Appendix I. Onderverdeling van vragen per Lt-motivatie bijdragende schaal

Ideal L2 Self

1. Ik heb deze taal nodig voor mijn toekomst (4)
2. Ik heb deze taal nodig om later een goede baan te krijgen (8)
3. Als ik mijn toekomst voor me zie, dan zie ik mijzelf deze taal gebruiken (18)
4. Als ik naar mijn toekomst kijk, dan stel ik me voor dat ik deze taal goed spreek (26)
5. Ik heb deze taal nodig voor mijn toekomstige carrière (27)
6. Als ik later wil studeren, dan moet ik deze taal spreken (40)
7. Ik heb deze taal nodig voor het werk dat ik later wil doen (41)

Internationale attitude

8. Ik kan me voorstellen ooit in een land te wonen waar deze taal gesproken wordt (7)
9. Deze taal is belangrijk omdat het een internationale taal is (17)
10. Ik kan me voorstellen dat ik deze taal ooit spreek met mensen uit andere landen (24)
11. Deze taal kan je overal ter wereld spreken (28)
12. Ik heb deze taal nodig om internationale contacten op te doen (36)
13. Ik bewonder mensen die deze taal goed spreken (38)
14. Ik wil graag kunnen communiceren met mensen die deze taal spreken (42)

Aan de taal gehecht belang

15. Ik vind deze taal handig (12)
16. Ik vind deze taal zinloos (15)
17. Ik leer deze taal alleen omdat het verplicht is op school (16)
18. Ik vind deze taal belangrijk (32)

Culturele taal- en reisinteresse

19. Televisieprogramma's in deze taal zijn stom (3)
20. Ik heb deze taal nodig om te kunnen reizen (5)
21. Ik vind het belangrijk om meer te weten over de cultuur van deze taal (20)
22. Ik wil tijdschriften, boeken of kranten in deze taal lezen (22)
23. Ik vind deze taal interessant (23)

Externe invloed

24. Mijn vrienden vinden het belangrijk dat ik deze taal leer (6)
25. Mijn ouders vinden het belangrijk dat ik deze taal goed leer (11)
26. Ik wil goed zijn in deze taal omdat andere mensen dat belangrijk vinden (13)
27. De mensen in mijn omgeving vinden dit een belangrijke taal (19)
28. Mijn ouders stimuleren me om deze taal te leren (25)
29. Als ik deze taal niet goed beheers, zou dat negatieve gevolgen hebben voor mij (29)
30. Als ik deze taal niet goed spreek, vinden anderen mij misschien wel dom (33)
31. Mijn ouders zouden teleurgesteld zijn als ik deze taal niet goed leer (35)

L2-leerervaring

32. Ik vind het leren van deze taal leuk (1)
33. Ik vind het leren van deze taal interessant (10)
34. Ik vind dit schoolvak interessant (31)
35. Ik zou het leuk vinden om vaker les te krijgen in deze taal (39)

L2-zelfvertrouwen

36. Ik vind dat ik goed ben in deze taal (2)
37. Ik word zenuwachtig als ik deze taal moet spreken in de les (9)
38. Ik heb altijd het gevoel dat iedereen deze taal beter kan dan ik (14)
39. Ik vind het makkelijk om deze taal te gebruiken (30)
40. Ik zou zenuwachtig worden als iemand mij aanspreekt in deze taal (34)
41. Ik zou het moeilijk vinden om met iemand in deze taal te spreken (37)

Enkele tweetaligheid

42. Als ik goed ben in deze taal, zijn andere talen niet belangrijk (21)

Geautomatiseerde beoordeling van schrijfvaardigheid:

de (on)mogelijkheden voor teksten van beginnende schrijvers

Hiske Feenstra, Karen Keune, Henk Pander Maat, Theo Eggen & Ted Sanders

In deze exploratieve studie is de relatie tussen tekstcomplexiteit en tekstkwaliteit onderzocht, en is verkend wat de mogelijkheden zijn om maten van tekstcomplexiteit in te zetten binnen een schrijfvaardigheidsmeting in het primair onderwijs. De resultaten van deze studie laten zien dat de validiteit van verschillende maten van tekstcomplexiteit beïnvloed wordt door het feit dat schrijfproducten van basisschoolleerlingen nog van gebrekkige kwaliteit zijn. Niettemin kan geautomatiseerde tekstanalyse dienen als een objectieve en informatieve meetmethode, waarmee een uitgebreide analyse mogelijk wordt van verscheidene linguïstische kenmerken die gerelateerd zijn aan tekstkwaliteit. De resultaten benadrukken daarnaast de noodzaak om nader onderzoek uit te voeren en zo de bruikbaarheid van geautomatiseerde tekstanalyse in het primair onderwijs verder te onderbouwen.

Taaltechnologische ontwikkelingen maken het mogelijk een groot aantal tekstkenmerken automatisch te analyseren. Applicaties die gebouwd zijn met behulp van *machine learning*-technieken kunnen woordsoorten benoemen, morfologische analyses uitvoeren en de structuur van zinnen bepalen. Op die manier zijn verscheidene tekstkenmerken te extraheren die kunnen dienen als voorspeller van de complexiteit van een tekst. Deze complexiteitskenmerken kunnen bijvoorbeeld gebruikt worden bij het beoordelen van de leesbaarheid van een tekst (onder de aanname dat een hoge tekstcomplexiteit het lezen bemoeilijkt) (Pander Maat & Kraf, 2009), of bij het bepalen van de vaardigheid van een schrijver (onder de aanname dat een hoge tekstcomplexiteit blijk geeft van een hoge schrijfvaardigheid).

Dit artikel doet verslag van een exploratieve studie naar de inzet van taaltechnologische applicaties bij het analyseren van opstellen geschreven door beginnende schrijvers; onderdeel van een proefschrift over het meten van schrijfvaardigheid in het primair onderwijs (Feenstra, 2014).[1] De huidige applicaties zijn tot nu toe voornamelijk ingezet voor teksten van (jong)volwassen schrijvers. Teksten van beginnende schrijvers bevatten namelijk relatief veel fouten in spelling en zinsbouw, bovendien is vaak is weinig tot geen interpunctie toegepast. Deze onvolkomenheden brengen enkele moeilijkheden in de analyse met zich mee, die verderop in dit

artikel besproken zullen worden. Het doel van dit onderzoek was (1) het verkennen van de relatie tussen maten van tekstcomplexiteit en tekstkwaliteit, en (2) de mogelijkheden om geselecteerde maten in te zetten binnen een schrijfvaardigheidsmeting in het primair onderwijs. Hiervoor is gebruik gemaakt van T-Scan, een programma voor geautomatiseerde complexiteitsanalyse van Nederlandse teksten (Pander Maat et al., 2014).

Het beoordelen van tekstkwaliteit

In schrijfvaardigheidsmetingen wordt tekstkwaliteit gewoonlijk geïnterpreteerd als maat voor schrijfvaardigheid, onder de veronderstelling dat vaardige schrijvers teksten van hoge kwaliteit produceren (Deane & Quinlan, 2010; McNamara, Crossley, & McCarthy, 2010; Witte & Faigley, 1981). Kandidaten schrijven een tekst op basis van een gerichte opdracht, zodat de schrijfproducten vergelijkbaar zijn. Vervolgens beoordelen een of meer beoordelaars de kwaliteit van deze schrijfproducten op basis van een beoordelingsmodel. Verschillende studies laten echter zien dat het betrouwbaar en valide beoordelen van tekstkwaliteit een notoir lastige taak is (Breland, Camp, Jones, Morris, & Rock, 1987; Godshalk, Swineford, & Coffman, 1966; Cushing Weigle, 2002; Knoch, 2011), omdat de oordelen die beoordelaars geven over de kwaliteit van een tekst door verschillende factoren onterecht worden beïnvloed.

Ten eerste beperkt de subjectiviteit van beoordelaars de consistentie van de oordelen over tekstkwaliteit (zie Meuffels, 1994). De score die een beoordelaar geeft, wordt niet alleen bepaald door de vaardigheid van de schrijver, maar ook beïnvloed door bijvoorbeeld de individuele weging van bepaalde tekstkenmerken, de striktheid waarmee het beoordelingsmodel wordt toegepast, en de omstandigheden tijdens het beoordelen. Daarnaast zijn effecten gevonden van zowel de specifieke schrijftaak als van de aspecten van tekstkwaliteit die beoordeeld worden (Schoonen, 2005; Van den Bergh, De Maeyer, Van Weijen, & Tillema, 2012). Al deze variabelen zorgen voor ongewenste variantie tussen de oordelen van verschillende beoordelaars, en binnen een reeks oordelen van een enkele beoordelaar.

Ondanks de beperkte betrouwbaarheid van de hierboven beschreven methode van schrijfvaardigheidsmeting, wordt het afnemen van schrijfopdrachten en het beoordelen van de schrijfproducten als een meer authentieke en directe methode gezien dan bijvoorbeeld het afnemen van een (meerkeuze) toets. Met andere woorden: het bevordert de validiteit van de meting (Bachman & Palmer, 1996). Om deze reden wordt het beoordelen van geproduceerde tekst essentieel geacht bij het meten van schrijfvaardigheid (Blood, 2012).

Om de effecten van beoordelaars, schrijftaken en beoordelingsmodellen te beperken, dienen binnen een schrijfvaardigheidsmeting per kandidaat meerdere taken afgenomen te worden die door meerdere beoordelaars worden beoordeeld (Schoonen, 2005, Van den Bergh, De Maeyer, Van Weijen, & Tillema, 2012), met als nadeel dat de toetsduur en de beoordelingslast aanzienlijk zijn. Het geautomatiseerd beoordelen van schrijfproducten (*automated essay scoring*, AES) kan bijdragen aan een objectieve beoordeling doordat taaltechnologische technieken worden ingezet om – zonder tussenkomst van een beoordelaar – tekstkenmerken te identificeren die indicatief zijn voor tekstkwaliteit. De ontwikkeling van geautomatiseerde scoring begon in de jaren 60 met *Project Essay Grade* (Page, 1966). Inmiddels zijn voor de Engelse taal verschillende AES-applicaties beschikbaar, waaronder *e-rater* (Burstein et al., 1998), *Intellimetric* (Elliot, 2003) en *Intelligent Essay Assessor* (Foltz, Landaur, & Laham, 1999).

Geautomatiseerde beoordeling van schrijfvaardigheid

De validiteit van het geautomatiseerd beoordelen van schrijfvaardigheid staat sinds de opkomst ervan ter discussie. In deze discussie zijn drie algemene bezwaren te onderscheiden (Page & Petersen, 1995). Ten eerste zullen computers nooit in staat zijn om taal op dezelfde manier te begrijpen en te waarderen als mensen dat doen. Bovendien zijn AES-applicaties wellicht niet in staat opstellen te detecteren die geschreven zijn met de bedoeling het systeem te ondermijnen door in te spelen op de beoordelingscriteria, en die daardoor een onterecht hoge (of lage) score zullen krijgen. Ten slotte wordt als bezwaar opgeworpen dat AES-applicaties voornamelijk oppervlaktekenmerken van een tekst kunnen beoordelen, waarmee voorbij wordt gegaan aan de kenmerken die er werkelijk toe doen bij het beoordelen van tekstkwaliteit (o.a. stijl, doel- en publiekgerichtheid).

In de loop der jaren zijn vele studies uitgevoerd met als doel het gebruik van AES te valideren. De meest gebruikte methode hierbij is het vergelijken van de overeenstemming tussen menselijke beoordelaars onderling met de overeenstemming tussen mens en machine (Attali & Burstein, 2006). In deze studies wordt over het algemeen een hoge overeenstemming gevonden tussen menselijke oordelen en geautomatiseerde oordelen over tekstkwaliteit, wat impliceert dat de kenmerken die gemeten worden door AES-applicaties inderdaad onderdelen van het construct schrijfvaardigheid zijn (Deane & Quinlan, 2010).

Echter, een hoge overeenstemming met menselijke oordelen is geen afdoende bewijs voor de validiteit van een geautomatiseerd oordeel. Zo blijkt tekstlengte bijvoorbeeld zeer hoog samen te hangen met oordelen over tekstkwaliteit, terwijl de lengte van een tekst op zichzelf geen indicator is van tekstkwaliteit en daarmee dus geen valide maat voor schrijfvaardigheid (Attali & Burstein, 2006). Bovendien zijn zelfs experts het vaak niet eens over de kwaliteit van een tekst, wat het gebruik van een menselijk oordeel als (enige) criterium onwenselijk maakt (Clauser, Kane, & Swanson, 2002).

Naast de overeenkomst met menselijke oordelen zijn daarom sinds begin deze eeuw andere manieren van validatie in opkomst. Yang, Buckendahl, Juszkiewicz en Bhola (2002) bespreken twee alternatieve validatiemethoden. In de eerste methode wordt de relatie tussen geautomatiseerde schrijfvaardigheidsscores en andere maten van schrijfvaardigheid onderzocht. In de tweede benadering ligt de focus op de interpreteerbaarheid van de scores die AES-applicaties produceren, en het belang om de onderliggende scoringsprocessen te begrijpen, in plaats van slechts te streven naar een hoge overeenstemming met menselijke oordelen. Ondanks hun grote waarde bij het onderbouwen van de validiteit van AES-scores, worden deze alternatieven slechts weinig toegepast in validatiestudies rondom het geautomatiseerd beoordelen van schrijfvaardigheid.

In de hier beschreven exploratieve studie zijn de hierboven genoemde alternatieve benaderingen van validiteitsonderzoek wel toegepast, om een zo compleet mogelijk beeld te krijgen van de inzetbaarheid van geautomatiseerde analyse voor teksten van beginnende schrijvers. Er zijn vooraf geen assumpties over de relatie tussen tekstcomplexiteit en schrijfvaardigheid aangenomen: alle maten van tekstcomplexiteit waren verkiesbaar als indicator van schrijfvaardigheid. De uiteindelijke selectie van maten is gebaseerd op een kwantitatieve studie naar de relatie met schrijfvaardigheid enerzijds, en een kwalitatieve studie naar de interpreteerbaarheid als maat van tekstkwaliteit anderzijds.

Methode

Dataverzameling

Op vijf verschillende basisscholen kregen begin 2009 in totaal 438 leerlingen uit groep 5 tot en met 8 de opdracht een persuasief briefje te schrijven. Om de invloed van handschrift op oordelen over tekstkwaliteit uit te schakelen en om de logistiek te vereenvoudigen, zijn de opstellen overgetypt. Uit de verzameling opstellen zijn op basis van expertoordelen ankeropstellen geselecteerd volgens de methode zoals beschreven in Van den Bergh & Rijlaarsdam (1986). Met deze ankeropstellen werden beoordelingsschalen samengesteld die bij de beoordeling van de opstellen als referentie dienden (Feenstra, 2012). Elk opstel werd door minstens twee ervaren beoordelaars van een oordeel voorzien op de aspecten Inhoud, Structuur en Taalverzorging.

Pilot met T-Scan

Op dit moment is er voor Nederlandstalige teksten geen applicatie beschikbaar die ontwikkeld is om schrijfvaardigheid mee te beoordelen. In deze studie is daarom gebruik gemaakt van het Nederlandse programma T-Scan (Pander Maat et al., 2014) om de linguïstische kenmerken van teksten mee te analyseren. T-Scan is ontworpen om de complexiteit van teksten te analyseren en wordt bijvoorbeeld ingezet om de leesbaarheid van teksten te voorspellen. Binnen T-Scan (versie maart 2013) worden 147 kenmerken van complexiteit bepaald, geclassificeerd in acht verschillende categorieën (zie tabel 1).

Om de specifieke moeilijkheden bij analyse van teksten van beginnende schrijvers te identificeren, werd eerst een beperkte steekproef (n=50) van opstellen geanalyseerd. Uit deze analyse bleek dat spelfouten en gebrek aan interpunctie een negatief effect hebben op de validiteit van de complexiteitsscore.

Ten eerste krijgen woorden die fout gespeld zijn in sommige gevallen een foutieve woordsoort toegewezen. In dit geval worden alle overige maten die gebaseerd zijn op de

CATEGORIE	N MATEN	GEMETEN TEKSTKENMERKEN (VOORBEELDEN)
Woordcomplexiteit	24	woordlengte, woordfrequentie
Zinscomplexiteit	30	zinslengte, zinscomplexiteit
Informatiedichtheid	12	o.a. type/token-ratio, inhoudswoorden, bijwoorden
Coherentie	17	connectieven, argumentoverlap, onbepaalde naamwoordgroepen
Concreetheid	12	concrete zelfstandig naamwoorden. en bijvoeglijk naamwoorden
Persoonlijkheid	23	o.a. persoonlijke voornaamwoorden, actiewerkwoorden, vraagwoorden
Woordsoorten	10	o.a. voorzetsels, bijwoorden, tussenwerpsels
Diversen	19	o.a. tegenwoordige tijd, koppelwerkwoorden, infinitieven

Tabel 1. Maten van tekstcomplexiteit in T-Scan (2013)

benoeming van woordsoorten (bijv. het aantal inhoudswoorden) beïnvloed. In andere gevallen wordt een fout gespeld woord als 'onbekend' gezien, wat invloed heeft op maten als woordfrequentie en concreetheid.

Ten tweede leidt het gebrek aan interpunctie ertoe dat zinnen als lang en complex worden beschouwd, omdat T-Scan zinseindetekens (punt, uitroepteken en vraagteken) gebruikt om de zinsgrenzen te identificeren. Naast een evident effect op de score voor zinscomplexiteit, beïnvloedt gebrekkige interpunctie ook alle maten waarin de zin wordt gebruikt als analyse-eenheid (i.e. 28 van de 147 maten).

Als gevolg van het bovenstaande wordt de validiteit van vrijwel elke maat in T-Scan negatief beïnvloed door onvolkomenheden in de te analyseren teksten. Om deze reden is ervoor gekozen de fouten in spelling en de ontbrekende interpunctie te herstellen, voorafgaand aan de analyse met T-Scan. Hiertoe werden alle opstellen gecontroleerd door een taalkundige en waar nodig verbeterd. In totaal werden 470 spelfouten verbeterd. Bij 110 opstellen ontbrak de interpunctie vrijwel volledig, en is het minimale aantal zinsgrenzen toegevoegd dat nodig was om de tekst begrijpelijk te maken.

Kwantitatieve analyse: selectie van complexiteitsmaten

Na de eventuele voorbewerking zijn alle opstellen geanalyseerd met behulp van T-Scan. De output van T-Scan bestaat per opstel uit waardes voor elke beschikbare maat van tekstcomplexiteit (zie tabel 1). Om te bepalen welke maten van tekstcomplexiteit mogelijk bruikbaar zijn als indicator van schrijfvaardigheid, is de samenhang tussen tekstcomplexiteit en zowel het leerjaar van de schrijver als het menselijk oordeel over het opstel onderzocht.

Verband tussen complexiteitsscore en leerjaar

Van de gehele opstelverzameling zijn de gemiddelde waarden per leerjaar berekend voor elke maat in T-Scan. Op basis van deze waarden zijn 35 maten geselecteerd waarvoor de waarden samenhang leken te vertonen (i.e. systematisch leken te stijgen of dalen) met het leerjaar, zie Appendix A. Van deze maten is vervolgens de statistische kwaliteit geëvalueerd. Eerst is de distributie van de uitkomsten per maat bekeken. Maten waarvoor te weinig observaties werden gevonden zijn uitgesloten van verdere analyse ('A' in Appendix A). Voor maten waarbij de score-distributie beïnvloed

SCORENIVEAU PERCENTIELEN	1 ≤p25	2 >p25 - p50	3 >p50 - p75	4 >p75	totaal
N opstellen groep 5	32	16	13	5	66
N opstellen groep 6	19	22	13	23	77
N opstellen groep 7	8	18	25	20	71
N opstellen groep 8	2	2	9	9	22
N opstellen totaal	**61**	**58**	**60**	**57**	**236**

Tabel 2. Opstellen per scoreniveau

werd door de afwezigheid van deze maat in een grote proportie opstellen, zijn alle opstellen met een waarde '0' uitgesloten van verdere analyse ('B' in Appendix A). Daarna is de Shapiro-Wilk-toets gebruikt om per maat te bepalen of de waarden normaal verdeeld waren. Wanneer geen normale verdeling werd gevonden, is de non-parametrische Wilcoxon-toets toegepast om significantie te toetsen ('C' in Appendix B). Voor alle andere maten werd de significantie getoetst met behulp van een t-toets.

Verband tussen complexiteitsscore en menselijke beoordeling schrijfvaardigheid

Van 236 opstellen uit de verzameling waren scores voor aspecten van schrijfvaardigheid (i.e. inhoud, structuur en taalverzorging) beschikbaar, gegeven door twee onafhankelijke beoordelaars uit een groep getrainde beoordelaars. Op basis van de totaalscores voor schrijfvaardigheid zijn percentielscores berekend en zijn de opstellen ingedeeld in vier vaardigheidsgroepen – onafhankelijk van leerjaar, zie tabel 2. Voor de 24 maten die na de analyse van samenhang met leerjaar overbleven, is de samenhang met de opstelscore bepaald, zie Appendix B. Ook voor deze maten is de statistische kwaliteit geëvalueerd door eerst de score-distributie te bekijken (en waar nodig opstellen uit te sluiten) en vervolgens de significantie te toetsen.

Resultaten

Appendix A geeft de samenhang van 35 T-Scan-maten met leerjaar weer. Per maat worden de gemiddelde waarden gegeven voor elk leerjaar, gevolgd door het significantieniveau voor het verschil tussen leerjaar 5 en 8. De codes onder 'Opmerking' geven eventuele bijzonderheden per maat aan. Op basis van deze uitkomsten zijn 11 complexiteitsmaten uitgesloten van verdere analyse ('^' in Appendix A). Appendix B geeft op gelijke wijze de samenhang van de overgebleven 24 T-Scan-maten met opstelscore weer.

Op basis van de samenhang tussen complexiteitsmaten uit T-Scan en zowel leerjaar als opstelscore, zoals weergegeven in Appendix A en B, zijn 13 complexiteitsmaten geïdentificeerd die valide indicatoren van schrijfvaardigheid lijken te zijn. In tabel 3 zijn de geselecteerde maten gepresenteerd, samen met een omschrijving en de veronderstelde relatie met schrijfvaardigheid. Deze veronderstelde relatie is nader onderzocht in een kwalitatieve analyse, die hierna wordt beschreven.

Kwalitatieve analyse: evaluatie van geselecteerde complexiteitsmaten

Van de geselecteerde maten van tekstcomplexiteit (tabel 3) wordt op basis van hun relatie met leerjaar en opstelscore aangenomen dat ze indicatief zijn voor schrijfvaardigheid. Op basis van de kwantitatieve analyse worden de volgende relaties tussen tekstcomplexiteit en schrijfvaardigheid verondersteld:

a. Woordcomplexiteit en woordrijkdom nemen toe met toenemende schrijfvaardigheid.
b. Zinscomplexiteit neemt toe met toenemende schrijfvaardigheid.
c. Het gebruik van conceptueel en syntactisch eenvoudige cohesieve elementen neemt af met toenemende schrijfvaardigheid.

Om te bepalen of de veronderstelde relaties tussen tekstcomplexiteit en schrijfvaardigheid valide zijn, zijn de uitkomsten van de geselecteerde maten op drie manieren nader onderzocht. Ten eerste zijn voor de maten die samengesteld zijn uit een specifieke set woorden (i.e. voorzetsels, verwijzende voornaamwoorden en connectieven) het aantal voorkomens per woord in kaart gebracht, en is de distributie per leerjaar geanalyseerd. Ten tweede zijn opstellen met extreme waarden

	MAAT	OMSCHRIJVING	VERONDERSTELD VERBAND MET SCHRIJFVAARDIGHEID[1]
WOORD			
Lexicale complexiteit	woordlengte	aantal letters per woord (gem.)	+ neemt toe met vaardigheid
	woordcomplexiteit	aantal morfemen per woord (gem.)	+ neemt toe met vaardigheid
	woordfrequentie	proportie woorden uit 50% meest frequente woorden uit frequentielijst[2]	- neemt af met vaardigheid
Lexicale rijkdom	bijwoordelijke bepalingen	aantal bijwoordelijke bepalingen per deelzin (gem.)	+ neemt toe met vaardigheid
	type/token-ratio	aantal versch. exemplaren (tokens) per totaal aantal lemma's (types) (gem.)	- neemt af met vaardigheid
	voorzetsels	aantal voorzetsels (per 1000 w.)	+ neemt toe met vaardigheid
ZIN			
Zinscomplexiteit	ondergeschikte bijzinnen	aantal ondergeschikte bijzinnen (per 1000 w.)	+ neemt toe met vaardigheid
	afhankelijkheidslengte	afstand tussen zinsdelen (gem.)	+ neemt toe met vaardigheid
	deelzinnen	aantal deelzinnen (per 1000 w.)	- neemt af met vaardigheid
	zinslengte	aantal woorden per zin (gem.)	+ neemt toe met vaardigheid
TEKST			
Coherentie	argumentoverlap	overlap tussen argumenten[3] in voorafgaande 10 woorden (per 1000 w.)	- neemt af met vaardigheid
	verwijzende voornaamwoorden	aantal persoonlijke/bezittelijke voornaamwoorden in 3e persoon en aanwijzende voornaamwoorden (per 1000 w.)	- neemt af met vaardigheid
	connectieven[4]	aantal connectieven (uit versch. categoriën[5]) (per 1000 w.)	- neemt af met vaardigheid

1. toename/afname met toenemende schrijfvaardigheid; 2. Staphorsius, 1994; 3. argumenten: voornaamwoorden (excl. aanwijzende); namen; zelfstandig naamwoorden; hoofdwerkwoorden; 4. samengestelde maat: T-Scan geeft waarden per categorie; 5. temporeel; opsommend; contrastief, comparatief, causaal

Tabel 3. Geselecteerde complexiteitsmaten voor een geautomatiseerde beoordeling van schrijfvaardigheid

POTENTIËLE BEDREIGINGEN	UITLEG EN MOGELIJKE OPLOSSINGEN
Taakafhankelijkheid	**Beïnvloedt woordcomplexiteit:** Woorden die gegeven worden in de schrijftaak, kunnen de gemiddelde woordcomplexiteit positief of negatief beïnvloeden, terwijl ze geen verband houden met de vaardigheid van de schrijver. Door per taak een lijst woorden op te stellen die uitgesloten dienen te worden van de analyse, kan dit ongewenste taakeffect naar verwachting verkleind worden. **Beïnvloedt coherentie:** Het gebruik van (expliciete) coherentierelaties hangt deels af van het tekstdoel dat door de schrijftaak wordt opgedragen. Hierdoor kan het aan- of afwezig zijn van bepaalde relaties alleen juist beoordeeld worden met inachtneming van het tekstdoel. Door per taak slechts de relevante coherentiematen mee te nemen in de evaluatie, kan dit ongewenste taakeffect naar verwachting verkleind worden.
Gebrekkige spelling	**Beïnvloedt maten op woordniveau:** Fout gespelde woorden worden niet herkend in frequentielijsten en krijgen mogelijk een foutieve woordsoort toegewezen, met als mogelijk gevolg dat deze woorden als laagfrequent worden beschouwd en/of andere maten op woordniveau beïnvloeden. Door een voorbewerkingsmodule te gebruiken waarin spelfouten worden gedetecteerd en gecorrigeerd, kunnen deze effecten van spelfouten waarschijnlijk beperkt worden. **Beïnvloedt maten op zinsniveau:** Fout gespelde woorden worden mogelijk niet herkend door de grammaticale ontleedfunctie binnen T-Scan, waardoor maten gebaseerd op zinsontleding beïnvloed worden. Door een voorbewerkingsmodule te gebruiken waarin spelfouten worden gedetecteerd en gecorrigeerd, kan dit effect waarschijnlijk beperkt worden.
Gebrek aan interpunctie	**Beïnvloedt maten op zinsniveau** In teksten waarin (zo goed als) geen zinseindetekens worden gebruikt, zal de zinslengte extreem hoog uitvallen. Hierdoor worden alle maten die gebaseerd zijn op zinslengte beïnvloed. Door een voorbewerkingsmodule te gebruiken waarmee ofwel opstellen met een gebrek aan interpunctie worden opgespoord en aangepast, ofwel waarmee een gestandaardiseerde interpunctie wordt toegepast op alle opstellen, kan dit effect waarschijnlijk ondervangen worden.
Tekstlengte	**Beïnvloedt maten op woordniveau** In een korte tekst heeft het gebruik van enkele elementen die lexicale complexiteit (bijv. lange woorden) en/of lexicale rijkdom (bijv. voorzetsels) aanduiden een relatief grote invloed, aangezien deze maten gebaseerd zijn op dichtheid (aantal per 1000 woorden). Tegelijkertijd zal de kans op meerdere voorkomens (tokens) van hetzelfde woord (type) klein zijn, wat de type/token-ratio (een maat voor lexicale diversiteit) onterecht verhoogt. Het berekenen van gemiddelden per deelzin (in plaats van dichtheid) lijkt dit effect niet te verhelpen (Feenstra, 2014). Door de specifieke effecten van tekstlengte op woordmaten nader te onderzoeken en/of door het aantal woorden waarop de maten gebaseerd zijn aan te passen, kunnen deze effecten mogelijk beperkt worden. **Beïnvloedt maten op tekstniveau** In een korte tekst heeft het gebruik van enkele cohesieve elementen een relatief grote invloed op de coherentiematen, die gebaseerd zijn op dichtheid (aantal per 1000 woorden). Het berekenen van gemiddelden per deelzin (in plaats van dichtheid) lijkt dit effect niet te verhelpen (Feenstra, 2014). Door de specifieke effecten van tekstlengte op woordmaten nader te onderzoeken en/of door het aantal woorden waarop de maten gebaseerd zijn aan te passen, kunnen deze effecten mogelijk beperkt worden.

Tabel 4. Evaluatie van complexiteitsmaten

geanalyseerd om zo de kenmerken te identificeren die tot de extreme hoge of lage waarden leiden en daarmee potentiele bedreigingen zijn voor de betrouwbaarheid en validiteit. Ten derde is de interpreteerbaarheid van de maten geëvalueerd door – op basis van de geselecteerde maten uit tabel 3 – opstellen van verschillende complexiteitsniveaus te selecteren, om zo de ontwikkeling in schrijfvaardigheid te illustreren.

Resultaten

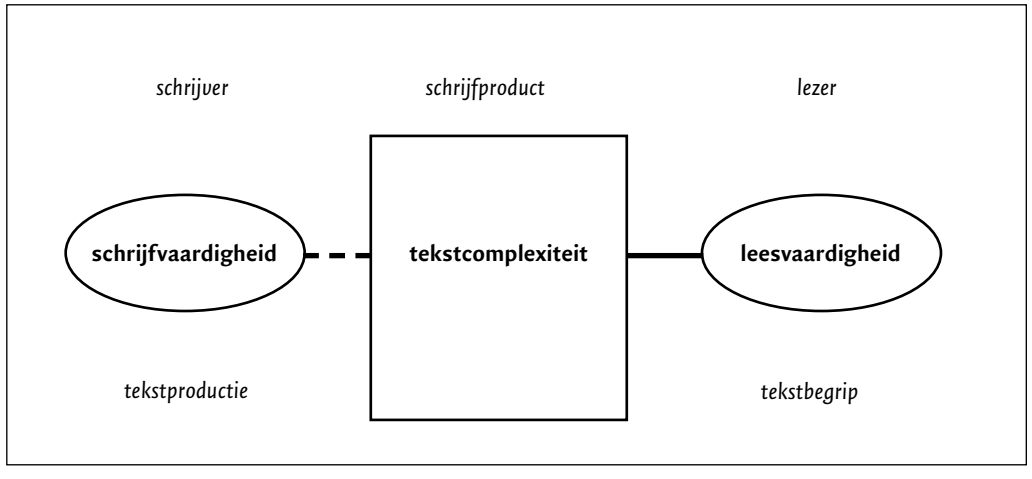
Op basis van bovenstaande drie evaluatiemethoden zijn voor alle geselecteerde maten de potentiële bedreigingen voor de betrouwbaarheid en validiteit geïdentificeerd en zijn mogelijke oplossingen benoemd. Tabel 4 geeft de resultaten van de evaluatie weer. De grote invloed van onvolkomenheden zoals ontbrekende interpunctie en gebrekkige spelling op de maten van tekstcomplexiteit was al in de pilot met T-Scan ontdekt, en is in deze studie ondervangen door het voorbewerken van de opstellen. Voor de volledigheid zijn deze kenmerken van teksten geschreven door beginnende schrijvers toch toegevoegd aan de evaluatie. In de resultaten valt verder op dat tekstlengte een ongewenste invloed heeft op maten op elk tekstniveau (woord, zin en hele tekst). Daarnaast wordt een groot aantal maten beïnvloed door de specifieke taak op basis waarvan de tekst geschreven is.

Discussie en conclusie

In de huidige studie is de mogelijkheid verkend om geautomatiseerde tekstanalyse te gebruiken bij een grootschalige schrijfvaardigheidsmeting in het primair onderwijs. Hiervoor is gebruik gemaakt van T-Scan, een programma voor geautomatiseerde complexiteitsanalyse van Nederlandse teksten. Om te bepalen welke kenmerken van tekstcomplexiteit indicatief zijn voor de schrijfvaardigheid van basisschoolleerlingen, is een kwantitatieve analyse uitgevoerd van de overeenstemming tussen 37 maten van tekstcomplexiteit uit T-Scan en twee maten van schrijfvaardigheid, te weten leerjaar en opstelscore. Op basis van de uitkomsten zijn 13 maten van tekstcomplexiteit geselecteerd die indicatief blijken voor schrijfvaardigheid (zie tabel 3). Om de interpreteerbaarheid en validiteit van deze 13 tekstkenmerken als maten van schrijfvaardigheid te evalueren is vervolgens een kwalitatieve analyse uitgevoerd.

De resultaten van deze kwalitatieve studie tonen ten eerste aan dat de gebrekkige kwaliteit van opstellen geschreven door beginnende schrijvers de geautomatiseerde analyse ernstig bemoeilijkt, aangezien voorbewerking van de schrijfproducten noodzakelijk is voor een betrouwbare meting en een valide interpretatie. Bovendien maken de resultaten duidelijk dat verschillende factoren (zoals tekstlengte en schrijfopdracht) de gemeten complexiteit van een tekst onrechtmatig beïnvloeden, wat inhoudt dat verscheidene maten voor tekstcomplexiteit aangepast dienen te worden om ze specifiek geschikt te maken voor de evaluatie van de schrijfvaardigheid. Het geautomatiseerd analyseren van tekstkenmerken is dus niet zonder meer inzetbaar voor de beoordeling van schrijfvaardigheid van beginnende schrijvers.

In deze studie is onderzocht in hoeverre kenmerken van tekstcomplexiteit kunnen dienen als indicatoren van schrijfvaardigheid. Op basis van leesbaarheidsonderzoek is al kennis opgedaan over de complexiteitskenmerken die het tekst*begrip* van de *lezer* beïnvloeden. Vanuit een ander perspectief zou de complexiteit van een tekst ook iets kunnen zeggen over het niveau van tekst*productie* van de *schrijver*, en daarmee over de vaardigheid van de tekstschrijver (zie figuur 1). Op basis van de huidige literatuur is echter nog weinig bekend over de relatie tussen tekstcomplexiteit en de vaardigheid van de schrijver in het



Figuur 1. Tekstcomplexiteit, leesvaardigheid en schrijfvaardigheid

algemeen, en over de relatie tussen coherentie en de ontwikkeling in schrijfvaardigheid in het bijzonder.

Hoewel coherentie gezien wordt als een belangrijk kenmerk van tekstkwaliteit, en daarmee als een wezenlijk onderdeel binnen een meting van schrijfvaardigheid (Sanders, 2005) blijkt het lastig op basis van taaltechnologie betrouwbare en valide maten voor geautomatiseerde beoordeling van coherentie te vinden. In dit onderzoek blijkt het gebruik van expliciete coherentie-elementen (connectieven) bijvoorbeeld af te nemen met een toenemende schrijfvaardigheid van leerlingen, terwijl onderzoek bij verder gevorderde schrijvers juist een ander beeld laat zien (Witte & Faigly, 1981; Sanders & Van Wijk, 1996). Een verklaring hiervoor zou kunnen zijn dat de afname van connectieven in deze studie vooral veroorzaakt wordt door een daling in het gebruik van de conceptueel en syntactisch eenvoudige connectieven als ‘maar’ en ‘en’. Het geautomatiseerd analyseren van coherentie wordt daarnaast bemoeilijkt door het feit dat coherentie in teksten voor een deel gebaseerd is op mentale samenhang, en dus impliciet kan blijven. Met andere woorden: voor een hogere samenhang is niet per definitie een hogere dichtheid aan connectieven nodig. Figuur 2 geeft enkele voorbeelden van de relatie tussen coherentie en het gebruik van connectieven.

Samen met het beoordelen van andere (min of meer) impliciete tekstkenmerken als

ik heb al tien repen **maar** ik weet niet waar de winkel is. kunnen jullie mij helpen? **maar** ik wil zo graag die telefoon. **maar** als ik die krijg dan ben ik heel blij.	Ik heb 8 punten en 2 wikkels opgestuurd, **omdat** ik nergens meer Smikkels kon vinden met punten. Ik wil zo graag die telefoon dat ik het maar zo heb gedaan. Ik heb alle winkels afgezocht en bij iedereen gevraagd. Ik kon die laatste twee punten niet meer krijgen. Ik wil die telefoon echt!	Ik ben een groot fan van uw repen, **en** ik zag, dat u een actie hield. Ik had acht repen, **maar** toen ik de volgende dag weer een reep ging kopen, zat er geen spaarmunt op! **Desondanks** heb ik **toch** jullie een brief gestuurd, met 8 spaarmunten, en 2 wikkels waar geen spaarmunten op zaten. **Toch** wil ik aan de actie meedoen, **dus** stuurde ik maar 2 wikkels. **Als** u zo vriendelijk wilt zijn, om toch een telefoon te sturen, zou ik u erg dankbaar zijn.
(A) lage coherentie, hoge connectieven-dichtheid (deels onjuist gebruik)	(B) hoge coherentie, lage connectieven-dichtheid	(C) hoge coherentie, hoge connectieven-dichtheid

Figuur 2. Verschillen in connectief-gebruik

lezersgerichtheid en stijl is dit een van de grootste onderzoeksuitdagingen op het terrein van geautomatiseerde beoordeling, ook internationaal gezien (Crossley & McNamara, 2012). Ook het gebruik van de type/token-ratio als maat van lexicale rijkdom behoeft nader onderzoek, aangezien deze maat onbetrouwbaar blijkt bij korte teksten: in een opstel met weinig woorden is minder kans op woordherhaling, wat resulteert in een onterecht hoge lexicale diversiteit. Bovendien is er op basis van de literatuur nog geen duidelijkheid over de relatie tussen type-token-ratio en coherentie.

De samenhang van de overige lexicale maten en van de maten voor syntactische complexiteit met schrijfvaardigheid (zie tabel 3) lijkt op basis van deze studie echter eenduidiger, wat het gebruik van deze maten veelbelovend maakt. Geautomatiseerde analyse biedt bovendien in de huidige vorm al wel een manier om op een efficiënte en consistente manier een verscheidenheid aan (expliciete) tekstkenmerken te evalueren, zowel voor individuele leerlingen als op groepsniveau. Op basis van hiervan kan inzicht verkregen worden in de ontwikkeling van taalgebruik en de mogelijke invloed van leerlingkenmerken hierop. Deze informatie kan de toetsing van schrijfvaardigheid verrijken en het schrijfonderwijs ondersteunen, bijvoorbeeld door het aanbieden van passende feedback op basis van de geanalyseerde tekstkenmerken. Daarnaast zijn de gegevens inzetbaar binnen grootschalig (peilings)onderzoek waarin prestaties van groepen leerlingen met elkaar vergeleken worden.

Al met al laten de resultaten van deze studie zien dat geautomatiseerde tekstanalyse kan dienen als een objectieve en informatieve meetmethode, waarmee een uitgebreide analyse van verscheidene linguïstische kenmerken die gerelateerd zijn aan tekstkwaliteit mogelijk wordt. De resultaten benadrukken daarnaast de noodzaak om nader onderzoek uit te voeren en zo de bruikbaarheid van geautomatiseerde tekstanalyse in het primair onderwijs verder te onderbouwen.

Met dank aan Rogier Kraf, die ons wegwijs maakte in de wereld van T-Scan.

Noot

1. Een digitale versie van het proefschrift met daarin het hier gerapporteerde onderzoek is in te zien via <www.hiskefeenstra.nl/dissertation>.

Literatuur

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater v. 2. *Journal of Technology, Learning, and Assessment*, 4(3). Retrieved from http://www.jtla.org

Bachman, L. F., & A S. Palmer (1996). *Language testing in practice*. Oxford, UK: Oxford University Press.

Bergh, H. van den, & Rijlaarsdam, G. (1986). Problemen met opstelbeoordeling? Een recept. *Levende Talen*, 413, 448–454.

Bergh, H. van den, De Maeyer, S., Van Weijen, D., & Tillema, M. (2012). Generalizability of text quality scores. In E. van Steendam, M. Tillema, G. C. W. Rijlaarsdam & H. van den Bergh (Eds.), *Measuring writing: Recent insights into theory, methodology and practices.* Leiden/Boston: Brill.

Blood, I. (2012). *Automated essay scoring: A literature review*. Retrieved from http://www.tc.columbia.edu/tesolalwebjournal

Breland, H., Camp, R., Jones, R. J., Morris, M. M., & Rock, D. A. (1987). *Assessing writing skill*. New York: College Entrance Examination Board.

Burstein, J., Braden-Harder, L., Chodorow, M., Hua, S., Kaplan, B., Kukich, K., Lu, C., & Wolff, S. (1998). *Computer analysis of essay content for automated score prediction: A prototype automated scoring system for GMAT analytical writing assessment essays.* ETS Research Report No. 98-15. Princeton, NJ: Educational Testing Service.

Clauser B. E., Kane M. T., & Swanson D. B. (2002). Validity issues for performance-based tests scored with computer-automated scoring systems. *Applied Measurement in Education*, 15(4), 413-432.

Crossley, S. A., & McNamara, D. S. (2012). Predicting second language writing proficiency: The role of cohesion, readability, and lexical difficulty. *Journal of Research in Reading*, 35(2), 115-135.

Cushing Weigle, S. (2002). Assessing writing. Cambridge: Cambridge University Press.

Deane, P. & Quinlan, T. (2010). What automated analyses of corpora can tell us about students' writing skills. *Journal of Writing Research*, 2(2), 151-177.

Elliot, S. (2003). Intellimetric: From here to validity. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 71-86). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Foltz, P. W., Landauer, T. K., & Laham, D. (1999). Automated essay scoring: Applications to educational technology. In *Proceedings of EdMedia '99*. Retrieved from ttp://www.psych.nmsu.edu/~pfoltz/reprints/Edmedia99.html

Feenstra, H. (2014). *Assessing writing ability in primary education. On the evaluation of text quality and text complexity.* (Proefschrift). Enschede: Universiteit Twente.

Feenstra, H. (2012). *De betrouwbaarheid van een beoordelingsschaal voor schrijfproducten.* Posterpresentatie tijdens de Onderwijs Research Dagen 2012, Wageningen.

Godshalk, F. I., Swineford, F., & Coffman, W. E. (1966). *The measurement of writing ability.* New York: College Entrance Examination Board.

Knoch, U. (2011). Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from? *Assessing Writing*, 16, 81-96.

McNamara, D., Crossley, S., & P. McCarthy (2010). Linguistic Features of Writing Quality. *Written Communication*, 2011(27), 57.

Meuffels, B. (1994). *De verguisde beoordelaar; Opstellen over opstelbeoordeling.* Amsterdam: Thesis Publishers.

Page, E. (1966). The imminence of grading essays by computer. *The Phi Delta Kappan.* 47(5), 238-243.

Page, E. B., & Petersen, N. S. (1995). The computer moves into essay grading: Updating the ancient test. *Phi Delta Kappan*, 76, 561-565.

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

Pander Maat, H. & Kraf, R. (2009). Leesbaarheidsonderzoek: oude problemen, nieuwe kansen. *Tijdschrift voor Taalbeheersing*, 31(2), 97-123.

Sanders, T. (2005). Tekst doordenken. Taalbeheersing als de studie van taalgebruik en tekstkwaliteit. *Tijdschrift voor Taalbeheersing*, 27(1), 58-74.

Sanders, T. & Van Wijk, C. (1996). Text analysis as a research tool: How hierarchical text structure contributes to the understanding of conceptual processes in writing. In M. Levy & S. Ransdell (Eds.), *The science of writing* (pp. 251-269). Mahwah NJ: Erlbaum.

Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modelling. *Language Testing*, 22, 1-30.

Witte, S., & Faigley, L. (1981). Coherence, cohesion, and writing quality: College composition and communication. *Language Studies and Composing*, 32(2). 189-204.

Yang, Y., Buckendahl, C., Juszkiewicz, P., & Bhola, D. (2002). A review of strategies for validating computer-automated scoring, *Applied Measurement in Education*, 15(4), 391-412.

Hiske Feenstra werkt als toetsdeskundige bij Cito waar ze de afgelopen jaren onder andere onderzoek deed naar het beoordelen van schrijfvaardigheid. In 2014 promoveerde ze op dit onderwerp aan de Universiteit Twente. Momenteel is ze verantwoordelijk voor de ontwikkeling van de adaptieve Centrale Eindtoets voor het primair onderwijs. Daarnaast is ze lid van een expertgroep die zich bezighoudt met de examinering van Nederlands op de Europese Scholen.
E-mail: <hiske.feenstra@cito.nl>.

Karen Keune is in 2012 gepromoveerd in de corpus- en computerlinguïstiek aan de Radboud Universiteit Nijmegen. Sinds 2011 is zij werkzaam bij Cito op de afdeling Psychometrie en Onderzoek waar zij als onderzoeker werkt aan het ontwikkelen en analyseren van diverse digitale taaltoetsen en -examens. Tevens is zij betrokken geweest bij verschillende onderzoeken naar het meten van schrijfvaardigheid en bij diverse nationale peilingsonderzoeken in het primair onderwijs (PPON).
E-mail: <karen.keune@cito.nl>.

Henk Pander Maat is senior docent-onderzoeker op het terrein van taal en communicatie aan de Faculteit Geesteswetenschappen van de Universiteit Utrecht. Hij houdt zich bezig met documentontwerp, begrijpelijkheidsonderzoek en automatische tekstanalyse.
E-mail: <h.l.w.pandermaat@uu.nl>.

Theo Eggen is als onderzoeker verbonden aan de afdeling Psychometrie en Onderzoek van Cito en is als bijzonder hoogleraar psychometrie werkzaam bij de afdeling Onderzoeksmethodologie, Meetmethoden en Data-analyse van de faculteit Behavioural, Management and Social Sciences van de Universiteit Twente. Hij is wetenschappelijk directeur van het Research Center voor Examinering en Certificering (RCEC). Zijn onderzoek heeft itemresponstheorie en computergestuurd adaptief toetsen als belangrijkste aandachtsgebieden.
E-mail: <theo.eggen@cito.nl>.

Ted Sanders is hoogleraar Taalbeheersing en hoofd van het Departement Talen, Literatuur en Communicatie van de Faculteit Geesteswetenschappen van de Universiteit Utrecht. Zijn onderzoek richt zich op coherentie: de manier waarop die in teksten in diverse talen tot uiting komt, de manier waarop lezers en schrijvers coherentie construeren, en de manier waarop kinderen leren coherentie uit te drukken. Bovendien is hij geïnteresseerd in de relatie met begrijpelijke taal en communicatie.
E-mail: <t.j.m.sanders@uu.nl>.

Appendix A. Het verband tussen leerjaar en 35 maten van tekstcomplexiteit in T-Scan

MAAT	Type[1]	GEMIDDELDE PER LEERJAAR 5 (n=101)	6 (n=113)	7 (n=113)	8 (n=94)	Significantie[2] 5 vs. 8	Opmerking[3]
WOORDCOMPLEXITEIT							
woordlengte (n letters)	g(w)	3,899	3,994	3,996	4,057	***	
woordcomplexiteit (n morfemen)	g(w)	1,194	1,224	1,236	1,242	***	
woordfrequentie (50%)	p	0,563	0,557	0,545	0,535	***	
ZINSCOMPLEXITEIT							
ondergeschikte bijzinnen	d	17,28	22,12	26,06	26,57	***	
relatieve bijzinnen^	d	0,578	1,061	1,857	2,214	n.v.t.	A
afhankelijkheidslengte (ow-ww)	g(t)	0,954	1,320	1,551	1,803	***	
afhankelijkheidslengte (totaal)	g(t)	1,355	1,539	1,676	1,798	***	
inhoudswoorden	g(z)	2,543	2,590	2,661	2,820	***	
deelzinnen	d	165,4	157,6	151,6	147,3	***	
zinslengte	g(t)	9,755	11,60	12,67	12,34	***	C
COHERENTIE							
contrastieve connectieven	d	39,79	34,52	33,01	26,07	***	
comparatieve connectieven	d	11,84	11,57	7,139	7,898	n.v.t.	A
temporele connectieven	d	3,849	5,327	5,254	7,601	n.v.t.	A
opsommende connectieven	d	45,11	41,44	35,89	35,03	**	
causale connectieven	d	47,39	42,12	44,72	38,08	-	B
verwijzende voornaamwoorden	d	71,42	69,40	66,29	56,50	**	C
argumentoverlap (lemmabuffer)	d	83,98	78,34	78,34	72,14	*	C
INFORMATIEDICHTHEID							
bijwoordelijke bepalingen	g(z)	0,871	0,937	1,022	0,983	*	
type/token-ratio	r	0,707	0,714	0,694	0,679	*	
PERSOONLIJKHEID							
concreetheid znw (strikt)	p	0,771	0,727	0,727	0,664	***	C
concreetheid znw (breed)^	p	0,798	0,748	0,762	0,691	***	C
persoonlijke verwijzingen ^	d	155,0	155,3	148,6	145,7	-	C
persoonlijke voornaamw (1e pers)^	d	113,4	108,8	105,6	102,7	**	
persoonlijke voornaamw (2e pers)^	d	131,0	129,3	125,6	127,0	-	
persoonlijke voornaamw (3e pers)	d	143,1	138,4	131,2	131,8	*	
OVERIG							
voorzetsels	d	47,82	59,88	62,87	80,37	***	
imperatieven^	d	8,342	6,671	5,703	5,105	n.v.t.	A
vraagzinnen^	d	9,330	9,059	5,737	4,775	n.v.t.	A
tegenwoordige tijd	d	148,7	134,5	125,9	117,7	***	
modale werkwoorden	d	52,48	46,26	44,67	39,41	***	B
hulpwerkwoorden van tijd^	d	0,130	0,159	0,203	0,212	n.v.t.	A
koppelwerkwoorden^	d	23,59	16,66	16,18	16,07	n.v.t.	A
tegenwoordige deelwoorden^	d	0,000	0,136	0,167	0,500	n.v.t.	A
infinitieven^	d	31,65	37,55	37,95	43,39	n.v.t.	A
voorspelbaarheid	m(z)	1,281	1,412	1,591	1,565	***	C

1. g=gem. per woord(w), (deel)zin (z), text (t); d=dichtheid (waarde per 1000 woorden); p=proportie, r=ratio
2. significantieniveau van verschil tussen groep 5 en groep 8: *** p < 0.001, ** p <0.01, * p<0.05, - p>0.05 (n.v.t.=te weinig observaties)
3. A: te weinig observaties; B: alle waarden 'o' verwijderd; C: geen normale verdeling, non-parametrische toets toegepast
^. maat is uitgesloten van verdere analyse

Appendix B: Het verband tussen schrijfscore en 24 complexiteitsmaten in T-Scan

MAAT	Type[1]	GEMIDDELDE PER SCORENIVEAU 1 (n=61)	2 (n=58)	3 (n=61)	4 (n=59)	Significantie[2] 1 vs. 4	Opmerking[3]
WOORDCOMPLEXITEIT							
woordlengte (n letters)	g(w)	3,848	4,004	4,019	4,067	***	
woordcomplexiteit (n morfemen)	g(w)	1,190	1,231	1,237	1,241	***	
woordfrequentie (50%)	p	0,562	0,560	0,548	0,532	*	
ZINSCOMPLEXITEIT							
ondergeschikte bijzinnen	d	18,37	19,05	25,84	27,15	**	D
afhankelijkheidslengte (ow-ww)	g(t)	0,993	1,341	1,509	1,734	***	D
afhankelijkheidslengte (totaal)	g(t)	1,394	1,543	1,690	1,799	***	D
inhoudswoorden	g(z)	2,615	2,628	2,649	2,806	*	D
deelzinnen	d	162,0	156,0	153,1	147,4	**	
zinslengte	g(t)	10,75	12,02	11,58	12,78	**	D
COHERENTIE							
contrastieve connectieven	d	46,17	46,23	38,60	31,52	***	B
comparatieve connectieven	d	17,28	6,893	7,875	7,589	n.v.t.	A
temporele connectieven	d	4,705	4,414	4,992	6,036	n.v.t.	A
opsommende connectieven	d	35,06	26,75	29,85	32,66	-	D
causale connectieven	d	32,80	24,04	32,56	29,34	-	D
verwijzende voornaamwoorden	d	33,49	32,43	32,98	28,33	-	D
argumentoverlap (lemmabuffer)	d	78,61	79,67	77,51	71,42	-	D
INFORMATIEDICHTHEID							
bijwoordelijke bepalingen	g(z)	0,899	0,974	0,958	1,055	*	
type/token-ratio	r	0,733	0,704	0,698	0,686	**	
PERSOONLIJKHEID							
concreetheid znw (strikt)	p	0,728	0,758	0,744	0,740	*	C
persoonlijke voornaamw (3e pers)	d	139,5	138,6	131,1	125,3	*	D
OVERIG							
voorzetsels	d	48,87	57,07	67,15	68,40	**	
werkwoord in tegenwoordige tijd	d	143,0	129,4	127,4	122,5	**	
modale werkwoorden	d	41,81	45,09	39,35	38,63	-	D
voorspelbaarheid	g(z)	1,346	1,441	1,505	1,615	***	D

1. g=gem. per woord(w), (deel)zin (z), text (t); d=dichtheid (waarde per 1000 woorden); p=proportie, r=ratio
2. significantieniveau van verschil tussen scoreniveaus 1 and 4: *** p < 0.001, ** p <0.01, * p<0.05, - p>0.05 (n.v.t.=te weinig observaties)
3. A: te weinig observaties; B: alle waarden '0' verwijderd; C: alle waarden '1' verwijderd; D: geen normale verdeling, non-parametrische toets toegepast

Hoe leren álle leerlingen in klassen van dertig goed moderne vreemde talen spreken?

Innovatief promotieonderzoek zoekt daarvoor oplossing

Hoe kun je feedback afstemmen op de individuele leerling en welke 'gedifferentieerde support' (Janssen e.a., 2015) in de vorm van uitleg en oefeningen kun je aanbieden om de gespreksvaardigheid in moderne vreemde talen van je leerlingen te verbeteren? En dat in reguliere klassen (waar dus sprake is van veel leerlingen)? Op deze vraag zoek ik antwoord in een vierjarig promotietraject in het kader van Dudoc-alfa <http://vakdidactiekgw.nl/dudoc-alfa/>.

Allerlei soorten fouten, niet willen spreken, spreekangst ...

Het lijkt een onmogelijke taak voor docenten in het voortgezet onderwijs om iedere leerling op maat feedback te geven om zijn gespreksvaardigheid in een vreemde taal te verbeteren. De ene leerling maakt fouten in de uitspraak, de ander gebruikt net de verkeerde woorden. Er zijn leerlingen die snel praten maar met heel veel grammaticale fouten, terwijl anderen aarzelen en lange denkpauzes nodig hebben omdat ze alles heel precies willen zeggen. En dan zijn er ook, misschien zelfs de meesten, die niet graag willen spreken, uit angst of desinteresse, en snel naar het Nederlands overgaan zodra de docent aan de andere kant van de klas staat ... De docent komt dus oren te kort tijdens een les gespreksvaardigheid.

Feedback: Wanneer, waarop, hoe? 'Tu es été à Paris?'

Docenten geven aan moeite te hebben met het geven van feedback op gespreksvaardigheid in moderne vreemde talen (Corda e.a., 2012). Wanneer geef je feedback en waarop, en hoe formuleer je de feedback opdat de leerling ervan leert en niet ontmoedigd raakt? Als de leerling net bezig is een zin te formuleren, wil je hem niet demotiveren door direct op zijn fouten te wijzen. Aan de andere kant moeten fouten ook niet inslijpen. Bovendien weet een leerling als een docent even wacht met feedback, vaak niet meer precies wanneer hij die fout maakte en wat hij toen wilde zeggen. Dat maakt dan dat hij er minder van leert. Uit onderzoek blijkt dat docenten om de communicatie niet te verstoren vaak zogenaamde recasts gebruiken (Lyster e.a., 2013). Ze herhalen wat de leerling zegt, maar dan verbeterd: 'Tu *as* été à Paris?'. Niet altijd heeft een leerling dan door dat hij een fout had gemaakt. En als hij het al door heeft, is het de vraag of hij de verbetering onthoudt. Uit onderzoek blijkt ook dat sommige leerlingen liever zelf willen nadenken voordat de docent verbetert (Yoshida, 2008). Daar leren ze veel meer van. De docent helpt dan beter door de fout te accentueren en bijvoorbeeld te vragen: 'Tu *es* été à Paris?'. Eventueel met extra hulp door eraan toe te voegen: 'C'est correct? Attention au verbe'. Maar weer andere leerlingen worden daar juist onzeker van, vooral als ze niet begrijpen waar de docent heen wil en zij zelf vooral bezig waren met de uitspraak of met het nadenken over wat ze eigenlijk wilden zeggen! Hoe moet de docent het goed doen met zoveel verschillen tussen leerlingen in een klas?