Grenzen aan het computationele lezen

Verhaar, P.A.F.

Grenzen aan het computationele lezen

In de afgelopen vijftig jaar zijn meer en meer functies op het gebied van de tekstoverdracht gedigitaliseerd. Door de komst van tekstverwerkers en van opmaakprogramma's, bijvoorbeeld, konden computers al in een vroeg stadium worden ingezet als een hulpmiddel bij de productie van boeken en tijdschriften. Fenomenen als *printing-on-demand* en online boekverkoop lieten hierna zien dat het digitale medium ook heel effectief kan worden gebruikt voor de distributie van teksten. Dankzij de ontwikkeling van browsers, *e-book readers* en *tablet computers* heeft de computer zich bovendien kunnen ontwikkelen tot een medium waarop teksten direct kunnen worden geconsumeerd. Hoewel het digitale medium in de laatste decennia dus al bijzonder ingrijpende gevolgen heeft gehad op een groot aantal aspecten van het boekenvak, werd het in de afgelopen jaren duidelijk dat de impact van het digitaliseringsproces zich nog veel verder kan uitstrekken. De teksten die via het digitale medium worden verspreid kunnen ook steeds beter door dit medium zelf worden gelezen. Computers kunnen steeds meer aspecten van het menselijke leesproces overnemen, dankzij de vrijwel voortdurende evolutie van algoritmes op het gebied van *Natural Language Processing*, *Machine Learning*, en kunstmatige intelligentie.[1]

Deze geautomatiseerde vorm van lezen, die ook wel wordt aangeduid met termen als *text mining* of *distant reading*,[2] kent momenteel veel toepassingen binnen de wetenschap. In vrijwel alle disciplines worden onderzoekers geconfronteerd met een overvloed aan publicaties, en computationele technieken kunnen vaak effectief worden ingezet bij het doorzoeken van zulke omvangrijke tekstcollecties. Digitale tools kunnen onderzoekers helpen bij het vinden van passages over specifieke onderwerpen, en ze kunnen binnen grote tekstcorpora eveneens trends en correlaties beschrijven. In zijn boek *Macroanalysis* betoogt Matthew Jockers dat computationele methoden ook op een productieve manier kunnen worden toegepast binnen de literatuurwetenschap. Het digitale medium stelt onderzoekers in staat om de schaal van analyses uit te breiden, en hierdoor kan de geschiedenis van literaire genres of van literaire perioden op een nieuwe manier worden onderzocht.[3] Omdat het gebruik van *text mining* technieken momenteel een hoge vlucht neemt is het van belang om zorgvuldig na te denken over de precieze kenmerken van deze methodologie en over de mogelijke gevolgen van deze vorm van analyse. Voor geesteswetenschappers die het geautomatiseerde lezen wellicht als een bedreiging zien is van belang om te benadrukken onderstrepen dat er momenteel nog belangrijke verschillen bestaan tussen het menselijke leesproces enerzijds en de manier waarop computers met tekstuele data omgaan anderzijds.

Een belangrijk verschil tussen de twee vormen van lezen is dat de focus bij het machinale lezen vaak beperkt blijft tot de formele of taalkundige kenmerken van teksten. Computers zijn heel goed in staat om woordfrequenties te berekenen, of om gegevens te generen over de grammaticale categorieën van deze woorden. Momenteel is voor computeralgoritmes echter nog lastig om data te produceren over de diepere betekenis van teksten. *Semantic Taggers* of tools die gebruik maken van Topic Modelling algoritmes

---

[1] Jurafsky, Daniel, and James H. Martin. *Speech and Language Processing*. Englewood Cliffs: Prentice Hall, 2008.

[2] Moretti, Franco. *Distant Reading*. London: Verso, 2013.

[3] Jockers, Matthew. *Macroanalysis : Digital Methods and Literary History*. Urbana: University of Illinois Press, 2013.

proberen wel een brug te slaan tussen de woorden die gebruikt zijn enerzijds en de betekenis van deze woorden anderzijds.[4] Een probleem bij deze huidige benaderingen is echter dat deze gebaseerd zijn op de aanname dat woorden een stabiele betekenis hebben. Deze aanname is uiteraard niet correct voor literaire teksten, aangezien de woorden in deze teksten vaak op creatieve en onvoorspelbare manieren zijn gebruikt. Zowel de spelling van woorden als de betekenissen die aan woorden zijn toegekend hebben in de loop van de geschiedenis vaak grote wijzigingen ondergaan, en voor computersoftware is het vaak nog lastig om op een goede manier met dit soort diachrone verschillen om te gaan. Ook voor de literatuurwetenschap bestaan er momenteel nog duidelijke beperkingen, omdat er geen bruikbare applicaties beschikbaar zijn voor de geautomatiseerde herkenning van stijlfiguren en van andere literaire technieken. Critici baseren hun analyses vaak op voorbeelden van alliteratie, rijm, metaforen, personificaties of connotaties, maar de herkenning van dit soort aspecten ligt nog steeds buiten het bereik van de huidige tools. Hetzelfde geldt voor de geautomatiseerde herkenning van fenomenen zoals ironie, sarcasme, zinspelingen of humor.

Op de tweede plaats is het ook van belang om vast te stellen dat *text mining* een non-responsieve en een context-onafhankelijke manier van analyseren inhoudt. Wanneer mensen teksten lezen heeft hun kennis van de sociale of historische context meestal een invloed op hoe specifieke woorden worden geïnterpreteerd en beoordeeld. Woorden en stijlfiguren worden bovendien gelezen in de context van een groter geheel, zoals een zin, een alinea of een paragraaf. Computer-gebaseerde tekstanalyses beginnen meestal met een proces waarmee lineaire teksten worden omgezet naar discrete data. De manier waarop deze data worden geproduceerd is op bepaalde formules gebaseerd, en de werking van deze formules staat vaak volledig los van het jaar waarin de tekst is geschreven, of van het geslacht of de sociale status van een auteur. De regels die zijn vastgelegd in een algoritme worden momenteel met een onbuigzame consistentie toegepast op alle teksten in een corpus, en wanneer die getallen eenmaal zijn geproduceerd komen ze los te staan van de oorspronkelijke fragmenten waar deze getallen betrekking op hebben. Over de algoritmes die woordfrequenties berekenen, bijvoorbeeld, wordt vaak aangegeven dat deze werken met het zogenaamde “bag of words” model. Een tekst wordt simpelweg gezien als een verzameling van losse woorden, en hierbij wordt de context waarin deze woorden zijn gebruikt volledig genegeerd.[5]

Het belangrijkste verschil tussen het menselijke lezen en het machinale lezen is waarschijnlijk dat computers zelf geen interpretatie kunnen toevoegen aan tekstanalyses. Digitale onderzoeksinstrumenten kunnen de gekwantificeerde aspecten van teksten op zeer geavanceerde manieren analyseren, via een veelheid aan complexe statistische procedures. Dit soort analyses kunnen bepaalde regelmatigheden, correlaties of uitzondelijkheden blootleggen. Ze bestaan uit logische en geformaliseerde gevolgtrekkingen, op basis van statistische en probabilistische regels die voorafgaand door mensen zijn opgesteld. Interpretatie houdt echter onder meer in dat de patronen die worden herkend ook worden verklaard. Er moet worden uitgelegd waarom bepaalde patronen überhaupt relevant of

---

[4] Zie, bijvoorbeeld, Blei, David M., Andrew Y. Ng, and Michael I. Jordan. “Latent Dirichlet Allocation”, in: *The Journal of Machine Learning Research* 3 (2003), p. 993–1022, en Paul Rayson. “From key words to key semantic domains”, in: *International Journal of Corpus Linguistics* 13:4 (2008), p. 519-549.

[5] Bilisoly, Roger. *Practical Text Mining with Perl*. Hoboken, N.J.: Wiley, 2008, p. 123.

verassend zijn. Hans-Georg Gadamer benadrukte verder dat literaire interpretatie zich primair richt op het vaststellen van de betekenis van een tekst. Deze betekenis staat niet vooral vast, en komt pas tot stand tijdens de interactie tussen de lezer en de tekst. De intenties, de verwachtingen en de kennis van de individuele lezer vormen een integraal onderdeel van de betekenis die aan een tekst wordt toegekend.[6] Een dergelijke literaire interpretatie, die een bepaalde intentionaliteit veronderstelt, kan uiteindelijk alleen door menselijke lezers worden uitgevoerd. Computer-gebaseerde en kwantitatieve analyses van teksten resulteren op de eerste plaats in nieuwe statistische artefacten, die, net als de oorspronkelijke teksten waar ze op zijn gebaseerd, moeten worden geïnterpreteerd.

Het is echter misleidend om te stellen dat de resultaten van computationele analyses volledig descriptief zijn. Veel van de gegevens die door computers worden geproduceerd zijn wel degelijk interpretatief van aard. Digitale onderzoeksinstrumenten zijn ontwikkeld door menselijk programmeurs, en deze programmeurs moeten vaak verschillende beslissingen nemen over de manier waarop een tool functioneert. Data moeten vaak worden opgeschoond, en bij het toepassen van algoritmes kunnen er vaak verschillende parameters worden toegepast. Hierbij spelen specifieke ideeën over de te volgen methodologie vaak een grote rol. Hoewel Franco Moretti beweerde dat het gebruik van computers ook een overgang impliceerde naar een meer wetenschappelijke en een meer objectieve benadering, is het duidelijk dat de resultaten van computer-gebaseerde onderzoeken, dankzij de vele subjectieve keuzes die genomen moeten worden, nog steeds kunnen worden betwist in onderzoek dat gebaseerd is op conventionele methoden.

Er kan, kortom, worden vastgesteld, dat computers lijden aan een vorm van dyslexie. Bepaalde aspecten van de tekst blijven bijna onvermijdelijk buiten het gezichtsveld van digitale tools, en dit soort omissies hebben ook negatieve gevolgen voor het tekstbegrip. Ondanks deze beperkingen kunnen computers ook helpen bij het doen van andersoortige observaties over teksten, onder meer door de consistentie waar ze mee werken, en door de schaalvergroting die ze mogelijk maken. Computationele methoden kunnen helpen bij het beschrijven van de diversiteit van de woordenschat van een auteur, op basis van een berekening van de *type-token ratio*, de verhouding tussen het aantal unieke woorden en het totaal aantal woorden. *Text mining* technieken kunnen helpen bij het bepalen van de complexiteit van een tekst, op basis van het gemiddeld aantal lettergrepen per woord, of het gemiddeld aantal woorden per zin. Computers kunnen patronen blootleggen in het gebruik van woordsoorten die heel frequent voorkomen in een tekst, zoals lidwoorden, voorzetsels of persoonlijke voornaamwoorden.

Al deze data kunnen effectief worden ingezet bij het vaststellen van verschillen en overeenkomsten tussen teksten en bij het contextualiseren van teksten. Hoewel digitale methoden de sociale en historische context van teksten vaak negeren, kunnen deze kwantitatieve analyses meestal wel inzichten opleveren over wat een individuele tekst onderscheidend of juist conventioneel maakt, binnen de context van het geanalyseerde corpus. De resultaten van deze stilometrische analyses kunnen vervolgens in verband worden gebracht met de inhoudelijke aspecten van de tekst. Wanneer duidelijk wordt dat twee teksten veel vormelijke kenmerken delen, kan vervolgens in een meer gerichte lezing worden verkend of die teksten ook thematische overeenkomsten vertonen. Er kan ook worden onderzocht of alle teksten die regelmatig zijn aangeprezen door critici ook specifieke vormelijke kenmerken delen. Statistische analyses kunnen, deels door de

[6] Gadamer, Hans-Georg. *Truth and Method*. New York: Seabury Press, 1975, p. 398.

fundamenteel andere manier waarop teksten worden benaderd, een nieuwe impuls geven aan het bestaande literatuuronderzoek.

Close reading en het computationele lezen vormen twee verschillende en complementaire methoden voor het produceren van kennis over teksten. De resultaten van deze twee methoden kunnen elkaar ondersteunen, maar ze kunnen elkaar ook tegenspreken. Omdat computers uiteindelijk alleen statistische formules toepassen op tekstuele data is er uiteindelijk altijd nog een menselijke lezer nodig om beslissingen te nemen over de juistheid van bepaalde bevindingen. De validiteit van een argumentatie of interpretatie kan momenteel nog niet worden berekend. In tegenstelling tot computers zijn menselijke lezers in staat om kritisch te reflecteren over resultaten, en hierdoor lijkt de computationele benadering nog steeds ondergeschikt aan de meer traditionele benaderingen. Hierbij kan echter wel de vraag worden gesteld of deze afhankelijkheidsrelatie ook in de toekomst gehandhaafd kan blijven. Dit artikel is ingegaan op de huidige tekortkomingen van het computationele lezen. De technologie ontwikkelt zich echter onverminderd door, en als gevolg van voortdurende technologische innovaties kunnen er meer en meer aspecten van het menselijke lezen worden nagebootst. Beslissingen die voorheen strikt menselijk leken kunnen in toenemende mate worden gedigitaliseerd. Op het moment lijkt de technologie nog gedienstig aan de menselijke onderzoeker, maar het streven om een superieure positie te behouden vormt wellicht een van de belangrijkste uitdagingen voor de *digital humanities* in de komende decennia.