De taal van de minister. Een exploratief-kwantitatief onderzoek

Ton van der Wouden & Jaap de Jong

In dit artikel doen we verslag van exploratief-kwantitatief onderzoek naar talige eigenschappen van ministeriële toespraken. We rapporteren over relevante taalkundige kenmerken van toespraken van ministers, en we vergelijken hun taal en taalgebruik met dat van 'gewone' taalgebruikers uit het Corpus Gesproken Nederlands. We gaan na of we via kwantitatief onderzoek de teksten van de ministeriële speechschrijvers kunnen typeren als meer spreektaal of meer schrijftaal. Wat betreft de onderzochte kenmerken blijkt de tekst van de ministeriële toespraak, anders dan sommige adviezen suggereren, meer op schrijftaal te lijken dan op spontane spreektaal.

1 Ministeriële toespraken

Het vak taalbeheersing heeft, op zijn minst ten dele, zijn wortels bij de toespraak (retorica). En nog steeds, of weer, staan toespraken volop in de belangstelling in het Nederlandse en Belgische taalbeheersingsonderzoek (zie bijv. Ensink & Sauer, 2003; Andeweg & De Jong, 2004; Van De Mieroop, 2005).

In het hier gepresenteerde onderzoek staat de ministeriële toespraak centraal. Intuïtief, pretheoretisch kunnen we de toespraak van ministers (en staatssecretarissen, hier in het algemeen aangeduid als 'de minister') omschrijven als een geformaliseerde situatie met onder meer de volgende kenmerken:

- De aanwezigheid van de minister is onder meer ceremonieel, en verhoogt de status van de gebeurtenis (opening van een school of congres, presentatie van een boek of een plan, uitreiking van een prijs, ...);
- Voordat hij zijn lintje doorknipt, zijn boek in ontvangst neemt, enz., snijdt de minister nog enkele beleidskwesties aan ten einde daarvoor enig draagvlak te verwerven;
- Het publiek doet alsof het iets interessants hoort, maar er is na afloop in het algemeen geen gelegenheid tot het stellen van vragen of geven van reacties;
- Kabinet-watchers en belanghebbenden kijken naar de inhoud, juist omdat het medium van de ministeriële toespraak onder meer wordt gebruikt om terloops zekere beleidsvoornemens naar voren te brengen of uit te proberen.

De meeste toespraken van Nederlandse ministers worden niet door henzelf geschreven, maar door professionele, in het algemeen anonieme, ghostwriters. Al nemen bewindslieden niet zelden de vrijheid meer of minder af te wijken van de geprepareerde tekst (door bijv. aan te sluiten op situatie of actualiteit, vergelijk Veltman, Andeweg & De Jong, 2003 en De Jong & Andeweg, 2004), we concentreren ons hier op de geschreven speechtekst.

Van de stijlmiddelen van het genre van de formele, geprepareerde toespraak, alsmede van hun effectiviteit, is inmiddels het nodige bekend (Andeweg & De Jong, 2004). Veel minder geldt dat voor de taalkundige eigenschappen ervan. Aan alle speechschrijvers wordt de eis gesteld een tekst te leveren die 'goed bekt'. In de adviesliteratuur treffen we adviezen aan als "Een speechtekst is een tekst die je

uitspreekt. Gebruik daarom geen schrijftaal." (Geel, 2004, p. 114) en "Een typische schrijftaaltekst [is] niet geschikt voor een toespraak, óók niet tijdens een symposium, conferentie of congres." (Korswagen, 1993, p. 117). Deze begrippen spreektaal en schrijftaal zijn niet bij voorbaat duidelijk. In dit artikel gaan we na of we via kwantitatief onderzoek de tekst van de professionele speechschrijvers van een aantal ministeries nader kunnen typeren als meer spreektaal of als schrijftaal. Onze bevindingen zijn gebaseerd op exploratief-kwantitatief onderzoek naar eigenschappen van een corpus speeches zoals ze "in de tas" zijn gegaan van de bewindspersonen, dan wel openbaar zijn gemaakt via officiële websites. We zijn sinds kort in de gelegenheid dit soort teksten te vergelijken met verschillende soorten meer en minder spontaan taalgebruik dankzij het Corpus Gesproken Nederlands (Van der Wouden, Hoekstra, Moortgat, Schuurman & Renmans 2002).

2 Kwantitatief onderzoek naar eigenschappen van teksten

Kwantitatief onderzoek naar eigenschappen van teksten heeft een lange traditie: al in 1851 suggereerde Augustus de Morgan dat kwesties rond de authenticiteit van bepaalde Bijbelboeken beslecht zouden kunnen worden door de gemiddelde woordlengtes van de brieven van Paulus te bestuderen (Kenny, 1982, p. 1). In de ruim anderhalve eeuw die er sindsdien zijn verstreken, heeft men nog vele andere dingen geteld en onderzocht en zijn er vele nieuwe technieken en algoritmen ontwikkeld; het moeizame rekenwerk kan tegenwoordig overgelaten worden aan de computer. Dat wil overigens nog helemaal niet zeggen dat heden ten dage automatisch en volstrekt betrouwbaar kan worden vastgesteld van wiens hand een onbekend of omstreden stuk is: op zijn best produceren de modernste technieken "encouraging results to what is acknowledged to be a difficult problem" (Holmes & Forsyth, 1995; vergelijk ook Van Dalen-Oskam, 2005 voor een toepassing in de Nederlandse letterkunde).

Anderzijds zijn er ook "gemakkelijke" problemen binnen dit veld: de taal waarin een bepaalde tekst is geschreven, is met relatief simpele middelen te raden (Cavnar & Trenkle, 1994). Een voorbeeld kan dit wellicht aannemelijk maken: een tekstfragment met veel korte woorden die de opeenvolging *th* bevatten, is met grote waarschijnlijkheid geschreven in het Engels (denk aan *this, that, those, they, them, thus,* en vooral *the*), terwijl *käyttöjärjestelmä ja sähköpostipalvelinohjelmisto ovat kaikki huomattavasti aikaisempaa tehokkaampia* beslist geen Engels is: het bevat dubbele klinkers en medeklinkers, zoals het Nederlands die ook heeft, maar de woorden zijn gemiddeld te lang en bevatten te weinig *e*'s en te veel umlauten om Nederlands te kunnen zijn, dus het zal wel Fins zijn.[1]

Kwantitatief onderzoek naar eigenschappen van tekstgenres is relatief zeldzaam, met Kessler, Nunberg & Schütze (1997) als een interessante uitzondering. Deze auteurs wijzen erop dat "genre" een notie is die niet gemakkelijk te operationaliseren is: "Genre is necessarily a heterogeneous classificatory principle, which is based among other things on the way a text was created, the way it is distributed, the register of language it uses, and the kind of audience it is addressed to."

Douglas Biber en zijn medewerkers hebben uitgebreid kwantitatief onderzoek gedaan naar eigenschappen van genres; Biber (1988) bijvoorbeeld rapporteert over onderzoek naar een zestigtal talige factoren, van bijwoorden van tijd tot zelfstandige naamwoorden, en van onpersoonlijke passief tot woordlengte. Interessant genoeg worden de resultaten van het onderzoek echter niet gegeven in termen van genres,

maar met behulp van dimensies als “informatief” vs. “betrokken” en “al of niet narratief”. Kessler et al. (1997) beschrijven daarentegen experimenten om teksten automatisch in genres in te delen. Hun conclusie is “that categorization decisions can be made with reasonable accuracy on the basis of surface cues.” In deze exploratieve bijdrage is het perspectief weer enigszins anders: hier zullen we trachten talige aspecten van ministeriële speeches te relateren aan andere eigenschappen van het genre, waarbij we ons ervan bewust zijn dat we de genrebenadering van Swales (1990) en Bahtia (1993) met tekstgrammaticale inslag (‘rhetorical moves’) buiten beschouwing laten. In paragraaf 4 besteden we aandacht aan lexicale frequenties, en in paragraaf 5 toetsen we onze bevindingen aan een aantal deelcorpora van het Corpus Gesproken Nederlands.

3 Onderzoeksopzet

Voor ons onderzoek hebben we een corpus samengesteld van 37 recente speeches van ministers en staatssecretarissen, deels afkomstig van de officiële sites van de departementen, deels onderhands verkregen. Tabel 1 biedt een overzichtje.

Tabel 1: Overzicht toesprakencorpus

Bewindspersoon	Aantal speeches	Gemiddelde lengte in woorden
Balkenende (minister-president)	9 (okt. 2002-maart 2005)	1334
De Graaf (minister voor Bestuurlijke vernieuwingen en Koninkrijksrelaties)	10 (maart-dec. 2004)	1214
Van der Hoeven (minister van Onderwijs)	4 (okt.-nov. 2005)	1556
Van der Laan (staatssecretaris van Cultuur)	5 (sept.-nov. 2005)	1494
Remkes (minister van Buitenlandse zaken)	9 (dec. 2003-nov. 2004)	1934

In totaal telt het corpus zo'n 56.000 woorden. De context waarvoor de verschillende toespraken waren bedoeld, is divers: van de officiële bekendmaking van het overlijden van Prinses Juliana tot de opening van een Academisch Jaar, van de opening van een congres tot het in ontvangst nemen van een boek.

Er zijn allerlei programmapakketten op de markt die het de onderzoeker mogelijk maken, snel corpusonderzoek te doen; ze verschillen in prijs, gebruiksgemak, en flexibiliteit. Voor het hier gerapporteerde onderzoek gebruikten we zowel het mild geprijsde, gebruiksvriendelijke WordSmith Tools (Scott, 2004) als het gratis, niet zo erg gebruiksvriendelijke, maar wel buitengewoon flexibele Ngram Statistics Package (NSP) (Banerjee & Pedersen, 2003).

4 Significante woorden en woordcombinaties

Wie in een tekst *brieke benen* tegenkomt, weet dat hij met een werk van Thomas Rosenboom van doen heeft, en aan de woorden *lederopdracht* en *verrekijk* alleen al herkent men de hand van Gerard Reve. Ook bij het identificeren van tekstgenres zou specifiek woordgebruik wel eens zijn nut kunnen afwerpen: Stubbs (2001, p. 7) geeft het voorbeeld van *warm front* (*warmtefront*) dat kenmerkend lijkt te zijn voor het teksttype weerbericht. Laten we daarom onze exercitie beginnen met een onderzoek naar lexicale frequenties. We werken van klein naar groot.

4.1 Woordfrequenties

Tabel 2 geeft een overzicht van de meest frequente woorden uit de toespraken. Ter vergelijking nemen we een kolom op met de meest frequente woorden uit het oudste corpus van het Nederlands (Uit den Boogaart, 1975).[2]

Tabel 2: Hoogfrequente woorden in toespraken en in het Corpus Eindhoven

	Toesprakencorpus	Corpus Eindhoven (1975)
N	Woord	Woord
1	de	de
2	en	van
3	het	een
4	van	het
5	een	in
6	in	en
7	dat	is
8	is	te
9	te	op

De voornaamste conclusie die we uit deze tabel kunnen trekken is, dat de taal van de toespraken in elk geval in dit opzicht niet op een opvallende manier verschilt van de taal van het oudere corpus, hoewel dat voornamelijk bestaat uit geschreven taal – het subcorpus gesproken Nederlands telt slechts ca. 150.000 woorden, oftewel een zesde van het totaal van het Corpus Eindhoven. Net als bij vergelijkbare tellingen in veel andere talen bestaat de kop van de frequentietabel uit korte functiewoorden (Kenny, 1982). Normaal duurt het wel tot positie 100 of nog later voor het eerste zelfstandig naamwoord of inhoudswerkwoord verschijnt.

4.2 Tweewoordcombinaties

Laten we daarom onze aandacht richten op opvallende combinaties van woorden. In tabel 3 staan de meest frequente combinaties bestaande uit twee woorden.

Tabel 3: Hoogfrequente tweewoordcombinaties

Rangorde	Cluster	Frequentie
1	van de	544
2	in de	327
3	voor de	146
4	en de	142
5	in het	142
6	van het	138
7	dat de	109
8	met de	108
9	aan de	107

De informatie die tabel 3 ons biedt is andermaal niet erg interessant: hoogfrequente combinaties van hoogfrequente functiewoorden zoals *van de*, *in de*, *voor de* en *van het* zijn hoogfrequent in de meeste tekstcorpora, dus daarin onderscheiden toespraken zich niet van andere soorten teksten.

Het heeft ook niet veel zin om lager in de tabel te gaan kijken, omdat de getallen dan snel te klein worden om betrouwbare conclusies te kunnen trekken. De resultaten zouden evenwel wel eens interessanter kunnen worden als we de frequentie van een combinatie konden relateren aan de kans op die combinatie. Dat is overigens geen nieuwe gedachte, sterker nog, er zijn inmiddels allerlei meer of minder complexe statistische significantietoetsen op de markt die elk hun voor- en nadelen, en hun voor- en tegenstanders hebben (Evert, 2004). Tamelijk populair zijn (varianten van) log-likelihood en tscore (Manning & Schütze, 1999).[3] Die toetsen gaan verschillend om met zeldzame data en scheve verdelingen, vandaar dat ze verschillende opvattingen hebben over wat de meest significante combinaties zijn (zie tabel 4) (een hoge score in deze tabel geeft aan dat de kans erg klein is dat deze combinatie resultaat is van toeval). We geven wederom alleen de top van de tabel, omdat ook hieronder de frequenties te laag worden om van betekenis te zijn.

Tabel 4: Meest significante tweewoordcombinaties

Combinatie	Log-likelihood	Combinatie	t-score
van de	1251.45	van de	19.43
dames en	725.12	in de	14.75
en heren	691.28	dames en	9.80
in de	667.28	en heren	9.80
publieke omroep	461.94	voor de	9.33
Den Haag	337.65	in het	9.13
de overheid	320.37	de overheid	8.60
ik ben	286.95	aan de	8.31
gekozen burgemeester	282.48	over de	8.23

We zien nu een dramatisch verschil met tabel 3, maar toch staat ook hier de combinatie *van de* bovenaan, als meest significant volgens de beide toetsen, terwijl ook *in de* hoog scoort. Dit zou erop kunnen wijzen dat het genre van de ministeriële toespraak meer dan gemiddeld gebruik maakt van complexe zelfstandig-naamwoordgroepen van het type *de ontdekking van de hemel*, maar dat zou nader onderzoek moeten uitwijzen, waarbij het telwerk nog eens herhaald wordt aan een syntactisch ontlede versie van het corpus, die er nu nog niet is.

De combinaties *dames en* en *en heren* scoren ook heel hoog, maar die zijn overduidelijk deel van een groter geheel, dus daar zullen we pas zometeen bij de driewoordcombinaties meer over zeggen. Verder zien we dat ook *de overheid* volgens beide toetsen een belangrijke combinatie is. En ten slotte bevestigt deze tabel wat al uit de literatuur bekend was, namelijk dat t-score hoge frequentie van een combinatie hoger waardeert dan log-likelihood. Wie dit soort statistiek wil gebruiken om automatisch belangrijke begrippen uit een tekst te destilleren, is duidelijk beter uit met log-likelihood: die test signaleert *publieke omroep*, *Den Haag*, *de overheid*, en *gekozen burgemeester*. Zonder te weten waar de toespraken over gingen, kunnen we ons toch, op basis van onze kennis van de Haagse politiek, heel goed voorstellen dat dat interessante thema's zijn waar bewindslieden in een gelegenheidstoespraak best eens iets over te berde zouden kunnen willen brengen.

4.3 Driewoordcombinaties

Geïnspireerd door dit succesje kijken we vervolgens naar combinaties bestaande uit 3 woorden. Tabel 5 geeft de meest frequente.[4]

Tabel 5: De meest frequente trigrammen in het toesprakencorpus

	Cluster	Frequentie
1	dames en heren	102
2	de publieke omroep	31
3	van de overheid	30
4	het gebied van	20
5	een van de	19
6	op het gebied	19
7	van de Europese	18
8	het is een	16
9	op dit moment	16
10	de aanpak van	15
11	en dat is	15
12	de gekozen burgemeester	15

Ook deze tabel biedt weer interessant materiaal: kennelijk waren *de publieke omroep* en *de gekozen burgemeester* belangrijke thema's in de tijd dat deze toespraken geschreven zijn. En ongeacht het onderwerp komen bouwstenen als *van de overheid*, *op*

het gebied van, *op dit moment* en *de aanpak van* altijd van pas in een krachtdadig politiek betoog. Opvallender wellicht is echter dat de combinatie *dames en heren* verreweg de meest voorkomende driewoordcombinatie is in het hele toesprakencorpus.

Een voor de hand liggende tegenwerping is, dat het juist helemaal niet zo verrassend is dat *dames en* heren zo vaak voorkomt: toespraken horen immers ergens mee te beginnen: met een opening. In het geval van de ministeriële toespraak is dat kennelijk meestal de standaardformule *dames en heren*. Anders dan bijvoorbeeld *geachte ingenieurs* of *vrienden van de Vara*, past deze opening op vrijwel iedere situatie. *Dames en heren* mag dan wel niet origineel zijn, nuttig is hij wel: op deze manier weet iedereen waar hij aan toe is. De minister maakt met zijn openingscliché duidelijk dat hij begint te spreken, en dat er opgelet, of op zijn minst gezwegen moet worden.

We willen hier echter twee kanttekeningen bij maken. In de eerste plaats blijkt *dames en heren* niet alleen gebruikt wordt als opening van de toespraak: we hebben slechts 39 toespraken en maar liefst 102 voorkomens van de combinatie. Nadere beschouwing van de teksten van de speeches wijst uit dat de combinatie ook gebruikt wordt ter inleiding van een samenvatting of conclusie of iets dergelijks (vergelijk voor het gebruik van deze frase De Jong & Andeweg, 2006). Het volgende voorbeeld illustreert dit soort gebruik:

> *Veel Amerikaanse universiteiten vinden dat verspreiding en exploitatie van kennis gewoon behoort tot hun kerntaak. Daardoor zijn ze sterk gericht op de samenleving en de samenleving ook op hen.*
> *Dames en heren,*
> *Ik heb u een aantal aangrijpingspunten genoemd voor het Innovatieplatform.*

In de tweede plaats blijkt ons corpus bovendien een aantal toespraken te bevatten die toch niet beginnen met *dames en heren*. Interessant genoeg gaat het hier steeds om radiotoespraken van premier Balkenende, en betreft het zonder uitzondering speeches bij gelegenheid van geboorte of overlijden van een lid van de koninklijke familie. We geven twee voorbeelden (alleen de openingszinnen):

> *Vanmiddag, om 1 minuut na 5 uur, zijn de Prins van Oranje en Prinses Máxima de trotse en gelukkige ouders geworden van een dochter.*
> *Omringd door haar directe familie, is Prinses Juliana thuis op Paleis Soestdijk overleden.*

We constateren dat alle onderzochte "gewone" toespraken (met één uitzondering, die met *geachte aanwezigen* begint) openen met *dames en heren* (eventueel voorafgegaan door *Koninklijke Hoogheid* of iets dergelijks), terwijl geen van de radiotoespraken ter gelegenheid van een gebeurtenis in het Koninklijk Huis deze opening gebruikt. Dit feit alleen al suggereert dat deze geboorte- en overlijdenstoespraken een eigen genre vormen, dat in elk geval in dit opzicht verschilt van "gewone" ministeriële toespraken. Een andere mogelijkheid is, dat de minister-president voor een andere opening kiest omdat die beter geschikt wordt geacht voor het medium radio (en alleen daarom al zou dit soort toespraken een ander genre kunnen vormen – vergelijk het citaat van Kessler et al. hierboven). In elk geval is het bovendien zo, dat een officiële mededeling van geboorte of overlijden, anders dan een opening of een boekpresentatie, nauwelijks geschikt is om beleidsvoornemens en dergelijke zaken naar voren te brengen of uit te proberen.

4.4 Langere combinaties

Voor de volledigheid bespreken we ook nog de vier- en meerwoordcombinaties. Bij de vierwoordcombinaties springt alleen de voorzetseluitdrukking *op het gebied van* eruit, met maar liefst 19 voorkomens, en bij de langere alleen nog *alleen het gesproken woord geldt*, dat als standaard-disclaimer bovenaan veel van de speechteksten staat afgedrukt maar vanzelfsprekend niet bedoeld is om uitgesproken te worden (Veltman et al., 2003). De overige vier- en meerwoordcombinaties zijn veel minder frequent en zo op het oog ook minder interessant.

Is *op het gebied van* dan een verbinding die typerend is voor ministeriële speeches (zoals *het kan* (*toch*) *niet zo zijn dat* (Pardoen, 1994) dat lijkt te zijn voor het Haagse debat)? Om dat na te gaan hebben we de distributie van de uitdrukking binnen het toesprakencorpus onderzocht. Van de 19 voorkomens bleken er maar liefst 14 afkomstig zijn uit de toespraken van minister Remkes. Kennelijk heeft hij, of waarschijnlijk eerder zijn tekstschrijver, een bijzondere voorkeur voor *op het gebied van* – of de verantwoordelijke heeft nagelaten deze zinswending te schrappen bij het tot toespraak verwerken van bouwstenen uit ambtelijke rapporten. De andere bewindslieden gebruiken de vaste verbinding op zijn hoogst twee keer in al hun toespraken bij elkaar, dus er is geen reden haar te beschouwen als kenmerkend voor het genre als geheel.[5] Als clichés de eerste woorden voor eerste gedachten zijn (Burger & De Jong, 1997), dan blijken speechschrijvers in staat hun opwellingen in tweede instantie te vervangen door een meer precieze boodschap in frissere bewoordingen.

Ter afsluiting van deze paragraaf over exploratief-kwantitief onderzoek naar woordcombinaties kunnen we concluderen dat deze combinaties ons lijken te kunnen helpen bij het vinden van (sub-)genres. Ministeriële toespraken die met *dames en heren* beginnen, blijken een andere functie te hebben, of voor een ander medium te zijn gemaakt, dan toespraken zonder die opening. Bovendien blijkt een vaak voorkomende vaste woordverbinding soms te kunnen verraden van welk departement of uit wiens tekstverwerker een toespraak afkomstig is.

5 Vergelijking met het Corpus Gesproken Nederlands

In de voorgaande paragraaf hebben we in het toesprakencorpus een paar eigenaardigheden gevonden die weleens kenmerkend voor het genre van de ministeriële toespraak zouden kunnen zijn. Om na te gaan of dat inderdaad het geval is, en om andere eigenaardigheden van de ministeriële toespraak op het spoor te komen, vergelijken we in deze paragraaf een aantal eigenschappen van ons toesprakencorpus met teksten uit het Corpus Gesproken Nederlands (CGN) (Oostdijk, 2000). Dat onlangs (2004) voltooide corpus is bedoeld als een representatieve steekproef van het hedendaagse gesproken Nederlands. Het bevat zo'n 9 miljoen woorden (900 uur) spraak van volwassen moedertaalsprekers van verschillende achtergrond, uit verschillende regio's uit Nederland en Vlaanderen, in een aantal verschillende spreeksituaties. We richten ons hier alleen op sprekers uit Nederland, die verantwoordelijk zijn voor ongeveer twee derde van het corpus, oftewel zo'n 5,7 miljoen woorden. In deze vergelijking beperken we ons tot het woordniveau.

We kijken eerst naar het woordgebruik in de beide corpora. Tabel 6 geeft een overzicht van de meest frequente woorden in de beide corpora, zoals berekend door Wordsmith Tools.

Tabel 6: De meest frequente woorden in het toesprakencorpus en in het CGN

	Toespraken		CGN	
N	Woord	%	Woord	%
1	de	5.75	ja	3.28
2	en	3.24	de	2.72
3	het	3.16	dat	2.66
4	van	3.08	uh	2.063
5	een	2.41	en	2.41
6	in	2.05	ik	2.13
7	dat	1.84	een	1.90
8	is	1.51	't	1.67
9	te	1.21	je	1.54
10	voor	1.13	die	1.51
		25.38		19.83

De verschillen tussen de twee corpora zijn dramatisch: het meest frequente woord van het CGN, *ja*, ontbreekt bijvoorbeeld in de top 10 bij de toespraken (het woordje staat daar met 7 voorkomens op een gedeelde 866e plaats). En de cijfers suggereren ook een verschil in lexicale rijkdom: waar de tien meest frequente woord-types samen verantwoordelijk zijn voor een kleine 20% van de woord-tokens in het CGN, corresponderen de tien meest frequente woord-types bij de toespraken met maar liefst 25% van alle woord-tokens. Nadere analyse van die lexicale rijkdom (een problematische notie, vergelijk Tweedie & Baayen, 1998) voert buiten het kader van dit exploratieve onderzoek.

Overigens moeten we hier meteen de kanttekening bij maken dat het CGN verre van homogeen is. Tabel 7 geeft een overzicht van de frequentie van *ja* in de grootste subcorpora van het CGN, en van het meest frequente woord in elk van die subcorpora.

Tabel 7: Ja *en andere frequente woorden in het Corpus Gesproken Nederlands*

Subcorpus	# *ja*	%	nr. 1	%
dialoog (N=1815735)	1	4.52	ja	4.52
telefoon (N=1286962)	1	5.96	ja	5.96
radio (N=1001366)	13	1.16	de	4.57
Voorgelezen (N=558543)	135	0.09	de	4.78
les (N=307876)	8	2.00	dat	2.91
interview (N=264621)	4	2.86	uh	3.98
(politiek) debat (N=220094)	35	0.47	de	5.02
televisie (N=196865)	17	0.92	de	4.02
simulated business (N=140349)	3	3.56	de	4.33
presentatie (N=63492)	97	0.13	uh	3.85

Deze tabel spreekt bijna voor zichzelf: dat *ja* het meest frequente woord is in het CGN komt doordat het het meest frequente woord is in de grootste subcorpora. Daar kunnen we aan toevoegen dat de twee grootste subcorpora van het CGN dialogen bevatten. Met enkele slagen om de arm kunnen we stellen dat *ja* in het huidige gesproken Nederlands vooral een dialoogwoord is, dat onder meer gebruikt wordt ter regulering van beurten (Van der Wouden, 2006). De ministeriële toespraak daarentegen is bij uitstek een monoloog: geen samenspraak maar een toespraak, vandaar dat *ja* nauwelijks voorkomt in het toesprakencorpus.

Eenzelfde redenering gaat op voor *uh*: dat is hoogfrequent in spontaan mondeling taalgebruik, en het komt ook veel voor in uitgesproken toespraken (vergelijk Van De Mieroop, 2005), maar het komt natuurlijk niet voor in de geschreven versies.

Als we alleen kijken naar de meest frequente woorden, dan vertoont de ministeriële toespraak nogal wat overeenkomsten met (voorgelezen) schrijftaal enerzijds en met de taal van het parlementair debat anderzijds. De toespraak heeft weinig met spontane spreektaal te maken. Of, om met Jelle de Vries te spreken, "*Schrijftaal* is niet per definitie 't zelfde als *geschreven taal*, want dat kan ook opgeschreven spreektaal zijn. Zo is *spreektaal* wat anders dan *gesproken taal*. Dat laatste is immers, zeker in taalgebruik voor openbare consumptie, vaak voorgelezen schrijftaal" (De Vries, 2001, p. 11, vergelijk ook Jansen, 1981).

In kwantitatief onderzoek kan men sinds de al genoemde De Morgan, en zeker sinds de leesbaarheidsformule van Rudolf Flesch (1960), moeilijk heen om maten als woord- en zinslengte. In tabel 8 staan de woordlengtes van de diverse speeches, geordend op gemiddelde woordlengte, in vergelijking met een aantal interessante subcorpora van het CGN.

Tabel 8: Gemiddelde woordlengtes

Wie	Aantal woorden	Gemiddelde woordlengte	Standaarddeviatie
Balkenende overige speeches	8943	5.19	3.27
Remkes	17430	5.17	3.52
Wie	Aantal woorden	Gemiddelde woordlengte	Standaarddeviatie
Vd Laan	7526	5.16	3.24
De Graaf	13189	5.15	3.34
Balkenende Koninklijk Huis	3138	5.04	3.03
Vd Hoeven	6284	4.98	3.15
Toespraken totaal	56510	5.14	3.33
CGN NL voorlees	558543	4.64	2.69
CGN NL debat	22094	4.58	3.17
CGN NL presentatie	63492	4.54	3.07
CGN NL dialoog	1815735	3.72	2.23

De verschillen tussen de speeches onderling zijn behoorlijk wat kleiner dan die met de andere genres in de tabel. Net als bij de factor frequente woorden, lijken speechteksten ook wat betreft de gemiddelde woordlengte meer op voorleestaal, dus op geschreven taal, dan op andere vormen van gesproken taal; de dialoog staat ook in dit opzicht het verst van de speechtaal.

Zinslengte tot slot is een buitengewoon problematische notie bij gesproken taal. Het vaststellen van zinsgrenzen in mondelinge spraak is problematisch en subjectief. Het corpus bevat wel punten, maar die zijn een artefact van de orthografische transcriptie: mensen spreken geen punten uit. Sterker nog: mensen praten niet in zinnen, dat wil zeggen, de zin is een schrijftaaleenheid en geen spreektaaleenheid (Miller & Weinert, 1998, p. 71). Kortom, we kunnen de speeches op deze factor niet vergelijken met het corpus.

6 Afsluitende opmerkingen

Met dit kwantitatief-exploratieve onderzoek hebben we geprobeerd enkele talige eigenschappen van ministeriële toespraken scherper in beeld te krijgen. Wat hebben deze exploraties opgeleverd? De tekst van de ministeriële toespraak lijkt, anders dan sommige adviezen suggereren, meer op schrijftaal dan op spontane spreektaal. Dit uit zich zowel in woordkeuze als gemiddelde woordlengte. Ondertussen nemen bewindslieden niet zelden de vrijheid meer of minder af te wijken van de geprepareerde tekst, door woordherhalingen, aarzelingen (*uh*), kleine verhaspelingen enzovoort (vergelijk Veltman et al., 2003 en De Jong & Andeweg, 2004).

Speechschrijvers schrijven een taal die een stuk formeler (in de zin van de factoren die we hier gemeten hebben) is dan spontane spreektaal, en zelfs dan gemiddelde schrijftaal (de voorgelezen taal van het CGN bevat nogal wat bellettrie). Ondertussen is die taal al sterk verbeterd ten opzichte van die van de ambtelijke stukken

waar de inhoudelijke gedeeltes van de rede doorgaans op gebaseerd zijn – speechschrijvers zijn wel vakmensen. Van de taal van de speechschrijver maakt de minister dan in zijn *actio* een iets spontaner klinkende taalvariant, maar het blijft een formele, geprepareerde toespraak, die niet in Jip-en-Janneketaal gesteld hoort te zijn.

Kwantitatief onderzoek naar taalverschijnselen kan ons helpen typerende eigenschappen van teksten in het algemeen, en van de ministeriële toespraak in het bijzonder, op het spoor te komen. Dat geldt des te meer sinds de voltooiing van het Corpus Gesproken Nederlands, dat een multi-dimensionaal referentiekader vormt van verschillende varianten van het hedendaags gesproken Nederlands. Bijna even boeiend zijn de doodlopende wegen die we tijdens onze exploraties zijn tegengekomen. Fraaie politieke clichés (*naar de ... toe, het kan* (*toch*) *niet zo zijn dan* en zeker *een situatie waarin vergaande participatie van en interactie met de getroffen bewoners voorop stond*) komen, afgezien van *op het gebied van*, met de hier gehanteerde methodes niet bovendrijven, om de simpele reden dat ze weliswaar opvallend zijn, maar niet frequent genoeg om in een corpus van deze omvang boven te komen drijven. Een uitdrukking als *het kan* (*toch*) *niet zo zijn dan* komt namelijk wel een aantal malen voor in het onderdeel "debate" van het CGN. Dat betekent wellicht dat dat soort politieke clichés tot het vaste improvisatie-instrumentarium van Haagse politici behoort, waar professionele tekstschrijvers goed genoeg zijn om ze te vermijden.

Noten

1. G. van Noord stelt een implementatie van het algoritme van Cavnar & Trenkle beschikbaar via http://odur.let.rug.nl/~vannoord/TextCat/. Het programma herkent deze noot als Nederlands, en het Finse zinnetje in de hoofdtekst (afkomstig van http://www.cs.helsinki.fi) als Fins of Estisch. Tekstverwerkers bevatten dikwijls een vergelijkbaar algoritme, zodat de tekstverwerker "weet" wanneer Engelse, en wanneer Nederlandse of Finse spellingscorrectie of afbreekregels moeten worden toegepast.
2. Onze tellingen zijn niet geheel vergelijkbaar met die van Uit den Boogaart, omdat die de woorden opsplitst naar woordsoort. Ter illustratie: als je de verschillende vormen van *dat* (voegwoord, aanwijzend voornaamwoord en betrekkelijk voornaamwoord) van tabel C (p. 426 en volgende) bij elkaar optelt, staat *dat* ook in de top 10.
3. Uit documentatie NSP (V0.71): "The log-likelihood ratio measures the deviation [sic] between the observed data and what would be expected if <word1> and <word2> were independent. The higher the score, the less evidence there is in favor of concluding that the words are independent." "The T-score is defined as a ratio of difference between the observed and the expected mean to the variance of the sample."
4. Er zijn nog nauwelijks statistische toetsen voor de significantie van drie- en meerwoordcombinaties (Villada Moirón, 2005). Gelukkig zijn de ruwe cijfers op zichzelf in dit geval al veelzeggend.
5. In het Corpus Gesproken Nederlands komt de voorzetseluitdrukking *op het gebied van* relatief vaker voor in het subcorpus "debat", dat veel materiaal uit de Haagse politiek bevat. De uitdrukking is dan misschien niet typerend voor het genre ministeriële speech, maar mogelijk wel voor de taal van Den Haag.

Literatuur

Andeweg, B., & Jong, J. de (2004). *De eerste minuten: attentum, benevolum en docilem parare in de inleiding van toespraken*. Den Haag: Sdu Uitgevers.

Banerjee, S., & Pedersen, T. (2003). The Design, Implementation, and Use of the Ngram Statistics. *Proc. Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, Mexico City.

Biber, D. (1988). *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Burger, P., & De Jong, J. (1997). *Handboek stijl. Adviezen voor aantrekkelijk schrijven*. Groningen: Martinus Nijhoff.

Cavnar, W.B., & Trenkle, J.M. (1994). N-Gram-Based Text Categorization. *Proc. Third Annual Symposium on Document Analysis and Information Retrieval* (161-175), Las Vegas, NV: UNLV Publications/Reprographics.

Dalen-Oskam, K. van (2005). De list van het lexicon. Auteursonderscheiding met behulp van computer-ondersteunde woordenschatanalyse. *Nederlandse Letterkunde, 10,* 212-233.

Ensink, T., & Sauer, C. (2003). *The Art of Commemoration, Fifty Years after the Warsaw Uprising*. (Discourse Approaches to Politics, Society and Culture). Amsterdam: John Benjamins.

Evert, S. (2004). *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Diss. Institut für maschinelle Sprachverarbeitung, Stuttgart.

Flesch, R. (1960). *How to write, speak and think more effectively*. New York: Harper & Row.

Geel, R. (2004). *Speech! Speech! Schrijf een succesvolle toespraak.* Bussum: Coutinho.

Holmes, D.I., & Forsyth, R.S. (1995). The Federalist Revisited: New Directions in Authorship Attribution. *Literary and Linguistic Computing, 10*, 111-127.

Jansen, F. (1981). *Syntaktische konstrukties in gesproken taal*. Diss. Leiden.

Jong, J. de, & Andeweg, B. (2004). *Speeches van OCenW. Retorische analyse en ontvangst bij publiek en pers.* Interne publicatie Universiteit Leiden / TU Delft.

Jong, J. de, & Andeweg, B. (2006). De problematische peroratio. Deze bundel.

Kenny, A. (1982), *The computation of style*. Oxford [etc.], Pergamon Press.

Kessler, B., Nunberg, G., & Schütze, H. (1997). Automatic Detection of Text Genre. *Proceedings ACL/EACL* 1997, Madrid, 32-38.

Korswagen, C.J.J. (Red.) (1993). *Drieluik mondelinge communicatie. I Doeltreffend spreken, presenteren en instrueren*. 2e dr. Houten/Zaventem: Bohn Stafleu Van Loghum.

Manning, C., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, Mass.: The MIT Press.

Miller, J., & Weinert, R. (1998). *Spontaneous spoken speech. Syntax and Discourse*. Oxford: Clarendon.

Oostdijk, N. (2000). Het Corpus Gesproken Nederlands. *Nederlandse Taalkunde,* 5, 280-284.

Pardoen, J. (1994). Het kan (toch) niet zo zijn dat ..., een pragma-syntactische benadering. In R. Boogaart & J. Noordegraaf (Red.), *Nauwe betrekkingen. Voor Theo Janssen bij zijn vijftigste verjaardag* (pp. 203-212). Amsterdam & Münster, Stichting Neerlandistiek VU & Nodus Publikationen.

Scott, M. (2004). *WordSmith Tools version 4*. Oxford: Oxford University Press.

Stubbs, M. (2001). *Words and phrases. Corpus studies of lexical semantics*. Oxford [etc.]: Blackwell.

Swales, J. (1990). *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge University Press.

Tweedie, F.J., & Baayen, R.H. (1999). How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities, 32*, 323-352.

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

Vries, J. de (2001). *Onze Nederlandse Spreektaal*. Den Haag: SDU Uitgevers.

Veltman, W., Andeweg, B., & Jong, J. de (2003). Alleen het gesproken woord telt: In hoeverre en waarom sprekers afwijken van speechteksten geschreven door professionele speechschrijvers. In L. van Waes, P. Cuvelier, G. Jacobs, & I. de Ridder (Red.), *Studies in taalbeheersing 1* (pp. 463-475). Assen: Koninklijke Van Gorcum.

Van De Mieroop, D. (2005). *Identiteitsconstructie in informatieve speeches. Een multi-methodologische analyse*. Diss. Universiteit Antwerpen.

Villada Moirón, B. (2005). *Data-driven Identification of Fixed Expressions and their Modifiability*. Diss. Groningen.

Wouden, T. van der, Hoekstra, H., Moortgat, M., Schuurman, I., & Renmans, B. (2002). Syntactische annotatie voor het Corpus Gesproken Nederlands (CGN). *Nederlandse Taalkunde, 7*, 335-352.

Wouden, T. van der (2006). On the phraseology of stop words. Ms. Leiden, onder review.