0,000008%	0,0061%
129	REISBAGAGEVERZEKERINGMAATSCHAPPIJ	33	1	0,000008%	0,0061%
130	RIJKSLANDBOUWVOORLICHTINGSDIENST	32	4	0,000031%	0,0122%
131	RIOOLWATERZUIVERINGSINSTALLATIE	31	11	0,000086%	0,0183%
132	RIOOLWATERZUIVERINGSINSTALLATIES	32	2	0,000016%	0,0061%
133	RUILVERKAVELINGSVERWACHTINGSWAARDE	34	1	0,000008%	0,0061%
134	SAMENLEVINGSGOEDERENGEMEENSCHAP	31	2	0,000016%	0,0061%
135	SCHADELOOSSTELLINGSOVEREENKOMST	31	1	0,000008%	0,0061%
136	SCHADEVERGOEDINGSVERPLICHTINGEN	31	1	0,000008%	0,0061%
137	SCHADEVERZEKERINGSMAATSCHAPPIJEN	32	2	0,000016%	0,0122%
138	SCHEEPSVERZEKERINGSMAATSCHAPPIJ	31	2	0,000016%	0,0061%
139	SLACHTAFVALVERWERKINGSINSTALLATIE	33	1	0,000008%	0,0061%
140	STANDAARDARBEIDSVOORWAARDENOVEREENKOMST	39	1	0,000008%	0,0061%

Nr.	Woordtype	Lengte	Abs. freq.	Rel. freq.	In % doc.
141	STANDAARDVERZORGINGSOVEREENKOMST	32	1	0,000008%	0,0061%
142	STRASZENVERKEHRSSICHERUNGSPFLICHT	33	1	0,000008%	0,0061%
143	TERBESCHIKKINGSTELLINGSOVEREENKOMST	35	6	0,000047%	0,0122%
144	TRANSPORTVERZEKERINGMAATSCHAPPIJ	32	1	0,000008%	0,0061%
145	TRANSPORTVERZEKERINGSMAATSCHAPPIJ	33	3	0,000023%	0,0183%
146	TRANSPORTVERZEKERINGSOVEREENKOMST	33	1	0,000008%	0,0061%
147	TWEEDUIZENDVIERHONDERDZEVENENDERTIG	35	1	0,000008%	0,0061%
148	UNBEDENKLICHKEITSBESCHEINIGUNGEN	32	10	0,000078%	0,0183%
149	UNFALLVERSICHERUNGSGESELLSCHAFT	31	1	0,000008%	0,0061%
150	UNTERHALTSPFLICHTIGENVERBLEIBENDE	33	1	0,000008%	0,0061%
151	VANWAARDEVERKLARINGSDAGVAARDING	31	1	0,000008%	0,0061%
152	VENNOOTSCHAPSBELASTINGVERPLICHTINGEN	36	1	0,000008%	0,0061%
153	VERKEERSONGEVALLENAANSPRAKELIJKHEID	35	1	0,000008%	0,0061%
154	VERLIESCOMPENSATIEMOGELIJKHEDEN	31	1	0,000008%	0,0061%
155	VERPLICHTINGSTELLINGSBESCHIKKING	32	2	0,000016%	0,0122%
156	VERSICHERUNGSAKTIENGESELLSCHAFT	31	2	0,000016%	0,0122%
157	VERTEGENWOORDIGERSWERKZAAMHEDEN	31	2	0,000016%	0,0061%
158	VERTEGENWOORDIGHEIDSBEVOEGDHEID	31	2	0,000016%	0,0122%
159	VERTEGENWOORDIGINGSACTIVITEITEN	31	1	0,000008%	0,0061%
160	VERTEGENWOORDIGINGSBEVOEGDHEDEN	31	3	0,000023%	0,0061%
161	VERTEGENWOORDIGINGSMOGELIJKHEID	31	2	0,000016%	0,0122%
162	VERTEGENWOORDIGINGSOVEREENKOMST	31	8	0,000062%	0,0183%
163	VERTEGENWOORDIGINGSOVEREENKOMSTEN	33	2	0,000016%	0,0061%
164	VERTEGENWOORDIGINGSVERHOUDINGEN	31	1	0,000008%	0,0061%
165	VERVOERDERSAANSPRAKELIJKHEIDSVERZEKERING	40	1	0,000008%	0,0061%
166	VERVOERSAANSPRAKELIJKHEIDSVRAGEN	32	1	0,000008%	0,0061%
167	VIERDUIZENDHONDERDDRIEENVEERTIG	31	1	0,000008%	0,0061%
168	VIERDUIZENDVIJFHONDERDNEGENENZEVENTIG	37	1	0,000008%	0,0061%
169	VIERHONDERDACHTENVEERTIGDUIZEND	31	1	0,000008%	0,0061%
170	VIJFENVIJFTIGDUIZENDZEVENHONDERDZESTIG	38	1	0,000008%	0,0061%
171	VIJFHONDERDNEGENENDERTIGDUIZENDVIERHONDERDVIJF	46	1	0,000008%	0,0061%
172	VIJFHONDERDZEVENENVIJFTIGDUIZEND	32	1	0,000008%	0,0061%
173	VOETGANGERSOVERSTEEKPLAATSLICHT	31	1	0,000008%	0,0061%
174	VOLLSTRECKBARKEITSBESCHEINIGUNG	31	2	0,000016%	0,0122%
175	VORENOVERWOGENEDEONRECHTRAATIGHEID	34	1	0,000008%	0,0061%
176	WERKLOOSHEIDSVERZEKERINGSPREMIES	32	1	0,000008%	0,0061%
177	ZELFBEDIENINGSGROOTHANDELSBEDRIJVEN	35	1	0,000008%	0,0061%
178	ZESDUIZENDNEGENHONDERDNEGENENDERTIG	35	1	0,000008%	0,0061%
179	ZESDUIZENDNEGENHONDERDVIERENNEGENTIG	36	1	0,000008%	0,0061%
180	ZESMILJOENNEGENHONDERDZESENNEGENTIGDUIZENDVIJF-HONDERDENELF	58	1	0,000008%	0,0061%
181	ZEVENHONDERDDRIEENDERTIGDUIZEND	31	1	0,000008%	0,0061%
182	ZIEKTEKOSTENVERZEKERINGSBEDRIJF	31	1	0,000008%	0,0061%
183	ZIEKTEKOSTENVERZEKERINGSFONDSEN	31	2	0,000016%	0,0061%
184	ZIEKTEKOSTENVERZEKERINGSOVEREENKOMST	36	1	0,000008%	0,0061%

Corpus Algemeen Nederlands

Nr.	Woordtype	Lengte	Abs. freq.	Rel. freq.	In % doc.
1	AANSPRAKELIJKHEIDSVERZEKERINGEN	31	1	0,000008%	0,2463%
2	ARBEIDSMARKTINFORMATIEVOORZIENING	33	1	0,000008%	0,2463%
3	ARBEIDSONGESCHIKTHEIDSUITKERING	31	2	0,000016%	0,2463%
4	ARBEIDSONGESCHIKTHEIDSVERZEKERING	33	6	0,000047%	0,7389%
5	ARBEIDSONGESCHIKTHEIDSVOORZIENINGEN	35	1	0,000008%	0,2463%
6	ARBEIDSPLAATSKWALITEITSVERBETERENDE	35	1	0,000008%	0,2463%
7	ARBEIDSPLAATSOVEREENKOMSTENSTRATEGIE	36	1	0,000008%	0,2463%
8	ARBEIDSVOORWAARDENVERBETERINGEN	31	1	0,000008%	0,2463%
9	AUSBILDUNGSPLATZFöRDERUNGSGESETZ	32	1	0,000008%	0,2463%
10	AUTOMATISERINGSSAMENWERKINGSVERBAND	35	1	0,000008%	0,2463%
11	BEDRIJFSVEILIGHEIDSFUNKTIONARISSEN	34	1	0,000008%	0,2463%
12	BEENMERGTRANSPLANTATIEPROGRAMMA	31	1	0,000008%	0,2463%
13	BEJAARDENZIEKENFONDSVERZEKERING	31	1	0,000008%	0,2463%
14	BELANGENBEHARTIGINGSORGANISATIES	32	3	0,000023%	0,4926%
15	BRANDSTOFKOSTENBEHEERVERGOEDING	31	1	0,000008%	0,2463%
16	CREDIETVERZEKERINGSMAATSCHAPPIJ	31	1	0,000008%	0,2463%
17	DRINKWATERLEIDINGMAATSCHAPPIJEN	31	1	0,000008%	0,2463%
18	EMANCIPATIESUBSIDIEMOGELIJKHEDEN	32	1	0,000008%	0,2463%
19	GEWELDUITOEFENINGSMOGELIJKHEDEN	31	1	0,000008%	0,2463%
20	GEZONDHEIDSVOORLICHTINGSORGANISATIE	35	1	0,000008%	0,2463%
21	HAEMAGGLUTINATIEREMMINGSREACTIE	31	1	0,000008%	0,2463%
22	HANDELSBESCHERMINGSOVERWEGINGEN	31	1	0,000008%	0,2463%
23	HERSTRUCTURERINGSMAATSCHAPPIJEN	31	1	0,000008%	0,2463%
24	HUWELIJKSVRUCHTBAARHEIDSCIJFERS	31	2	0,000016%	0,4926%
25	INFORMATIEVERSTREKKINGSPROBLEEM	31	1	0,000008%	0,2463%
26	INFORMATIEVERWERKINGSFACILITEITEN	33	1	0,000008%	0,2463%
27	INKOMENSREALISATIEMOGELIJKHEDEN	31	1	0,000008%	0,2463%
28	INVALIDITEITSPENSIOENGERECHTIGDEN	33	2	0,000016%	0,2463%
29	KANDIDAATVOLKSVERTEGENWOORDIGER	31	1	0,000008%	0,2463%
30	KINDERBESCHERMINGSGROEPSLEIDSTER	32	1	0,000008%	0,2463%
31	KONSTITUENTENSTRUKTUURGRAMMATIKA	32	1	0,000008%	0,2463%
32	LEVENSVERZEKERINGSMAATSCHAPPIJEN	32	1	0,000008%	0,2463%
33	LUCHTVERONTREINIGINGSPROBLEMATIEK	33	1	0,000008%	0,2463%
34	MARKTVERZADIGINGSVERSCHIJNSELEN	31	1	0,000008%	0,2463%
35	NEEMDANWATDRUPPELSVOORDEZENUWENVANDEREFORM	42	1	0,000008%	0,2463%
36	ONDERNEMERSVERANTWOORDELIJKHEID	31	3	0,000023%	0,4926%
37	ONGEVALLENINZITTENDENVERZEKERING	32	2	0,000016%	0,2463%
38	OPLEIDINGSEMANCIPATIEVERSCHIJNSELEN	35	1	0,000008%	0,2463%
39	OPPERVLAKTEBEHANDELINGSINDUSTRIE	32	1	0,000008%	0,2463%
40	OPPERVLAKTEWATERVERONTREINIGING	31	3	0,000023%	0,2463%
41	PERSONEELFUNCTIONARISSENDICHTHEID	33	1	0,000008%	0,2463%
42	PERSONEELPLANNINGSDOELSTELLINGEN	32	1	0,000008%	0,2463%
43	PERSONEELPLANNINGSPROGRAMMATUUR	31	1	0,000008%	0,2463%
44	PERSONEELVOORZIENINGSMOGELIJKHEDEN	34	1	0,000008%	0,2463%
45	PERSONEELVOORZIENINGSVOORSTELLEN	32	1	0,000008%	0,2463%
46	PLANTENZIEKTEBESTRIJDINGSMIDDELEN	33	1	0,000008%	0,2463%
47	PROFESSIONALISERINGSVERSCHIJNSEL	32	2	0,000016%	0,2463%
48	PROJECTONTWIKKELINGSMAATSCHAPPIJ	32	3	0,000023%	0,2463%
49	PROJECTONTWIKKELINGSMAATSCHAPPIJEN	34	1	0,000008%	0,2463%
50	RIJKSLANDBOUWVOORLICHTINGSDIENST	32	1	0,000008%	0,2463%
51	RIOOLWATERZUIVERINGSINSTALLATIE	31	1	0,000008%	0,2463%
52	RIOOLWATERZUIVERINGSINSTALLATIES	32	1	0,000008%	0,2463%
53	SAMENLEVINGSVERANTWOORDELIJKHEID	32	1	0,000008%	0,2463%
54	TELECOMMUNICATIEDIENSTVERLENINGEN	33	1	0,000008%	0,2463%
55	VENNOOTSCHAPSBELASTINGPLICHTIGE	31	1	0,000008%	0,2463%
56	VERANTWOORDELIJKHEIDSONTWIKKELING	33	1	0,000008%	0,2463%
57	WERKGEVERSBEDRIJFSTAKORGANISATIE	32	1	0,000008%	0,2463%

BIJLAGE 3 - FREQUENTIELIJSTEN (EERSTE 300 RANGNUMMERS)

In het nu volgende gedeelte zijn de 300 woordtypen met de hoogste frequenties uit elk van de drie (volledige) corpora opgenomen. De lijst van elk corpus beslaat daarbij telkens 2 naast elkaar gelegen pagina's.

Corpus wetgevingsteksten

Rangnr.	Woordtype	Abs. freq.	Rel. freq.	In % doc.
1	DE	899333	7,0218%	98,21%
2	VAN	759046	5,9265%	98,79%
3	HET	437151	3,4132%	95,44%
4	IN	316151	2,4685%	94,82%
5	EN	285753	2,2311%	94,13%
6	EEN	236772	1,8487%	85,03%
7	OF	157635	1,2308%	79,06%
8	VOOR	140172	1,0944%	83,18%
9	OP	129700	1,0127%	83,41%
10	IS	115722	0,9035%	80,69%
11	MET	114068	0,8906%	85,03%
12	TE	108198	0,8448%	76,56%
13	ART	101843	0,7952%	82,09%
14	AAN	98641	0,7702%	78,99%
15	LID	96108	0,7504%	66,98%
16	ARTIKEL	92743	0,7241%	70,87%
17	WORDEN	92634	0,7233%	80,42%
18	DIE	91737	0,7163%	73,78%
19	1	86409	0,6747%	83,88%
20	WORDT	85716	0,6693%	74,58%
21	BIJ	85185	0,6651%	78,48%
22	ZIJN	85099	0,6644%	69,48%
23	DOOR	82302	0,6426%	69,13%
24	DAT	79173	0,6182%	73,56%
25	TOT	75861	0,5923%	70,90%
26	2	71135	0,5554%	80,56%
27	NIET	59673	0,4659%	62,37%
28	DEZE	59344	0,4633%	64,89%
29	ALS	56037	0,4375%	63,87%
30	WET	53805	0,4201%	49,76%
31	EERSTE	52825	0,4124%	58,17%
32	MINISTER	52053	0,4064%	55,75%
33	TEN	51294	0,4005%	59,97%
34	BEDOELD	46458	0,3627%	54,01%
35	DAN	46330	0,3617%	54,27%
36	INDIEN	45425	0,3547%	52,70%
37	DER	45375	0,3543%	50,19%
38	3	44630	0,3485%	68,25%
39	ONDER	41964	0,3276%	54,91%
40	DIT	37869	0,2957%	52,96%
41	KAN	37854	0,2956%	53,71%
42	A	37705	0,2944%	52,92%
43	ONZE	36202	0,2827%	38,62%
44	GEMEENTE	35580	0,2778%	14,38%
45	B	34734	0,2712%	51,01%
46	BESLUIT	34240	0,2673%	39,38%
47	TOEPASSING	32289	0,2521%	48,01%
48	4	29428	0,2298%	57,49%
49	TWEEDE	28855	0,2253%	47,35%
50	STB	27163	0,2121%	37,93%
51	HEEFT	26615	0,2078%	41,01%
52	WELKE	25900	0,2022%	41,84%
53	NA	25639	0,2002%	46,51%
54	BEPAALDE	25407	0,1984%	42,91%
55	DEN	25255	0,1972%	10,89%
56	BEDOELDE	24786	0,1935%	38,95%
57	TER	23536	0,1838%	42,69%
58	HEBBEN	23325	0,1821%	49,26%
59	UIT	23103	0,1804%	41,00%
60	C	22676	0,1771%	42,04%
61	HAND	22040	0,1721%	15,95%
62	5	21918	0,1711%	48,05%
63	RAAD	21487	0,1678%	31,17%
64	ZIJ	21016	0,1641%	37,42%
65	ANDERE	20277	0,1583%	39,11%
66	NAAR	19604	0,1531%	33,86%
67	I	19234	0,1502%	26,47%
68	HIJ	18875	0,1474%	26,60%
69	BETREKKING	18764	0,1465%	34,65%
70	II	18535	0,1447%	22,58%
71	KRACHTENS	18462	0,1441%	31,99%
72	ZAL	17892	0,1397%	39,55%
73	MEER	17316	0,1352%	32,39%
74	MOET	17247	0,1347%	24,95%
75	OVER	17161	0,1340%	31,81%

Rangnr.	Woordtype	Abs. freq.	Rel. freq.	In % doc.
76	F	16944	0,1323%	26,49%
77	6	16760	0,1309%	40,53%
78	KUNNEN	16703	0,1304%	35,19%
79	ALGEMENE	16466	0,1286%	30,65%
80	ZAKEN	15796	0,1233%	30,18%
81	ARTIKELEN	15670	0,1223%	34,53%
82	DERDE	15338	0,1198%	32,90%
83	DAG	14794	0,1155%	30,38%
84	BINNEN	14727	0,1150%	28,59%
85	ENZ	14426	0,1126%	24,35%
86	WEL	14346	0,1120%	30,94%
87	GEEN	14079	0,1099%	31,42%
88	BEDRAG	13741	0,1073%	18,39%
89	NR	13471	0,1052%	15,74%
90	OM	13441	0,1049%	28,94%
91	WAAROP	13385	0,1045%	30,29%
92	DATUM	13322	0,1040%	20,45%
93	ZOVER	13291	0,1038%	29,04%
94	BESTUUR	13234	0,1033%	21,30%
95	AANZIEN	13057	0,1019%	30,00%
96	JAAR	13041	0,1018%	22,23%
97	D	12920	0,1009%	29,08%
98	7	12879	0,1006%	34,98%
99	HEM	12809	0,1000%	24,84%
100	INGEVOLGE	12604	0,0984%	26,58%
101	LEDEN	12523	0,0978%	18,69%
102	WIJ	12497	0,0976%	27,06%
103	WAARIN	12402	0,0968%	33,65%
104	GEMEENTEN	12281	0,0959%	9,68%
105	GRENS	12236	0,0955%	3,55%
106	ZICH	12127	0,0947%	26,99%
107	VERSTAAN	12045	0,0940%	35,20%
108	10	11689	0,0913%	29,64%
109	DAARVAN	11606	0,0906%	30,57%
110	ALLE	11603	0,0906%	31,27%
111	8	11282	0,0881%	31,70%
112	BEPALINGEN	11100	0,0867%	29,51%
113	PLAATS	11014	0,0860%	25,81%
114	GROND	10920	0,0853%	24,37%
115	ALSMEDE	10862	0,0848%	29,56%
116	SECTIE	10813	0,0844%	2,31%
117	WAARVAN	10749	0,0839%	29,86%
118	MOETEN	10516	0,0821%	18,35%
119	HUN	10481	0,0818%	22,17%
120	VERZOEK	10437	0,0815%	19,34%
121	JANUARI	10253	0,0801%	18,92%
122	STELLEN	10209	0,0797%	27,02%
123	WERKING	10163	0,0794%	31,95%
124	BEHANDELING	9988	0,0780%	18,71%
125	PER	9974	0,0779%	16,55%
126	STATENGENERAAL	9932	0,0775%	13,71%
127	COMMISSIE	9911	0,0774%	10,08%
128	UITVOERING	9899	0,0773%	32,23%
129	STATE	9894	0,0773%	22,15%
130	DIENST	9809	0,0766%	13,15%
131	HOOFDSTUK	9772	0,0763%	27,11%
132	E	9745	0,0761%	22,79%
133	GEBIED	9682	0,0756%	10,71%
134	BESCHIKKING	9590	0,0749%	17,37%
135	TUSSEN	9523	0,0744%	15,86%
136	BEKEND	9518	0,0743%	11,05%
137	INGANG	9513	0,0743%	26,12%
138	GENOEMDE	9436	0,0737%	23,13%
139	REGELING	9433	0,0737%	17,72%
140	PERSONEN	9393	0,0733%	18,02%
141	GEVAL	9334	0,0729%	23,90%
142	9	9326	0,0728%	27,41%
143	TIJDSTIP	9322	0,0728%	20,49%
144	ADVIES	9292	0,0726%	16,74%
145	VOORSCHRIFTEN	9274	0,0724%	19,81%
146	STAATSBLAD	9252	0,0722%	26,55%
147	GEGEVEN	9231	0,0721%	29,06%
148	VORIGE	9166	0,0716%	19,49%
149	RECHT	9135	0,0713%	15,26%
150	WIJZE	9093	0,0710%	24,61%

Corpus wetgevingsteksten

Rangnr.	Woordtype	Abs. freq.	Rel. freq.	In % doc.
151	BEVOEGD	9066	0,0708%	17,89%
152	GENOMEN	8950	0,0699%	28,16%
153	DOEN	8924	0,0697%	27,12%
154	DECEMBER	8819	0,0689%	17,61%
155	AANGEWEZEN	8789	0,0686%	20,52%
156	KOSTEN	8725	0,0681%	15,42%
157	VERLEEND	8715	0,0680%	19,17%
158	ONDERWIJS	8688	0,0678%	8,57%
159	OVEREENKOMSTIGE	8624	0,0673%	22,90%
160	MINSTE	8537	0,0667%	16,10%
161	OVEREENKOMSTIG	8536	0,0666%	21,79%
162	GEPLAATST	8504	0,0664%	30,81%
163	VASTGESTELD	8497	0,0663%	23,08%
164	12	8441	0,0659%	24,46%
165	HOOGSTE	8431	0,0658%	17,11%
166	GESTELD	8409	0,0657%	22,56%
167	JUSTITIE	8250	0,0644%	26,04%
168	TWEE	8212	0,0641%	21,52%
169	AMBTENAREN	8200	0,0640%	19,89%
170	15	8145	0,0636%	22,29%
171	GEHOORD	8128	0,0635%	30,01%
172	WAAR	8100	0,0632%	15,91%
173	11	8029	0,0627%	24,01%
174	AANTAL	8013	0,0626%	14,18%
175	ZIE	7957	0,0621%	12,24%
176	ZO	7955	0,0621%	24,64%
177	STATEN	7953	0,0621%	9,72%
178	ZULLEN	7940	0,0620%	20,07%
179	ONDERSCHEIDENLIJK	7901	0,0617%	16,06%
180	PENSIOEN	7881	0,0615%	5,16%
181	JULI	7836	0,0612%	16,35%
182	BETROKKEN	7659	0,0598%	16,85%
183	DAARBIJ	7651	0,0597%	22,45%
184	GEBRUIK	7566	0,0591%	16,33%
185	BEROEP	7551	0,0590%	11,41%
186	PERCEEL	7545	0,0589%	2,39%
187	GEDURENDE	7539	0,0589%	18,41%
188	BIJL	7412	0,0579%	5,78%
189	GOEDEREN	7409	0,0578%	7,54%
190	GEMAAKT	7392	0,0577%	20,74%
191	NIEUWE	7385	0,0577%	10,05%
192	WE	7348	0,0574%	23,71%
193	TREEDT	7344	0,0573%	29,64%
194	TEGEN	7328	0,0572%	18,13%
195	REGELEN	7301	0,0570%	17,76%
196	BIJZONDERE	7297	0,0570%	17,45%
197	DAARTOE	7292	0,0569%	22,02%
198	NEDERLANDSE	7248	0,0566%	18,87%
199	MAATREGEL	7241	0,0565%	13,53%
200	13	7170	0,0560%	21,66%
201	N	7116	0,0556%	8,56%
202	VIERDE	7082	0,0553%	19,24%
203	VOORZITTER	6980	0,0545%	8,88%
204	TERMIJN	6975	0,0545%	15,26%
205	STAAT	6963	0,0544%	15,80%
206	VOLGENDE	6915	0,0540%	21,42%
207	14	6912	0,0540%	21,27%
208	MAG	6851	0,0535%	12,75%
209	BLZ	6846	0,0535%	7,21%
210	ONDERZOEK	6839	0,0534%	11,57%
211	MEDE	6837	0,0534%	20,61%
212	KADASTRAAL	6810	0,0532%	1,84%
213	UITKERING	6805	0,0531%	6,49%
214	BEPALEN	6792	0,0530%	21,03%
215	BEPAALD	6704	0,0523%	20,16%
216	BURGEMEESTER	6689	0,0522%	8,17%
216	DIEN	6689	0,0522%	19,30%
218	SLECHTS	6654	0,0520%	18,88%
219	BETREFT	6628	0,0518%	17,51%
220	HAAR	6559	0,0512%	14,53%
221	RICHTING	6556	0,0512%	2,49%
222	G	6522	0,0509%	16,47%
223	KENNIS	6481	0,0506%	15,70%
224	20	6461	0,0504%	18,67%
225	OMTRENT	6456	0,0504%	17,50%

Rangnr.	Woordtype	Abs. freq.	Rel. freq.	In % doc.
226	MAANDEN	6452	0,0504%	15,16%
227	BETREFFENDE	6426	0,0502%	18,60%
228	VOLGENS	6423	0,0501%	16,53%
229	VERVALLEN	6397	0,0499%	13,79%
230	J	6376	0,0498%	20,14%
231	KAMERSTUKKEN	6311	0,0493%	5,75%
231	WANNEER	6311	0,0493%	14,09%
233	16	6305	0,0492%	19,14%
234	GEVEN	6281	0,0490%	17,97%
235	WIJZIGING	6265	0,0489%	15,11%
236	JAREN	6240	0,0487%	13,42%
237	MOGELIJK	6213	0,0485%	17,37%
238	SOCIALE	6208	0,0485%	9,84%
239	VERGUNNING	6199	0,0484%	6,94%
240	GEVALLEN	6133	0,0479%	17,56%
241	ZONDER	6111	0,0477%	17,77%
242	DAGEN	6064	0,0473%	14,10%
243	GEDEPUTEERDE	6051	0,0472%	7,34%
244	GEZAG	6040	0,0472%	8,21%
245	BESLISSING	6006	0,0469%	11,97%
246	AMBTENAAR	6004	0,0469%	7,82%
247	GEACHT	5965	0,0466%	15,58%
248	GEGEVENS	5954	0,0465%	11,58%
249	18	5933	0,0463%	18,11%
250	GESCHIEDT	5917	0,0462%	17,99%
251	DIENT	5912	0,0462%	14,41%
252	BLADZ	5911	0,0462%	6,42%
253	VAST	5863	0,0458%	17,85%
254	DEEL	5844	0,0456%	15,00%
255	M	5818	0,0454%	11,13%
256	VERGOEDING	5813	0,0454%	9,95%
257	GELEGEN	5803	0,0453%	12,19%
258	OVERLEG	5791	0,0452%	19,12%
259	GELIJK	5788	0,0452%	20,57%
260	JUNI	5724	0,0447%	13,96%
261	ONS	5680	0,0443%	14,71%
261	OPENBAAR	5680	0,0443%	12,50%
263	GEDEELTE	5662	0,0442%	11,65%
264	TEVENS	5654	0,0441%	18,59%
265	VERBAND	5637	0,0440%	16,86%
266	17	5620	0,0439%	18,10%
267	31	5566	0,0435%	15,10%
268	ELK	5565	0,0435%	15,41%
269	TENZIJ	5546	0,0433%	16,75%
270	30	5545	0,0433%	16,34%
271	AFDELING	5531	0,0432%	15,06%
272	BIJLAGE	5520	0,0431%	9,46%
273	H	5519	0,0431%	16,61%
274	BUITEN	5508	0,0430%	15,56%
275	BEHOEVE	5503	0,0430%	15,37%
276	ZODANIG	5495	0,0429%	15,49%
277	DRIE	5491	0,0429%	15,40%
278	AFSCHRIFT	5483	0,0428%	19,10%
279	MEI	5476	0,0428%	13,01%
280	GEDAAN	5455	0,0426%	15,44%
281	ALGEMEEN	5444	0,0425%	13,30%
282	SCHIP	5443	0,0425%	3,46%
283	PUNT	5441	0,0425%	3,79%
284	APRIL	5406	0,0422%	13,51%
285	ANDER	5390	0,0421%	16,50%
286	WERKZAAMHEDEN	5337	0,0417%	11,02%
287	NEDERLANDEN	5334	0,0416%	24,26%
288	DEGENE	5333	0,0416%	12,23%
289	25	5329	0,0416%	16,82%
290	GESTELDE	5286	0,0413%	15,08%
291	HOUDENDE	5276	0,0412%	20,03%
292	VERPLICHT	5270	0,0411%	12,26%
293	SCHRIFTELIJK	5236	0,0409%	13,74%
294	OOK	5208	0,0407%	16,33%
295	ZAKE	5207	0,0407%	14,26%
296	ZULKS	5180	0,0404%	20,76%
297	KAMER	5163	0,0403%	6,02%
298	BEDRAAGT	5158	0,0403%	12,16%
299	VASTSTELLING	5103	0,0398%	15,00%
300	GOEDKEURING	5093	0,0398%	10,95%

Corpus jurisprudentieteksten

Rangnr.	Woordtype	Abs. freq.	Rel. freq.	In % doc.
1	DE	2597537	6,4631%	97,57%
2	VAN	1957452	4,8705%	98,59%
3	HET	1406008	3,4984%	95,90%
4	IN	986401	2,4543%	96,23%
5	DAT	917971	2,2841%	95,02%
6	EN	707918	1,7614%	96,06%
7	EEN	649203	1,6153%	94,97%
8	TE	587032	1,4606%	97,79%
9	IS	507448	1,2626%	95,02%
10	OP	476258	1,1850%	94,69%
11	NIET	433890	1,0796%	94,56%
12	HEEFT	374749	0,9324%	92,90%
13	AAN	303584	0,7554%	92,58%
14	DOOR	301730	0,7508%	92,95%
15	ZIJN	300133	0,7468%	92,81%
16	MET	288083	0,7168%	92,26%
17	VOOR	283602	0,7056%	91,83%
18	DIE	252974	0,6294%	90,47%
19	TOT	241589	0,6011%	92,41%
20	ALS	236583	0,5887%	91,51%
21	BIJ	222475	0,5536%	91,55%
22	WORDEN	219272	0,5456%	92,99%
23	ART	197784	0,4921%	84,39%
24	DEZE	191315	0,4760%	90,10%
25	OF	187144	0,4656%	86,74%
26	HOF	174874	0,4351%	82,40%
27	DIT	149037	0,3708%	86,53%
28	RB	144535	0,3596%	74,47%
29	OOK	141002	0,3508%	85,34%
30	TEN	139606	0,3474%	87,78%
31	ZOU	124392	0,3095%	79,65%
32	KAN	117889	0,2933%	85,93%
33	OM	117551	0,2925%	83,85%
34	UIT	113632	0,2828%	83,38%
35	GEEN	109878	0,2734%	84,36%
36	NAAR	108143	0,2691%	82,33%
37	DAN	107088	0,2665%	81,68%
38	1	106313	0,2645%	80,97%
39	WAS	102549	0,2552%	74,25%
40	HAAR	101906	0,2536%	68,44%
41	BEROEP	101386	0,2523%	73,25%
42	ZICH	101142	0,2517%	83,40%
43	ONDER	95841	0,2380%	75,87%
44	HIJ	94897	0,2361%	72,90%
45	HEBBEN	91092	0,2267%	77,46%
46	ZIJ	90443	0,2250%	73,79%
47	TEGEN	88473	0,2201%	91,62%
48	WORDT	87868	0,2186%	77,74%
49	2	81905	0,2038%	77,30%
50	MIDDEL	76793	0,1911%	51,94%
51	TER	76085	0,1893%	77,72%
52	DER	75763	0,1885%	75,02%
53	LID	75586	0,1881%	62,51%
54	CASSATIE	74482	0,1853%	49,87%
55	F	73116	0,1819%	55,91%
56	VONNIS	68683	0,1709%	62,64%
57	KUNNEN	68493	0,1704%	74,25%
58	WELKE	66580	0,1657%	74,69%
59	HAD	65984	0,1642%	66,33%
60	MOET	65141	0,1621%	77,36%
61	IK	64407	0,1603%	50,30%
62	OVER	60925	0,1516%	68,44%
63	RECHT	60777	0,1512%	73,73%
64	ER	60732	0,1511%	69,60%
65	3	60440	0,1504%	70,83%
66	A	59341	0,1477%	68,20%
67	ZAL	57499	0,1431%	69,34%
68	GEVAL	57025	0,1419%	70,34%
69	GROND	57006	0,1418%	71,16%
70	EERSTE	56442	0,1404%	67,82%
71	O	56221	0,1399%	59,51%
72	ARREST	55966	0,1393%	45,83%
73	P	55786	0,1388%	47,83%
74	HEM	54346	0,1352%	65,43%
75	NOG	54028	0,1344%	69,55%

Rangnr.	Woordtype	Abs. freq.	Rel. freq.	In % doc.
76	MEER	53479	0,1331%	67,43%
77	NU	53478	0,1331%	74,96%
78	WEL	53307	0,1326%	70,19%
79	WET	52320	0,1302%	58,78%
80	NA	50989	0,1269%	67,81%
81	PP	50791	0,1264%	50,23%
82	TUSSEN	50110	0,1247%	62,45%
83	OORDEEL	49750	0,1238%	64,18%
84	NJ	48616	0,1210%	99,91%
85	VORDERING	48171	0,1199%	48,70%
86	4	46054	0,1146%	64,89%
87	BESCHIKKING	46017	0,1145%	35,23%
88	MR	45687	0,1137%	90,01%
89	GESTELD	45461	0,1131%	68,20%
90	VERDACHTE	44389	0,1104%	23,69%
91	HR	43626	0,1085%	48,76%
92	BESLISSING	43426	0,1081%	58,56%
93	OMDAT	42615	0,1060%	62,33%
94	B	42164	0,1049%	56,21%
95	MAAR	41547	0,1034%	58,15%
96	ANDERE	41311	0,1028%	61,91%
97	ZOALS	39692	0,0988%	63,24%
98	5	38859	0,0967%	61,74%
99	HOGER	38682	0,0962%	56,04%
100	MOETEN	38491	0,0958%	61,97%
101	BELANG	38355	0,0954%	57,86%
102	STAAT	38332	0,0954%	55,16%
103	GEGEVEN	38261	0,0952%	60,90%
104	WERD	37950	0,0944%	59,41%
105	VRAAG	37795	0,0940%	53,76%
106	INDIEN	37341	0,0929%	59,91%
107	BESTREDEN	37047	0,0922%	54,81%
108	ZONDER	36575	0,0910%	59,11%
109	SLECHTS	36484	0,0908%	62,48%
110	ONDERDEEL	36343	0,0904%	27,74%
111	6	36309	0,0903%	58,73%
112	ZAAK	35966	0,0895%	56,07%
113	DERHALVE	35892	0,0893%	68,12%
114	RO	35803	0,0891%	24,00%
115	ONDERHAVIGE	35279	0,0878%	59,28%
116	RAAD	35019	0,0871%	54,67%
117	OMSTANDIGHEDEN	34446	0,0857%	51,85%
118	WAARVAN	34194	0,0851%	64,00%
119	GEMEENTE	33963	0,0845%	17,35%
120	HIER	33759	0,0840%	56,73%
121	CONCLUSIE	33559	0,0835%	58,10%
122	ALTHANS	33540	0,0835%	50,08%
123	NR	33141	0,0825%	72,67%
124	DOEN	32955	0,0820%	60,18%
125	AL	32834	0,0817%	57,12%
126	REEDS	32168	0,0800%	56,60%
127	HETGEEN	31919	0,0794%	58,28%
128	S	31537	0,0785%	44,23%
129	DOCH	30841	0,0767%	59,91%
130	BEDOELD	30732	0,0765%	57,04%
131	ZIN	30531	0,0760%	50,64%
132	WAT	30490	0,0759%	53,91%
133	TWEEDE	30267	0,0753%	52,73%
134	GRIEF	29996	0,0746%	30,65%
135	HUN	29805	0,0742%	43,51%
136	DAARVAN	29216	0,0727%	57,38%
137	FEITEN	29204	0,0727%	48,30%
138	BETREKKING	28913	0,0719%	52,99%
139	SUB	28831	0,0717%	38,04%
140	ECHTER	28718	0,0715%	59,18%
141	OVEREENKOMST	28569	0,0711%	22,22%
142	BEDOELDE	28327	0,0705%	52,04%
143	GENOEMDE	28210	0,0702%	55,47%
144	ONRECHTE	27978	0,0696%	49,57%
145	IMMERS	27942	0,0695%	59,26%
146	RECHTER	27833	0,0693%	40,69%
147	FEIT	27827	0,0692%	50,37%
148	ZODAT	27723	0,0690%	61,33%
149	MEDE	27693	0,0689%	54,14%
150	BLIJKT	27535	0,0685%	56,23%

Corpus jurisprudentieteksten

Rangnr.	Woordtype	Abs. freq.	Rel. freq.	In % doc.
151	GEDING	27179	0,0676%	44,59%
152	JUNI	27173	0,0676%	51,66%
153	7	27135	0,0675%	52,74%
154	TERWIJL	27027	0,0672%	57,80%
155	BW	26826	0,0667%	29,85%
156	JAN	26527	0,0660%	46,25%
157	KOSTEN	26142	0,0650%	37,26%
158	ENOF	25903	0,0645%	40,64%
159	VERZOEK	25819	0,0642%	34,32%
160	8	25811	0,0642%	48,62%
161	KTR	25790	0,0642%	18,56%
162	ZO	25748	0,0641%	52,05%
163	ALDUS	25653	0,0638%	51,70%
164	ZOVER	25425	0,0633%	45,86%
165	WAARIN	25225	0,0628%	55,21%
166	MAART	25172	0,0626%	48,35%
167	MAN	25142	0,0626%	13,03%
168	EISER	24971	0,0621%	21,70%
169	GEBRUIK	24791	0,0617%	32,59%
170	AANZIEN	24737	0,0615%	54,12%
171	BEDRAG	24680	0,0614%	29,29%
172	MEI	24648	0,0613%	48,16%
173	VROUW	24604	0,0612%	11,89%
174	VERBAND	24588	0,0612%	48,12%
175	C	24501	0,0610%	47,44%
176	MIJ	24480	0,0609%	41,56%
177	J	24450	0,0608%	53,32%
178	ZULKS	24403	0,0607%	54,29%
179	10	24395	0,0607%	46,71%
180	GEDAAGDE	24359	0,0606%	24,14%
181	VOORTS	23961	0,0596%	54,66%
182	OVERWOGEN	23751	0,0591%	49,38%
183	THANS	23704	0,0590%	50,66%
184	PER	23524	0,0585%	30,95%
185	VERKLAARD	23485	0,0584%	51,81%
186	THE	23389	0,0582%	5,10%
187	BLIJKENS	23304	0,0580%	53,91%
188	OMTRENT	23175	0,0577%	47,44%
189	WAARBIJ	23156	0,0576%	57,52%
190	KOMEN	22944	0,0571%	49,69%
191	PLAATS	22934	0,0571%	49,33%
192	VERWEER	22891	0,0570%	39,33%
193	UITSPRAAK	22822	0,0568%	44,49%
194	TOEPASSING	22744	0,0566%	43,77%
195	APRIL	22708	0,0565%	45,39%
196	VOLGENS	22635	0,0563%	47,63%
197	DEC	22626	0,0563%	44,21%
198	NOV	22516	0,0560%	44,32%
199	SCHADE	22201	0,0552%	19,48%
200	ANDER	22094	0,0550%	48,80%
200	BEPAALDE	22094	0,0550%	49,31%
202	VERKLARING	22063	0,0549%	31,31%
203	AMSTERDAM	22006	0,0548%	35,59%
204	MEN	22003	0,0547%	37,14%
205	WAAROP	21989	0,0547%	51,86%
206	ZIE	21931	0,0546%	35,91%
207	9	21841	0,0543%	45,11%
208	WAREN	21777	0,0542%	40,23%
209	BETREFT	21692	0,0540%	48,37%
210	STRIJD	21613	0,0538%	40,87%
211	WIJZE	21360	0,0531%	45,97%
212	ANDERS	21331	0,0531%	49,81%
213	OKT	21255	0,0529%	41,75%
214	DIENT	21163	0,0527%	50,41%
215	EISERES	21153	0,0526%	20,41%
216	ALLE	21099	0,0525%	44,95%
217	DUS	20996	0,0522%	45,78%
218	ALLEEN	20984	0,0522%	47,50%
219	MAKEN	20974	0,0522%	45,11%
220	DD	20908	0,0520%	47,92%
221	DAAR	20706	0,0515%	49,45%
222	12	20593	0,0512%	42,33%
223	KOMT	20513	0,0510%	49,12%
224	STELLEN	20432	0,0508%	46,90%
225	SPRAKE	20398	0,0508%	44,77%

Rangnr.	Woordtype	Abs. freq.	Rel. freq.	In % doc.
226	WEG	20250	0,0504%	35,35%
227	VOLGT	20248	0,0504%	53,28%
228	KON	20245	0,0504%	41,69%
229	TWEE	20028	0,0498%	45,61%
230	FEBR	19905	0,0495%	40,79%
231	GEMAAKT	19790	0,0492%	46,18%
232	TOEN	19701	0,0490%	40,85%
233	GEHEEL	19575	0,0487%	43,51%
234	NEMEN	19515	0,0486%	44,98%
235	ACHT	19247	0,0479%	47,78%
236	GEDAAN	19233	0,0479%	46,17%
237	DAARBIJ	19207	0,0478%	47,66%
238	GRONDSLAG	19195	0,0478%	40,17%
239	FEITELIJKE	19142	0,0476%	36,97%
240	I	19053	0,0474%	35,81%
241	ENZ	19025	0,0473%	64,41%
242	REQ	18895	0,0470%	10,80%
243	11	18820	0,0468%	40,86%
244	HOFS	18810	0,0468%	26,02%
245	ZELF	18713	0,0466%	43,16%
246	GRONDEN	18706	0,0465%	41,90%
247	1983	18674	0,0465%	24,71%
248	AF	18651	0,0464%	46,98%
249	STELLING	18585	0,0462%	37,15%
250	1981	18572	0,0462%	24,43%
251	ZAKE	18483	0,0460%	45,16%
252	GENOMEN	18444	0,0459%	43,88%
253	WAAR	18433	0,0459%	42,73%
254	ONDERZOEK	18310	0,0456%	32,40%
255	NO	18269	0,0455%	44,49%
256	LEIDEN	18119	0,0451%	47,35%
257	DAGVAARDING	17998	0,0448%	29,93%
258	JULI	17930	0,0446%	36,03%
259	GESTELDE	17894	0,0445%	43,06%
260	GAAT	17888	0,0445%	41,43%
261	ZOUDEN	17860	0,0444%	39,48%
262	GEWEEST	17730	0,0441%	41,89%
263	VOLGENDE	17724	0,0441%	47,72%
264	15	17659	0,0439%	39,66%
265	UW	17620	0,0438%	33,61%
266	14	17617	0,0438%	39,45%
267	13	17602	0,0438%	39,12%
268	1980	17587	0,0438%	24,00%
269	M	17585	0,0438%	38,96%
270	1982	17539	0,0436%	23,63%
271	BEIDE	17520	0,0436%	41,49%
272	SEPT	17515	0,0436%	35,29%
273	DAARTOE	17444	0,0434%	46,48%
274	JUIST	17198	0,0428%	42,11%
275	GAAN	17107	0,0426%	40,83%
276	GEVEN	17103	0,0426%	41,92%
277	HOGE	17049	0,0424%	52,60%
278	JAAR	17011	0,0423%	31,62%
279	II	17009	0,0423%	33,38%
280	SV	17006	0,0423%	22,32%
281	TERECHTZITTING	16933	0,0421%	26,33%
282	NEDERLAND	16704	0,0416%	21,94%
283	1984	16691	0,0415%	23,11%
284	BEHANDELING	16649	0,0414%	36,34%
285	NAME	16644	0,0414%	42,17%
286	NOCH	16636	0,0414%	35,65%
287	AANGENOMEN	16574	0,0412%	41,56%
288	ENIGE	16510	0,0411%	43,64%
289	AANGEVOERD	16478	0,0410%	43,86%
290	GELET	16430	0,0409%	46,31%
291	WANNEER	16423	0,0409%	39,23%
292	AANGEZIEN	16405	0,0408%	41,11%
293	W	16383	0,0408%	35,46%
294	REDENEN	16332	0,0406%	36,50%
295	1979	16198	0,0403%	23,62%
296	BEPALING	16197	0,0403%	35,00%
297	SGRAVENHAGE	16164	0,0402%	37,05%
298	NEDERLANDSE	16162	0,0402%	25,43%
299	DEN	16098	0,0401%	30,10%
300	GEZIEN	16083	0,0400%	43,06%

Corpus algemeen Nederlands

Rangnr.	Woordtype	Abs. freq.	Rel. freq.	In % doc.
1	DE	313273	0,7795%	100,00%
2	VAN	170425	0,4240%	100,00%
3	HET	151236	0,3763%	100,00%
4	EEN	118626	0,2952%	100,00%
5	EN	114921	0,2859%	100,00%
6	IN	106200	0,2642%	100,00%
7	IS	60319	0,1501%	99,01%
8	DAT	60126	0,1496%	100,00%
9	TE	59392	0,1478%	100,00%
10	OP	48776	0,1214%	100,00%
11	ZIJN	47271	0,1176%	100,00%
12	DIE	46569	0,1159%	100,00%
13	VOOR	39504	0,0983%	100,00%
14	MET	37319	0,0929%	100,00%
15	NIET	37158	0,0925%	100,00%
16	ALS	32809	0,0816%	99,75%
17	AAN	29217	0,0727%	99,75%
18	OOK	26930	0,0670%	99,01%
19	OF	25273	0,0629%	98,28%
20	WORDEN	24959	0,0621%	95,57%
21	ER	23426	0,0583%	99,51%
22	DOOR	23344	0,0581%	98,77%
23	DAN	22915	0,0570%	99,51%
24	DEZE	22529	0,0561%	96,55%
25	BIJ	22482	0,0559%	99,01%
26	OM	22162	0,0551%	99,26%
27	MAAR	21567	0,0537%	99,75%
28	TOT	19420	0,0483%	99,26%
29	WORDT	17954	0,0447%	89,90%
30	HIJ	17949	0,0447%	91,38%
31	KAN	17538	0,0436%	95,07%
32	DIT	17312	0,0431%	97,04%
33	ZICH	16772	0,0417%	98,77%
34	MEN	15814	0,0393%	86,70%
35	NAAR	13760	0,0342%	98,03%
36	ZE	13752	0,0342%	85,22%
37	UIT	13408	0,0334%	99,01%
38	KUNNEN	12787	0,0318%	94,33%
39	OVER	12752	0,0317%	98,28%
40	HEEFT	12609	0,0314%	95,57%
41	IK	12371	0,0308%	72,91%
42	MEER	12350	0,0307%	98,52%
43	WAS	12339	0,0307%	90,64%
44	WAT	12178	0,0303%	96,80%
45	NOG	11647	0,0290%	98,52%
46	WE	11144	0,0277%	81,77%
47	HEBBEN	10757	0,0268%	96,55%
48	ZIJ	10341	0,0257%	89,66%
49	ZO	9938	0,0247%	98,52%
50	MOET	9741	0,0242%	92,61%
51	WEL	9697	0,0241%	98,77%
52	HUN	9494	0,0236%	91,87%
53	HAAR	9203	0,0229%	83,74%
54	ANDERE	8816	0,0219%	95,32%
55	GEEN	8656	0,0215%	97,54%
56	JE	8655	0,0215%	50,99%
57	AL	8545	0,0213%	95,07%
58	ZAL	8057	0,0200%	90,39%
59	ZOU	7307	0,0182%	94,33%
60	TUSSEN	7264	0,0181%	89,90%
61	U	7188	0,0179%	35,22%
62	VEEL	7033	0,0175%	91,87%
63	NU	6551	0,0163%	92,36%
64	WERD	6212	0,0155%	87,93%
65	MOETEN	6016	0,0150%	87,93%
66	ALLEEN	5855	0,0146%	91,13%
67	ZOALS	5782	0,0144%	87,93%
68	ONDER	5510	0,0137%	91,63%
69	EERSTE	5400	0,0134%	93,10%
70	EIGEN	5191	0,0129%	85,22%
71	MAKEN	5148	0,0128%	88,42%
72	ECHTER	5119	0,0127%	82,51%
73	WAAR	5118	0,0127%	92,12%
74	HAD	5095	0,0127%	75,37%
75	NA	5089	0,0127%	92,36%

Rangnr.	Woordtype	Abs. freq.	Rel. freq.	In % doc.
76	TEN	4961	0,0123%	84,73%
77	OMDAT	4848	0,0121%	89,41%
78	KOMEN	4737	0,0118%	90,64%
79	ALLE	4716	0,0117%	87,44%
80	HOE	4697	0,0117%	87,68%
81	GOED	4675	0,0116%	83,74%
82	GROTE	4649	0,0116%	88,92%
83	DUS	4632	0,0115%	84,48%
84	TEGEN	4631	0,0115%	86,21%
85	WIJ	4628	0,0115%	66,26%
86	GAAN	4600	0,0114%	91,87%
87	HEM	4594	0,0114%	71,92%
88	ONS	4545	0,0113%	80,05%
89	WEER	4483	0,0112%	87,68%
90	TOCH	4480	0,0111%	88,67%
91	HIER	4444	0,0111%	89,66%
92	NIEUWE	4436	0,0110%	82,02%
93	GAAT	4422	0,0110%	84,48%
94	WANNEER	4370	0,0109%	70,44%
95	TIJD	4368	0,0109%	88,18%
96	DAAR	4340	0,0108%	87,93%
97	VOORAL	4199	0,0104%	81,03%
98	MOGELIJK	4188	0,0104%	84,73%
99	PLAATS	4173	0,0104%	89,16%
100	STAAT	4068	0,0101%	83,25%
101	DOEN	4051	0,0101%	84,98%
102	TWEE	4014	0,0100%	88,18%
103	ZELF	3938	0,0098%	85,47%
104	KOMT	3915	0,0097%	84,98%
105	MENSEN	3835	0,0095%	73,65%
106	AANTAL	3808	0,0095%	74,14%
107	BINNEN	3803	0,0095%	83,00%
108	STEEDS	3755	0,0093%	81,28%
109	WAARIN	3738	0,0093%	84,98%
110	WIL	3635	0,0090%	83,50%
111	WAREN	3603	0,0090%	78,33%
112	VAAK	3602	0,0090%	73,15%
113	ELKAAR	3518	0,0088%	81,77%
114	IETS	3505	0,0087%	79,31%
115	LEVEN	3489	0,0087%	71,92%
116	BEPAALDE	3485	0,0087%	73,15%
117	GEVEN	3458	0,0086%	83,00%
118	ZONDER	3352	0,0083%	85,96%
119	ZEER	3328	0,0083%	76,85%
120	ÉÉN	3326	0,0083%	83,50%
121	VERSCHILLENDE	3279	0,0082%	75,12%
122	LATEN	3150	0,0078%	78,08%
123	JAREN	3140	0,0078%	78,33%
124	TOE	3083	0,0077%	87,44%
125	ZIEN	3065	0,0076%	81,28%
126	JAAR	3045	0,0076%	74,63%
127	GEVAL	3029	0,0075%	79,80%
128	WELKE	3020	0,0075%	71,43%
129	AF	3012	0,0075%	88,92%
130	ZULLEN	2974	0,0074%	75,62%
131	DER	2954	0,0074%	65,52%
132	MINDER	2940	0,0073%	74,63%
133	ALTIJD	2900	0,0072%	79,06%
134	EENS	2894	0,0072%	81,77%
135	MIJN	2860	0,0071%	56,16%
136	MENS	2814	0,0070%	51,23%
137	LAATSTE	2810	0,0070%	86,45%
138	TOEN	2803	0,0070%	68,97%
139	VRAAG	2798	0,0070%	79,06%
140	SLECHTS	2777	0,0069%	75,86%
141	WIJZE	2762	0,0069%	74,14%
142	DUIDELIJK	2745	0,0068%	80,05%
143	ONZE	2734	0,0068%	69,70%
144	ZELFS	2711	0,0067%	80,30%
145	BELANG	2575	0,0064%	66,50%
146	WERK	2521	0,0063%	68,97%
147	ONDERZOEK	2511	0,0062%	53,94%
148	BIJVOORBEELD	2502	0,0062%	59,36%
149	TWEEDE	2475	0,0062%	78,33%
150	BAL	2449	0,0061%	6,65%

Rangnr.	Woordtype	Abs. freq.	Rel. freq.	In % doc.
151	UW	2446	0,0061%	22,41%
152	VORM	2443	0,0061%	67,73%
153	VERDER	2431	0,0060%	80,79%
154	ONTWIKKELING	2399	0,0060%	61,08%
155	WERDEN	2398	0,0060%	67,98%
156	ENKELE	2380	0,0059%	79,80%
157	KON	2377	0,0059%	67,24%
158	DEEL	2369	0,0059%	71,18%
159	SOMS	2364	0,0059%	68,97%
160	ZIN	2357	0,0059%	70,20%
161	ZON	2356	0,0059%	70,69%
162	HEEL	2341	0,0058%	68,23%
163	ZEGGEN	2336	0,0058%	76,85%
164	GEHEEL	2331	0,0058%	74,63%
165	GEBRUIK	2321	0,0058%	63,55%
166	ALGEMEEN	2313	0,0058%	71,43%
167	JUIST	2284	0,0057%	73,40%
168	WAARBIJ	2274	0,0057%	67,24%
169	DAARBIJ	2267	0,0056%	73,40%
170	WANT	2266	0,0056%	64,53%
171	1	2264	0,0056%	60,84%
172	WIE	2261	0,0056%	70,44%
173	GOEDE	2242	0,0056%	65,52%
173	VORMEN	2242	0,0056%	69,21%
175	ALLES	2213	0,0055%	71,18%
176	GROOT	2208	0,0055%	76,35%
177	NODIG	2203	0,0055%	69,21%
178	ANDERS	2201	0,0055%	77,09%
179	WEINIG	2199	0,0055%	71,43%
180	ANDER	2185	0,0054%	74,14%
181	POLITIEKE	2183	0,0054%	46,55%
182	STAAN	2178	0,0054%	79,31%
183	HOUDEN	2175	0,0054%	74,14%
183	TERWIJL	2175	0,0054%	72,91%
185	ZOUDEN	2154	0,0054%	73,65%
186	SOCIALE	2120	0,0053%	50,74%
187	NEMEN	2114	0,0053%	73,65%
188	2	2091	0,0052%	58,87%
189	WEG	2074	0,0052%	77,09%
190	VANUIT	2073	0,0052%	67,98%
191	GEGEVEN	2067	0,0051%	74,38%
192	DENKEN	2065	0,0051%	64,04%
193	ONDERWIJS	2063	0,0051%	34,73%
194	REEDS	2055	0,0051%	63,05%
195	KRIJGEN	2052	0,0051%	68,72%
196	VINDEN	2046	0,0051%	72,17%
197	TER	2029	0,0050%	74,38%
198	HAND	2019	0,0050%	73,89%
199	MAG	2011	0,0050%	66,75%
200	GEMAAKT	2009	0,0050%	74,38%
201	ZOWEL	1995	0,0050%	68,47%
202	BETEKENIS	1992	0,0050%	61,58%
203	MEE	1984	0,0049%	72,91%
204	ROL	1959	0,0049%	65,27%
205	DERGELIJKE	1951	0,0049%	60,34%
206	DRIE	1947	0,0048%	72,91%
207	VRAGEN	1942	0,0048%	65,52%
208	RECHT	1939	0,0048%	56,16%
209	MAN	1936	0,0048%	44,83%
210	EERST	1933	0,0048%	72,17%
211	BUITEN	1920	0,0048%	74,63%
211	HEN	1920	0,0048%	63,05%
211	STELLEN	1920	0,0048%	72,91%
214	AANDACHT	1915	0,0048%	66,75%
215	BESTAAN	1914	0,0048%	71,43%
216	BRENGEN	1901	0,0047%	70,44%
217	VOLGENDE	1892	0,0047%	71,18%
217	WOORD	1892	0,0047%	56,16%
219	KINDEREN	1891	0,0047%	42,86%
220	ENIGE	1885	0,0047%	74,14%
221	BEPAALD	1871	0,0047%	67,24%
222	WILLEN	1870	0,0047%	67,49%
223	BESTAAT	1864	0,0046%	65,76%
224	WAAROP	1857	0,0046%	74,38%
225	BEIDE	1855	0,0046%	65,27%

Rangnr.	Woordtype	Abs.freq.	Rel. freq.	In % doc.
226	DAAROM	1853	0,0046%	67,98%
226	WERELD	1853	0,0046%	53,94%
228	VROUW	1841	0,0046%	37,68%
229	GROND	1828	0,0045%	68,97%
230	MEESTAL	1825	0,0045%	54,68%
231	KWAM	1819	0,0045%	65,02%
232	ZEKER	1795	0,0045%	77,09%
233	VELE	1778	0,0044%	63,55%
234	PER	1777	0,0044%	58,37%
234	VIA	1777	0,0044%	59,61%
236	MIJ	1769	0,0044%	52,22%
237	NIETS	1762	0,0044%	67,24%
238	IMMERS	1758	0,0044%	65,52%
239	BLIJVEN	1752	0,0044%	74,38%
240	MAAKT	1732	0,0043%	66,26%
241	INVLOED	1722	0,0043%	57,88%
242	KENNIS	1717	0,0043%	55,42%
243	GEEFT	1711	0,0043%	65,27%
244	GEBRUIKT	1710	0,0043%	53,20%
245	VOLGENS	1705	0,0042%	63,30%
246	STERK	1704	0,0042%	63,79%
247	LAAT	1694	0,0042%	73,15%
248	GEWEEST	1686	0,0042%	75,37%
248	VOORBEELD	1686	0,0042%	63,05%
250	WERKEN	1685	0,0042%	62,07%
251	NEDERLANDSE	1682	0,0042%	51,97%
252	PROBLEMEN	1681	0,0042%	60,34%
253	WETENSCHAP	1680	0,0042%	35,47%
254	SITUATIE	1673	0,0042%	61,82%
255	LIGT	1669	0,0042%	66,50%
256	NATUURLIJK	1664	0,0041%	65,02%
257	ZICHZELF	1661	0,0041%	61,33%
258	BETER	1646	0,0041%	66,01%
259	NOOIT	1644	0,0041%	66,75%
260	MATE	1641	0,0041%	65,27%
261	HEB	1637	0,0041%	49,01%
262	DEZELFDE	1633	0,0041%	62,07%
263	BOVENDIEN	1631	0,0041%	67,49%
264	EVEN	1628	0,0041%	64,78%
265	3	1627	0,0040%	57,64%
266	GEBIED	1626	0,0040%	56,90%
267	GEVALLEN	1623	0,0040%	54,93%
268	BELANGRIJKE	1609	0,0040%	63,79%
269	DOET	1607	0,0040%	63,79%
270	BELEID	1605	0,0040%	39,90%
271	VERBAND	1593	0,0040%	64,53%
272	LATER	1589	0,0040%	69,21%
272	SOORT	1589	0,0040%	62,56%
274	ALGEMENE	1584	0,0039%	58,62%
274	GING	1584	0,0039%	58,87%
276	OVERHEID	1577	0,0039%	35,22%
277	KIND	1560	0,0039%	32,76%
278	BLIJKT	1555	0,0039%	62,81%
279	TAAL	1549	0,0039%	37,68%
280	HADDEN	1547	0,0038%	61,33%
281	MOEST	1545	0,0038%	57,64%
282	MANIER	1543	0,0038%	53,94%
283	ZEI	1542	0,0038%	33,25%
284	WETEN	1536	0,0038%	63,30%
285	DAARMEE	1534	0,0038%	67,49%
286	NAAST	1522	0,0038%	69,95%
287	BLIJFT	1507	0,0037%	64,78%
288	GROEP	1506	0,0037%	51,97%
289	TERUG	1504	0,0037%	71,92%
290	MOGELIJKHEDEN	1502	0,0037%	57,64%
291	BETREFT	1501	0,0037%	63,55%
292	GEZIEN	1498	0,0037%	72,91%
293	OUDE	1495	0,0037%	56,40%
294	SPELEN	1494	0,0037%	54,68%
294	ZODAT	1494	0,0037%	57,64%
296	ZAKEN	1493	0,0037%	58,62%
297	PAS	1491	0,0037%	67,73%
298	DIENT	1488	0,0037%	60,84%
299	BELANGRIJK	1476	0,0037%	59,11%
300	INDIEN	1472	0,0037%	42,86%

BIJLAGE 4 - VERGELIJKINGSLIJST VAN TYPEN UIT DE KOP VAN DE FREQUENTIELIJSTEN

Deze bijlage bevat de vergelijkingslijst van de woordtypen in de kop van de frequentielijsten van de gereduceerde juridische corpora en het corpus algemeen Nederlands. De woordtypen staan in deze lijst op alfabetische volgorde, erachter staat hun rangnummer en relatieve frequentie in de drie corpora vermeld. '---' in een bepaalde kolom betekent dat het woordtype in het betreffende corpus niet voorkomt in de kop van de frequentielijst. Deze kop omvat bij het gereduceerde corpus wetgevingsteksten de eerste 120, bij het gereduceerde corpus jurisprudentieteksten de eerste 133 en bij het corpus algemeen Nederlands de eerste 155 rangnummers.

Woordtype	Wetgevingsteksten		Jurisprudentieteksten		Alg. Nederlands	
	Rangnr.	Rel. freq.	Rangnr.	Rel. freq.	Rangnr	Rel. freq.
1	19	0,6808%	38	0,2628%	---	---
2	26	0,5603%	49	0,1965%	---	---
3	36	0,3526%	66	0,1452%	---	---
4	48	0,2337%	87	0,1158%	---	---
5	62	0,1717%	100	0,0965%	---	---
6	75	0,1327%	113	0,0858%	---	---
7	95	0,1003%	---	---	---	---
8	109	0,0892%	---	---	---	---
10	107	0,0905%	---	---	---	---
A	40	0,3007%	67	0,1452%	---	---
AAN	14	0,7738%	13	0,7641%	17	0,6329%
AANTAL	---	---	---	---	106	0,0825%
AANZIEN	96	0,1002%	---	---	---	---
AF	---	---	---	---	129	0,0652%
AL	---	---	122	0,0808%	57	0,1851%
ALGEMENE	78	0,1252%	---	---	---	---
ALLE	106	0,0925%	---	---	79	0,1022%
ALLEEN	---	---	---	---	66	0,1268%
ALS	29	0,4407%	20	0,5841%	16	0,7107%
ALSMEDE	112	0,0865%	---	---	---	---
ALTHANS	---	---	124	0,0789%	---	---
ALTIJD	---	---	---	---	133	0,0628%
ANDERE	66	0,1537%	95	0,1019%	54	0,1910%
ARREST	---	---	75	0,1337%	---	---
ART	13	0,7969%	23	0,4909%	---	---
ARTIKEL	16	0,7259%	---	---	---	---
ARTIKELEN	79	0,1244%	---	---	---	---
B	44	0,2765%	90	0,1062%	---	---
BAL	---	---	---	---	150	0,0530%
BEDOELD	34	0,3703%	129	0,0760%	---	---
BEDOELDE	55	0,1954%	---	---	---	---
BEDRAG	88	0,1074%	---	---	---	---
BELANG	---	---	99	0,0987%	145	0,0558%
BEPAALDE	54	0,1955%	---	---	116	0,0755%
BEPALINGEN	111	0,0879%	---	---	---	---
BEROEP	---	---	42	0,2444%	---	---
BESCHIKKING	---	---	88	0,1148%	---	---
BESLISSING	---	---	94	0,1022%	---	---
BESLUIT	46	0,2713%	---	---	---	---
BESTREDEN	---	---	108	0,0922%	---	---
BESTUUR	100	0,0991%	---	---	---	---
BETREKKING	69	0,1480%	---	---	---	---
BIJ	21	0,6698%	21	0,5561%	25	0,4870%
BIJVOORBEELD	---	---	---	---	148	0,0542%
BINNEN	84	0,1145%	---	---	107	0,0824%
BW	---	---	133	0,0738%	---	---
C	60	0,1798%	---	---	---	---
CASSATIE	---	---	53	0,1849%	---	---
COMMISSIE	117	0,0830%	---	---	---	---
CONCLUSIE	---	---	119	0,0831%	---	---
D	98	0,0995%	---	---	---	---
DAAR	---	---	---	---	96	0,0940%
DAARVAN	109	0,0892%	---	---	---	---
DAG	82	0,1190%	---	---	---	---
DAN	35	0,3601%	39	0,2590%	23	0,4964%
DAT	24	0,6191%	5	2,2892%	8	1,3024%
DATUM	90	0,1051%	---	---	---	---
DE	1	7,0383%	1	6,5142%	1	6,7860%
DEN	56	0,1856%	---	---	---	---
DER	38	0,3491%	51	0,1876%	131	0,0640%
DERDE	83	0,1190%	---	---	---	---
DERHALVE	---	---	103	0,0932%	---	---
DEZE	27	0,4668%	24	0,4829%	24	0,4880%
DIE	18	0,7181%	18	0,6296%	12	1,0088%
DIT	41	0,2964%	28	0,3614%	32	0,3750%

Woordtype	Wetgevingsteksten		Jurisprudentieteksten		Alg. Nederlands	
	Rangnr.	Rel. freq.	Rangnr.	Rel. freq.	Rangnr	Rel. freq.
DOCH	---	---	127	0,0777%	---	---
DOEN	---	---	128	0,0776%	101	0,0878%
DOOR	23	0,6397%	14	0,7586%	22	0,5057%
DUIDELIJK	---	---	---	---	142	0,0595%
DUS	---	---	---	---	83	0,1003%
ECHTER	---	---	---	---	72	0,1109%
EEN	6	1,8764%	7	1,6138%	4	2,5696%
ÉÉN	---	---	---	---	120	0,0720%
EENS	---	---	---	---	134	0,0627%
EERSTE	31	0,4161%	71	0,1415%	69	0,1170%
EIGEN	---	---	---	---	70	0,1124%
ELKAAR	---	---	---	---	113	0,0762%
EN	5	2,2368%	6	1,7574%	5	2,4894%
ENZ	85	0,1116%	---	---	---	---
ER	---	---	65	0,1502%	21	0,5074%
F	81	0,1209%	55	0,1752%	---	---
GAAN	---	---	---	---	86	0,0996%
GAAT	---	---	---	---	93	0,0958%
GEEN	86	0,1110%	36	0,2747%	55	0,1875%
GEGEVEN	---	---	102	0,0944%	---	---
GEMEENTE	45	0,2750%	---	---	---	---
GEMEENTEN	108	0,0896%	---	---	---	---
GESTELD	---	---	89	0,1117%	---	---
GEVAL	---	---	68	0,1445%	127	0,0656%
GEVEN	---	---	---	---	117	0,0749%
GOED	---	---	---	---	81	0,1013%
GRIEF	---	---	126	0,0779%	---	---
GROND	112	0,0865%	70	0,1417%	---	---
GROTE	---	---	---	---	82	0,1007%
HAAR	---	---	37	0,2690%	53	0,1994%
HAD	---	---	60	0,1607%	74	0,1104%
HAND	61	0,1754%	---	---	---	---
HEBBEN	58	0,1845%	46	0,2269%	47	0,2330%
HEEFT	51	0,2092%	12	0,9342%	40	0,2731%
HEM	102	0,0956%	82	0,1283%	87	0,0995%
HET	3	3,4096%	3	3,4730%	3	3,2760%
HETGEEN	---	---	123	0,0798%	---	---
HIER	---	---	117	0,0837%	91	0,0963%
HIJ	68	0,1487%	44	0,2283%	30	0,3888%
HOE	---	---	---	---	80	0,1017%
HOF	---	---	26	0,4227%	---	---
HOGER	---	---	110	0,0917%	---	---
HR	---	---	91	0,1051%	---	---
HUN	118	0,0826%	131	0,0752%	52	0,2057%
I	65	0,1540%	---	---	---	---
IETS	---	---	---	---	114	0,0759%
II	70	0,1462%	---	---	---	---
IK	---	---	62	0,1575%	41	0,2680%
IN	4	2,4722%	4	2,4443%	6	2,3005%
INDIEN	37	0,3518%	106	0,0928%	---	---
INGEVOLGE	103	0,0953%	---	---	---	---
IS	10	0,9086%	9	1,2616%	7	1,3066%
JAAR	92	0,1011%	---	---	126	0,0660%
JANUARI	114	0,0858%	---	---	---	---
JAREN	---	---	---	---	123	0,0680%
JE	---	---	---	---	56	0,1875%
KAN	42	0,2963%	32	0,2913%	31	0,3799%
KOMEN	---	---	---	---	78	0,1026%
KOMT	---	---	---	---	104	0,0848%
KRACHTENS	73	0,1402%	---	---	---	---
KUNNEN	76	0,1319%	58	0,1666%	38	0,2770%
LAATSTE	---	---	---	---	137	0,0609%
LATEN	---	---	---	---	122	0,0682%
LEDEN	94	0,1006%	---	---	---	---
LEVEN	---	---	---	---	115	0,0756%

Woordtype	Wetgevingsteksten		Jurisprudentieteksten		Alg. Nederlands	
	Rangnr.	Rel. freq.	Rangnr.	Rel. freq.	Rangnr	Rel. freq.
LID	15	0,7586%	52	0,1866%	---	---
MAAR	---	---	96	0,1016%	27	0,4672%
MAKEN	---	---	---	---	71	0,1115%
MEER	74	0,1343%	80	0,1295%	42	0,2675%
MEN	---	---	---	---	34	0,3426%
MENS	---	---	---	---	136	0,0610%
MENSEN	---	---	---	---	105	0,0831%
MET	11	0,8883%	16	0,7119%	14	0,8084%
MIDDEL	---	---	54	0,1837%	---	---
MIJN	---	---	---	---	135	0,0620%
MINDER	---	---	---	---	132	0,0637%
MINISTER	32	0,4108%	---	---	---	---
MOET	72	0,1410%	59	0,1633%	50	0,2110%
MOETEN	115	0,0850%	101	0,0953%	65	0,1303%
MOGELIJK	---	---	---	---	98	0,0907%
MR	---	---	86	0,1209%	---	---
NA	52	0,2081%	85	0,1235%	75	0,1102%
NAAR	67	0,1523%	35	0,2783%	35	0,2981%
NIET	28	0,4627%	11	1,0654%	15	0,8049%
NIEUWE	---	---	---	---	92	0,0961%
NJ	---	---	83	0,1249%	---	---
NOG	---	---	74	0,1350%	45	0,2523%
NR	97	0,0998%	120	0,0817%	---	---
NU	---	---	73	0,1381%	63	0,1419%
O	---	---	63	0,1538%	---	---
OF	7	1,2321%	25	0,4555%	19	0,5475%
OM	93	0,1011%	33	0,2877%	26	0,4801%
OMDAT	---	---	92	0,1043%	77	0,1050%
OMSTANDIGHEDEN	---	---	118	0,0834%	---	---
ONDER	39	0,3315%	43	0,2304%	68	0,1194%
ONDERDEEL	---	---	114	0,0857%	---	---
ONDERHAVIGE	---	---	109	0,0920%	---	---
ONDERZOEK	---	---	---	---	147	0,0544%
ONS	---	---	---	---	88	0,0985%
ONTWIKKELING	---	---	---	---	154	0,0520%
ONZE	43	0,2857%	---	---	143	0,0592%
OOK	---	---	30	0,3491%	18	0,5833%
OORDEEL	---	---	84	0,1241%	---	---
OP	9	1,0158%	10	1,1814%	10	1,0566%
OVER	77	0,1286%	64	0,1502%	39	0,2762%
P	---	---	72	0,1388%	---	---
PLAATS	119	0,0823%	---	---	99	0,0904%
PP	---	---	77	0,1318%	---	---
RAAD	64	0,1650%	112	0,0859%	---	---
RB	---	---	27	0,3719%	---	---
RECHT	---	---	61	0,1603%	---	---
RECHTER	---	---	130	0,0758%	---	---
REEDS	---	---	125	0,0786%	---	---
RO	---	---	121	0,0809%	---	---
SLECHTS	---	---	107	0,0924%	140	0,0602%
STAAT	---	---	93	0,1029%	100	0,0881%
STB	50	0,2120%	---	---	---	---
STEEDS	---	---	---	---	108	0,0813%
STELLEN	120	0,0817%	---	---	---	---
TE	12	0,8381%	8	1,4772%	9	1,2865%
TEGEN	---	---	47	0,2209%	84	0,1003%
TEN	33	0,4094%	29	0,3500%	76	0,1075%
TER	57	0,1851%	50	0,1898%	---	---
TIJD	---	---	---	---	95	0,0946%
TOCH	---	---	---	---	90	0,0970%
TOE	---	---	---	---	124	0,0668%
TOEN	---	---	---	---	138	0,0607%
TOEPASSING	47	0,2538%	---	---	---	---
TOT	25	0,5770%	19	0,6121%	28	0,4207%
TUSSEN	---	---	79	0,1303%	60	0,1574%

Woordtype	Wetgevingsteksten		Jurisprudentieteksten		Alg. Nederlands	
	Rangnr.	Rel. freq.	Rangnr.	Rel. freq.	Rangnr	Rel. freq.
TWEE	---	---	---	---	102	0,0869%
TWEEDE	49	0,2250%	---	---	149	0,0536%
U	---	---	---	---	61	0,1557%
UIT	59	0,1830%	34	0,2852%	37	0,2904%
UW	---	---	---	---	151	0,0530%
VAAK	---	---	---	---	112	0,0780%
VAN	2	5,9512%	2	4,9012%	2	3,6917%
VEEL	---	---	---	---	62	0,1523%
VERDACHTE	---	---	98	0,0990%	---	---
VERDER	---	---	---	---	153	0,0527%
VERSCHILLENDE	---	---	---	---	121	0,0710%
VERSTAAN	105	0,0932%	---	---	---	---
VONNIS	---	---	56	0,1722%	---	---
VOOR	8	1,0967%	17	0,7048%	13	0,8557%
VOORAL	---	---	---	---	97	0,0910%
VORDERING	---	---	81	0,1289%	---	---
VORM	---	---	---	---	152	0,0529%
VRAAG	---	---	104	0,0930%	139	0,0606%
WAAR	---	---	---	---	73	0,1109%
WAARIN	99	0,0993%	---	---	109	0,0810%
WAAROP	89	0,1065%	---	---	---	---
WAARVAN	116	0,0845%	115	0,0851%	---	---
WANNEER	---	---	---	---	94	0,0947%
WAREN	---	---	---	---	111	0,0780%
WAS	---	---	41	0,2467%	43	0,2673%
WAT	---	---	---	---	44	0,2638%
WE	---	---	---	---	46	0,2414%
WEER	---	---	---	---	89	0,0971%
WEL	87	0,1092%	78	0,1316%	51	0,2101%
WELKE	53	0,1988%	57	0,1668%	128	0,0654%
WERD	---	---	105	0,0929%	64	0,1346%
WERDEN	---	---	---	---	155	0,0519%
WERK	---	---	---	---	146	0,0546%
WET	30	0,4198%	76	0,1319%	---	---
WIJ	101	0,0980%	---	---	85	0,1003%
WIJZE	---	---	---	---	141	0,0598%
WIL	---	---	---	---	110	0,0787%
WORDEN	17	0,7221%	22	0,5443%	20	0,5407%
WORDT	22	0,6678%	48	0,2181%	29	0,3889%
ZAAK	---	---	116	0,0843%	---	---
ZAKEN	80	0,1239%	---	---	---	---
ZAL	71	0,1427%	69	0,1442%	58	0,1745%
ZE	---	---	---	---	36	0,2979%
ZEER	---	---	---	---	119	0,0721%
ZELF	---	---	---	---	103	0,0853%
ZELFS	---	---	---	---	144	0,0587%
ZICH	104	0,0948%	40	0,2534%	33	0,3633%
ZIEN	---	---	---	---	125	0,0664%
ZIJ	63	0,1682%	45	0,2281%	48	0,2240%
ZIJN	20	0,6788%	15	0,7363%	11	1,0240%
ZIN	---	---	132	0,0741%	---	---
ZO	---	---	---	---	49	0,2153%
ZOALS	---	---	97	0,1009%	67	0,1252%
ZONDER	---	---	111	0,0877%	118	0,0726%
ZOU	---	---	31	0,3074%	59	0,1583%
ZOVER	91	0,1050%	---	---	---	---
ZULLEN	---	---	---	---	130	0,0644%

REGISTER VAN PERSOONSNAMEN

Aarts, J. 17, 18, 20, 21
Austin, W.B. 22

Baayen, R.H. 19, 23, 24, 26-30, 86
Bailey, R.W. 18, 21, 22, 32
Bakker, D. 17
Bal, P.L. 2
Batagelj, V. 225, 230, 237, 238
Battus, H. 95
Bax, C.J. 2, 11
Bennett, P.E. 21
Berckel, J.A.Th.M. van 20
Bergh, G.C.J.J. van den 1, 10, 11, 35
Bianchi, H. 13, 14
Biber, D. 39
uit den Boogaart, P.C 4, 20, 32, 56
Brandt Corstius, H. 4, 8, 16-20, 22, 23, 33, 227
Bren, M. 225, 230, 237, 238
Broekman, J.M. 1, 10, 11, 35
Butler, C.S. 22, 32

Carroll, J.B. 21, 65, 80
Chang-Rodriguez, E. 20, 65
Cherubini, W. 22
Chomsky, N. 18
Cluett, R. 22
Condon, E.U. 24
de la Court, J.F.H.A. 20, 32
Cross, G.R. 224

Davis, D. 39
Deville, G. 38
Dik, S.C. 4, 9, 16

Eemeren, F.H. van 14, 35, 151
Ehrenberg, A.S.C. 161
Ellegård, A. 22
Erikstad, O.M. 31, 84, 91-93, 122, 164, 165, 193, 219, 268, 270
Estoup, J.B. 24, 98

Finegan, E. 39
Francis, N.W. 20

Guiraud, P. 28-31, 89, 120

Haan, J.I. de 10
Hempel, C.G. 38
Hendrickx, R.L.L. 223
Herdan, G. 18, 26-28, 30, 31, 84-92, 117-120, 144, 145, 163, 190, 268, 270
Hofhuis, H.F.M. 2, 11, 35

Juilland, A.G. 20, 65

Kenny, A. 22
Kerkmeester, H.O. 6
Knottenbelt, B. 51
Kooij, J.G. 4, 9, 16
Kučera, H. 20
Kuo, S. 224

Landen, J.P. van der 243
Lesk, M.E. 224
Linschoten, J. 20
Loevinger, L. 6
Loth, M.A. 1, 2, 14

Maes, A.A. 12
Mandelbrot, B. 26, 28, 31
Martin, W. 16, 20, 21, 32, 36-38, 57, 58, 65, 93, 223
Meier, H. 20
Meijs, W. 17, 18, 20, 21, 32, 34
Mendenhall, T.C. 18, 21
Milic, L.T. 22
Morris, Ch.W. 9
Mosteller, R. 22
Mulder, R.V. De 1, 2, 6, 7, 11, 16, 39, 56, 67, 73, 160, 170, 176-182, 214, 217, 219, 235

Muller, Ch. 26
Naelten, M.A.M.G. van 26
Neijt, A. 17

Orlov, Ju.K. 28, 29
Oskamp, A. 2, 11, 39, 56, 67, 73, 170, 176-182, 214, 217, 219

Pas, E. ten 36-38

Rapoport, A. 24, 98
Reenen, P.Th. van 14, 19
Reinsma, M. 11, 151
Renkema, J. 2, 11, 12
Rescher, N. 38
Rijsbergen, C.J. van 225

Salton, G. 224
Sapir, E. 11, 13, 14
Scherpenzeel, R. 42
Sciarone, A.G. 20
Siertsema, B. 14
Smeaton, A.F. 225
Štejnfel'dt, E.A. 20
Sterkenburg, P.G.J. van 37, 38
Stiles, H.E. 224

Thompson, J.R. 21
Thomson, G.H. 21
Tops, G.A.J. 223
Torringa, R.A. 51
Traversa, V.P. 20

Wallace, D.L. 22
Waring 26
Whorf 11, 13, 14
Wildemast, C.A.M. 7
Willett, P. 225
Witteveen, W.J. 2, 11

Yule, G.U. 21, 27, 85-89, 117-119, 130, 144, 145, 163, 190, 268, 270

Zipf, G.K. 24-26, 28, 67, 73, 74, 98, 110, 135, 176-182, 218

ZAKENREGISTER

algebraïsche taalkunde 16, 17
ambtelijk 2, 11, 12
argumentatieleer 14

computer-taalkunde 16-18, 36
concept 7, 12, 38
conceptueel 7
conceptuele zoektechnieken 7
continue 160
continuüm 160
corpustaalkunde 17-19, 34
covariantie 92, 121, 147

databanken 6
discrete 160
dissimilariteit 225
documentatie 6, 228
documentatiesysteem 8
drempelwaarde 228, 239, 240, 253

Eindhovens corpus 12, 20, 32, 33, 56, 176, 214, 216, 217
empirisch 2, 6
empirische wetenschap 6
ervaringswereld 2, 6, 38
exemplarisch corpus 32, 57, 58

filosofie 2, 10

herhalingsfrequentie 27, 86, 119, 130, 145, 163, 268

jurimetrie 6

kwantitatieve aspecten 2, 3, 11, 12, 18, 19, 64, 266
kwantitatieve methoden 2, 227, 266, 267
kwantitatieve taalkunde 8, 16-24, 32, 34, 266

lemma 34, 61
lemmatisering 5
lettergebruik 4
linguïstiek 2, 16
linguïstische constanten 163, 165, 218, 270

morfologie 10

NLEX 42-44, 46, 266
nominale anaforen 12
normatief 6

parsers 5
Poisson-verdeling 27
pragmatiek 2, 9-11
pragmatisch 36, 37, 226

rechtsbronnen 1
rechtsfilosofie 2
rechtspraktijk 3, 11, 16, 35
rechtstaal 1, 10
rechtswetenschap 1, 3, 5, 6, 16
regressielijn 72-74, 91, 92, 109, 110, 121, 122, 134, 135, 146-148, 170, 171, 173, 174, 175-180, 182, 192, 193, 206, 213, 218
residuals 72, 91, 92, 121, 122, 147, 148, 165, 179-182, 213, 218

Sapir-Whorf hypothese 13, 14
schrijftaal 4, 34, 41, 56, 59, 214, 216
semantiek 2, 5, 9-11, 221, 222, 226
semantisch 4, 37, 38, 151, 213, 221, 263, 271
similariteits-score 237, 241-248
similariteitsrelatie 223, 245, 246, 265

STAIRS 43
standaarddeviatie 27, 50, 55, 59, 65, 86, 162, 164, 191, 206, 219, 250, 269
STATUS 43, 52
stigmatisering 12, 13
stijl 18, 19, 21, 22, 28, 34, 45-47, 180, 254
subcorpora 65, 216, 221
subtaal 19, 35-38, 266
syntactisch 221
syntaxis 2, 4, 9, 10

t-test 92, 122, 148
taalfilosofie 2
taalvariatie 57
taalvermogen 18, 19
taalwetenschap 1, 5, 16, 31
taggers 5
talige activiteit 1
token/type ratio 83-85, 89, 90, 106, 114, 116, 117, 119, 131, 133, 135, 136, 139, 141, 142, 143, 145, 166, 189, 190, 193, 219, 268, 270
type/token ratio 77, 84, 89, 120

vaktaal 19, 35-39, 58, 223, 266
variantie 73, 91, 93, 110, 121, 135, 147
verwachtingspatroon 99, 126
voortschrijdend gemiddelde 102-104, 126, 128, 129, 153-156, 200, 202-204, 269
vrijheidsgraden 92, 122, 148

Wet van Zipf 24, 26, 28, 67, 74, 110, 135, 176-179
Wetboek van Strafrecht 2, 214
Wetboek van Strafvordering 2, 214
woord-niveau 4
woordlemma 8, 9
woordsoorten 5, 105, 212
woordtoken 8, 9
woordtype 8, 9
woordvormen 4, 34, 38

Zipf-karakteristiek 67, 177, 182

Curriculum Vitae

Kees van Noortwijk werd geboren op 9 juni 1961 te Schiedam. Na zijn middelbare school opleiding studeerde hij rechten aan de Erasmus Universiteit Rotterdam. Sinds zijn afstuderen in 1984 is hij verbonden geweest aan het Centrum voor Informatica en Recht aan deze universiteit, aanvankelijk als wetenschappelijk onderzoeker, later als universitair docent. Hij verzorgt sinds een aantal jaren onderwijs op het gebied van juridische informatica (computer-toepassingen voor wetenschap en praktijk van het recht). Daarnaast is hij sinds 1985 betrokken bij de ontwikkeling van juridische computer advies systemen (het JURICAS project), en verricht hij sinds 1988 jurimetrisch onderzoek naar specifieke eigenschappen van juridisch tekstmateriaal.