

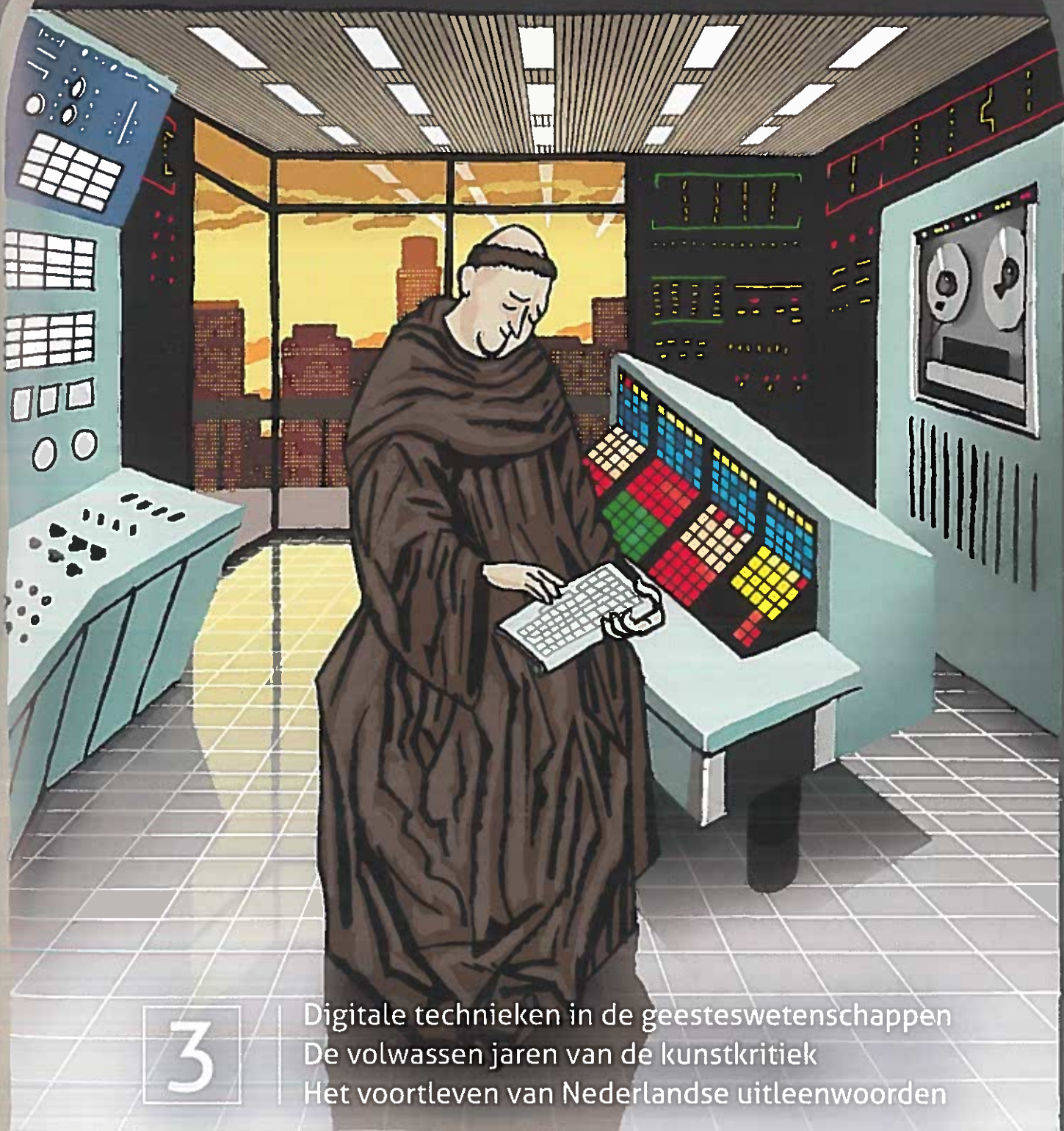
Neerlandia

Jaargang 120 / 2016

Nederlands-Vlaams tijdschrift
voor taal, cultuur en maatschappij

UBU Binnenstad
GW-TS

2^e SEP. 2016



3

Digitale technieken in de geesteswetenschappen
De volwassen jaren van de kunstkritiek
Het voortleven van Nederlandse uitleenwoorden

WETENSCHAPPEN

Beta humanities

Mike
Kestemont

De opmars van digitale technieken in de geesteswetenschappen

Alfa en bèta

De moderne wetenschap wordt doorgaans opgesplitst in twee grote kennisdomeinen, waarbij de alfa- of geesteswetenschappen in een sterke oppositie worden neergezet met de 'exacte' bètawetenschappen, zoals de fysica of chemie. De geesteswetenschappen of humanities omvatten traditioneel disciplines zoals de musicologie, de taal- en letterkunde en de geschiedenis. Deze wetenschapstakken hebben met elkaar gemeen dat zij de 'voortbrengselen' van de menselijke geest bestuderen. Dat is althans de brede definitie die Rens Bod hanteert in *De vergeten wetenschappen*, het fraaie overzichtswerk dat hij recent publiceerde over de geschiedenis van de humanities.

Wijnen Steve Jobs, de visionaire ondernemer achter *Apple* en *Pixar*, betwijfelde of het wel opportuun was om de werelden van alfa- en bètawetenschappen zo hermetisch van elkaar te scheiden. Hij liet zich geregeld ontvallen dat het niet louter de technologische vooruitgang was die zijn bedrijven groot heeft gemaakt: in Jobs' ogen kon slechts het 'huwelijk' van de geesteswetenschappen en computertechnologie de gewenste basis bieden voor succesvolle ondernemingen in de 21e eeuw. Inmiddels zoeken de alfa- en bètawetenschappen inderdaad volop toenadering: in de internationale beweging van de *digital humanities* verkent men enthousiast hoe moderne computertechnieken kunnen worden ingezet om het onderwijs en onderzoek uit de traditionele geesteswetenschappen te ondersteunen en te verbeteren met digitale hulpmiddelen. De *e-humanities* als *enhanced humanities*, zegt men wel eens.

De voortschrijdende digitalisering van bronnen in de geesteswetenschappen zorgt niettemin ook voor de nodige grieptijnen. Zo vormen intellectuele eigendomsrechten nog geregeld een struikelblok bij het digitaliseren en verspreiden van werken.

Het data-delugium

Enigszins verrassend is dat een vooruitstrevende Italiaanse jezuïet, Roberto Busa S.J. (1913-2011), vandaag de dag als geestelijke vader van de digital humanities mag gelden. In samenwerking met de Amerikaanse computer-gigant IBM (*the Big Blue*), heeft priester Busa in de vorige eeuw een uniek digitaliseringsproject opgezet dat zeker naar de toenmalige normen ongezien ambitieus was. Busa, die onderzoek deed naar het woordgebruik van de

middeleeuwse filosoof Thomas van Aquino, heeft IBM zo ver gekregen om gedurende drie decennia een groot team van Italiaanse typistes te financieren om de volledige werken van Aquino – 56 volumes in gedrukte vorm – geduldig op ponskaarten over te zetten om ze beter doorzoekbaar te maken. Busa beschouwde de computer als een geschenk van God en het wekt dan ook geen verbazing dat de jaarlijkse 'Nobelprijs' in de digital humanities naar de geliefde Italiaanse priester werd vernoemd.

Door de onstuitbare opkomst van het internet en de persoonlijke computer beschikken onderzoekers de laatste jaren over nog veel grotere hoeveelheden gegevens dan vroeger. Nieuwe boeken worden in ijntempo in nieuwe digitale formaten als ePub gepubliceerd, maar ook oudere werken worden op grote schaal opnieuw digitaal beschikbaar- en doorzoekbaar gemaakt via initiatieven als *Google Books*. Dichter bij huis is de Digitale Bibliotheek voor de Nederlandse Letteren (*dbnl.org*) een fantastisch online initiatief, dat een indrukwekkende digitale doorsnede biedt van de Nederlandstalige literatuurproductie. Dergelijke projecten bieden inmiddels voor veel lezers een volwaardig virtueel alternatief voor de fysieke dorpsbibliotheek of boekhandel.

De voortschrijdende digitalisering van bronnen in de geesteswetenschappen zorgt niettemin ook voor de nodige grieptijnen. Zo vormen intellectuele eigendomsrechten nog geregeld een struikelblok bij het digitaliseren en verspreiden van werken. Verschillende bibliotheken en musea gooien liever geen digitale versies van hun topstukken online te grabbel en ook veel wetenschappelijke uitgeverij zijn er niet happig op om hun uitgaves vrij toegankelijk op het web te publiceren. Dergelijke conservatieve krachten in het veld oogsten dan ook vaak onbegrip in de *openaccess* beweging. Die laatste stelt dat wetenschappelijke gegevens en



ook belangrijk cultureel erfgoed in alle openheid gedeeld zouden moeten kunnen worden, zeker in de geesteswetenschappen, waar onderzoek in de regel met geld van de gemeenschap wordt gefinancierd.

De Lage Landen

De Lage Landen doen het voor een relatief klein cultuur- en taalgebied lang niet slecht in de digital humanities. Instellingen zoals het Huygens Instituut (Den Haag) of de Koninklijke Academie voor Nederlandse Taal- en Letterkunde (Gent) zijn al sinds jaar en dag toonaangevende spelers in het internationale onderzoeklandschap. Het Huygens Instituut heeft indertijd bijvoorbeeld furore gemaakt met de prachtige online editie van de briefwisseling van Vincent van Gogh (*vangoghletters.org*).

Als editieproject is de Van Goghbriefwisseling een huzarenstuk, want anders dan in traditionele gedrukte leesuitgave kan de lezer spontaan wisselen tussen verschillende informatielagen. Zo kan men moeiteloos overstappen van een fotografische facsimilerpresentatie van een brief, naar een nauwkeurige transcriptie of zelfs de Engelstalige vertaling. Alle namen in de brieven zijn bovendien rijkelijk geannoteerd. Het dynamisch karakter van een dergelijk editieproject is natuurlijk cruciaal. Experts benadrukken namelijk dat een editie pas

echt volwaardig digitaal is als die niet meer gedrukt kan worden zonder dat een substantieel deel van de leesfunctionaliteit verloren gaat.

Het Van Goghproject is een voortreffelijk voorbeeld van een modern editieproject in de digitale geesteswetenschappen, maar staat zeker niet alleen in de Lage Landen. Het Instituut voor Nederlandse Lexicologie werpt bijvoorbeeld hoge ogen met de grote verzameling historische woordenboeken die het online gratis beschikbaar stelt (*gtb.inl.nl*). De diachrone dekingsgraad van deze kwaliteitsvolle lexica is internationaal ongezien – zeker voor een relatief kleine taal als het Nederlands. Met dergelijke hulpmiddelen kunnen bijvoorbeeld scholieren nu moeiteloos de betekenis



Roberto Busa S.J. (1913-2011) wordt beschouwd als de grondlegger van de digital humanities, vanwege zijn pionierswerk inzake de digitalisering van de Index Thomisticus in samenwerking met IBM. Op deze foto is een grote groep typistes voor Busa aan het werk in Italië.

Schermafdrak van de online editie van de briefwisseling van Vincent van Gogh op vangoghletters.org, een geslaagd voorbeeld van een dynamische teksteditie in de digital humanities

Galbraith ontmaskerd

In veel gevallen leven de digital humanities nog op gespannen voet met de traditionele geesteswetenschappers. Critici vragen zich af of computers wel klaar zijn om de kritische blik, laat staan het gezonde verstand van menselijke lezers al geheel te vervangen in het geesteswetenschappelijke onderzoek. Die argwaan is niet alleen gezond maar ook terecht: hoewel de artificiële intelligentie de laatste jaren met rasse schreden vooruitgang heeft geboekt, is er nog een lange weg te gaan vooraleer we in software het menselijk taalbegrip of leesvermogen volledig zullen kunnen reproduceren. Toch zijn er al heel wat taken waar de computer de menselijke onderzoeker kan ondersteunen – en zelfs overtreffen.

Een erg succesvolle applicatie van digital humanities is auteursherkenning in de stylometrie, een tak van de computationele geesteswetenschappen waarin men anonieme teksten op basis van unieke kenmerken in de schrijfstijl van de documenten aan bekende auteurs probeert toe te schrijven. Stylometrici hechten doorgaans veel belang aan het gebruik van functiewoorden: de kleine groep woorden in een taal die weinig betekenis in zich dragen en alleen een grammaticale rol in een taal spelen, zoals lidwoorden of voorzetsels. Dergelijke woorden zijn zo frequent dat auteurs ze onbewust gebruiken, waardoor ze moeilijk geïmiteerd kunnen worden. Bovendien beïnvloedt het onderwerp van een tekst nauwelijks de frequentie van bepaalde functiewoorden, zodat men vaak gelooft dat functiewoorden interessant zijn voor het herkennen van auteurs over de grenzen van verschillende tekstsoorten en onderwerpen heen.

Enkele jaren geleden nog boekte de stylometrie een belangwekkend resultaat aan de hand van computationele technieken voor auteursherkenning. De onbekende auteur Robert Galbraith publiceerde in 2013 zijn niet onverdiensteijk debuutroman *The Cuckoo's Calling* (Koekoeksjong). De Sunday Times kreeg na enkele maanden de geheimzinnige tip binnen dat Galbraith het pseudoniem zou zijn van J.K. Rowling, de befaamde romanschrijfster achter het grote Harry Pottersucces. Patrick Juola, een van de boegbeelden van de hedendaagse stylometrie, werd vervolgens ingeschakeld door de Sunday Times om te onderzoeken of *The Cuckoo's Calling* stilistisch inderdaad

of het uitzicht van een woord traceren van de Oud-nederlandse periode tot nu. Dat benadrukt hoe de digital humanities ook een belangrijke maatschappelijke meerwaarde hebben. In tegenstelling tot de dure, gespecialiseerde standaardwerken die men vroeger moest aanschaffen, hebben leerlingen nu gewoon online toegang tot kwalitatief hoogstaande leermiddelen. Dat komt de democratisering van het onderwijs natuurlijk ten goede.

Panoramisch lezen

De duizelingwekkende hoeveelheid aan digitale gegevens waarover onderzoekers beschikken, zet klassieke analysemodellen momenteel flink onder druk in de geesteswetenschappen, zoals de traditionele hermeneutiek of interpretatieleer. Vroeger was het niet ongebruikelijk dat een letterkundige zich maandenlang met een handvol boeken terugtrok voor een tijdrovende, nauwgezette lectuur van deze werken. Een dergelijk kleinschalige close reading is volgens velen niet meer van deze tijd. In *Macroanalysis: Digital Methods and Literary History*, een frisse monografie over de opkomst van digitale methodes in de literatuurstudie, stelde Matt Jockers nogal polemisch dat onze literaire *big data* de dag van vandaag close reading niet alleen onpraktisch maar zelfs ontoereikend maakt in de letterkunde. Want wie krijgt het allemaal nog gelezen?

Tegenover deze *close reading* plaatsen onderzoekers in de digital humanities nu steeds vaker *distant reading*, een vorm van panoramisch lezen waarbij literaire teksten op grote schaal met computers worden geanalyseerd. Op die manier kunnen veel meer teksten in het onderzoek worden betrokken en hoeven onderzoekers hun werkzaamheden niet op voorhand te beperken tot een klein, willekeurig geselecteerd groepje canonteksten waarvan zij hopen dat die interessant zullen blijken te zijn. Volgens Jockers valt *distant reading* als een soort macroletterkunde nog het best te vergelijken met macro-economie. In plaats van het consumeergedrag van kleine groepjes mensen (micro-economie) te modelleren, tracht *distant reading* grove regelmatigheden te ontdekken in het literaire systeem op een veel grotere schaal. In de taalkunde is een dergelijk grootschalig onderzoek, op basis van enorme databanken, overigens al veel langer gebruikelijk en worden woordstatistieken op basis van tekstcorpora met triljoenen woorden berekend.

toegeschreven kon worden aan de befaamde Harry Potterauteur. Juola slaagde er opvallend genoeg in de nodige stylometrische argumenten te verzamelen voor die onverwachte attributie, zodat Rowling al snel moest toegeven dat het om een pseudoniem ging.

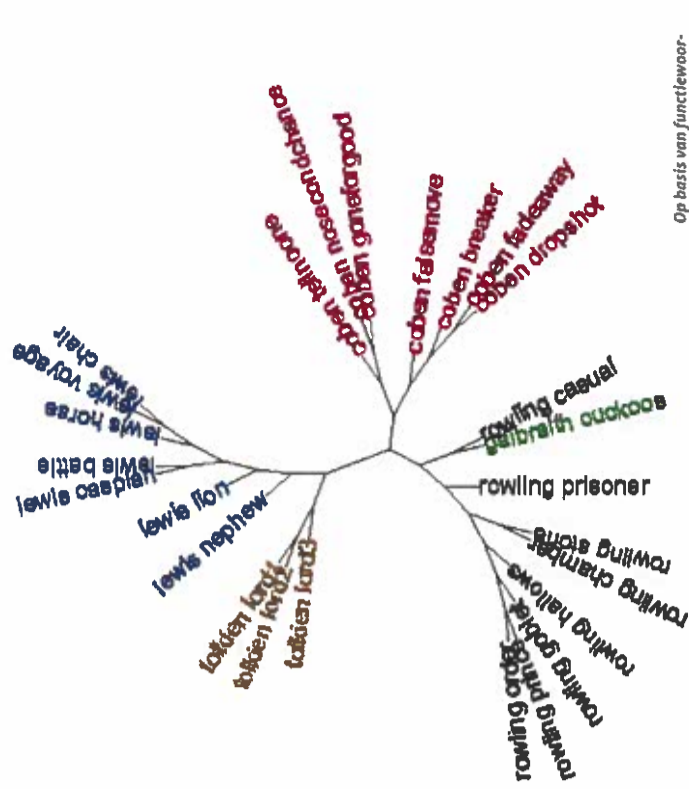
Dergelijke technieken worden inmiddels ook gretig toegepast op de Nederlandstalige literatuur. Voor de middeleeuwse periode is er bijvoorbeeld uitvoerig onderzoek geweest naar de bekende *Roman van Walewein*. Dat onderzoek heeft ertoe bijgedragen dat we nu de bijdrage beter kunnen ontwarren van Penninc en Vostaert, de twee auteurs die in het slot van de tekst vermeld worden. Bovendien is op basis van auteursherkenningstechnieken gebleken dat de auteur van de iconische ridderroman *Karel ende Elegast* waarschijnlijk nog twee andere Middelenlandse romans heeft geschreven. Het spannende aan dergelijk onderzoek is dat men dit eeuwenoude literaire erfgoed nieuw leven in kan blazen door er de nieuwste analysetechnieken op toe te passen.

Een mooie toekomst

In de Lage Landen mogen we best trots zijn op enkele mooie verwezenlijkingen in de digital humanities. Momenteel loopt zelfs een enorm onderzoeksproject in Nederland (*nederland.nl*), waarin men al het reeds beschikbaar gemaakte materiaal en tools voor digitaal onderzoek in de geesteswetenschappen aan elkaar probeert te koppelen en aan te bieden vanachter één geïntegreerde gebruikersinterface. Dat inspirerende project zal gebruikers er zo toe in staat stellen met luttele muisklikken ontzaglijke aantallen teksten uit de cultuurgeschiedenis van de Nederlanden te analyseren. Dat betekent echter jammer genoeg niet dat er geen werk meer aan de winkel is.

Op onderwijsgebied ontwikkelt het hoger onderwijs in zowel Vlaanderen als Nederland zich nog te traag. Hoewel steeds meer universiteiten in de Benelux mooie masterprogramma's ontwikkelen op het gebied van digital humanities, worden studenten vooral in de eerste jaren van het hoger onderwijs nog te weinig bewust gemaakt van de (on)mogelijkheden van hedendaagse computertechnieken. Op die manier worden studenten dan ook nog onvoldoende voorbereid om een geloofwaardig antwoord te bieden op de digitale uitdagingen waarvoor onze snel

Galbraith Bootstrap Consensus Tree



Op basis van Junctiewoorden kunnen statistische vergelijkingsmethodes de schrijfstijl van verschillende auteurs onderscheiden. Dit boomdiagram vergelijkt het betwiste werk van 'Galbraith' (Cuckoo's) met werk van Rowling, Coben en het bekende paar fantasyauteurs Lewis en Tolkien. De infografiek toont duidelijk dat The Cuckoo's Calling (in het groen) relatief goed aansluit bij Rowlings schrijfstijl in de roman The Casual Vacancy uit 2012.

100-3000 MFW Cuffed @ 0% Classic Delta distance Consensus 0.5

veranderende maatschappij en arbeidsmarkt hen zal plaatsen. Ook in de geesteswetenschappen geldt immers het adagium: beter zelf programmeren dan zelf geprogrammeerd worden.

Toch ziet de toekomst er mooi uit voor de digitale geesteswetenschappen: in de lijn van Steve Jobs' reïdening is er eigenlijk zelden zo veel belangstelling geweest voor humanities, ook in exotische gebieden als de informatica. De computerlinguïstiek bijvoorbeeld – een segment van de arbeidsmarkt dat op sterven na dood leek na het debacle rond Lernout en Hauspie – heeft duidelijk een tweede adem gevonden en mag momenteel misschien wel als een van de meest opwindende wetenschapsgebieden tout court beschouwd worden. Zelfs de oplossing voor enkele bijzonder lastige taken, zoals machinevertaling of spraakherkenning, lijken sinds kort opnieuw wat dichterbij.

Mike Kestemont doceert digitale tekstanalyse aan het departement letterkunde van de Universiteit Antwerpen.

Contact: mike.kestemont@uantwerpen.be