TIJDSCHRIFT VOOR TAALBEHEERSING
www.tijdschriftvoortaalbeheersing.nl
Uitgave: Amsterdam University Press

Tekstgenres analyseren op lexicale complexiteit met T-Scan

Henk Pander Maat & Nick Dekker

TT 38 (3): 263–304
DOI: 10.5117/TVT2016.3.PAND

Abstract

Using T-Scan to analyse the lexical complexity of text genres

T-Scan is a tool for the automatic analysis of Dutch text. This paper presents the first large-scale corpus analysis with T-Scan, focusing on lexical complexity. A collection of nearly 1000 text specimens was assembled, containing ten genres: travel blogs, celebrity news features, novels, textbooks for vocational secondary schools, textbooks for general secondary schools, news reports, opinion pieces, political programs, medical advice texts and research articles. The lexical complexity features in the analysis include morphology, word frequency, various word concreteness indices, personal pronouns, names and verb tense. Systematic genre differences are found, such that a genre detection model comprising 18 T-Scan features correctly identifies 83 percent of the corpus texts. Most lexical features differentiating genres intuitively relate to text topic complexity. A closer analysis is offered of the contrast between the two textbook samples in the corpus, which differ only in the educational levels they cater for. Again, topic variation seems a more important factor than stylistic variation. We demonstrate a new method to examine stylistic variation, which consists of within-genre comparisons using the genre prediction; more specifically, 'deviant' texts are compared to 'typical' members of their genre.

Keywords: lexical complexity, corpus research, automatic text analysis, readability, stylistic variation

1 Inleiding

T-Scan is een tool voor automatische tekstanalyse,[1] [2] ontwikkeld door een team van onderzoekers en programmeurs uit Utrecht, Tilburg en Nijmegen (Pander Maat e.a. 2014; Pander Maat e.a. 2016). De tool is toegankelijk voor onderzoeksdoeleinden (zie https://webservices-lst.science.ru.nl). Dit artikel demonstreert de mogelijkheden van T-Scan wat betreft het beschrijven van lexicale complexiteit. We laten in een corpusanalyse op bijna 1000 teksten zien hoe genres verschillen in complexiteit. De corpusanalyse vindt plaats in drie fasen. Eerst worden tien genres vergeleken, en wordt een genrevoorspellingsmodel gebouwd met lexicale complexiteitskenmerken. Vervolgens wordt ingezoomd op twee verwante genres, om na te gaan in hoeverre complexiteitsverschillen inhoudelijk dan wel stilistisch van aard zijn. Ten slotte demonstreren we een nieuwe manier om de stilistische complexiteit van individuele teksten te beschrijven, namelijk door gebruik te maken van het genrevoorspellingsmodel.

2 Genres en tekstcomplexiteit

Onder een genre verstaan we een klasse van interacties of boodschappen die cultureel herkenbaar is op basis van een specifieke combinatie van contextuele en tekstuele kenmerken (zie Trosborg 1997, Lee 2001 en Biber & Conrad 2009). In eerste instantie kunnen genres contextueel worden gedefinieerd met behulp van de volgende kenmerken:

- het maatschappelijke domein waarbinnen het genre functioneert;
- het typerende thema van de boodschappen;
- het medium waarin de genreboodschappen verzonden worden;
- het communicatieve doel van de boodschappen;
- de doelgroep van de boodschappen.

In tweede instantie zijn genres herkenbaar aan conventionele opbouwprincipes en aan centrale teksthandelingen (Pander Maat 2002), in de Engelse literatuur ook wel aangeduid als ‘text types’ (Trosborg 1997) of ‘text forms’ (Werlich 1982). Ten slotte kennen genres een karakteristiek taalgebruik en veelal ook een typerende vormgeving. De taalgebruikscomponent wordt vaak ‘register’ genoemd (Biber & Conrad 2009), waarbij register staat voor situationeel bepaalde kenmerken van taalgebruik. Dit artikel gaat over lexicale registerverschillen.

In Tabel 1 geven we karakteristieken van de tien genres die in ons corpus zijn opgenomen, zonder in te gaan op taalgebruik en vormgeving.

Tabel 1 Tien genres omschreven in termen van context, thema en structuur

	Genre	Domein	Thema	Medium	Communicatiedoel	Doelgroep	Teksthandelingen (T) Structuur (S)
1	***Reisverslag***	Vrije tijd	Reisbelevenis	Digitaal	Informatief	Bekenden en aspirant-reizigers	T: Uiteenzetten, beschrijven, vertellen S: vaak chronologisch of geografisch
2	***Celebrity-nieuwsbericht***	Media	Belevenis beroemdheid	Papier / digitaal	Entertainment	Human-interestlezers	T: Vertellen, berichten S: ligt niet vast
3	***Roman***	Cultuur	Fictief verhaal	Papier / digitaal	Entertainment	Fictielezers	T: Vertellen S: ligt niet vast
4	***Schoolboek vmbo***	Onderwijs	Leerstof	Papier	Educatief	Leerling onderbouw vmbo	T: Uiteenzetten, beschrijven, vertellen S: verschilt per vak
5	***Schoolboek havo/vwo***	Onderwijs	Leerstof	Papier	Educatief	Leerling bovenbouw havo / vwo	T: Uiteenzetten, beschrijven, vertellen S: verschilt per vak
6	***Nieuwsbericht***	Media	Nieuwsfeit	Papier / digitaal	Informatief	Geïnteresseerde burger	T: Meedelen S: Kern-uitwerking
7	***Opiniestuk in krant***	Media	Nieuwsfeit	Papier / digitaal	Persuasief	Geïnteresseerde burger	T: Argumenteren, aanbevelen S: ligt niet vast
8	***Medisch advies***	Zorg	Medische klacht	Papier / digitaal	Informatief / instructief	Patiënten	T: Uiteenzetten, instrueren S: probleem-oplossing
9	***Verkiezingsprogramma***	Politiek	Plannen politieke partij	Papier / digitaal	Persuasief	Geïnteresseerde burger	T: Uiteenzetten, argumenteren S: opsomming beleidsterreinen
10	***Onderzoeksartikel***	Wetenschap	Onderzoek	Papier / digitaal	Informatief / persuasief	Onder-zoekers	T: Rapporteren S: inleiding – kader – resultaten – conclusie

Als we de variatie in de verschillende kolommen van Tabel 1 bezien, stellen we vast dat er een behoorlijke variatie bestaat in domeinen (7 stuks), thema's (8), doelgroepen (8) en communicatiedoelen (informatief, instructief, educatief, persuasief, entertainment). Hetzelfde geldt voor teksthandelingen en principes van tekstopbouw. Veel minder variatie is er in media, omdat het corpus geheel bestaat uit papieren of digitale geschreven teksten.

Tabel 1 bevat twee 'minimale paren': paren van genres die vooral op één contextuele dimensie verschillen. Nieuwsbericht en opiniestuk verschillen wat betreft communicatiedoel, en vmbo- en havo-schoolboeken wat be-

treft doelgroep. Dit soort paren geeft ons de kans om effecten van afzonderlijke dimensies op taalgebruik na te gaan, en dat zullen we hieronder ook doen voor de schoolboeken.

In de literatuur over genres is niet veel aandacht voor de doelgroepdimensie. Toch is deze dimensie van belang. Laten we haar buiten beschouwing in onze genredefinitie, dan zouden kinderboeken en romans als één genre gezien worden, net als onderzoeksartikelen en populairwetenschappelijke artikelen. Die consequentie lijkt ons niet acceptabel. Genres zijn mede gedefinieerd in termen van de discourse-gemeenschap die zij bedienen (Fahnestock 1986; Swales 1990).

Een nieuwe tak in de literatuur over genrestijlen vormt het computationele werk over genredetectie (zie bijvoorbeeld Kanaris & Stamatatos 2009, Stamatatos et al. 2000a, Stamatatos et al. 200b, Ashegi et al. 2014). Maar in die literatuur vindt men geen kwalitatieve informatie over de complexiteit van verschillende tekstgenres. Men probeert genres te onderscheiden op basis van kenmerken als de frequentie van opeenvolgingen van 4 letters, de frequentie van de 10 of 50 meest gebruikte woorden (veelal functiewoorden), de frequentie van woordsoorten, opeenvolgingen van woordsoorten, de lengte van de tekst, de lengte van de zinnen en de lengte van de woorden. Geen van die kenmerken wordt nader geïnterpreteerd. Omdat veel studies gaan over genres van webdocumenten, is er daarnaast ook aandacht voor niet-tekstuele kenmerken als URL's of HTML-tags. Verder is er in deze studies veel aandacht voor de prestaties van specifieke machineleer-algoritmes. Dat is begrijpelijk gezien het doel van dit onderzoek, te weten het vinden van een handige en effectieve manier om genres te onderscheiden.

Voor dit artikel is deze literatuur minder goed bruikbaar, omdat ons doel heel anders is: op basis van automatische analyse de complexiteit van teksten beschrijven. We zullen hieronder een genredetectie-analyse doen, maar die analyse speelt in ons onderzoek een heel andere rol dan in de computationele literatuur. We zullen er een genrevoorspellingsmodel uit afleiden dat als referentiepunt kan dienen bij het bespreken van individuele teksten op complexiteit. Dat betekent ten eerste dat we ons beperken tot kenmerken die plausibel te relateren zijn aan tekstcomplexiteit. Ten tweede gaat het ons niet om het optimaliseren van de voorspelling door verschillende algoritmes uit te proberen. Wel moet voorspellende waarde van een het model behoorlijk zijn, wil het gebruikt kunnen worden als referentiepunt.

Dit brengt ons op de volgende onderzoeksvragen:

1 Welke kenmerken biedt T-Scan voor het beschrijven van lexicale complexiteit?
2 In hoeverre verschillen tekstgenres in lexicale complexiteit?
3 Zijn deze verschillen tussen genres groot genoeg om te komen tot een genrevoorspellingsmodel?
4 Is dit model bruikbaar om afwijkende teksten te identificeren?

3 Corpusbronnen

De tien genres zijn hierboven algemeen geïntroduceerd. We gaan nu verder in op de bronnen die gebruikt zijn bij de samenstelling van het corpus.[3] Het streven was dat elk genre vertegenwoordigd zou zijn met 100 teksten of tekstfragmenten van minimaal 300 woorden, waarbij zo veel mogelijk auteurs per genre zijn opgenomen. Hieronder wordt de corpusverzameling per genre verder toegelicht. Tabel 2 geeft een overzicht van de aantallen teksten, hun lengte en de verschijningsjaren.

Reisverslagen
Het gaat hier om reisverslagen van 'amateurs' van websites als waarbenjij.nu en Kras. De teksten beschrijven belevenissen van individuele reizigers. De teksten zijn ingekort tot ongeveer 300 woorden. Elk werelddeel is met een aantal reisbestemmingen vertegenwoordigd in de teksten.

Celebrity-nieuws
We kozen 100 teksten van roddelwebsites zoals die van RTL Boulevard (20), ze.nl (12), Vrouwonline (11), en uit roddelrubrieken in AD.nl (15) en AD.nl (5); deze teksten zijn niet persoonlijk ondertekend, zodat we niet zeker zijn over het aantal auteurs dat erbij betrokken is. Bij de ondertekende teksten dragen enkele journalisten meer dan een bericht bij. Enkele representatieve titels zijn *Whitney Houston is platzak* en *Liefde op de set*. De teksten zijn volledig overgenomen, waardoor de lengte varieert.

Romans
Het gaat hier om 100 fragmenten uit naoorlogse Nederlandse romans voor volwassenen, waarbij de term 'roman' ruim is opgevat. Er komen gevestigde literaire auteurs in het corpus voor, zoals Hermans, Reve en Zwagerman, maar er is ook plaats voor 'chicklit' en thrillers. De 100 teksten zijn van de hand van 87 auteurs; enkele auteurs komen dus meer dan eens voor. Van elke roman zijn de eerste 1000 woorden gekozen. Wanneer dit

aantal ergens midden in een alinea bereikt werd, is de hele alinea meegenomen.

Nieuwsberichten
De nieuwsberichten komen uit het de binnenland- (50) en buitenlandrubrieken (50) van het Algemeen Dagblad (50) en NRC Handelsblad (50). Net als de opinieteksten zijn de nieuwsberichten volledig overgenomen, waardoor de lengte varieert. De nieuwsberichten zijn (op één geval na) geheel overgenomen, waarbij de titels werden weggelaten. Enkele journalisten droegen aan meer dan een bericht bij, en een vijftal berichten verscheen met als bron 'AD redactie'.

Opiniestukken in kranten
De opiniërende teksten komen uit De Volkskrant (50) en het NRC Handelsblad (50). De onderwerpen lopen sterk uiteen; van politiek en recht tot sport, sociologie en cultuur. De Volkskrant-teksten komen van het internet, de teksten uit de NRC zijn overgenomen uit de fysieke krant. De teksten zijn in hun geheel overgenomen; de lengte varieert dan ook sterk (zie tabel 2).

Folders met medisch advies
Bij vijf van de teksten ondervonden we technische problemen, zodat er 95 overblijven. Daaronder zijn 14 online versies van folders over veelvoorkomende aandoeningen, zoals die ook te vinden zijn bij apotheken en ziekenhuizen, en 81 soortgelijke teksten van websites van apothekers. Het gaat in beide gevallen om teksten voor leken over onderwerpen zoals hoofdpijn, botbreuken, en brandend maagzuur. De folders en webteksten zijn volledig overgenomen, waardoor de lengte verschilt. Hoewel de teksten uit dezelfde bron (bijvoorbeeld de 13 teksten uit gezondheidsplein.nl en de 12 uit Kringapotheek.nl) wellicht dezelfde redacteur hebben, is de eerste versie van de teksten waarschijnlijk door telkens andere inhoudsspecialisten geschreven.

Verkiezingsprogramma's
Het gaat hier om fragmenten uit verkiezingsprogramma's voor een drietal Tweede Kamerverkiezingen (89) en gemeenteraadsverkiezingen (27). De fragmenten zijn gekozen rond verschillende vaste thema's zoals wonen, economie en onderwijs; van dat soort teksten wordt de eerste versie meestal geschreven door inhoudsspecialisten. In totaal komen 16 politieke partijen aan bod. Van elk onderwerp zijn de eerste 1000 woorden overgenomen.

Als de tekst korter dan 1000 woorden is, is de hele tekst gebruikt. Dat geldt ook voor teksten waarin de 1000 woorden midden in een alinea bereikt werden. Alle teksten komen van partijwebsites; meestal zijn ze overgenomen uit pdf-versies van het verkiezingsprogramma.

Onderzoeksartikelen
Het gaat hier om fragmenten uit het theoretisch kader van onderzoeksartikelen. Daaruit zijn telkens de eerste 300 woorden genomen. De teksten komen uit het *Tijdschrift voor taalbeheersing* (50) en *Mens en maatschappij* (50).

Schoolboeken vmbo-onderbouw
Het gaat om teksten uit vmbo-studieboeken voor de vakken aardrijkskunde, geschiedenis, economie en Nederlands. Alle teksten zijn bedoeld voor de onderbouw (leerjaar 1 en 2). Er zijn alleen teksten voor de niveaus kader- en basisberoepsgericht (de twee laagste niveaus) gebruikt, en dus geen teksten voor de gemengde en theoretische leerweg. We verzamelden de methodes die er voor elk vak waren, en namen uit iedere methode een aantal teksten met als doel 25 teksten per vak. In Bijlage 1 staat een tabel met de gebruikte methodes. We zochten telkens naar stukken aaneengesloten proza. Sommige fragmenten zijn uit online versies van schoolboeken gehaald (bijvoorbeeld van de website van de uitgever). De meeste teksten zijn echter overgetypt. De teksten zijn ongeveer 300 woorden lang (de minimale lengte is 270 woorden). Wanneer de 300 woorden midden in een alinea bereikt werden, is de hele alinea overgenomen.

Schoolboeken havo-bovenbouw
De havo-schoolteksten zijn op dezelfde manier gekozen en overgenomen als de vmbo-schoolteksten. Alle havo-teksten zijn bedoeld voor de bovenbouw (leerjaar 4 en 5). Er zijn alleen boeken gebruikt die specifiek voor havo-leerlingen bedoeld zijn, dus geen teksten voor vmbo/havo of havo/vwo.

Het bleek voor beide niveaus niet eenvoudig om 100 fragmenten van voldoende lengte te verzamelen. Niet alleen zijn er niet al te veel verschillende methodes voor een vak op een bepaald niveau, ook zijn de fragmenten samenhangend proza in deze boeken vaak korter dan 300 woorden. Een andere complicatie was dat in schoolboeken Nederlands alleen aaneengesloten proza te vinden was in de voorbeeldteksten uit literaire bronnen. Omdat het hier meestal gaat om romanfragmenten en dus niet om schoolboekteksten, zijn deze fragmenten naderhand buiten beschouwing

gelaten. We hebben ons dus beperkt tot de 'zaakvakken'. Uiteindelijk kwam hiermee het aantal teksten voor vmbo op 76 en voor havo op 85 (op dit niveau bleken meer verschillende geschiedenismethodes te bestaan, vandaar dit hogere aantal). De in totaal 27 gebruikte methodes worden vermeld in Bijlage 1.

Hoewel het bij de overblijvende teksten altijd gaat om 'zaakvakken', is de thematische diversiteit in dit deelcorpus groter dan elders. Een andere reden voor behoedzaamheid is dat er uit ieder schoolboek meerdere fragmenten zijn gekozen. We zullen de schoolboekteksten meer in detail analyseren in paragraaf 5.

Tabel 2 Gegevens over het tekstcorpus

Genre	Aantal teksten	Min. lengte	Max. lengte	Gemiddelde lengte (sd)	Totaal aantal woorden	Aantal auteurs	Periode van verschijnen
Reisverslag	100	287	317	301 (6)	30.140	100	2012
Celebrity-nieuws	100	252	899	384 (139)	38.380	Onbekend	2012
Roman	100	903	2.692	1.104 (199)	110.435	87	1946-2012
Vmbo-schoolboek	76*	240	507	346 (47)	25.765	Bijlage 1	2000-2013
Havo-schoolboek	85*	267	635	358 (68)	30.212	Bijlage 1	1998-2012
Nieuwsbericht	100	279	1.236	605 (207)	60.481	85	Najaar 2012
Opiniestuk	100	386	2.617	823 (325)	82.292	100	Najaar 2012
Medisch advies	95	212	5.506	1.010 (809)	95.928	Onbekend	2005-2012
Verkiezingsprogramma	116	92	3.294	834 (494)	96.719	Onbekend	2002-2012
Onderzoeksartikel	100	272	324	304 (9)	30.406	100	2006-2012
Totaal	972	92	5.506	606 (441)	600.758		

* Exclusief het schoolvak Nederlands

4 Tien genres vergeleken op lexicale complexiteit

4.1 Op welke lexicale kenmerken is onze genrevoorspelling gebaseerd?

Dit artikel gaat over de lexicale complexiteit van tekstgenres en teksten. We beginnen met het vergelijken van genres, en eindigen met het vergelijken van teksten met andere teksten in zijn genre. Voor die vergelijking hebben we een genrevoorspellingsmodel nodig. We beschrijven eerst met behulp van welke kenmerken we gezocht hebben naar voorspellingsmodellen. In het vervolg van de paragraaf bespreken we voor een aantal belangrijke kenmerken afzonderlijk hoe zij scoren per genre. Aan het eind

van de paragraaf presenteren we onze de voorspellingsmodellen waarin de kenmerken gecombineerd worden.

Een preliminaire vraag bij de kenmerkselectie is of we uitgaan van afzonderlijke kenmerken of van kenmerkgroepen die zijn afgeleid uit een factoranalyse, zoals Biber & Conrad in hun dimensionele corpusanalyses doen (Biber 1992; Biber & Conrad 2009). Wij hebben gekozen voor afzonderlijke kenmerken. De voornaamste reden daarvoor is dat het kenmerkenpalet in dit artikel minder breed is dan dat in de genoemde studies. Wij beperken ons tot lexicale complexiteit, en gaan daarop veel dieper in dan Biber, die zich in 1992 wat betreft woordmoeilijkheid beperkt tot woordlengte en nominalisaties. Mogelijk is een factoranalyse interessant wanneer we lexicale en grammaticale kenmerken gaan combineren. Maar omdat we hier een meer gedetailleerd beeld schetsen van een deelaspect, brengen we de aard en de bijdragen van de afzonderlijke kenmerken liefst zo duidelijk mogelijk in beeld.

Hoe zijn nu de lexicale kenmerken gekozen? Het eerste uitgangspunt bij de kenmerkselectie is dat onze kenmerken moeten gaan over de complexiteit van afzonderlijke woorden. Dat betekent allereerst dat zinsbouw- en coherentiekenmerken van teksten buiten beschouwing blijven. Daarbij hebben we ook dichtheden van woordsoorten (nomina, adjectieven enz.) terzijde gelaten. Immers, deze kenmerken reflecteren eerder keuzes wat betreft de zinsbouw dan woordkeuzes: iemand die veel werkwoorden gebruikt, kiest voor andere constructies dan iemand die veel naamwoorden gebruikt (vgl. bijvoorbeeld Biber en Gray (2010) over de informatiedichte constructies die zorgen voor het hoge aantal nomina in wetenschappelijke artikelen).

Ten tweede moeten de lexicale kenmerken intuïtief plausibel te verbinden zijn met woordcomplexiteit. Daaronder verstaan we simpelweg de kans dat een lezer het woord niet probleemloos kan interpreteren. Lexicale T-Scankenmerken die vooral pragmatisch interessant zijn, hebben we daarom buiten beschouwing gelaten. Dat geldt ook voor kenmerken als de type-token-ratio; hoewel dit kenmerk vaak in verband wordt gebracht met lexicale diversiteit, is het ook nogal gevoelig voor de opbouw van de tekst (zie ook Pander Maat et al. 2014).

Een derde uitgangspunt bij onze kenmerkkeuze is dat tekstkenmerken interessanter zijn naarmate ze grotere verschillen tussen genres laten zien. Dat spreekt enigszins vanzelf, omdat we genrevoorspellers zoeken; maar het heeft ook een inhoudelijke reden. Het taalgebruik in een tekst is functioneel gemotiveerd, en daarom gevoelig voor medium, thema, doelgroep, doel, teksthandeling en structuur van de tekst. Met uitzondering van de

factor medium is er in ons corpus een ruime variatie op deze contextuele en tekststructurele parameters. Wanneer tekstkenmerken verschillen tussen genres, mogen we aannemen dat die tekstkenmerken gevoelig zijn voor die contextuele parameters; en daarmee zijn het interessante registerkenmerken. In het eerste deel van dit artikel zullen we genreverschillen rapporteren zonder veel interpretatie. In het tweede deel zullen we proberen iets dieper in te gaan op twee parameters (thema en doelgroep) die in een bepaald contrast aan het werk zijn.

We hebben de discriminerende kracht van kenmerken op twee manieren gedefinieerd. Ten eerste de waarde van η^2 wanneer het kenmerk gebruikt wordt als afhankelijke variabele in een variantieanalyse met genre als onafhankelijke variabele. Omdat we later genrevoorspellingen proberen te doen, hebben we daarnaast een effectgroottemaat gebruikt die daarbij past: de Nagelkerkes R^2 die het kenmerk oplevert wanneer het als enige genrevoorspeller gebruikt wordt in een multinomiale regressie (De η^2 en deze R^2 correleren overigens zeer hoog met elkaar: $r = .95$). Kenmerken die op beide waarden boven de .15 scoren, zijn voorlopig meegenomen.

Een vierde criterium bij de kenmerkselectie betreft het vermijden van conceptuele overlap in de verzameling. T-Scan biedt voor allerlei tekstaspecten meerdere kenmerken, die subtiel verschillen in definitie. Zo kan de woordfrequentie zowel met als zonder eigennamen genomen worden. In zulke gevallen is de kenmerkvariant gekozen die genres het beste discrimineert. Verder geeft T-Scan som-kenmerken, gebaseerd op het samennemen van groepen specifiekere kenmerken; dat is bijvoorbeeld het geval voor woordconcreetheid. Voor deze kenmerken moet ofwel het globale kenmerk ofwel een groep goed discriminerende kenmerken worden gekozen.

Ten slotte kijken we naar statistische overlap. Om te beginnen elimineren we telkens het minst discriminerende kenmerk van twee kenmerken die hoger correleren dan .70. Daarbij is een uitzondering gemaakt voor woordlengte. Dat kenmerk correleert -.85 met woordfrequentie, maar voorspelt, op zich genomen, genres wat beter. Toch hebben we gekozen voor woordfrequentie omdat dit kenmerk directer te relateren is aan complexiteit dan woordlengte; we gaan daar hieronder nog wat verder op in. De overblijvende set kenmerken is geanalyseerd op collineariteit, waarbij de kenmerken met VIF-waarden van 10 of hoger werden verwijderd.

We zullen de resultaten van onze genrevoorspelling aan het eind van deze paragraaf bespreken. Om de lezer vertrouwder te raken met de T-Scankenmerken, bespreken we eerst voor een aantal kenmerken hoe zij scoren in onze tien genres.

4.2 Woordlengte en morfologische kenmerken

Woordlengte is het eenvoudigst te meten tekstkenmerk, en het heeft dan ook een lange traditie in leesbaarheidsonderzoek (Dale & Chall 1948, Flesch 1948, Bormuth 1966; zie Staphorsius 1994 voor een overzicht). Empirisch onderzoek laat echter zien dat woordlengte slechts een kleine invloed heeft op het leesgemak bij normale lezers; die invloed is er vooral bij infrequente woorden, en neemt verder af met toenemen van de leeftijd en de leesvaardigheid (Barton et al. 2014). In hoeverre woordlengtes verschillen tussen genres is niet zo duidelijk. Toch is het interessant om woordlengteverschillen te analyseren, vooral in relatie met morfologische kenmerken.

Een paar opmerkingen over de getallen in de tabellen die volgen. We zullen daarin de genregemiddelden rapporteren, en daarop variantieanalyses doen; hoewel niet alle kenmerken normaal verdeeld zijn binnen genres, zijn variantieanalyses bij deze aantallen waarnemingen voldoende robuust tegenover normaliteitsschendingen (Field 2010). Verder maken we in elke kolom de hoogste en de laagste waarde vet, zodat de uitersten van de 'genreschaal' goed zichtbaar zijn.

Tabel 3 geeft gemiddelden voor woordlengtes. T-Scan biedt de optie om bij woordlengtes namen buiten beschouwing te laten; die optie is hier gekozen, omdat de lengte van een naam meer van toeval afhankelijk lijkt dan die van een gewoon woord. Het blijkt dat de woordlengtes systematisch verschillen per genre. Wanneer we woordlengtes meten in letters, wordt maar liefst 68% van de variantie in woordlengtes verklaard door genreverschillen ($F\ [9,962] = 228.00$, $p < .001$, $\eta^2 = .68$). De narratieve genres (reisverslag, celebrity-nieuws, roman) hebben de kortste woorden, verkiezingsprogramma's en onderzoeksartikelen de langste.

We kunnen woordlengtes ook uitdrukken in morfemen. In T-Scan gebeurt dat op basis van Frog, met een accuratesse van minimaal 98% (Van den Bosch 2007). Frog geeft bv. *gebruiksklaar* een lengte van drie morfemen: *gebruik, -s* en *klaar; gelijkwaardige* heeft er vier: *gelijk, waarde*, *-ig* en *-e*; en *taalbeheersing* bestaat uit *taal, be-, heers* en *-ing*. Ook woordlengtes in morfemen verschillen sterk tussen genres ($F\ [9,962] = 151.03$, $p < .001$, $\eta^2 = .59$). De vraag is vervolgens hoeveel van het woordlengteverschil in letters toe te schrijven is aan verschil in aantal morfemen. Het blijkt dat ook wat betreft morfeemlengte (aantal letters gedeeld door het aantal morfemen) de genreverschillen nog behoorlijk zijn ($F\ [9,962] = 109.17$, $p < .001$, $\eta^2 = .51$), en in dezelfde richting wijzen als die voor woordlengtes. Met andere woorden, genres met langere woorden hebben zowel meer morfemen per woord als meer letters per morfeem.

Tabel 3 Genreverschillen in woord- en morfeemlengte

Genre	Woord-lengte in letters	SD	Woord-lengte in morfemen	SD	Morfeem-lengte in letters	SD
Reisverslag	**4.49**	.25	1.33	.05	**3.38**	.10
Celebrity-nieuws	4.61	.21	**1.32**	.05	3.49	.11
Roman	4.59	.26	1.33	.05	3.46	.10
Vmbo-schoolboek	4.80	.27	1.40	.07	3.44	.11
Havo-schoolboek	5.33	.33	1.48	.08	3.60	.12
Nieuwsbericht	5.13	.25	1.44	.06	3.57	.10
Opiniestuk	5.28	.31	1.45	.06	3.64	.12
Medisch advies	5.28	.31	1.46	.06	3.62	.14
Verkiezingsprogramma	5.58	.34	**1.54**	.08	3.64	.10
Onderzoeksartikel	**5.77**	.36	1.52	.08	**3.79**	.15
Totaal	5.10	.52	1.43	.10	3.57	.16

We kunnen corrigeren voor morfemen, maar wellicht zijn vooral combinaties van vrije morfemen interessant: samenstellingen. Het is denkbaar dat de woordlengte sterk wordt beïnvloed door het aantal samenstellingen in de tekst. Daarbij interesseren ons vooral transparante samenstellingen, dat wil zeggen samenstellingen waarin de woorddelen steun bieden bij de interpretatie van het geheel (een *schooldirecteur* is een *directeur* van een *school*). Het is denkbaar dat in een tekst met veel transparante samenstellingen de woordlengte een overtrokken beeld geeft van de complexiteit. Aan de andere kant wijst een hoog aantal transparante samenstellingen erop dat een tekst verwijst naar vrij specifieke concepten: immers, *inkomstenbelasting* is een meer gespecialiseerd concept dan *belasting*.

T-Scan geeft alleen informatie over nominale samenstellingen, maar het merendeel van de samenstellingen is ook nominaal. We moeten onderscheid maken tussen twee factoren die kunnen bijdragen aan woordlengte: het aantal nomina in de tekst (naamwoorden zijn langer dan andere woorden) en de proportie samenstellingen per naamwoord. En inderdaad correleert woordlengte .72 met het aantal nomina per duizend woorden (verder 'dichtheid') en .47 met de proportie samenstellingen op de nomina. We zien in Tabel 4 dat de narratieve genres minder nomina hebben, en schoolboeken, gezondheidsteksten, politieke teksten en onderzoeksartikelen juist meer ($F\ [9,962] = 116.47$, $p < .001$, $\eta^2 = .52$). De verschillen in de proportie samenstellingen onder die nomina zijn minder groot, hoewel ook hier de narratieve genres het laagst scoren ($F\ [9,962] = 23.93$, $p < .001$, $\eta^2 = .18$). De meeste samenstellingen, en dus de meest specifieke concepten, vinden we in gezondheidsteksten, politieke teksten en havo-schoolboeken.

Tabel 4 Genreverschillen wat betreft nomina en samenstellingen

Genre	Dichtheid* nomina	SD	Proportie samen-stellingen	SD	Woord-lengte	SD	Woordleng-te gecorr. voor sam.	SD
Reisverslag	171	31	.13	.074	**4.49**	.25	**4.42**	.19
Celebrity-nieuws	**157**	26	**.12**	.055	4.61	.21	4.59	.18
Roman	174	22	.13	.043	4.59	.26	4.50	.26
Vmbo-schoolboek	217	28	.14	.077	4.80	.27	4.69	.23
Havo-schoolboek	233	34	.20	.089	5.33	.33	5.05	.23
Nieuwsbericht	202	23	.19	.059	5.13	.25	4.97	.22
Opiniestuk	202	25	.18	.060	5.28	.31	5.08	.25
Medisch advies	**237**	25	**.22**	.074	5.28	.31	4.97	.28
Verkiezingsprogramma	233	25	.20	.066	5.58	.34	5.27	.25
Onderzoeksartikel	228	31	.15	.071	**5.77**	.36	**5.52**	.30
Totaal	205	39	.17	.074	5.10	.52	4.91	.42

* Dichtheid = aantal per duizend woorden; gecorr. voor sam. = gecorrigeerd voor samenstellingen)

T-Scan geeft ons sinds kort de mogelijkheid de lengte en de frequentie van die samenstellingen te vergelijken met die van enkelvoudige nomina. Daartoe is een lijst van 46.000 nomina ingebouwd met informatie over hun samenstellingskarakter (zie Pander Maat et al. 2016, Bijlage L).

Samenstellingen zijn inderdaad veel langer. Een gemiddeld nomen is 7.9 letters lang, een gemiddeld ongeleed nomen 7.0 letters, en een gemiddelde samenstelling 12.6 letters. We kunnen de woordlengte ook corrigeren voor de invloed van samenstellingen. T-Scan biedt namelijk een woordlengte waarin de lengte van samenstellingen bepaald wordt aan de hand van het basiswoord. In die gecorrigeerde woordlengte is een nomen nog maar 6.8 letters lang, en Tabel 4 laat zien dat de gemiddelde lengte van alle tekstwoorden daalt van 5.1 naar 4.9 letters. Maar het belangrijkste is dat we ook voor die gecorrigeerde woordlengte een groot genre-effect vinden: ($F[9,962] = 214.90$, $p < .001$, $\eta^2 = .67$). De rangorde van genres naar woordlengte blijft daarbij grotendeels gelijk. Dat betekent dat teksten met lange woorden niet alleen meer samenstellingen hebben, maar dat hun samenstellingen ook langere basiswoorden bevatten.

We hebben nu vastgesteld dat het woordlengteverschil tussen genres niet puur een kwestie is van aantal morfemen per woord of van aantal vrije morfemen per woord (dus aantal samenstellingen). Teksten met langere woorden hebben meer morfemen en meer samenstellingen, maar ook langere morfemen. Het woordlengte-effect geldt dus zowel voor woorden als voor woorddelen.

Nu leren woord- en morfeemlengteverschillen ons weinig over verschil-

len in complexiteit. Wel houden we het aantal transparante samenstellingen vast als kenmerk voor onze genrevoorspelling, omdat ze een indicatie zijn van een meer gespecialiseerde tekstinhoud. En we gaan nu in op kenmerken die directer in verband te brengen zijn met tekstcomplexiteit: woordfrequentie, abstractheid en persoonlijkheid.

4.3 Woordfrequenties

Zipf (1936) heeft al geopperd dat er een relatie is tussen woordlengte en woordfrequenties. En woordfrequenties zijn in principe interessanter dan woordlengtes, omdat frequenties opgevat kunnen worden als schatting van de kans dat de lezer een woord goed kent. Breland (1996) vond dan ook behoorlijke frequentie-effecten op het beantwoorden van vragen naar woordbetekenissen (zie ook Ryder & Slater 1988). En naarmate lezers meer woorden uit een tekst kennen, stijgt hun begrip (Schmitt et al. 2011).

Laten we eerst duidelijk maken hoe woordfrequenties in T-Scan worden gemeten. T-Scan biedt twee soorten woordfrequentiematen. Het eerste type geeft de exacte frequentie aan per woord, waarbij we ons beperken tot inhoudswoorden (naamwoorden, namen, adjectieven, bijwoorden en 'gewone werkwoorden', dat wil zeggen werkwoorden die geen hulpwerkwoord of koppelwerkwoord zijn of kunnen zijn). Daarbij is de logaritme (grondtal 10) genomen van de frequentie en niet de ruwe frequentie. Daardoor scoort een woord met een frequentie van een miljoen niet duizend keer zo hoog als een woord met een frequentie van 1.000, maar drie keer zo hoog. Om woordfrequenties op basis van verschillende corpora te kunnen vergelijken, nemen we verder niet de logaritme van de absolute maar die van de relatieve frequentie. Op voorstel van Van Heuven et al. (2014) standaardiseren we daarbij de frequentie op een miljard woorden. Dat heeft het voordeel dat ook bij lage frequenties nog onderscheid gemaakt kan worden. Bijvoorbeeld: wanneer een woord eenmaal voorkomt in een corpus van 1 miljoen woorden, bedraagt de gestandaardiseerde frequentie 1.000, en de logaritme daarvan (verder 'freqlog') dus 3. Namen worden volgens Camblin et al. (2007) anders verwerkt dan 'gewone woorden'. Daarom zijn ook freqlog-varianten gedefinieerd waarbij namen worden overgeslagen.

Het tweede type woordfrequentiegegevens geeft aan welke proportie tekstwoorden als frequent kan worden gedefinieerd. Daarbij wordt gewerkt met hedendaagse corpora met 'volwassen' taalgebruik. Het al of niet frequent zijn wordt bepaald met frequentie-ranglijsten. Bij bijvoorbeeld Freq1000 wordt simpelweg gekeken hoeveel van de tekstwoorden horen tot de 'top1000' van de frequentielijst. Bij Freq1000_inhwrd gaat het

om dezelfde proportie, maar dan wordt alleen gekeken naar inhoudswoorden.

T-Scan biedt daarbij twee opties voor het onderliggende corpus:

– SoNaR totaal (Oostdijk et al. 2013; voor onderzoekers is dit corpus toegankelijk op https://portal.clarin.inl.nl/opensonar_whitelab);
– Subtlex (Keuleers et al. 2010).

Bij SoNaR gaat het vooral om schriftelijk taalgebruik, waarbij informele genres qua omvang in de minderheid zijn. Subtlex daarentegen is een corpus met Nederlandse ondertitels voor films en series, en bevat met name alledaagse (zij het niet-spontane) conversatie.

Elke frequentielijst kent slordigheden en eigenaardigheden. De slordigheden (bijvoorbeeld niet-bestaande woorden, verkeerd gespelde en verkeerd ingelezen woorden) zijn handmatig verwijderd uit de lijsten die gebruikt worden voor de Freq1000 tot 20000. Wat betreft eigenaardigheden: de herkomst van het corpus bepaalt wat frequent is. Het Subtlex-corpus bestaat uit ondertitels bij Engelse en Amerikaanse films, documentaires en series. Daarom bevat het duizenden Engelse namen (vooral persoons-, maar ook geografische namen en evenementnamen zoals *Thanksgiving*), Engelse en Spaanse aanspreekvormen (*mrs., signor*) en ook onvertaald gebleven Engelse woorden. Deze niet-Nederlandse elementen zijn handmatig verwijderd uit de lijst van de 20.000 meest frequente woorden.

Voor de 20.000-woordenlijst gebaseerd op het SoNaR-totaalcorpus is een opschoning in twee stappen uitgevoerd. Eerst zijn niet-bestaande en buitenlandse woorden verwijderd. Vervolgens is gekeken naar de duizenden namen (plaatsnamen, persoonsnamen, organisatienamen). Omdat Sonar voor bijna 80% bestaat uit Vlaamse teksten, zijn die namen nogal Zuid-Nederlands gekleurd. We achten het aannemelijk dat die namen geen goede indruk geven van de vertrouwdheid van het tekstvocabulaire voor Noord-Nederlanders. Maar ook principieel is het kwestieus of namen horen bij het basisvocabulaire. Voor persoonsnamen lijkt dat sowieso niet het geval. Voor geografische en organisatienamen zijn wellicht een handvol namen van nationale betekenis (Nederland, België, Amsterdam, Brussel). Voor andere namen geldt dat het nogal toevallig is welke ervan in de tekst voorkomen. Daarom zijn ook uit de Sonar-totaallijst van 20.000 woorden de namen vervangen door 'gewone woorden' verderop in de frequentielijst.

Na alle correcties en opschoningen hebben we de aard van beide corpora verkend door de eerste 1.000 woorden te vergelijken. Een korte samenvatting van deze vergelijking is als volgt:

- 626 woorden komen in beide top-1000 lijsten.
- 185 van de 374 woorden die alleen in de SoNaR top-1000 voorkomen gaan over hoeveelheden (*4, procent*) tijden (*april, 2003*), politiek (*premier, burgemeester*), financiën of economie (*jaarbasis, directeur, euro, winst*), sport (*titel, finale*), of plaatsen (*nationale, buitenlandse*). Daarnaast treffen we informele spellingvarianten aan als *ni, nie* en *'k*, digitaal jargon als *spam* en de Zuid-Nederlandse pronomina *gij* en *ge*.
- 139 van de 374 woorden die alleen in de Subtlex top-1000 voorkomen verwijzen naar personen (aanspreekvormen als *meneer*, nomina als *kerel*, pronomina als *je* en *mezelf*), intieme of familierelaties (*liefje, broer, seks, trouwen*), alledaagse interjecties (*alsjeblieft, welterusten, ja*), evaluaties en emoties (*spijt, geweldig, kwaad, klootzak*) en misdaad (*vermoord, agent, drugs*).

SoNaR draagt duidelijk de sporen van het grote aandeel van nieuwsteksten; daarin is immers precieze informatie van belang over hoeveelheden, plaatsen en tijden, en worden thema's als politiek, economie en sport veel besproken. Subtlex daarentegen is duidelijk conversationeel van aard: in gewone gesprekken gaat het vaak over personen en evaluaties en komen nogal wat interjecties voor. Een wat minder wenselijke bias in dit corpus is dat er in films en series erg veel aandacht is voor misdaad.

We gaan nu na hoe de woordfrequenties verschillen per genre, en hoe die verschillen afhangen van het gebruikte corpus. We concentreren ons daartoe op het eerste type gegeven: de logaritme van de relatieve frequentie (verder 'freqlog').

In de eerste vier kolommen van Tabel 5 vergelijken we de freqlogs voor de tien genres tussen de corpora. Om te beginnen valt op dat de SoNaR-frequenties voor de schrijftaal in ons corpus een stuk hoger zijn dan de Subtlex-frequenties: op een logaritmische schaal correspondeert het gemiddelde verschil tussen 4.61 en 4.37 met frequenties die zo'n 75% hoger zijn. Een tweede belangrijk verschil is dat het genre-effect in de SoNaR-frequenties minder groot is ($F\ [9{,}960] = 71.01$, $p < .001$, $\eta^2 = .40$) dan dat in de Subtlex-frequenties ($F\ [9{,}962] = 153.29$, $p < .001$, $\eta^2 = .59$).

Nu zegt dat allemaal nog niet veel over de vraag welke frequentielijst de voorkeur verdient wanneer het erom gaat om in te schatten hoe alledaags de woorden in een tekst zijn. Om die vraag te beantwoorden, kijken we hoe tekstgenres in beide lijsten ten opzichte van elkaar scoren. In dat levert SoNaR enkele contra-intuïtieve resultaten. Zo zijn voor SoNaR de woorden in verkiezingsprogramma's even frequent als die in romans; en zijn de woorden in onderzoeksartikelen bijna even frequent als die in medische

adviezen. In dit opzicht levert Subtlex meer plausibele resultaten op, waarin de frequenties van woorden in verkiezingsprogramma's en onderzoeksartikelen als aanzienlijk minder frequent worden geschat. We werken dan ook verder met Subtlex-frequenties.

We zeiden al dat de genreverschillen in woordfrequentie groot zijn. Een betere indruk daarvan krijgen we als we de logaritmes even terugvertalen in gewone getallen. De hoogste gemiddelde woordfrequentie vinden we in roddelberichten; een logaritme van 4.87 komt overeen met een frequentie van ongeveer 74.000 op een miljard woorden, oftewel 74 op een miljoen. Woorden in onderzoeksartikelen hebben een gemiddelde frequentie van zo'n 22 per miljoen. Het verschil tussen vmbo- en havo-schoolboeken is ook nog behoorlijk (63 versus 33 per miljoen); daarop komen we later terug.

Tabel 5 Genreverschillen in woordfrequenties

Genre	Freqlog Sonar	SD	Freqlog Subtlex	SD	Freqlog Subtlex, nomina	SD	Freqlog Subtlex nomina, gecorr. *	SD
Reisverslag	4.82	.25	4.77	.30	4.12	.29	4.38	.22
Roddelbericht	**4.87**	.17	**4.80**	.22	4.26	.25	4.51	.20
Roman	4.57	.21	4.56	.26	4.06	.23	4.33	.19
Vmbo-schoolboek	4.80	.23	4.73	.25	**4.31**	.32	**4.58**	.26
Havo-schoolboek	4.52	.24	4.14	.26	3.64	.31	4.09	.22
Nieuwsbericht	4.65	.15	4.35	.22	3.83	.30	4.22	.24
Opiniestuk	4.58	.18	4.25	.24	3.67	.23	4.06	.17
Medisch advies	4.40	.22	4.24	.24	3.62	.26	4.03	.23
Verkiezingsprogramma	4.59	.18	4.09	.26	3.61	.29	4.06	.20
Onderzoeksartikel	**4.34**	.21	**3.91**	.23	**3.51**	.27	**3.86**	.28
Totaal	4.61	.26	4.37	.39	3.85	.39	4.20	.31

* Voor samenstellingen is de frequentie van het basiswoord genomen in plaats van die van het hele woord.

Ook bij woordfrequenties kunnen we onderscheid maken tussen samenstellingen en andere nomina. Immers, psycholinguïstisch onderzoek lijkt erop te wijzen dat woorddelen van transparante samenstellingen zelfstandig semantisch verwerkt worden. Zo vond Zwitserlood (1994) dat het transparante *kerkorgel* meer de gedachte aan een priester oproept dan het niet-transparante *drankorgel* de gedachte oproept aan bier (hoewel het laatste woord hierover nog niet is gezegd, zie Pollatsek & Hyöna 2005). Maar misschien is dit soort onderzoek voor ons minder relevant, omdat het gaat over bekende samenstellingen. Voor ons is een andere overweging

belangrijker. In T-Scan gebruiken we woordfrequenties als een schatting van de kans dat een woord überhaupt te interpreteren is voor een lezer. Wanneer de woordfrequenties van samenstellingen laag zijn, levert dat voor transparante samenstellingen wellicht een onderschatting op van de bekendheid van de interpreteerbaarheid.

De gemiddelde Subtlex-freqlog voor nomina is 3.85 (sd .39). Maar er is een groot verschil tussen de freqlog voor ongelede nomina (4.19, sd .32) en die voor samenstellingen (2.19, sd .47). Samenstellingen zijn dus 100 maal zo zeldzaam als ongelede nomina! Als we de freqlog voor nomina corrigeren door bij samenstellingen de frequentie van het basiswoord mee te laten tellen, stijgt hij naar 4.20 (zie Tabel 5). Maar belangrijk is dat ook op deze gecorrigeerde maat de verschillen tussen genres groot blijven ($F[9,962] = 102.50$, $p < .001$, $\eta^2 = .49$), en in dezelfde richting gaan als de ruwe verschillen.

Het is vooralsnog een open vraag of de frequentiecorrectie voor samenstellingen inderdaad een betrouwbaarder beeld geeft van woordcomplexiteit. Enerzijds laat een hoog aantal samenstellingen zien dat er een gespecialiseerd vocabulaire gehanteerd wordt: transparante samenstellingen vormen lexicale pakketjes met nogal specifieke informatie. Anderzijds slagen de meeste lezers er waarschijnlijk in om de pakketjes ter plekke uit te pakken. Een laagfrequente samenstelling is daarom eenvoudiger te interpreteren dan een ongeleed woord met dezelfde frequentie. Voorlopig kiezen we voor onze genreanalyse hieronder de niet-gecorrigeerde woordfrequentie.

4.4 Woordconcreetheid

Veel auteurs hebben woordcomplexiteit in verband gebracht met woordconcreetheid, meestal gedefinieerd als de mate waarin het woord gelinkt is aan een zintuiglijk waarneembaar concept (Sadoski et al. 2000, Spooren et al. 2015, Keulaars et al. 2014). Toch kan concreetheid nog op verschillende manieren worden geoperationaliseerd. Men kan woorden door proefpersonen laten scoren op concreetheid, zoals Spooren et al. (2015) en Brysbaert et al. (2014) doen; of men kan werken met woordenlijsten die door experts geannoteerd zijn.

T-Scan gebruikt de tweede benadering, en werkt met geannoteerde lijsten nomina, adjectieven en werkwoorden. De eerste versies van de nomina- en adjectievenlijsten zijn afkomstig uit het Referentie Bestand Nederlands (Martin & Maks et al. 2005). Naderhand zijn deze lijsten anders ingedeeld, handmatig nagekeken en aangevuld. Nomina en adjectieven worden relatief fijnmazig semantisch geclassificeerd. Zo zijn de nomina

ingedeeld in vijftien klassen (zie Tabel 6; voor details zie Pander Maat et al. 2016, Bijlage D). Bij wijze van proef hebben de auteurs 1388 woorden onafhankelijk gecodeerd; 87.5% daarvan kreeg dezelfde code, en Cohens Kappa was .72.

Het label 'concreet' wordt daarbij gedefinieerd in termen van de subklassen. Als strikt-concreet worden gezien de klassen 1-7 uit Tabel 6, als ruim-concreet de klassen 1-10. Deze definitie maakt een belangrijk voordeel duidelijk van het operationaliseren via expertannotaties: het is duidelijker wat er in die annotaties verstaan wordt onder concreetheid. Immers, striktconcreet zijn bij T-Scan de woorden verwijzend naar bepaalde ontologische categorieën. Wie dat wil, kan zich concentreren op bepaalde categorieën, of die juist buiten beschouwing laten. Een tweede voordeel is dat meerduidige woorden ongedefinieerd kunnen worden gelaten. Polyseme en ambigue woorden als *kant* en *poot* krijgen in de lijst van Brysbaert et al. (2014) één concreetheidswaarde, terwijl ze op heel verschillende manieren in teksten gebruikt worden.

Onze 'ontologische' definitie van concreetheid heeft als nadeel dat concreetheidsverschillen binnen klassen worden verwaarloosd. Dat geeft vooral problemen bij hyperoniemen. Zo zijn de persoonswoorden *mens* en *voetballer* intuïtief niet even concreet. Omdat het aantal hyperoniemen per definitie niet groot is, is dit probleem wellicht in de praktijk minder zwaarwegend.

Tabel 6 Semantische klassen voor naamwoorden in T-Scan

Klasse	Voorbeelden
1. Persoon	*leraar, schreeuwlelijk*
2. Plant en dier	*mus, eik*
3. Gebruiksvoorwerp	*stoel, weefgetouw*
4. Waarneembare substantie	*modder, kerrie*
5. Voeding en verzorging	*melk, sigaret, bruistablet*
6. Concreet niet-dynamisch overig	*galblaas, vulkaan*
7. Waarneembare gebeurtenis	*aai, ademhaling*
8. Plaats	*Amsterdam, voorkamer*
9. Tijd	*feestdag, periode*
10. Maat	*euro, dB*
11. Niet-waarneembare substantie	*fosfor, splijtstof*
12. Niet-waarneembaar gebeuren	*crisis, loonverlaging*
13. Organisatie	*werkgeversorganisatie*
14. Abstract overig	*christendom, motto*
15. Ongedefinieerd	*kant, poot*

De gegevens voor de afzonderlijke klassen bieden interessante inzichten in thematische verschillen tussen genres; we zullen er een aantal opnemen in genrevoorspellingsmodel 1 hieronder. Maar omwille van de ruimte beperken we ons in Tabel 7 tot de strikt-concrete en ruim-concrete superklassen.

Tabel 7 Genreverschillen in concreetheid van naamwoorden

Genre	Proportie strikt-concrete nomina	SD	Proportie ruim-concrete nomina	SD	Dichtheid abstracte nominalisaties	SD	Proportie algemene nomina	SD
Reisverslag	.35	.11	**.67**	.09	**15**	9	**.05**	.03
Celebrity-nieuws	.37	.09	.53	.10	18	9	.09	.05
Roman	**.46**	.09	.63	.10	**15**	7	.06	.04
Vmbo-schoolboek	.31	.15	.54	.18	24	16	.07	.06
Havo-schoolboek	.20	.12	.38	.15	43	22	.14	.06
Nieuwsbericht	.28	.10	.47	.11	36	15	.11	.05
Opiniestuk	.21	.08	.33	.10	42	16	.19	.07
Medisch advies	.34	.14	.42	.14	52	20	.13	.06
Verkiezingsprogramma	**.12**	.05	.25	.08	62	19	.19	.05
Onderzoeksartikel	.13	.08	**.19**	.11	**65**	25	**.39**	.15
Totaal	.28	.15	.43	.19	38	25	.14	.12

Het genreverschil is groot, zowel voor strikt-concrete ($F [9,962] = 111.22$, $p < .001$, $\eta^2 = .51$) als voor ruim-concrete nomina ($F [9,962] = 190.50$, $p < .001$, $\eta^2 = .64$). Voor beide kenmerken zijn verkiezingsprogramma's en onderzoeksartikelen duidelijk de minst concrete genres. Wanneer we plaats- en tijdwoorden (maat-woorden komen bijna niet voor) als abstract beschouwen, zijn romans de meest concrete teksten. Wanneer we deze woorden bij de concrete groep rekenen, dan zijn reisblogs de meest concrete teksten: die bevatten namelijk relatief veel plaats- en tijdwoorden. In Tabel 7 valt wederom het verschil op tussen schoolboeken voor de vmbo-onderbouw en de havo-bovenbouw. Daarop komen we later terug.

Hoewel abstracte nominalisaties relatief weinig voorkomen (gemiddeld 38 maal per 1.000 woorden) laten zij duidelijke genreverschillen zien ($F [9,962] = 129.16$, $p < .001$, $\eta^2 = .55$). Die verschillen gaan gelijk op met die voor de strikt-concrete naamwoorden, vooral wat betreft de abstractheid van verkiezingsprogramma's en onderzoeksartikelen en de relatieve concreetheid van de drie narratieve genres.

T-Scan biedt ook informatie over een subklasse uit de abstracte woorden, namelijk om de woorden die met willekeurig welke inhoud kunnen worden gecombineerd; bijvoorbeeld *idee, opvatting, methode, resultaat,*

probleem en *discussie*. Flowerdew en Forrest (2015) noemen deze woorden 'signalling nouns', in T-Scan spreken we van algemene nomina. De software bevat een lijst van bijna 1.300 zulke nomina (zie verder Pander Maat et al. 2016, Bijlage J). De proportie van deze algemene nomina laat grote genreverschillen zien (F [9,962] = 209.29, $p < .001$, $\eta^2 = .66$). Zoals te verwachten is, blinken wetenschappelijke artikelen uit in het gebruik van deze woorden, met verkiezingsprogramma's en opiniestukken op gepaste afstand als tweede. De algemene nomina maken het dus mogelijk om binnen de abstracte teksten verder onderscheid te maken tussen meer en minder algemene vocabulaires.

T-Scan geeft ook informatie over de concreetheid van adjectieven en werkwoorden. De semantische indeling van adjectieven in T-Scan is deels gericht op waarneembaarheid maar gaat ook in op het al of niet evaluatieve karakter van het woord, zie Tabel 8.

Tabel 8 Semantische klassen voor adjectieven in T-Scan

Klasse	Voorbeelden
1. Direct waarneembare kenmerken van personen	*doodsbleek, dwergachtig*
2. Emotionele kenmerken en sociaal gedrag	*gegriefd, goedgelovig*
3. Direct waarneembare kenmerken van dingen	*flanellen, geel*
4. Niet-direct waarneembare kenmerken	*teerarm, kiemvrij*
5. Tijd	*voorbijgaand, vrijdags*
6. Plaats	*binnenlands, Gelders*
7. Specifieke evaluatie (positief/negatief)	*onverslijtbaar; lawaaierig*
8. Algemene evaluatie (positief/negatief/zonder richting)	*mooi; verwerpelijk; aanmerkelijk*
9. Epistemische evaluatie (positief/negatief)	*steekhoudend; onzinnig*
10. Overige (niet-evaluatieve) abstracte adjectieven	*aanverwant, aandachtig*
11. Ongedefinieerd	*belastbaar, druk, smal*

Het meest genre-onderscheidende adjectiefkenmerk is de proportie neutrale abstracte adjectieven (F [9,962] = 91.07, $p < .001$, $\eta^2 = .46$), zie Tabel 9. In de genrevoorspelling hieronder zullen we daarnaast de proportie striktconcrete adjectieven gebruiken, die bestaat uit de groepen 1 tot en met 3 in Tabel 8. Ook deze proportie is een behoorlijke genre-onderscheider, waarbij de hoge score voor romans opvalt (F [9,962] = 69.05, $p < .001$, $\eta^2 = .39$).

Tabel 9 **Genreverschillen wat betreft adjectieven en werkwoorden; het gaat telkens om proporties**

Genre	Neutrale abstracte adjectieven	SD	Strikt-concrete adjectieven	SD	Concrete werkw.	SD	Algemene werkw.	SD
Reisverslag	**.15**	.09	.09	.08	**.10**	.06	**.06**	.04
Celebrity-nieuws	.22	.10	.10	.07	.03	.03	.10	.04
Roman	.18	.06	**.17**	.06	**.10**	.04	.09	.03
Vmbo-schoolboek	.26	.15	.04	.05	.04	.04	.07	.04
Havo-schoolboek	.36	.18	.03	.05	.02	.03	.14	.06
Nieuwsbericht	.28	.10	.05	.05	.03	.03	.12	.04
Opiniestuk	.34	.10	.05	.05	.02	.02	.15	.05
Medisch advies	.25	.10	.08	.07	.06	.04	.16	.05
Verkiezingsprogramma	.34	.10	**.02**	.04	**.01**	.01	.13	.04
Onderzoeksartikel	**.53**	.14	**.02**	.03	**.01**	.01	**.25**	.07
Totaal	.29	.15	.07	.07	.04	.05	.13	.07

Ten slotte worden ook werkwoorden geanalyseerd op concreetheid, zij het minder verfijnd. Ze worden globaal ingedeeld in concreet (*kwetteren, lassen, vriezen, ruiken*), abstract (*aanbesteden, frustreren, verschaffen*) en ongedefinieerd (*leeglopen, verfrissen*). De proportie concrete werkwoorden verschilt behoorlijk tussen genres ($F\ [9,962] = 99.60$, $p < .001$, $\eta^2 = .48$). Daarnaast zijn ook voor werkwoorden algemene woorden apart gezet in een lijst (*interesseren, categoriseren, argumenteren*). De proportie algemene werkwoorden vertoont nog sterkere genreverschillen ($F\ [9,962] = 133.55$, $p < .001$, $\eta^2 = .56$). Tabel 9 laat zien dat zowel voor abstracte adjectieven als voor algemene werkwoorden de onderzoeksartikelen een klasse apart zijn. Verder laten de schoolboekniveaus voor algemene werkwoorden hetzelfde verschil zien als voor algemene naamwoorden. De havo-boeken zijn consequent academischer geschreven dan de vmbo-boeken. Een laatste kenmerk van werkwoorden dat we hieronder als genrevoorspeller zullen gebruiken, is het aantal modale werkwoorden per deelzin, dat ook opgevat kan worden als een maat voor abstract taalgebruik ($F\ [9,962] = 45.16$, $p < .001$, $\eta^2 = .30$). Genres met veel modale werkwoorden, zoals medische adviezen en verkiezingsprogramma's, leggen immers meer nadruk op mogelijkheden en wenselijkheden.

4.5 Persoonlijke voornaamwoorden

Sinds Flesch (1948) is aangenomen dat teksten over personen makkelijker te lezen zijn dan andere teksten. De eenvoudigste manier om naar personen te verwijzen is het gebruiken van persoonlijke voornaamwoorden.

Omdat die woorden in principe verwijzen naar personen die in de context al bekend geacht worden, zijn persoonlijke voornaamwoorden eenvoudig te interpreteren.

De dichtheid daarvan verschilt sterk tussen genres ($F [9,962] = 141.29$, $p < .001$, $\eta^2 = .57$). Het is interessant om verder onderscheid te maken tussen eerste, tweede en derde personen. Ook die subklassen vertonen behoorlijke genreverschillen, vooral pronomina van de eerste persoon ($F [9,962] = 107.93$, $p < .001$, $\eta^2 = .50$) en derde persoon ($F [9,962] = 115.16$, $p < .001$, $\eta^2 = .52$), en in iets mindere mate de tweede-persoon ($F [9,962] = 61.63$, $p < .001$, $\eta^2 = .37$). Zie Tabel 10, waarin blijkt dat de eerste persoon kenmerkend is voor het reisverslag, de tweede persoon voor vmbo-schoolboeken en medisch advies en de derde persoon voor roddelberichten en romans. Uiteraard zijn er ook behoorlijk wat eerste personen in romans aanwezig. De negatieve correlatie van -.56 tussen de dichtheden van eerste en derde personen in romans (elders is die negatieve correlatie afwezig of een stuk lager) bevestigt het idee dat in een roman vaak slechts één van beide personen dominant aanwezig is.

Tabel 10 Genreverschillen wat betreft persoonlijke voornaamwoorden

Genre	Dichtheid vnw. 1e persoon	SD	Dichtheid vnw. 2e persoon	SD	Dichtheid vnw. 3e persoon	SD	Dichtheid alle pers. vnw.	SD
Reisverslag	**48**	18	6	7	11	13	65	22
Celebrity-nieuws	24	19	6	7	46	18	75	26
Roman	34	29	9	9	**58**	25	**101**	29
Vmbo-schoolboek	3	7	19	21	29	20	51	25
Havo-schoolboek	3	5	3	7	20	17	26	16
Nieuwsbericht	8	9	3	4	30	17	40	21
Opiniestuk	9	9	5	10	20	11	34	19
Medisch advies	**1**	2	**27**	19	**7**	6	36	20
Verkiezingsprogramma	15	16	**1**	3	11	7	27	19
Onderzoeksartikel	5	7	**1**	4	14	12	**21**	14
Totaal	16	21	8	13	24	22	48	33

1e-pers. vnw. = persoonlijk voornaamwoord van de eerste persoon

We hebben nu drie soorten van lexicale complexiteit behandeld die belangrijke genreverschillen aan het licht brengen: woordfrequentie indiceert de specialistische aard van de behandelde concepten, concreetheid de waarneembaarheid ervan, en teksten met persoonlijke voornaamwoorden relateren concepten aan veelal constant blijvende persoonlijke perspectieven.

4.6 Namen en werkwoordstijden

Ten slotte wijzen we op twee andere lexicale kenmerken die wat minder evident gerelateerd zijn aan complexiteit. T-Scan maakt onderscheid tussen persoonsnamen, organisatienamen, plaatsnamen, productnamen en evenementnamen. Omdat intuïtief persoonsnamen de tekst persoonlijker maken en organisatienamen de tekst abstracter maken, lijken deze namen in ieder geval van belang. Verder valt te betogen dat plaatsnamen bijdragen aan de concreetheid van de tekst. Omdat productnamen en evenementnamen een minder duidelijke relatie met complexiteit hebben en bovendien bijzonder infrequent zijn, laten we deze kenmerken verder buiten beschouwing. In Tabel 11 beperken we ons tot persoonsnamen (genomen op het totaal van namen en nomina); deze namen blijken het sterkste verband met genre te vertonen ($F [9,962] = 118.92$, $p < .001$, $\eta^2 = .53$). Daarbij scoren, niet verbazend, de celebrity-berichten veruit het hoogst, gevolgd door nieuwsberichten(zie Tabel 11).

Een ander lexicaal kenmerk is de tijd van het werkwoord. T-Scan telt de proportie verleden tijden per vervoegd werkwoord, en de dichtheid van hulpwerkwoorden van tijd, die gepaard gaan met voltooide tijden. We vinden deze kenmerken interessant omdat ze verband houden met narrativiteit, en omdat narrativiteit in ander onderzoek wel gerelateerd is aan complexiteit (Graesser et al. 2011). Tabel 11 laat voor de proportie verleden tijden een spectrum zien met roddelteksten en romans aan de ene kant en medische adviezen en verkiezingsteksten aan de andere ($F [9,962] = 53.66$, $p < .001$, $\eta^2 = .33$). De voltooide tijden onderscheiden ook redelijk tussen genres ($F [9,962] = 48.40$, $p < .001$, $\eta^2 = .31$) en vertonen in de uitersten hetzelfde beeld. In de middenwaarden zien we verschillen tussen verleden en voltooide tijden. Zo bevatten schoolboeken behoorlijk wat verleden tijden (met name vanwege de geschiedenisboeken, die .77 scoren in vmbo-boeken en .85 voor havo-boeken), maar scoren ze bij voltooide tijden juist relatief laag (ook in de geschiedenisboeken: de dichtheden in vmbo- en havo-boeken zijn resp. 9 en 14).

De verhouding tussen verleden en voltooide tijden per genre laat een verschil zien tussen twee soorten teksten. Als een genre relatief veel verleden tijden en relatief minder hulpwerkwoorden van tijd bevat, mogen we zeggen dat een tekst ons 'verplaatst naar het verleden': dat zien we in geschiedenisboeken en in wat mindere mate in romans, die extreem scoren in verleden tijden en minder extreem in hulpwerkwoorden van tijd. Als beide kenmerken behoorlijk scoren, wordt het verleden ook besproken vanuit zijn relevantie voor het heden; dat is tenminste een belangrijke semantische interpretatie van de voltooide tijd (Mittwoch 2008). Als die

interpretatie algemene geldigheid heeft, dan vinden we zo'n oriëntatie op relevantie voor het heden in reisverslagen, roddelberichten, nieuwsberichten en opiniestukken.

Tabel 11 Genreverschillen wat betreft namen en werkwoordstijden

Genre	Proportie persoonsnamen	SD	Prop. ww. in verl. tijd	SD	Dichtheid hww. van tijd	SD
Reisverslag	.04	.04	.33	.29	19	11
Celebrity-nieuws	**.15**	.06	.35	.21	**24**	11
Roman	.06	.04	**.58**	.30	22	7
Vmbo-schoolboek	.04	.04	.33	.37	10	6
Havo-schoolboek	.02	.04	.41	.40	13	8
Nieuwsbericht	.08	.05	.29	.18	18	7
Opiniestuk	.04	.04	.19	.13	16	7
Medisch advies	**.01**	.01	**.02**	.03	**8**	5
Verkiezingsprogramma	**.01**	.02	.03	.05	10	8
Onderzoeksartikel	.04	.03	.14	.19	11	8
Totaal	.05	.06	.26	.29	15	10

Prop. = proportie; ww. = werkwoorden; verl. tijd = verleden tijd; hww. = hulpwerkwoord

4.7 Genrevoorspelling

We hebben nu een indruk van de kenmerken die T-Scan biedt voor het beschrijven van lexicale complexiteit (vraag 1), en een eerste idee van genreverschillen op die maten (vraag 2). We snijden nu vraag 3 aan: zijn deze verschillen groot genoeg om een behoorlijk genrevoorspellingsmodel te maken?

Aan het begin van deze paragraaf beschreven we hoe we kenmerken kozen voor onze voorspellingsmodellen. De selectie resulteerde in een set van 29 kenmerken, waarvan een deel hierboven al de revue is gepasseerd. De set valt uiteen in acht subgroepen (zie ook Tabel 12):

1 het aantal samenstellingen onder de nomina (1);
2 de woordfrequentie (2);
3 de concreetheid van nomina (3-16);
4 de concreetheid van adjectieven (17-18);
5 de concreetheid van werkwoorden (19-21);
6 persoonlijke voornaamwoorden (22-24);
7 verschillende soorten namen (25-27);
8 verleden en voltooide werkwoordstijden (28-29).

Het geheel van de acht groepen definieert een ruime opvatting van begrip complexiteit; daarmee hebben we ons eerste voorspellingsmodel gebouwd.

Maar er zijn twee discussiepunten die ons ertoe hebben gebracht ook kleinere kenmerksets te testen.

Ten eerste kan betoogd worden dat het grote aantal specifieke werkwoordklassen (kenmerken 5-15) wel de genreherkenning verbetert, maar geen goede indicaties levert voor complexiteit; immers, complexiteit hangt niet af van de specifieke semantische klasse, maar van het concreet dan wel abstract zijn van het nomen. Daarom zijn in model 2 de afzonderlijke klassen vervangen door het somkenmerk 'strikt-concrete nomina'; dat kenmerk was juist buiten model 1 gelaten.

Ten tweede kan er twijfel over bestaan of de werkwoordstijden wel een indicatie zijn van complexiteit. Wel komen verleden en voltooide tijden vaker voor in narratieve genres, en zijn narratieve genres wellicht eenvoudiger dan niet-narratieve genres. Om na te gaan wat het gewicht is van dit aspect, zijn de tijden in model 3 buiten beschouwing gebleven.

Aldus zijn de drie sets voorspellers gemotiveerd voor een genredetectieanalyse. Uitgaande van normaal verdeelde intervalvariabelen, zou voor die analyse een discriminantanalyse in aanmerking komen. Maar omdat de normaliteitsassumpties voor een discriminantanalyse strenger zijn dan die voor enkelvoudige variantieanalyses die we tot dusver gebruikten (Field 2010), hebben we gewerkt met een multinomiale logistische regressie. Die methode biedt ook het voordeel dat de bijdrage van afzonderlijke voorspellers helder blijft. In Tabel 12 worden de modellen schematisch weergegeven met samenvattende modelinformatie. Gedetailleerde informatie per model is te verkrijgen bij de eerste auteur.

Tabel 12 Drie genrevoorspellingsmodellen

	Kenmerk	Model 1	Model 2	Model 3
1	Proportie samenstellingen op nomina	n.s.	**	**
2	Woordfrequentie zonder namen (logaritme)	***	***	***
3	Proportie algemene naamwoorden	***	***	***
4	Dichtheid van abstracte nominalisaties	***	***	***
5	Proportie strikt-concrete naamwoorden		**	n.s.
6	Proportie nomina over personen	***		
7	Proportie nomina over gebruiksvoorwerpen	***		
8	Proportie nomina over voeding en verzorging	n.s.		
9	Proportie nomina over concrete substanties	*		
10	Proportie overige concrete niet-dynamische nomina	***		
11	Proportie nomina over concrete gebeurtenissen	***		
12	Proportie nomina over plaatsen	***		
13	Proportie nomina van tijd	***		

	Kenmerk	Model 1	Model 2	Model 3
14	Proportie nomina over abstracte substanties	***		
15	Proportie nomina over abstractie gebeurtenissen	n.s.		
16	Proportie nomina over organisatie	***		
17	Proportie neutraal abstracte adjectieven	***	**	**
18	Proportie strikt-concrete adjectieven	**	***	***
19	Proportie algemene werkwoorden	***	***	***
20	Proportie concrete werkwoorden	*	***	***
21	Aantal modale werkwoorden per deelzin	***	***	***
22	Dichtheid van voornaamwoorden 1e persoon	***	***	***
23	Dichtheid van voornaamwoorden 2e persoon	***	***	***
24	Dichtheid van voornaamwoorden 3e persoon	***	***	***
25	Proportie persoonsnamen op nomina plus namen	***	***	***
26	Dichtheid van organisatienamen	***	***	***
27	Dichtheid van plaatsnamen	**	***	***
28	Proportie werkwoorden in verleden tijd	***	***	
29	Aantal hulpwerkwoorden van tijd per deelzin	***	***	
	Cox & Snell R^2	.981	.973	.967
	Nagelkerke R^2	.991	.983	.977
	Proportie correct voorspeld	.898	.832	.789
	Aantal significante voorspellers	25	18	15

*** = p < .001; ** = p < .01; * = p < .05

Model 1 voorspelt voor 90% van de 972 teksten het correcte genre. Onder de significante voorspellers in dit model vinden we een behoorlijk aantal (9 van de 25) specifieke nominale klassen. Vervangen we die door het algemene kenmerk 'strikt-concrete' nomina, dan wordt nog 83% voorspeld, met 18 kenmerken. Laten we vervolgens de werkwoordstijden uit het model, dan daalt de voorspelling naar 79%.

Bij deze getallen moet vermeld worden dat er geen kruisvalidatie heeft plaatsgevonden. Wel hebben we twee 'baseline' modellen gedraaid, om na te gaan of onze semantisch gedefinieerde kenmerken (woordsoorten, concreetheid, verschillende soorten namen) meerwaarde hebben boven modellen met eenvoudiger gedefinieerde kenmerken. Als eerste baseline hebben we een model met alleen woordlengte als voorspeller gebruikt (Cox & Snell R^2 = .679, Nagelkerke R^2 = .686, 32% correct voorspeld), als tweede baseline een model met woordlengte en woordfrequentie (Cox & Snell R^2 = .734 Nagelkerke R^2 = .742, 35% correct voorspeld). Daaruit blijkt dat de rijkere modellen in Tabel 12 aanzienlijk beter presteren.

Voor de analyses in paragraaf 6 hieronder is model 2 als referentiepunt gekozen: het model dus waarin niet de specifieke nominaklassen, maar wel de werkwoordstijden zijn opgenomen. Uit de confusiematrix voor dat

model blijkt dat opinieteksten het lastigst te voorspellen zijn (60% correct; zij worden veel verward met havoteksten en nieuwsteksten); daarnaast worden havoteksten (68% correct) regelmatig verward met nieuwsberichten en verkiezingsprogramma's. Zoals gezegd is de genrevoorspelling in dit artikel geen doel, maar een middel om te komen tot stilistische uitspraken over afzonderlijke teksten; zie verder paragraaf 6.

5 Verschillen tussen vmbo- en havo-schoolboeken

Tot dusver hebben we een belangrijke vraag niet gesteld: waar komen de taalgebruiksverschillen tussen genres eigenlijk vandaan? Omdat we hierboven zagen dat onze tien genres variatie vertonen wat betreft thema, doel, doelgroep, teksthandelingen en structuur, komen al die factoren in aanmerking als verklaring. Toch is onze indruk dat thematische verschillen de hoofdrol spelen bij de kenmerken die hierboven genoemd zijn. De meer of minder specialistische aard van het thema raakt de woordfrequenties, de abstractheid van het thema raakt de concreetheidsscores, de rol van personen in het thema raakt het aantal persoonlijke voornaamwoorden en namen, en de werkwoordstijden zijn minstens deels afhankelijk van de vraag of de tekst gaat over gebeurtenissen in het verleden of niet.

Tot zover gaat de analyse over de complexiteit van registers, waarmee gedoeld wordt op alle situationeel bepaalde taalgebruikskenmerken. Stel nu dat we ons willen beperken tot stilistische verschillen, waarbij we stijl 'dualistisch' opvatten, uitgaande van een onderscheid tussen vorm en inhoud (Leech & Short 1981). Een stilistische keuze is dan een keuze tussen alternatieve formuleringen voor globaal dezelfde informatie, of om het in cognitief-grammaticale termen te formuleren, als een keuze tussen verschillende 'construals' van dezelfde conceptuele inhoud (Langacker 2013, 55-89). Waar in ons corpus zouden we dan stilistische variatie kunnen vinden?

Hier is een van onze minimale genreparen interessant. Zoals gezegd beschikken we over een 'minimaal genrepaar': verschillen tussen vmbo- en havo-schoolboeken lijken vooral te kunnen worden toegeschreven aan verschillen in voorkennis en taalvaardigheid tussen leerlingen in de vmbo-onderbouw en in de havo-bovenbouw. In termen van de Expertgroep Doorlopende Leerlijnen (2008) bevinden we ons in de vmbo-onderbouw tussen niveau 1F en 2F, en bij de havo-bovenbouw op niveau 3F. Nu zijn er in dit kleine deelcorpus (76 vmbo- en 85 havo-teksten) drie schoolvakken vertegenwoordigd: economie, aardrijkskunde en geschiedenis. Dit maakt

het mogelijk om de invloed van doelgroepniveau te vergelijken met die van het schoolvak. De verschillen tussen schoolvakken zijn per definitie inhoudelijk, maar wellicht zijn de verschillen tussen schoolniveaus mede stilistisch van aard.

We hebben hierboven vastgesteld dat de observaties in het schoolboekencorpus niet geheel onafhankelijk verzameld zijn: er bestaat per vak en niveau een beperkt aantal schoolboeken, waaruit we meerdere fragmenten hebben gekozen. Daarom presenteren we hieronder een multilevel-regressieanalyse, waarbij we de effecten van niveau en schoolvak schatten in een model waarin gemiddelden per schoolboek kunnen variëren ('random intercepts'), net als de omvang van de niveau- en vakeffecten ('random slopes'). Het blijkt dat de random effecten nergens het model significant verbeteren. Dat betekent dat de afhankelijkheid in de waarnemingen geen grote gevolgen heeft en dat er gegeneraliseerd mag worden over schoolboeken. We zien hieronder af van rapportages van modellen die geen verbetering bleken.

In de regressies zijn vakeffecten geschat met dummyvariabelen per vak; ook interacties tussen vakdummy's en niveau zijn bekeken. Om de interpretatie van wat volgt te vergemakkelijken, geven we hieronder eerst de gemiddelde scores voor de zes cellen (twee niveaus maal drie schoolvakken).

Tabel 13 Woordfrequenties en samenstellingen in schoolboeken

	Freqlog	SD	Proportie sam./nw.	SD	Freqlog ong. nw.	SD	Freqlog sam. nw.	SD
AK-vmbo	4.77	.28	.16	.09	4.65	.30	2.78	.60
AK-havo	4.10	.29	.25	.09	4.03	.24	2.00	.28
EC-vmbo	4.80	.21	.16	.07	4.63	.27	2.63	.58
EC-havo	4.08	.28	.22	.08	3.99	.21	1.76	.31
GE-vmbo	4.63	.23	.12	.06	4.48	.19	2.75	.59
GE-havo	4.21	.22	.14	.06	4.10	.24	2.16	.37

Sam. = samenstellingen; ong. = ongelede; nw. = naamwoorden. N = 25 voor elke cel, behalve voor GE-vmbo (n = 26) en GE-havo (n = 35)

Tabel 13 geeft de gemiddelden voor woordfrequenties en samenstellingen. De regressieanalyse vindt voor de totale woordfrequentie alleen een effect van niveau: op de havo worden minder frequente woorden gebruikt ($b = -.60$, $t(155.04) = -14.44$, $p < .001$). Hetzelfde niveau-effect is er bij de frequentie van ongelede woorden ($b = -.54$, $t(147.84) = -13.50$, $p < .001$). Voor de woordfrequentie van samenstellingen is er naast het niveau-effect ($b = -.74$, $t(157) = -9.94$, $p < .001$) een effect van schoolvak. De positieve coëfficiënten

voor aardrijkskunde (b = .19, t (8.32) = 2.08, p < .05) en geschiedenis (b = .26, t (8.88) = 2.96, p < .01) laten zien dat de woordfrequentie van de samenstellingen in die beide vakken hoger is dan van die in economieboeken.

Ook het aantal samenstellingen per nomen is hoger in havo- dan in vmbo-boeken (b = .06, t (154.15) = 4.69, p < .001). Daarnaast hebben economieboeken (b = .06, t (10.49) = 3.78, p < .01) en aardrijkskundeboeken (b = .08, t (10.74) = 4.48, p < .01) meer samenstellingen dan geschiedenisboeken. Dit alles wijst erop dat dat havo-boeken meer gespecialiseerde concepten behandelen dan vmbo-boeken, en dat de vaktermen hun piek bereiken in samenstellingen in economieboeken.

Tabel 14 Concreetheid van nomina en persoonlijke voornaamwoorden in schoolboeken

	Prop. alg. nw.	SD	Prop. conc. nw. ruim	SD	Dichth. vnw2	SD	Dichth. vnw3	SD	Dichth. pers. nm.	SD
AK-vmbo	.06	.04	.58	.14	25	22	17	20	.013	.020
AK-havo	.14	.07	.41	.11	6	8	10	7	.030	.036
EC-vmbo	.11	.07	.37	.16	26	25	34	21	.067	.048
EC-havo	.17	.06	.23	.12	5	10	14	12	.001	.003
GE-vmbo	.05	.04	.66	.11	8	10	34	14	.010	.021
GE-havo	.11	.05	.46	.10	0	2	32	18	.052	.048

Prop. = proportie; alg. = algemene; nw. = zelfstandig naamwoord; dichth. = dichtheid; vnw2/3 = voornaamwoord van de tweede/derde persoon; pers.nm. = persoonsnamenVoor de aantallen waarnemingen zie Tabel 13

Tabel 14 geeft allereerst gemiddelden voor de concreetheid van nomina en voor persoonlijke voornaamwoorden. Voor de proportie algemene naamwoorden vindt de regressieanalyse hogere scores voor de havo (b = .07, t (153.10) = 7.43, p < .001); daarnaast scoren aardrijkskunde (b = -.04, t (12.23) = –3.86, p < .01) en geschiedenis (b = -.06, t (11.94) = -5.55, p < .001) lager dan economie. Voor de proportie ruim-concrete naamwoorden vinden we dezelfde effecten in omgekeerde richting: lagere scores voor de havo (b = -17, t (154.70) = -8.96, p < .001), en daarnaast hogere scores voor aardrijkskunde (b = .20, t (14.03) = 6.68, p < .001) en geschiedenis (b = .25, t (14.62) = 8.97, p < .001).

Voor de dichtheid van persoonlijke voornaamwoorden geeft Tabel 14 een complexer beeld. De tweede-persoonsvoornaamwoorden komen minder voor op de havo (b = -20.1, t (155.43) = -6.92, p < .001), en meer bij aardrijkskunde (b = 30.6, t (117.05) = 3.84, p < .01) en economie (b = 30.8, t (111.29) = 3.86, p < .001) dan bij geschiedenis; daarnaast is er een interactie die teruggaat op het wat kleinere niveauverschil bij geschiedenis (voor de

interactieterm 'geschiedenis*niveau' geldt: b = 13.1, t (149.07) = 2.73, p < .01). Ook bij de derde-persoonsvoornaamwoorden vinden we een hoofdeffect van niveau dat wordt genuanceerd door een interactie. Op havoniveau zijn de scores lager (b = -12.7, t (155.71) = -3.92, p < .001), maar geschiedenisteksten scoren hoog op beide niveaus (voor de interactieterm geldt: b = 11.5, t (154.18), t = 2.12, p < .05). We concluderen dat aardrijkskunde- en economieboeken op het vmbo duidelijk meer persoonlijke voornaamwoorden gebruiken dan op de havo.

Een ander persoonlijkheidskenmerk is het aantal persoonsnamen, gemeten als proportie op het totaal van de naamwoorden en namen (daarom zijn de proporties vrij klein). Voor die proportie vinden we hoofdeffecten voor niveau en vak. De vmbo-teksten bevatten meer namen: b = .016, t (157.00) = 2.85, p < .01). Daarnaast bevatten, begrijpelijkerwijs, de aardrijkskundeteksten veel minder namen dan de geschiedenisteksten (b = -.053, t (157.00) = -7.89, p < .001), en datzelfde geldt voor de economieteksten (b = -.039, t (157.00) = -5.92, p < .001).

Vatten we deze resultaten samen, dan zien we niveau-effecten voor woordfrequentie; bij de andere variabelen spelen zowel niveau als schoolvak een rol: het aantal samenstellingen, de woordfrequentie, de woordconcreetheid, de persoonlijke voornaamwoorden en de persoonsnamen. Het niveau-effect voor persoonlijke voornaamwoorden en persoonsnamen sluit overigens aan bij de resultaten van Evers-Vermeul en Holtermann (2013), die vaststellen dat vmbo-examenteksten meer aansprekingen en personages bevatten dan examenteksten voor havo en vwo.

Onze vraag hierboven was in hoeverre deze lexicale kenmerken inhoudelijke dan wel stilistische verschillen reflecteren. Het is duidelijk dat vakeffecten teruggaan op inhoudelijke verschillen. Bij niveauverschillen is dat minder duidelijk. Enerzijds stelt het leerplan eisen aan de inhouden die op de verschillende niveaus moeten worden aangeboden, dus zijn ook niveauverschillen deels inhoudelijk van aard. Anderzijds is voorstelbaar dat dezelfde thema's meer of minder persoonlijk en concreet worden besproken.

Laten we daarom kijken naar tekstvoorbeelden. Fragment 1, 2 en 3 zijn afkomstig uit economieboeken voor de vmbo-onderbouw. Fragment 1 kent hoogfrequente woorden en veel derde-persoonsvoornaamwoorden. Fragment 2 heeft geen enkele samenstelling en kiest het generieke tweedepersoonsperspectief; dit laatste geldt algemeen voor de tweede personen in schoolboeken: de leerlingen worden niet individueel aangesproken. In fragment 3 zien we hoe abstracte concepten als *inkomen* en *werkgelegenheid* in ontwikkelingslanden worden gecombineerd met voorbeelden van producten uit die landen, wat een relatief hoge concreetheidsscore ople-

vert. Inhoudelijk vallen deze fragmenten op doordat ze economische kwesties benaderen vanuit het dagelijks leven van consumenten.

(1) Brenda heeft een taak in huis. Ze doet een paar lichte huishoudelijke klusjes en ze zorgt voor de hond (afbeelding 1). Ze krijgt daarvoor een extraatje bij haar zakgeld.
De meeste jongeren krijgen zakgeld. Zakgeld is een deel van hun inkomsten in geld. Als ze er niets voor hoeven te doen, hoort dat geld bij hun inkomsten zonder tegenprestatie. Als ze er wel iets voor moeten doen, hoort dat geld bij hun inkomsten met tegenprestatie. Of je beide soorten ontvangsten bij je zakgeld telt, maak je zelf uit.

(2) Een budget is bedoeld om aan vast te houden. Als dat niet lukt en je gaat over je budget heen, moet je budgetteren. Anders kunnen er financiële problemen ontstaan. Je moet je uitgaven opnieuw afstemmen op je inkomsten.
Als je een budget wilt verhogen, moet je budgetteren. Want je houdt dan minder geld over voor de andere uitgaven. Je moet de andere budgetten verlagen, of minder gaan sparen. Budgetteren is ook nodig als je meer wilt gaan sparen, want er is dan minder geld beschikbaar voor je uitgaven.

(3) De meeste mensen geven tegenwoordig veel geld uit aan hun tuin. Ze leggen een terras aan en kopen planten, verlichting en tuinmeubels. Vroeger waren die meubels meestal van plastic, maar tegenwoordig zijn ze steeds vaker van teakhout. Sommige mensen vinden dat mooi, maar andere mensen kopen juist teakhout om de ontwikkelingslanden te helpen. Door producten uit de ontwikkelingslanden te kopen, zorg je voor inkomen en werkgelegenheid in die landen.

De fragmenten 4 en 5 komen uit economieboeken voor de havo-bovenbouw. Fragment 4 bevat enkele extreem laagfrequente samenstellingen (*kapitaalmarkt, termijndeposito*). Fragment 5 kent geen concrete woorden, en evenmin persoonlijke voornaamwoorden. Inhoudelijk staan in deze fragmenten institutionele kanten van de economie centraal (de geldmarkt, bedrijfsfusies).

(4) Alle vermogenstitels die een looptijd hebben van meer dan twee jaar, behoren tot de kapitaalmarkt. Is de looptijd korter dan jaar, dan worden de vermogenstitels tot de geldmarkt gerekend.

Gezinnen en bedrijven die geld storten op kortlopende termijndeposito's of kortlopende spaarrekeningen maken het mogelijk dat banken dit geld kunnen aanbieden op de geldmarkt.

(5) Bedrijven kunnen op verschillende manieren groeien. Als een groot bedrijf een kleiner bedrijf koopt, spreken we meestal van een overname. Een andere mogelijkheid is een fusie: het gaat dan om de samenvoeging van twee min of meer gelijkwaardige bedrijven. Vaak komt een fusie tot stand doordat één onderneming de uitstaande aandelen opkoopt van de andere onderneming. Soms is het resultaat van een fusie, dat de overgenomen onderneming een onderdeel van het grote geheel wordt, maar het is ook mogelijk dat een nieuwe onderneming wordt opgericht.

De variatie in deze voorbeelden duiden we veelal inhoudelijk, en soms stilistisch. De keuze van een individueel consumentenperspectief is in fragment 1 en 2 het gevolg van de besproken thema's (zakgeld, budgetteren). In fragment 3 worden voorbeelden van consumentenproducten gebruikt ter illustratie van macro-economische begrippen; omdat die voorbeelden ook achterwege kunnen blijven, zou je die keuze stilistisch kunnen noemen. In fragment 4 ligt de nadruk zozeer op institutionele thema's dat een dergelijke concretisering minder voor de hand ligt. In fragment 5 is een voorbeeld van een fusie denkbaar, maar daarmee wordt de tekst waarschijnlijk niet veel concreter, noch persoonlijker.

6 Teksten vergelijken binnen genres

De conclusie is voorlopig dat ook in ons schoolboekencorpus verschillen in taalgebruik vooral teruggaan op thematische verschillen. Om te bepalen of de lexicale T-Scankenmerken gevoelig zijn voor stilistische variatie, is een beter gecontroleerde tekstkeuze nodig. We zouden bijvoorbeeld herschrijvingen van teksten voor verschillende leesniveaus kunnen vergelijken, en we zouden teksten kunnen vergelijken met teksten in hetzelfde genre. We demonstreren nu een nieuwe variant van de laatste benadering.

We kunnen namelijk onze genrevoorspelling hierboven gebruiken om teksten binnen een genre met elkaar te vergelijken. Deze analyse laat namelijk niet alleen zien welke kenmerken helpen bij het onderscheiden van genres en dus verschillen per genre, maar geeft ook informatie over afzonderlijke teksten, te weten een 'genrewaarschijnlijkheid' voor elke tekst. Een

eenvoudig te voorspellen tekst heeft een hoge waarschijnlijkheid op het juiste genre en lage waarschijnlijkheden op de andere genres. Maar er zijn ook teksten die fout voorspeld worden. En juist omdat we beschikken over behoorlijk presterende voorspelmodellen, worden foute voorspellingen interessant. Een sterk voorspelmodel laat zien dat we redelijk kunnen omschrijven wat gebruikelijk is binnen een genre. En een fout voorspelde teksten wijkt kennelijk af van deze genrespecifieke norm.

Ter illustratie volgen hieronder fragmenten uit twee havo-aardrijkskundeteksten. De tekst waaruit fragment 6 komt, wordt in de genrevoorspelling via model 1 (Tabel 12) gezien als een 'typische' havo-tekst: de kans dat het gaat om een havo-tekst wordt geschat op 89%, de kans dat het gaat om een vmbo-tekst op 0%. De tekst waaruit fragment 7 komt, wordt daarentegen fout geclassificeerd; de kans dat het een vmbo-tekst is, wordt geschat op 98% en de kans dat het een havo-tekst is op slechts 1%. Nu is de genrevoorspelling voor havo-teksten nog niet geweldig betrouwbaar (68% correct), maar deze enorme verschillen in waarschijnlijkheden doen toch vermoeden dat er iets bijzonders aan de hand is met de tekst van fragment 7.

(6) Nederland is lange tijd beschouwd als emigratieland. Vooral na de Tweede Wereldoorlog was het vertrek groot. Het hoogtepunt was in 1952 toen 50 000 landgenoten emigreerden. De emigratie werd zelfs door de Nederlandse overheid begeleid. Als instrument tegen de toen heersende werkloosheid wilde de overheid actief meehelpen de bevolkingsomvang terug te dringen. Andere argumenten om het land te verlaten waren angst voor een Russische invasie, angst voor overbevolking, een aangenamer klimaat en bestaande relaties met mensen in het bestemmingsland. Toch hebben de aantallen jaarlijkse immigranten de bevolkingsontwikkeling in Nederland maar beperkt beïnvloed. Bovendien is gebleken dat de relatie tussen emigratie en overbevolking zich anders heeft ontwikkeld. Door naoorlogse wederopbouw trok de werkgelegenheid namelijk weer aan, ondanks een groeiende bevolking. De economische groei heeft zelfs tot belangrijke immigratie geleid.

(7) Globalisering gaat om het proces van het ontstaan van steeds meer samenhang in de wereld. Je staat er niet bij stil, maar bij de meeste dingen die je dagelijks gebruikt zijn vele landen in de wereld betrokken! Het beeldscherm van bijvoorbeeld je laptop komt waarschijnlijk uit Zuid-Korea, de batterij ervan uit Japan en de harde schijf uit Thailand. Andere onderdelen van je laptop komen uit China, de Ver-

> enigde Staten, Taiwan, Mexico, de Filipijnen, Singapore, Ierland, Duitsland en Israël. Slechts één product en al twaalf landen die erbij betrokken zijn! Er zijn dus mondiale stromen van goederen en grondstoffen. Door globalisering zijn wij zo betrokken bij tal landen in de wereld. Wij hebben invloed op de ontwikkeling van de mensen die er wonen. Omgekeerd hebben mensen en bedrijven in andere delen van de wereld invloed op ons.

Vier kenmerken vertonen behoorlijke verschillen tussen de teksten achter fragment 6 en 7:

- de woordfrequentie (4.00 voor fragment 6 versus 4.72 voor fragment 7);
- de proportie samenstellingen op nomina (32% versus 9%; vergelijk *bevolkingsomvang* en *bestemmingsland*);
- de proportie ruim-concrete nomina (37% versus 48%); de tekst achter fragment 6 heeft weliswaar meer persoonlijke nomina (*landgenoten*) dan die achter fragment 7, maar meer abstracte gebeurteniswoorden (*emigratie*), organisatiewoorden (*overheid*) en andere abstracte woorden (*werkgelegenheid*); fragment 7 daarentegen noemt meer plaatsen (*China*);
- en de dichtheid van persoonlijke voornaamwoorden (6 versus 23).

Deze verschillen laten de tekst van 7 eruitzien als een vmbo-tekst. En omdat het in beide gevallen gaat om abstracte havo-aardrijkskunde (migratie, globalisering), lijkt de kans groter dat we hier te maken hebben met stilistische variatie. En inderdaad kunnen wij ons een minder complexe variant voorstellen van fragment 6, en een complexere variant van fragment 7.

7 Conclusies en discussie

We hebben in dit artikel allereerst laten zien welke kenmerken interessant kunnen zijn bij het automatisch analyseren van lexicale complexiteit. Daarbij biedt T-Scan interessante nieuwe mogelijkheden. Zo is het begrip woordlengte hierboven niet alleen beschreven in termen van letters, maar ook in termen van morfemen en vrije morfemen. Daarbij is duidelijk geworden dat genreverschillen in woordlengtes ook gelden voor woorddelen. Verder hebben we laten zien dat woordlengte een minder leerzame maat is dan het aantal samenstellingen, dat immers als een indicatie voor specificiteit kan worden opgevat.

Verder is gedemonstreerd hoe woordfrequenties afhangen van de gebruikte corpora. Voorts zijn verschillende nieuwe operationaliseringen van concreetheid getoond, die een veel gedifferentieerder inzicht geven in genreverschillen dan mogelijk is door alle woorden een enkelvoudige graad van concreetheid toe te kennen. Ten slotte zijn verschillen getoond wat betreft persoonlijke voornaamwoorden, persoons-, organisatie- en plaatsnamen, en werkwoordstijden. De genreverschillen in werkwoordstijden ondersteunen de gedachte dat zij een indicatie zijn van narrativiteit.

In eerste instantie zijn tien genres gekarakteriseerd, vervolgens zijn we nader ingegaan op twee verwante genres: vmbo- en havo-schoolboeken. Daarbij doet zich de vraag voor in hoeverre complexiteitsverschillen inhoudelijk dan wel stilistisch van aard zijn. Die vraag is theoretisch van belang, omdat onderwerpkeuzes echt iets anders zijn dan taalgebruikskeuzes. Maar praktisch is het belang van deze kwestie even groot, omdat stilistische verschillen belangrijker zijn voor tekstontwerpers dan inhoudelijke. In dit kader demonstreren we een nieuwe manier om de lexicale complexiteit van individuele teksten te beschrijven; deze werkt met een genrevoorspellingsmodel dat ons in staat stelt 'normale' en 'afwijkende' teksten te onderscheiden.

Welke perspectieven biedt T-Scan voor toegepast onderzoek? Om te beginnen is de tool interessant voor academische tekstonderzoekers die willen controleren of de teksten in hun experimenten verschillen wat betreft lexicale complexiteit. Maar belangrijker voor de praktijk zijn de perspectieven voor het op complexiteit beoordelen (of kiezen) van teksten buiten de academie.

Op dat punt is de belangrijkste les dat individuele teksten alleen zinvol gekarakteriseerd kunnen worden binnen hun genre, uitgaande van een genrevoorspellingsmodel. Dat onderzoeksartikelen moeilijker zijn dan romans, is immers iedereen duidelijk. Interessanter is de vraag welke range aan complexiteitsniveaus we aantreffen binnen een genre. Een antwoord op die vraag vinden we bijvoorbeeld door 'normale' en 'afwijkende' exemplaren van het genre met elkaar te vergelijken. Zulke vergelijkingen vragen om corpusonderzoek als het hier gepresenteerde, waarin op basis van een zorgvuldige tekstverzameling wordt bepaald wat gebruikelijke waarden zijn voor taalgebruikskenmerken binnen een genre. Zulke peilingen zijn overigens voor genretheoretici even interessant als voor praktijkmensen.

De aanname dat teksten bij voorkeur binnen hun genre moeten worden bezien klinkt misschien niet sensationeel, maar hij is zeker geen gemeengoed. Het is immers gebruikelijk geworden om teksten, ongeacht hun genre, van bepaalde 'taalniveaus' te voorzien (Kraf, Lentz & Pander Maat

2011 bespreken enkele instrumenten die zulke niveaus leveren). Maar uit de kracht van onze genrevoorspellingsmodellen blijkt dat genres redelijk homogeen zijn qua taalgebruik; met andere woorden, lang niet alles is mogelijk binnen een bepaald genre. Daarom is het niet zo leerzaam is het taalgebruik van een schoolboek te vergelijken met dat van een reisverslag. Veel interessanter is het om een schoolboek te vergelijken met andere schoolboeken, zoals hierboven is geïllustreerd.

Ten slotte merken we, wellicht ten overvloede, op dat tekstgerichte analyses op complexiteit uiteindelijk moeten worden gecombineerd met effectonderzoek onder lezers. Dat is ook de intentie achter de ontwikkeling van T-Scan. Maar hierboven is gebleken dat de tool ook los van zulk lezersonderzoek kan helpen om interessante tekstobservaties te doen.

Noten

1. Rogier Kraf heeft een essentiële rol gespeeld in de eerste fase van de ontwikkeling van T-Scan (2009-2012). Meer informatie over betrokkenen bij de ontwikkeling is te vinden in Pander Maat et al. (2016).
2. Een deel van de ontwikkeling van T-Scan is gefinancierd in het kader van het NWO-project *Toward a validated reading level tool for Dutch*.
3. Wij zijn veel dank verschuldigd aan Geeske Brummel. Zij verzamelde in het kader van een Utrechtse onderzoekstage in 2012 een deel van ons corpus, en demonstreerde een allereerste genrevoorspelling met T-Scan (Brummel 2012).

Bibliografie

Asheghi, N.R., Markert, K., & Sharoff, S. (2014). Semi-supervised graph-based genre classification for web pages. *TextGraphs, 9,* 39-47.

Barton, J.S.F., Hanif, H.M., Björnström, L.E. & Hills, C. (2014). The word-length effect in reading: A review. *Cognitive Neuropsychology, 31*(5-6), 378-412.

Bosch, A. van den, Busser, G.J., Canisius, S., and Daelemans, W. (2007). An efficient memory-based morphosyntactic tagger and parser for Dutch. In P. Dirix, I. Schuurman, V. Vandeghinste, and F. Van Eynde (eds.), *Computational Linguistics in the Netherlands 2006:* Selected Papers of the Seventeenth CLIN Meeting. Utrecht: LOT, 191-206.

Biber, D. (1992). On the complexity of discourse complexity: a multidimensional analysis. *Discourse Processes 15*(1), 133-163.

Biber, D. & Conrad, S. (2009). *Register, genre and style*. Cambridge: Cambridge University Press.

Biber, D., & Gray, B. (2010). Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes, 9*(1), 2-20.

Bormuth (1966). Readability: a new approach. *Reading Research Quarterly, 79*-132.

Breland, H.M. (1996). Word frequency and word difficulty: a comparison of counts in four corpora. *Psychological Science, 7*(2), 96-99.

Brummel, G. (2012). *The usability of T-Scan for automatic genre classification*. Internship Paper UiL-OTS, Utrecht University.

Brysbaert, M., Stevens, M., De Deyne, S., Voorspoels, W., & Storms, G. (2014). Norms of age of acquisition and concreteness for 30,000 Dutch words. *Acta Psychologica, 150,* 80-84.

Camblin, C., Ledoux, K., Boudewyn, M., Gordon, P.C. & Swaab, T.Y. (2007). Processing new and repeated names: Effects of coreference on repetition priming with speech and fast RSVP. *Brain Research, 1146,* 172-184.

Dale, E. & J.S. Chall (1948). A formula for predicting readability. *Educational Research Bulletin 27,* 37-54.

Expertgroep Doorlopende Leerlijnen (2008). *Over de drempels met taal en rekenen*. Enschede: SLO.

Evers-Vermeul, J. & Holtermann, M. (2013). Doorlopende leerlijnen: implicaties voor leveling van leer- en examenteksten voor het middelbaar onderwijs. *Tijdschrift voor Taalbeheersing, 25*(1), 1-24.

Fahnestock, J. (1986). Accommodating Science. The Rhetorical Life of Scientific Facts. *Written Communication 3*(3), 275-296.

Field, A. (2010). *Discovering statistics using SPSS*. Third Edition. Sage Publications.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology, 32*(3), 221-233.

Flowerdew, J. & Forest, R.W. (2015). *Signalling nouns in English. A corpus-based approach*. Cambridge: Cambridge University Press.

Graesser, A.C., McNamara, D.S., & Kulikowich, J.M. (2011). Coh-Metrix: providing multilevel analyses of text characteristics. *Educational researcher, 40*(5), 223-234.

Heuven, W.J. van, Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology, 67*(6), 1176-1190.

Kanaris, I., & Stamatatos, E. (2009). Learning to recognize webpage genres. *Information Processing & Management, 45*(5), 499-512.

Keuleers, E., Brysbaert, M. & New, B. (2010). SUBTLEX-NL: A new frequency measure for Dutch words based on film subtitles. *Behavior Research Methods, 42*(3), 643-650.

Kraf, R., Lentz, L. & Pander Maat, H. (2011). Drie Nederlandse instrumenten voor het automatisch voorspellen van begrijpelijkheid. Een klein consumentenonderzoek. *Tijdschrift voor Taalbeheersing, 33*(3), 249-265.

Langacker, R. (2013). *Essentials of cognitive grammar*. Oxford University Press.

Lee, D. (2001). Genres, registers, text types, domains, and styles: clarifying the concepts and navigating a path through the BNC jungle. *Language Learning and Technology, 5,* 37-72.

Leech, G. & Short, M. (1981/2007). *Style in fiction. A linguistic introduction to English fictional prose*. Second Edition. Harlow, UK: Pearson Longman.

Martin, W. & Maks, I. (2005). *Referentie Bestand Nederlands*. Met medewerking van S. Bopp en M. Groot.

Mittwoch, A. (2008). The English resultative perfect and its relationship to the experiential perfect and the simple past tense. *Linguistics and Philosophy, 31*(3), 323-351.

Oostdijk, N., Reynaert, M. Hoste, V. & Heuvel, H. van den (2013). *SoNaR User Documentation*. Version 1.0.4. 2005).

Pander Maat, H. (2002). *Tekstanalyse. Wat teksten tot teksten maakt*. Bussum: Coutinho.

Pander Maat, H., Kraf, R. & Dekker, N. (2016). Handleiding T-Scan. Geraadpleegd op 12 januari 2016, https://github.com/proycon/tscan/blob/master/docs/tscanhandleiding.pdf.

Pander Maat, H., Kraf, R., Bosch, A. van den, Dekker, N., Gompel, M. van, Kleijn, S., Sanders, T.J.M. & Sloot, K. van der (2014). T-Scan: a new tool for analyzing Dutch text. *Computational Linguistics in the Netherlands Journal, 4,* 53-74.

Pollatsek, A., & Hyönä, J. (2005). The role of semantic transparency in the processing of Finnish compound words. *Language and Cognitive Processes, 20*(1-2), 261-290.

Ryder, R.J. & W.H. Slater (1988). The relationship between word frequency and word knowledge. *Journal of Educational Research, 81*(5), 312-317.

Sadoski, M., E. Goetz & M. Rodriguez (2000). Engaging texts: effects of concreteness on comprehensibility, interest, and recall in four text types. *Journal of Educational Psychology, 92*(1), 85-95.

Schmitt, N., Jiang, X. & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal, 95,* 26-43.

Spooren, W., Hustinx, L., Aben, J. & Turkenburg, E. (2015). Concreetheid onder de loep. In M. Boogaard, B. van den Bogaerde, S. Bacchini, M. Curcic, N. de Jong, E. le Pichon & L. Rasier (Eds.), *Proceedings of de achtste Anéla Conferentie Toegepaste Taalwetenschap 2015* (pp. 97-110). Delft: Eburon.

Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (2000a). Automatic text categorization in terms of genre and author. *Computational Linguistics, 26* (4), 471-495.

Stamatatos, E., Fakotakis, N., & Kokkonakis, G. (2000b). Text genre detection using common word frequencies. In *Proceedings of the international conference on computational linguistics* (COLING2000) (pp. 808-814).

Staphorsius, G. (1994). *Leesbaarheid en leesvaardigheid. De ontwikkeling van een domeingericht meetinstrument*. Cito, Arnhem.

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

Trosborg, A. (1997). Text typology: Register, genre and text type. *BENJAMINS TRANSLATION LIBRARY, 26,* 3-24.

Werlich, E. (1982). *A text grammar of English*. Heidelberg: Quelle & Meyer.

Zipf, G. (1936/2002). *The Psychobiology of language. An introduction to dynamic philology*. Oxon: Routledge.

Zwitserlood, P. (1994). The role of semantic transparency in the processing and representation of Dutch compounds. *Language and Cognitive Processes, 9*(3), 341-368.

Over de auteurs

Henk Pander Maat is senior-onderzoeker bij het Utrecht Institute of Linguistics OTS aan de Universiteit Utrecht.

Nick Dekker studeerde in Utrecht Nederlands en Communicatie en organisatie; hij is nu webredacteur bij Vitens.

Bijlage 1 Het schoolboekencorpus

Overzicht

(Tussen haakjes staan de aantallen fragmenten die per methode zijn gebruikt)

Niveau	Vak	Methode
vmbo (76)	**Aardrijkskunde (25)**	Atlantis (5)
		BuiteNland (5)
		De Geo (5)
		Terra (5)
		Wereldwijs (5)
	Economie (25)	Praktische economie (6)
		Economisch bekeken (7)
		Pincode (7)
		Index (5)
	Geschiedenis (26)	De Geschiedeniswerkplaats (5)
		Memo (6)
		Pharos (5)
		Sfinx (5)
		Sprekend verleden (5)
havo (85)	**Aardrijkskunde (25)**	Terra (5)
		Wereldwijs (7)
		De Geo (8)
		Atlantis (5)
	Economie (25)	Economie in balans (10)
		Percent (5)
		Praktische economie (5)
		Transactie (5)
	Geschiedenis (35)	De Geschiedeniswerkplaats (6)
		Feniks (7)
		Memo (12)
		Pharos (5)
		Sprekend verleden (5)

Titelbeschrijvingen

Vmbo, aardrijkskunde

Dieleman, E., Kroeze, B., van der Pol, R. & van de Ven, M. (Eindredactie: Janssen, M.) (2008). *BuiteNland VMBO-KGT Aardrijkskunde voor de basisvorming* (tweede druk, tweede oplage). Groningen: Wolters-Noordhoff.

Hooghuis, F., Nijnatten, H., Peenstra, T., Schouten, M. & Wanrooij, B. van (Eindredactie: Hooghuis, F., Nijnatten, H. & Peenstra, T.) (2003). *Atlantis vmbo kgt, deel 1*. Amersfoort: Thieme-Meulenhof.

Kunnen, L., Nonnekes, H., Reichard, A. & J. Remmers-Kamp (Redactie: Nonnekes, H. & Reichard, A. (2003). *Terra vmbo kgt 1 informatieboek* (tweede editie). Groningen: Wolters-Noordhoff.

Ten Brinke, W.B., Broeke, J.L., Groen, H., De Jong, C. & Klauw, van der, E. (2006). *De Geo VMBO KGT lesboek 1*. Amersfoort: Thieme-Meulenhof.

Van der Berg. G., Bloothoofd, T., de Boer, M., Botter, H. & van Oorschot, F. (red.) (2007). *Wereldwijs vmbo kgt, deel 1* (vierde druk). Malmberg.

Vmbo, economie

Adriaansen, P. (2004). *Praktische Economie 2 vmbo kgt*. 's-Hertogenbosch: Malmberg.

Huitema, J., Peters, L., Vaart, I. van der (2004). *Economisch bekeken, basisvorming vmbo-kgt handboek* (zesde druk). 's-Hertogenbosch: Malmberg.

Kruis, M. (2006). *Pincode onderbouw vmbo kgt leerboek*. Groningen: Noordhoff.

Scholte, P., Janssen, K., Kuijpers, M., Voorend, P., Wevers, F. (2003). *Index vmbo kgt voor de basisvorming* (eerste druk). Amersfoort: ThiemeMeulenhoff.

Vmbo, geschiedenis

Berents, D., Bos, J., Rombouts, G. & Veldkamp, M. (2000). *Sfinx leesboek 1 vbo* (druk 1). Zutphen: Thieme.

Bon, J. van, Hendriks, S., Lelieveld, J., Spree, M., Stroo, R., Veldkamp, M., Venner, J., Voogt, A., & van Voorst, A. (2003). *Pharos kgt deel 1*. Amersfoort: Thieme-Meulenhoff.

Bruns, D. (red.) (2012). *De Geschiedeniswerkplaats informatieboek deel 1*. Groningen: Noordhoff.

Buskop, H., Dalhuisen, L., van der Geest, R., Steegh, F. & van der Waal, C. (2004). *Sprekend verleden handboek 1* (druk 4). Amersfoort: Nijgh Versluys.

Schrover, W. & Boxtel, C. (2+012). *Memo handboek KGT deel 2*. Groningen: Noordhoff.

Havo, aardrijkskunde

Gerits, G. (Eindredactie: I. Hendriks) (2004). *Atlantis Havo tweede fase* (tweede druk). Amersfoort: Thieme-Meulenhoff.

J.H. Bulthuis (redactie: J. Bos, J. Hofker) (1999). *De Geo havo bovenbouw. Informatieboek: Regionale beeldvorming*. Amsterdam: Meulenhoff.

Lentjes, W. (2003). *Wereldwijs Bladerboek Havo*. 's-Hertogenbosch: Malmberg.

Lentjes, W., Palings, H., Saverkouls, T., Schuiringa, J., Terlingen, M. & Teune, P. (Eindredactie:

Terlingen, M.) (2003). *Wereldwijs Handboek Havo: regionale beeldvorming* (tweede druk). 's-Hertogenbosch: Malmberg.

Lentjes. W., Palings, H., Savelouls, T., Terlingen, M., Weidema, B. & De Wolf, M. (2012). *Wereldwijs Havo tweede fase Klimaat en landschapszones: Aarde 1.* 's-Hertogenbosch: Malmberg.

Palings, H. (2003). *Wereldwijs Zakboek Bladerboek Havo*. 's-Hertogenbosch: Malmberg.

Steenbakkers, G., Ariaens, D., Oedekerk, F., van Zijl, M. & Zwiestra, A. (1999). *Terra bovenbouw Havo. Domein Migratie en vervoer*. Groningen: Wolters Noordhoff.

Havo, economie

Duijm, H., Gorter, G.F. (2002). *Percent havo, Totaalvak 1. Theorieboek*. Amersfoort: Thieme-Meulenhoff.

Haak, J.K. van den (2002). *Economie in Balans, 2e fase totaalvak havo, theorieboek 1.* Amersfoort: Nijgh Versluys.

Hinloopen, J., Adriaansen, P., Zuiderwijk, A. (2009). *Praktische economie, havo voor de 2e fase* (5e druk). Malmberg.

Kentson, J., Janssen, K., van de Donk, W. (1998). *Transactie havo voor de 2e fase deel 2*. Amsterdam: Meulenhoff Educatief.

Havo, geschiedenis

Buskop, H., Dalhuisen, L., van der Geest, R. & Steegh, F., Bastiaans, C. & de Waal, C. (1998). *Sprekend verleden 2e fase Handboek A* (druk 2, oplage 1). Amersfoort: Nijgh Versluys.

Hageraats, B., van der Heyden, C., van Oudheusden, J., van de Pol, L., Raaijmakers, J., Rongen, W., Salman, J., Schuitemaker, P., van Voorst, A. & van der Kaap, A. (1998). *Pharos themaboek voor de tweede fase* (druk 1). Amsterdam: Meulenhoff Educatief.

Kreek, R. de, Verberne, L., Veldkamp, M., Venner, J., Bosch, A. J., Voorst, A. van, Oudheusden, J. van, Boonstra, R., Heijden C. van der, Haperen M. van (2007). *Feniks, Havo tweede fase, overzicht v/d geschiedenis*. Amersfoort: Thieme-Meulenhoff.

Memo (2012). *Memo. Geschiedenis voor de tweede fase*. 's-Hertogenbosch: Malmberg.

Verkuil, D., Hijstek, B., van der Geugten, T. & Betten, E. (2006). *De Geschiedeniswerkplaats*. Groningen: Wolters-Noordhoff.