Taal in Uitvoering

INAUGURELE REDE DOOR PROF. DR. ANTAL VAN DEN BOSCH

Radboud Universiteit Nijmegen

INAUGURELE REDE

PROF. DR. ANTAL VAN DEN BOSCH

De computationele linguïstiek heeft zich ontwikkeld tot een extreem datagedreven wetenschapsgebied met een uiterst dwarse blik op taal. Een systeem als *Google Translate* vertaalt intussen meer dan zestig talen zonder enige taalkundige kennis. Wie onder de motorkap kijkt ziet een agnostisch systeem dat, ongeacht het taalpaar, steeds dezelfde soort analogieën maakt: 'deze zin lijkt op die zinnen in mijn enorme geheugen, dus bouw ik de vertaling op uit de vertalingen van die zinnen.' De opkomst van deze methode valt samen met de groeiende stortvloed aan beschikbare digitale informatie op internet. Onderzoekers zwemmen in de gegevens, en bouwen systemen die steeds beter teksten vertalen, corrigeren en samenvatten, maar ook teksten koppelen aan de echte wereld: tweets aan gebeurtenissen, teksten aan sociale netwerken. Is er buiten dit technologisch positivisme ruimte voor reflectie? Wat heeft de taalkunde aan deze 'kijk ma, zonder handen'-bravoure? Antal van den Bosch beargumenteert in zijn oratie dat de taalkunde juist veel kan leren van de interne werking van deze systemen.

Prof. dr. Antal van den Bosch (Made, 1969) is hoogleraar Example-based language modelling aan de Radboud Universiteit Nijmegen. Hij studeerde computerlinguïstiek (1992, Tilburg), promoveerde in de informatica (1997, Maastricht) en bouwde vervolgens in Tilburg met Walter Daelemans (Antwerpen) aan een methode voor automatische geheugengebaseerde taalverwerking. Sinds 2011 is Van den Bosch hoogleraar aan de Radboud Universiteit, waar hij werkt aan methoden voor automatisch vertalen, spellingcorrectie, en aan text analytics-methoden voor het extraheren van informatie en kennis uit grootschalige hoeveelheden teksten. Van den Bosch is lid van de Koninklijke Academie voor de Wetenschappen (KNAW).

Radboud Universiteit Nijmegen

TAAL IN UITVOERING

Deze oratie is opgedragen aan mijn vader, Jan van den Bosch.

'Die Wörter und Wortgruppen, die wir in der Rede verwenden, erzeugen sich nur zum Teil durch blosse gedächtnismässige Reproduktion des früher Aufgenommenen. Ungefähr eben so viel Anteil daran hat eine kombinatorische Tätigkeit, welche auf der Existenz der Proportionengruppen basiert ist. Die Kombination besteht dabei gewissermassen in der Auflösung einer Proportionengleichung, indem nach dem Muster von schon geläufig gewordenen analogen Proportionen zu einem gleichfalls geläufigen Worte ein zweites Proportionsglied frei geschaffen wird. Diesen Vorgang nennen wir Analogiebildung.' (Paul 1920: 110)

Taal in Uitvoering

Rede uitgesproken bij de aanvaarding van het ambt van hoogleraar Example-based language modelling aan de Faculteit der Letteren van de Radboud Universiteit Nijmegen op vrijdag 9 november 2012

door Antal van den Bosch

Vormgeving en opmaak: Nies en Partners bno, Nijmegen
Fotografie omslag: Bert Beelen
Drukwerk: Van Eck & Oosterink

ISBN -13 : 978 -90 -9027 -167 -5

Mijnheer de rector magnificus, geachte aanwezigen,
U
weet
wat
ik
ga
zeggen.

Toen ik bij *ga* was, vertelde de verwachtingenmachine in uw hoofd dat *zeggen* een hele goede kans maakte; *doen* was ook mogelijk. We kunnen er niets aan doen: als we luisteren, dan voorspellen we voortdurend wat volgt. Van die onstopbare machine is een eenvoudig computermodel te maken. Het model weet welke woorden hoe vaak kunnen voorkomen na welke een, twee, of drie woorden ervoor, omdat het heel veel teksten heeft gezien (in dit geval vier miljard woorden aan Nederlandse tekst) en kan op basis daarop verwachtingen uitspreken. Het werkt als volgt.

Na het zien van het eerste woord *U* kunnen enkele duizenden woorden volgen. *Weet* staat op nummer 6 van de ranglijst van meest voorkomende volgende woorden. Zodra we *weet* horen, kunnen de verwachtingen worden aangescherpt. De computer kent maar 110 mogelijke vervolgen; *U weet wel* en *U weet dat* zijn de nummers 1 en 2; *U weet wat* staat op nummer 8. Dan wordt het eenvoudiger: de optie *ik* is nummer 6 van de 12 volgende mogelijke woorden. En dan openen de mogelijkheden zich weer enigszins; tussen enkele tientallen mogelijkheden verwacht de computer het woord *wil* en *bedoel* als de meest waarschijnlijke vervolgen; ga staat op nummer 9. Tenslotte is *zeggen* het tweede meest waarschijnlijke woord, na *doen*.

Van dit korte zinnetje werd ieder woord door het model voorzien, hoewel steeds niet als het meest waarschijnlijke vervolg. Het model ziet dus in grote lijnen aankomen wat er gaat gebeuren. Is dat handig? Ja. Ik geef drie voorbeelden. Stel, u zit in deze zaal en wil snel iets twitteren. Het model dat ik net illustreerde zit in uw telefoon, en kijkt mee met wat u aan het typen bent. Het model probeert met u mee te voorspellen. U typt *Van den Bosch verkoopt grote. Aantallen? Delen?* Zodra u de eerstvolgende letter

Figuur 1. Verwachtingen van een statistisch taalmodel over het volgende woord na ieder van de woorden *U*, *weet*, *wat*, *ik*, en *ga*. De relatieve grootte van een woord weerspiegelt de waarschijnlijkheid die het model aan dat woord hecht.

intypt vervallen alle mogelijkheden die met andere letters beginnen, en weet het model met redelijke zekerheid welk woord u aan het typen bent (*onzin*). Als u de goede suggesties volgt kunt u, zoals we in experimenten hebben aangetoond (Van den Bosch, 2011), 40 procent of meer van uw toetsaanslagen overslaan. Dat ons model nog niet in iedere telefoon of tekstverwerker is ingebouwd ligt aan zijn relatieve logheid; het model houdt nu eenmaal een enorme berg gegevens bij. Het is niet meer maar ook niet minder dan een hele grote inhoudsopgave van het Nederlands.

Een tweede voorbeeld: spellingcorrectie. Een model dat in grote lijnen weet wat er verwacht kan worden, kan wellicht ook zien dat de schrijver het ene woord opschreef, terwijl hij een ander word bedoelde. Het foute woord lijkt op het bedoelde woord, en blijkbaar heeft de schrijver even niet opgelet. Neem de Troonrede van 1963[1]. Troonredes zijn toonbeelden van correct taalgebruik, maar toch viel het niemand op (of misschien viel het koningin Juliana op, bij het voorlezen) dat er een prachtige fout in de tekst staat: *wetenschapelijk*, met één p. De volledige zin luidt *Om deze redenen zal de Regering de uitbreiding van de instellingen voor wetenschapelijk onderwijs en onderzoek krachtig bevorderen*. Behalve de spelfout een prachtige boodschap. Als een tekst al helemaal af is, zoals deze, kan ons model naast de linkerwoorden ook de rechterwoorden meenemen. Het model doorloopt de tekst woord voor woord, dekt ieder word denkbeeldig af, en haalt een verwachtingenlijstje uit zijn geheugen op, op basis van de context de *instellingen voor* aan de linkerkant, en *onderwijs en onderzoek* aan de rechterkant.

Figuur 2: Moment tijdens het voorlezen van de troonrede door H.M. Koningin Juliana op 17 september 1963. Van links naar rechts H.K.H. Prinses Irene, Z.K.H. Prins Bernhard, H.M. de Koningin, H.K.H. Prinses Beatrix en H.K.H. Prinses Margriet. *Historisch Archief ANP Foundation.*

Binnen enkele milliseconden vindt het model zeven mogelijke kandidaten voor de positie waar nu wetenschapelijk staat: de functiewoorden *van* en *het*, en de inhoudswoorden *voortgezet*, *theoretisch*, *hoger*, *universitair*, en *wetenschappelijk*, met twee p's. Dit laatste woord lijkt verdacht veel op *wetenschapelijk*. Omdat het tussen de verwachte woorden staat, kunnen we veronderstellen dat de schrijver waarschijnlijk het woord met twee p's bedoelde. Op dezelfde manier detecteert en corrigeert deze methode ook verwarringen tussen bestaande woorden, zoals *word* en *wordt* of *zei* en *zij*. We hebben deze methode ingebouwd in *Valkuil.net*[2], een automatische spellingcorrector voor het Nederlands die vrij beschikbaar is op internet. Inmiddels heeft onze Nederlandse spellingcorrector *Valkuil.net* een Engels zusje gekregen: *Fowlt.net*[3].

... de instellingen voor **wetenschapelijk** onderwijs en onderzoek ...

... de **salarissen** voor	het	onderwijs en **overheidspersoneel** ...
... het **bijzonder** voor	het	onderwijs en **de** ...
... **meer aandacht** voor	technisch	onderwijs en **verbetering** ...
... **van scholen** voor	speciaal	onderwijs en **de** ...
... **zevenhonderd scholen** voor	voortgezet	onderwijs en **beroepsonderwijs** ...
... **het huisvestingsbudget** voor	voortgezet	onderwijs en **beroepsonderwijs** ...
... de **cao** voor	het	onderwijs.
... de **bres** voor	het	onderwijs.
... de **deelnemers** voor	theoretisch	onderwijs **een of** ...
... de **zorg** voor	het	onderwijs **gekenmerkt door** ...
... de **kosten** voor	het	onderwijs **e.d. omlaag** ...
... **dat** instellingen voor	hoger	**beroepsonderwijs** en **universiteiten** .
... **kwaliteit van het**	universitaire	onderwijs en onderzoek ...
... **nieuwe centrum moet**	het	onderwijs en onderzoek ...
... **doen met wat**	wetenschappelijk	onderwijs en onderzoek ...
... **uitspraken over zaken**	van	onderwijs en onderzoek ...
... **op het terrein**	van	onderwijs en onderzoek ...

Figuur 3: De spelfout *wetenschapelijk* in zijn context uit de Troonrede van 1963, en zeventien vergelijkbare contexten met overlappende woorden. De correctie *wetenschappelijk* is een van de mogelijke woorden in de centrale positie.

COMPOSITIONALITEIT

Een derde toepassing van hetzelfde model is vertalen. Het model kijkt dan niet meer naar een missend of tijdelijk weggelaten woord om het te voorspellen, maar probeert voor ieder woord een ander woord te vinden dat hetzelfde betekent – in een andere taal. Het model blijft grotendeels hetzelfde: vind een variant van een woord in zijn lokale context. Als dit lukt, dan is de helft van het vertaalprobleem opgelost. De andere helft van het probleem is het vinden van een goede volgorde waarin de vertaalde woorden moeten staan. Voor dat laatste probleem kan weer het eerste model gebruikt worden dat verwachtingen kan uitspreken over woorden in context: dat model krijgt bijvoorbeeld

alle mogelijke volgordes te zien en mag dan die volgorde kiezen die als geheel het best aan zijn verwachtingen voldoet. Dit is een minimale beschrijving van een statistisch vertaalsysteem, zoals *Google Translate*[4].

Een belangrijk inzicht dat bijvoorbeeld in *Google Translate* is toegepast, is dat het onverstandig is om alleen woorden in hun context te vertalen. We weten van het werk van onderzoekers als Philipp Koehn (Koehn, Och, en Marcu, 2003) dat het veel beter is om bij deze vertaaltaak niet losse woorden te vertalen, maar ook langere reeksen van woorden; sterker nog: hoe langer, hoe beter. Als je model een lang stuk tekst succesvol kan vertalen in een lang stuk tekst in de andere taal, dan heeft hij twee vliegen in een klap geslagen: hij heeft de goede vertalingen van een aantal woorden tegelijk gevonden, en ze staan ook nog eens in de juiste volgorde.

Met welke frasen een statistisch vertaalsysteem werkt, hangt af van het vertaalde materiaal waarop het systeem wordt getraind. Bij iedere training wordt er een nieuw vertaalfrasenboek gemaakt dat naast algemene frasenparen als *in het algemeen – generally*, frasen bevat die enigszins of heel erg typisch zijn voor de teksten waarop getraind is. Een vertaalsysteem dat getraind is op teksten van het Europese Parlement, in het Nederlands en het Engels, heeft bijvoorbeeld opgeslagen dat beroep op vaak in vertaalde zinnen samen voorkomt met *call on*, *appeal to*, *call upon* en *urge*. Het is goed om dat te weten, want beroep kan ook nog vertalen naar *profession* en *occupation* – maar, en dat is hier essentieel, niet in de context van deze frase; niet met *op* erna.

Met frasen als *beroep op*, *een eigen huis* en *vrij verkeer* (*free movement*) zijn we op een niveau van woordgroepen die wel eens collocaties worden genoemd. Ze bestrijken een spectrum van volkomen versteende uitdrukkingen (*wis en waarachtig*) tot combinaties van woorden die gewoon bij voorkeur naast elkaar staan (*sterke koffie*). Wat ze allemaal gemeen hebben is dat ze als eenheid duidelijk refereren naar iets in de echte of denkbeeldige wereld, en die verwijzing is uniek voor deze combinatie van woorden. Een eigen huis is niet alleen maar een huis waarin je woont; je hebt dat huis gekocht, niet gehuurd. Met een technische term kun je zeggen dat deze frase een deels niet-compositionele betekenis draagt. De betekenis is niet puur een optelsom van de betekenis van zijn delen, zoals Frege het bedoelde met zijn compositionaliteitsprincipe, maar ook niet puur non-compositioneel; de betekenis van *een eigen huis* is deels te raden uit de betekenissen van zijn onderdelen. Alleen het aspect van 'gekocht, niet gehuurd' is niet te raden en is dus het niet-compositionele stuk van de betekenis.

In het frasenboek van een statistisch vertaalsysteem vind je juist deze frasen, precies omdat ze niet-compositionele stukken betekenis bevatten. Statistische vertaalsystemen bieden dus zomaar een antwoord (ik zeg niet het antwoord, maar een antwoord) op een klassieke en zwaar betwiste taalkundige vraag: in hoeverre taal compositioneel is. En dat terwijl ze daar niet eens voor gemaakt zijn, laat staan dat de oorspronkelijke ontwerpers van statistische vertaalsystemen in deze vraag geïnteresseerd waren; het ging hen om het maken van zo goed mogelijke systemen. Het gaat

die systemen, die autonoom werken en zelfstandig proberen zo dicht mogelijk bij de referentievertalingen te komen van hun oefenmateriaal, om het behouden van betekenis bij het vertalen – dat ze daarbij betekenisdragende bouwstenen vinden van een of meerdere woorden lang is omdat ze dat nodig hebben om de taak zo goed mogelijk te volbrengen.

vrij verkeer	free movement
uiterst moeilijk	extremely difficult
in uitvoering	in progress
in dienst van	in the service of
beseffen dat de	recognize that the
tragische gebeurtenissen	tragic events
aan de andere kant	on the other hand
van ons allemaal	of all
diegenen onder ons	those of us
juist op dit	precisely this
echt nodig	absolutely necessary
evenals voor de	as well as for the
dit heeft geleid	this led

Figuur 4. Voorbeelden van vaak aangetroffen paren van meerwoordsfrasen in een Nederlands-Engels parallel corpus.

DATA EN THEORIE

Methode X, aangewend om doel Y te bereiken, produceert een onbedoeld maar belangwekkend bijproduct Z. Dit type toevalstreffer, dit serendipiteitsmotief lijkt voorbehouden te zijn aan de legendes van de exacte wetenschappen: aspartaam was een kandidaat-medicijn voor maagzweren; teflon ontstond bij het brouwen van een koelvloeistof in een zilveren pot. Vergelijkbare verhalen liggen aan de basis van insuline, penicilline, gevulkaniseerd rubber, röntgenstraling en microgolven. Ik mik misschien hoog met mijn vergelijkingen, maar de analogie is er: een datagedreven methode die gebruik maakt van computers, genereert onbedoeld een antwoord op een taalkundige vraag die in het verleden vooral vanuit een andere hoek, theoretisch, werd benaderd.

En zo sluit ik aan bij de eeuwige dans tussen datagedreven en theoriegedreven methoden van onderzoek, de twee grote drijfveren achter de wetenschap die elkaars tegenpolen zijn en niet zonder elkaar kunnen. Data en theorie zijn de echte polen van de wetenschap, niet de overschatte termen alfa en bèta. Wie wil begrijpen dat dit zogenaamd diepgewortelde verschil eigenlijk gebaseerd is op een misverstand, kan ik het recent verschenen boek *De zin van de ommezijde* van taalkunstenaar Frank van Pamelen (2012) aanraden. In dit boek voor jong en oud beschrijft Van Pamelen de op wantrouwen gebaseerde strijd tussen alfa's en bèta's: de alfa's vinden dat ze worden weggecijferd door de bèta's, de bèta's waarschuwen de alfa's dat ze op hun tellen moeten passen.

Zonder al te veel te verklappen wordt duidelijk dat de twee meer gemeen hebben dan ze denken.

De komst van de computer als onderzoeksinstrument is in dit opzicht bijzonder boeiend: de computer werkt nivellerend. Wie ziet nog het verschil tussen het bureau van een alfa- of bètawetenschapper? Er staat in beide gevallen een computer op. Het apparaat heft langzaam maar zeker het resterende verschil tussen alfa en bèta op, terwijl de strijd tussen data en theorie er met de computer eindeloos veel slagveldruimte bij krijgt.

De datagedreven methode heeft aan de computer in ieder geval een goede kameraad. Een perfect geheugen, groot genoeg om meer tekst te bevatten dan een mens ooit in zijn of haar leven zal lezen, spreken, luisteren of schrijven, uitgerust met informatieverwerkende motortjes die miljarden elementaire manipulaties van informatie per seconde kunnen uitvoeren, bijvoorbeeld om dat enorme geheugen door te spitten op zoek naar vertalingen van stukken zin. Het simpele krachtige idee van de datagedreven methode heeft zich uitbetaald in de automatische spraakherkenning en in automatisch vertalen, en bijvoorbeeld ook in het zoeken op het web (het koppelen van korte stukjes tekst, *queries*, aan gerelateerde langere stukken tekst, webpagina's). Geen van deze problemen is volledig opgelost, verre van dat, maar de computationele datagedreven methoden voor spraak herkennen en vertalen hebben de stand van de techniek inmiddels een stuk verder gebracht dan hun computationele voorgangers.

IMPLICIT LINGUISTICS

Ik sloot in 2011 mijn *Vici-project Implicit Linguistics*[5] af. De naam verwees naar mijn onderzoeksagenda: op weg naar een nieuwe computationele taalwetenschap met een extreem datagedreven methode, die technologie voortbrengt waar, op een hele andere manier dan normaal, taalkundige inzichten aan ontleend kunnen worden. Die agenda wil ik onverkort voortzetten en ik wil daarbij samenwerking zoeken met natuurlijke partners in de brede taalwetenschappen die data als onlosmakelijk onderdeel zien van hun onderzoek. Het concept van de betekenisdragende meerwoordsfrasen weerklinkt in het concept van de constructie in *construction grammar*, die meerdere woorden kan omspannen en waarvan de betekenis niet volledig compositioneel is (Goldberg, 2006). Daarnaast is er recent psycholinguïstisch werk (Bannard en Matthews, 2008; Arnon en Snider, 2010) dat laat zien dat frequente meerwoordsfrasen net zulke reactietijden teweeg brengen bij proefpersonen als losse woorden; je zou dus kunnen veronderstellen dat ze net zo goed op een of andere manier opgeslagen lijken te zijn in ons geheugen. We lijken hier een brug te kunnen slaan tussen computationele modellen van taal en andere taalkundige gebieden.

De titel *Implicit Linguistics* suggereerde natuurlijk een contrast met *explicit linguistics*, het expliciet gebruik maken van linguïstische abstracties als woordsoorten, grammaticale relaties, of semantische rollen, en structuren om die abstracties in te

ordenen. Aan dit idee wordt nog altijd vastgehouden, vaak met de boodschap dat die expliciete modellering noodzakelijk is, een sine *qua non*. Jan Odijk schrijft in het pas verschenen rapport *Het Nederlands in het Digitale Tijdperk* (Odijk, 2012) wat hij de 'typische toepassingsarchitectuur voor tekstverwerking' noemt: een inputtekst doorloopt een proces van voorverwerking, grammaticale analyse, semantische analyse, en taakspecifieke modules. Een rode draad die door al mijn onderzoek heen loopt, vanaf de bescheiden eerste stappen twintig jaar geleden, is om te proberen die hele interne architectuur te versimpelen, en experimenteel te testen of expliciete abstracties helpen of niet bij het leren van natuurlijke taalverwerkingstaken. Toen ik in mijn promotieonderzoek computers trainde om nieuwe woorden uit te spreken op grond van verschillende theorieën, moest ik concluderen dat de computer met minder theorie tot betere resultaten kwam (Van den Bosch, 1997). Bij sommige taken, zoals ontleden, ontkom je er niet aan om gebruik te maken van linguïstische abstracties, al was het maar omdat het eindproduct, een ontleding, eruit bestaat. Tekst-naar-teksttaken zoals vertalen en spellingcorrectie zijn in hun invoer noch uitvoer afhankelijk van expliciete taalkundige concepten – tekst in, tekst uit – en zijn daarom een mooi testterrein voor *Implicit Linguistics*.

VAN TEKST NAAR VRAAG

Toch heeft tekst-naar-tekstverwerking ook zijn beperkingen. Je stopt er tekst in, en er komt tekst uit. Knap aan de buitenkant, maar het is natuurlijk het innerlijk dat telt. Teksten hebben iets te vertellen. Ze bevatten informatie en kennis; ze zijn door iemand geschreven die daarmee een zekere bedoeling had, ze zijn voor iemand geschreven. Schrijvers en lezers zijn personen in situaties die soms dicht bij elkaar liggen, soms ver van elkaar. Teksten reflecteren denkbeelden over de wereld, of denkbeeldige werelden, en doen dat in een oneindige variëteit. Geef tien mensen een tekst te vertalen van het Engels naar het Nederlands, en je krijgt tien vergelijkbare maar verschillende teksten met hier en daar dezelfde zinnen en constructies. Vraag dezelfde tien mensen een gebeurtenis te beschrijven, en je eindigt met tien volstrekt verschillende verhalen waarvan misschien niet één zin hetzelfde is. Ieder aspect van de gebeurtenis dat wordt opgenomen in de tekst, kan weer op duizend-en-een manieren in woorden worden gevat. Toch gaat het om dezelfde gebeurtenis, dezelfde kern. Taal is een ongekend veelzijdig basismateriaal. Taal, dat is de glitter in de caleidoscoop van de wereld.

Iedere tekst die je ziet is dus het resultaat van een grotendeels onbewust keuzeproces van de schrijver uit een hele grote berg van mogelijke andere teksten die net zo goed gegenereerd hadden kunnen worden. In de klassieke benadering van computationele analyse van teksten met expliciet taalkundige abstracties als ontledingen wordt iedere tekst gezien als *de* tekst, waarvan de betekenis compositioneel kan worden berekend. Terwijl het maar *een* tekst is, een toevallige tussenstap tussen schrijver en wereld. Nu kun je geïnteresseerd zijn in *de* tekst en zijn taalkundige eigenschappen, dat is een volkomen valide interesse, maar deze opvatting van een tekst als *de* tekst helpt je niet als het je te

doen is om vragen over de schrijver en de wereld. Wat hier nodig is, is een flexibelere opstelling, en hier zie ik een duidelijke rol weggelegd voor de impliciete methode. Laat ik mijn punt illustreren aan de hand van een aantal voorbeelden. Neem het volgende korte bericht, gepost op Twitter.com[6]:

> @Ramoen1 die kérol laat me niet zomaar de snelweg op euy.. ik rij redelijk goed, dus agge nog mee wil als ik rijbewijs heb; lief doen >.<

Waar woont deze twittergebruiker? In de gebruikersgegevens van deze waarschijnlijke mannelijke twitteraar zegt hij dat hij uit Hoogerheide komt en een 'Trotse Brabander' is. De Brabanders onder ons zijn waarschijnlijk niet verbaasd. Hoe weten we dat? Het ene woord *agge* was genoeg om de gok te wagen. Is het mogelijk om op basis van teksten als tweets eigenschappen van de spreker te bepalen, zoals herkomst, geslacht en de leeftijd van de schrijver te bepalen? Dit soort vragen zijn relevant voor sociolinguïsten, communicatiewetenschappers, en rechercheurs van politie. We kunnen de computer miljoenen tweets geven met de herkomst van de twitteraar, en de computer de opdracht geven: zoek eens uit welke aanwijzingen je nodig hebt om die herkomst zo accuraat mogelijk te voorspellen. Een woord als *agge* gaat hoog scoren als een indicator voor Brabants.

Een tweede voorbeeld: een tekst uit een veldboek van *Naturalis*, het Nationaal Natuurhistorisch Museum in Leiden. De tekst staat in een logboek opgetekend tijdens een expeditie in 1968 in Suriname, en beschrijft de vondst van een kikker:

> Lithodytes lineatus, Brownsberg, aan voet, onder stuk rot hout, 13.07.1968, 8.45 u., RMNH 26076 Dorsolateraal strepen heldergeel, tekening op dijen vuurrood, veel feller als bij P. femoralis.

U bent geen bioloog, maar u weet in welk stuk van deze tekst de plek wordt beschreven waar de kikker is gevonden. Het correct identificeren van dit soort gegevens is belangrijk voor onderzoek naar biodiversiteit, naar onderzoek van verdwijnende diersoorten en de relatie met veranderende biotopen door bijvoorbeeld houtkap. We trainden de computer om deze taak te verrichten (Van den Bosch e.a., 2009) en hij knapte dat klusje aardig op, met een hele redelijke foutenmarge, zonder gebruik te maken van enige taalkundige kennis, maar zoekend naar woorden als onder, struik, en bos.

Nog een tekst: een dagboekaantekening[7] die beschrijft hoe een Nijmeegs gezin na twee bange dagen en nachten op 19 september 1944 naar buiten durft te komen in een bevrijde stad.

> toen het om 6 uur licht werd. was Nijmegen vrij we waren de plunderaars kwijt. voorzichtig gingen we naar buiten steeds meer mensen kwamen buiten en we konden de gevolgen zien van de gevechten in den afgelopen dag en nacht. op de hoek van onze straat lag stuk geschoten een kleine Amerikaanse tank. de bemanning 3 Amerikanen en een Hollander lagen er dood naast. allen met opgeheven handen. in de richting van de brug. vermoedelijk hebben ze zich nog willen overgeven.

De '3' lijkt een '2' geweest te zijn die gecorrigeerd is tot een '3'. Wie iets weet van de verschrikkingen van de dagen in september 1944 toen gestreden werd om deze streek en de stad Nijmegen, kent waarschijnlijk de geschiedenis van verzetsheld Jan van Hoof, die een dag nadat hij – naar verluidt – Duitse springladingen onder de Waalbrug onschadelijk had gemaakt, op 19 september 1944 werd doodgeschoten op de hoek van de Nieuwe Markt en de Lange Hezelstraat. Hij was twee Britten in een klein pantservoertuig de weg door het centrum aan het wijzen. Is de Hollander in het ooggetuigenverslag Jan van Hoof, is de tank het pantservoertuig, en zijn de 3 (of 2) Amerikanen de twee Britten van de Royal Engineers die Jan van Hoof de weg aan het wijzen was? Op hoeveel manieren kunnen ooggetuigen dezelfde gebeurtenis beschrijven, en kan de computer helpen bij het vinden van die overlappende maar verschillende beschrijvingen? Dit is een nog openstaande vraag in een projectvoorstel dat nu beoordeeld wordt.

Figuur 5. Een panoramisch overzicht van de stad Nijmegen en de Waalbrug, 28 september 1944.

Een krantenbericht dan, van 2 april 1935, in *De Tribune*. Deze scan van de originele krantenpagina, waarop alle letters en woorden zijn herkend met behulp van optische letterherkenning, OCR, is te vinden op de website van de Koninklijke Bibliotheek in Den Haag waarop miljoenen Nederlandse krantenpagina's uit bijna vier eeuwen terug te vinden zijn.[8]

> Zaterdag heeft te Den Haag een bespreking plaats gehad tussen de partijen, die betrokken waren bij het dreigend conflict aan de N.V. Zachts Stoomschoenfabriek te Kaatsheuvel. De bespreking stond onder leiding van den rijksbemiddelaar en had tot resultaat, dat de directie van bedoelde fabriek, zwichtend voor de stakingsdreiging van het personeel, besloot het door de arbeiders verlengde contract te tekenen, zij het met enkele wijzigingen. Waarin die wijzigingen bestaan, deelt het persbericnt, waaraan wij een en ander ontlenen, niet mede, zodat het dus niet onmogelijk is, dat ook de vakbondsleiders belangrijke concessies aan de directie hebben gedaan. In elk geval schijnt het "gevaar" van een staking thans opgeheven te zijn.

Een begrijpendlezenvraag zou nu kunnen luiden: is de staking nu doorgegaan of niet? En: vanuit welk perspectief werd dit artikel geschreven? Als u heeft meegelezen, dan begrijpt u wat er met de staking is gebeurd. Dat dit verhaal vanuit een communistisch perspectief is geschreven is misschien minder duidelijk; de *quotes* rondom "*gevaar*" zijn hier misschien wel de mooiste *clue*. We vinden in de database van de Koninklijke Bibliotheek overigens een handvol artikelen over deze stakingsdreiging, vanuit verschillende perspectieven. Een historicus die de geschiedenis van de sociale beweging onderzoekt, is geïnteresseerd in de terugkerende oorzaken die ertoe kunnen leiden dat een stakingsdreiging afgewend wordt. In een studie uit 2010 ontwikkelden we een automatisch classificatiesysteem voor het detecteren van krantenartikelen over de dreiging van stakingen (Van den Hoven, Van den Bosch, en Zervanou, 2010), iets wat met mensenogen en in mensentijd mogelijk is, maar ontstellend veel werk. Het systeem vond zelfstandig uit dat woorden als *rijksbemiddelaar*, *vakbondsleiders* en *partijen*, en ook woorden als *conflict* en *dreigend*, zeker in combinatie met elkaar, in hoge mate wijzen op berichtgeving over stakingsdreiging. Dit soort nog vrij simpele, ruwe filters, een soort omgekeerd spamfilter voor historisch onderzoek, zouden de basis kunnen zijn voor grootschalig onderzoek naar gebeurtenissen die uiteindelijk niet hebben plaatsgevonden; *counterfactual history*. *Unhistory*, of ongeschiedenis, was de suggestie van gast-postdoc en historicus Marten Düring. *Danke schön*, Marten.

Nog een dan: een volksverhaal, verzameld in het Noord-Brabants Sagenboek van J.R.W. Sinninghe, uitgegeven in 1933 (Sinninghe, 1933).

> Een kolenbrander, die in 't Mastbosch te Ginneken woonde, had zeven vreemdelingen, die van den baan naar Breda waren afgedwaald, en 's nachts bij zijn hut

> hadden aangeklopt, vermoord. De achtste ontkwam en ijlde naar de stad, naar de heeren van den gerechte. Het proces was kort; voor den moordenaar werd bij zijn hut een galg opgericht. Hij bleef echter ontkennen. Nog toen hij onder de galg stond, zwoer hij een vreeselijken eed: "Als ik 't gedaan heb, mogen de dooden tegen mij getuigen en de duivel mij meevoeren naar de eeuwigheid." En zie, daar opende zich de aarde. Zeven gebalde vuisten staken uit den grond op. Zoo ontstonden de Zeven Heuveltjes[9]. Een pekzwarte hond vloog uit de bosschen, greep hem aan en voerde hem mee naar de eeuwigheid – door het Eeuwige Laantje.

Er zijn een paar mensen hier in deze zaal die dit verhaal hiervoor al eens hoorden, bijvoorbeeld in de langere horrorvariant waarbij het Antwerpse studenten zijn die door een nachtelijk noodweer aankloppen bij het hutje, maar voor iedereen zal het verhaal een paar herkenningspunten bevatten, zeker als je wel vaker andere volksverhalen hebt gelezen of gehoord. Twee duidelijke terugkerende verhaalelementen, of motieven, zijn de duivelse hond die iemand wegvoert (naar de hel), en het naamgevingsmotief, vaak aan het eind van een tekst, waar wordt samengevat dat het dus zo kwam dat de plek de naam kreeg die het nu nog heeft.

Opnieuw de vraag: op hoeveel manieren kun je deze twee motieven verwoorden? Hoe zie je om te beginnen dat twee van die parafrases hetzelfde motief beschrijven? De KNAW heeft ons de middelen gegeven om dat te onderzoeken (Karsdorp, Van Kranenburg, Meder, Trieschnigg en Van den Bosch, 2012); we doen dat op het Meertens Instituut in Amsterdam waar een prachtige volksverhalendatabase is bijeengebracht door Theo Meder.

Dit waren vijf voorbeelden van teksten die we tegenkomen in onze verschillende projecten met partners uit andere disciplines binnen en buiten de geesteswetenschappen. Uiterst uiteenlopende werelden. Zojuist stond u even aan de voet van een berg in Suriname, en tilde een rot stuk hout op waaronder een kikker zich verscholen hield. Wennend aan het eerste licht van de dag stond u bij een gevallen verzetsstrijder naast een kapotgeschoten legervoertuig in het centrum van Nijmegen. U zag zeven gebalde vuisten uit graven oprijzen in een bos. Intense filmscènes, onvermoeibaar gegenereerd door onze interne machinerie. Met die werelden in uw hoofd was het niet moeilijk om mee te denken met de vragen die ik erbij stelde en de patronen die ik erin aanwees. Maar hoe krijgen we de computer zo ver?

Het beste antwoord dat ik daarop kan geven, op basis van alles wat we in deze verschillende projecten tot nu toe geleerd hebben, is dat je naar de echte vraag moet luisteren, de gegevens moet nemen zoals ze zijn, en geen enkel vooropgezet idee moet hebben over hoe je die vraag gaat beantwoorden. Geef de computer alleen de ruwe data, de tekst, en stel je vraag. De methoden die wij gebruiken, en die de impliciete route volgen, vragen vervolgens allereerst een aantal (liefst heel veel) voorbeelden van teksten waarbij het antwoord al is gegeven. Die antwoorden zijn er vaak al omdat er al veel data

door mensenhanden geanalyseerd zijn voor onderzoek, of omdat de antwoorden al deels meegegeven zijn met de tekst, zoals de woonplaats van twitteraars. *Hashtags* in tweets kunnen er soms perfect voor gebruikt worden. Of aantallen reacties, aantallen sterren bij beoordelingen, enzovoort.

En dan komt de computer soms met de simpelste en mooiste oplossingen. Het enige wat hij doet is gevallen vergelijken – nieuwe gevallen met bestaande gevallen in zijn geheugen – en daarbij de aanwezigheid van sommige woorden of combinaties van woorden zwaarder laat wegen dan van andere. Neem *zin in*: als een tweet deze woordencombinatie bevat, dan gaat hij over een toekomstige gebeurtenis. Je zou niet meteen denken aan *zin in* en eerder denken aan meer directe verwijzingen naar tijd, zoals *morgen* of *vanavond*, die ook zwaar wegen natuurlijk. Of denk aan een woord als *agge* dat in zijn eentje een hele sterk filter is op Brabantse tweeps; daar hoef je verder niet veel aanwijzingen meer bij te hebben.

Mijn werkhypothese is dat wij mensen op heel veel manieren tot hetzelfde antwoord kunnen komen vanwege onze uitstekende kennis van de wereld om ons heen, terwijl de computer, die die wereldkennis grotendeels nog ontbeert, toch ook heel slim kleine, subtiele aanwijzingen kan oppikken, als hij maar op het probleem wordt gericht en wordt gedwongen dat probleem zo precies mogelijk op te lossen, met hulp van voorbeelden. Dit is in technische termen het verschil tussen een generatief model, een nooit aflatende generieke verwachtingenmachine (een metafoor die ik eerder al associeerde aan hoe het in het mensenbrein werkt, en die je terugziet in onze tekst-naar-tekst-modellen), en een discriminatief model, dat maar een ding kan, zoals Brabanders eruit pikken – maar dan ook meteen goed. Zolang we wereldkennis nog niet goed in één groot computermodel gepropt krijgen, blijken die discriminatieve modellen een prima alternatief, en moeten we er gewoon zo veel maken als er vragen zijn.

Het idee van voorbeeldgedreven modellen, van voorbeeldgebaseerd redeneren, is een blauwdruk voor een oplossing. Er is geen enkel systeem dat alles gaat oplossen; iedere vraag, gecombineerd met voldoende voorbeelden, levert weer een nieuw systeem op.

OPEN ONDERZOEK

We bedrijven misschien een nieuw soort taalkunde of een heel datagedreven soort kennistechnologie, maar we maken ook gewoon systemen die het goed doen. Veel van onze ideeën kunnen we kwijt in projecten met een valorisatiecomponent; zelfs al interesseert het ons niet, dan levert ons onderzoek nog software op die soms interessant is voor anderen. Het klinkt alsof ons vakgebied goed in de markt ligt. Schuiven we daarmee ook langzaam naar de commerciële sector? Het gebeurt inderdaad dat onderzoekers en studenten uit ons midden ondernemingen starten en succesvol taal- of spraaktechnologie gaan aanbieden als een commercieel product, en het is een plezier om contact te blijven houden met deze bedrijven, bijvoorbeeld om studenten naar door te kunnen verwijzen. Maar toch blijft een contingent van ons bewust in de academi-

sche context. Als wij systemen bouwen dan doen we dat om onze hypothesen te toetsen, of om met een andere academische discipline te onderzoeken welke kennis we uit hun materiaal kunnen halen. Als dat systemen oplevert die het doen, dan is dat een mooie aanwijzing dat onze ideeën de juiste kant in gingen.

Soms maken we die systemen beschikbaar, als demonstratie, of als webservice, een internetdienst, die niet alleen door mensen maar ook door software te gebruiken is. Als de onderliggende software goed genoeg is, bijvoorbeeld omdat wetenschappelijk programmeurs de software van onderzoekers hebben gefatsoeneerd, dan brengen we hem uit als *open source* software: het staat iedereen vrij om de software te downloaden, te gebruiken, te veranderen, en in te bouwen in weer andere systemen, mits dezelfde voorwaarden blijven gelden. Onze meest gebruikte software (Daelemans, Zavrel, Van der Sloot en Van den Bosch, 2010), met bijdragen van een lange reeks van mensen van de universiteiten van Tilburg, Antwerpen en nu ook Nijmegen, is als pakket[10] meegeleverd met varianten van het *Linux operating system*, dat zelf ook *open source* is, en wordt dus niet meer alleen door ons gedistribueerd. Onze wetenschappelijke publicaties waarin deze software beschreven wordt krijgen misschien enkele tientallen tot soms een paar honderd verwijzingen in de literatuur, maar onze software is door de jaren heen duizenden keren gedownload.

Je software delen met andere wetenschappers is goed voor de wetenschappelijke integriteit (om resultaten te kunnen repliceren), maar ook goed voor de wetenschappelijke impact van het idee achter de software. Dat zijn twee daverende redenen voor het tot *open source* verklaren van software die in wetenschappelijke projecten is ontstaan, en die goed genoeg is om aan de buitenwereld gegeven te worden. Datzelfde geldt ook voor het delen van tekstuele data en andere onderzoeksgegevens zoals annotaties, maar hier hebben grote overheidsgesteunde projecten waarin teksten voor onderzoek werden verzameld, bijvoorbeeld het Corpus Gesproken Nederlands (Oostdijk *et al.*, 2002) en SoNaR (Reynaert *et al.*, 2010), ons geleerd dat rechten wel gerespecteerd dienen te blijven. In bestaande rechten wil ik dan ook niet treden, maar ik hoop wel dat steeds meer mensen en instellingen het idee omarmen van vrije deling van teksten onder licenties als *Creative Commons*[11] dat bij Wikipedia wordt gebruikt[12], of volledig vrij zoals de Europese Unie dat doet (The European Parliament and the Council, 2001).

De wetenschap zelf heeft ook nog wel wat harde noten te kraken als het gaat om het delen van software en onderzoeksgegevens. Het kan beter. Het moet beter. Iedere wetenschapper moet bij zichzelf te rade gaan als hij of zij op data of software zit en die niet vrijgeeft als het onderzoek gedaan is en de publicaties geschreven zijn. Er schuilt ook een probleem in de tijdelijkheid van financiering van het meeste van wat we doen. Of misschien is die tijdelijkheid juist helemaal niet het probleem, en is het aan de creativiteit van de onderzoekers om steeds nieuwe wegen te vinden om goede ideeën te laten overleven; daar worden goede ideeën alleen maar sterker van.

DANKWOORD

En toen was ik in Nijmegen.

De kans om mijn onderzoek in de context van de Nijmeegse letterenfaculteit voort te zetten heb ik met beide handen aangegrepen. Ik kreeg een bijzonder warm welkom, en kreeg de gelegenheid om me aan te sluiten bij en vorm te geven aan nieuwe initiatieven, zoals de vorming van een focusgroep rondom *e-humanities*, en van een onderzoekseenheid binnen het Centre for Language Studies, de PI Group Language and Speech Technology. Het lidmaatschap van het departement Bedrijfscommunicatie nodigt uit tot samenwerking en toepassingen van taaltechnologie in domeinen als gezondheidscommunicatie en journalistiek. Ik heb in Nijmegen ook buiten de faculteit de ene na de andere aansluiting gevonden. Met hulp van de Faculteit Natuurwetenschappen, Wiskunde en Informatica bouwden we bij hen op zolder een computerlab. Met Wessel Kraaijs Information Foraging Lab in die faculteit gloort een mooie samenwerking, evenals met The Language Archive van het Max Planck Instituut, en het Donders Institute. Het belang dat zowel het college van bestuur van deze universiteit als het bestuur van de Faculteit der Letteren hecht aan onderzoek op het gebied van taal en communicatie, en aan de rol van technologie daarin, is uniek in dit land, en ik zie grote mogelijkheden in het uitbouwen en combineren van expertises en disciplines. Ik wil het college van bestuur en het faculteitsbestuur, met name Paul Sars en Theo Engelen, danken voor het getoonde vertrouwen.

Vertrouwen, visie, zin in onderzoek doen, dat is wat me in Nijmegen direct opviel toen ik mijn nieuwe collega's leerde kennen. Iedereen van de twee onderzoeksinstituten van de letterenfaculteit, het CLS en HLCS, met wie ik het genoegen heb gehad kennis te maken dank ik voor de stimulerende gesprekken en ideeën voor nieuw onderzoek. Als het vandaag niet gebeurt, dan morgen wel. Paula Fikkert dank ik bijzonder voor de ongecompliceerde positieve steun. Ik vind het fantastisch om met Nelleke Oostdijk, Suzan Verberne, Hans van Halteren en de andere taal- en spraaktechnologen rondom het Centre for Language and Speech Techology van Henk van den Heuvel een nieuwe onderzoeksgroep op te starten, en het is ook een voorrecht om op ongeregelde momenten binnen te kunnen vallen voor een dosis inzicht en goede raad bij Lou Boves.

Het departement Bedrijfscommunicatie heeft me ook veel geschonken: Hans Hoeken heeft me wegwijs gemaakt, Margot van Mulken, Béryl Hilberink en de ''meisjes van het secretariaat'' regelden alles wat los en vast zat. Dank aan iedereen van het departement voor een warm welkom. De club jonge onderzoekers die ik heb kunnen recruteren voor de nieuwe projecten die het afgelopen jaar zijn begonnen dank ik voor hun inzet en aanstekelijk enthousiasme: Maarten, Florian, Ali, Kalliopi, Marten, Iris, Diana, Wessel, en Sebas. Ik ben ook heel blij dat ik naast deze bewoners van het Erasmusgebouw ook nog mag samenwerken in projecten buiten de deur met PhD-studenten Folgert Karsdorp, Peter Berck, Matje van de Camp, Sander Wubben, en

Véronique Verhagen, en postdoc Erik Tjong Kim Sang, en de begeleidende teams van hun projecten op het Meertens Instituut , de Universiteit Twente, Tilburg University en het Netherlands eScience Center.

De groep mensen die niet meer mijn collega's zijn is radicaal groter geworden toen ik vorig jaar zomer in Tilburg vertrok. Ik dank Jaap van den Herik, Fons Maes, Harry Bunt en Marc Swerts voor alle mooie dingen die we samen hebben bereikt. Jaap, ook nu we niet op dezelfde universiteit werken maak ik mee hoe belangrijk je rol is in het verwezenlijken van een infrastructuur voor *e-humanities* in Nederland door je onvermoeibare werk voor het NWO CATCH-programma. Ik werk door met Emiel Krahmer, Ad Backus, Ko van der Sloot, Martin Reynaert en Eric Postma aan de projecten die we samen in gang zetten in Tilburg. Op die universiteit heb ik zowel gestudeerd als een lange periode gewerkt. Een van de eerste docenten die ik daar ontmoette was Walter Daelemans. Ik sprak mijn bewondering en dank vier jaar geleden in mijn Tilburgse oratie al eens uit, maar ik herhaal het graag: Walter, je grandioze inzichten in taal en kunstmatige intelligentie en je hoge standaarden in het doen van onderzoek zijn nog altijd modellen voor mij. De kameraadschap en vriendschap die je biedt, zijn me zeer dierbaar.

Walter werkte een aantal jaren hier in Nijmegen. Als een van de plaatsen in Nederland waar al vroeg computerlinguïstiek werd bedreven, door de groepen van Jan van Bakel en Jan Aarts, werd op het toenmalige NICI ook al heel vroeg de link gelegd tussen taaltechnologie en psycholinguïstiek. De wegen die Gerard Kempen en Ton Dijkstra verkenden en nog altijd verkennen liggen voor me om te volgen. Een aanwijzing dat ik dat al aan het doen ben is de titel van deze oratie, *Taal in Uitvoering*, de titel van een boek uit 1984 van deze twee hooggeleerde collega's (Dijkstra en Kempen, 1984).

Mijn dank geldt ook, zonder namen te noemen, maar de goede verstaander weet genoeg, voor iedereen in Nederland en daarbuiten met wie ik heb mogen samenwerken, en nog steeds samenwerk, aan onderzoek en de organisatie van onderzoek.

Terug naar de basis. Ik haal veel kracht uit het gevoel van thuis zijn in het midden van mijn gezin, waar ik ook ben. Dat gevoel, en die basis zelf, heb ik meegekregen van mijn lieve ouders. Lieve papa en mama, wat zijn jullie sterk met z'n tweeën, en wat is het fijn dat jullie erbij zijn.

Lieve Liam en Merijn, jullie zijn in meerdere opzichten mijn jeugd van tegenwoordig. Ik schud mijn hoofd om dezelfde dingen die ik vroeger deed, zoals al mijn tijd achter de computer zitten. Jullie zijn grote schatten en ga vooral zo door.

Liefste Anne-Marie, aan het eind van de dag ben jij er met een lieve glimlach en een knuffel. Jij bent mijn aarde.

Ik heb
gezegd.

NOTEN

1 http://www.troonredes.nl/2010/troonrede-17-september-1963/
2 http://valkuil.net
3 http://fowlt.net werd gelanceerd op 18 januari 2013.
4 http://translate.google.com
5 http://ilk.uvt.nl/il
6 http://twitter.com
7 http://www.noviomagus.nl/Gastredactie/Breukelen/Cat/cwdata/07.html
8 http://kranten.kb.nl/
9 Ik besefte pas bij het opnemen van dit verhaal in deze oratie dat we als gezin van de Zeven Heuveltjes naar de Zevenheuvelen zijn verhuisd.
10 http://packages.debian.org/source/sid/science/timbl
11 http://creativecommons.nl/
12 http://en.wikipedia.org/wiki/Wikipedia:Copyrights

BIBLIOGRAFIE

- Arnon, I. and Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, **62:1**, pp. 67–82.
- Bannard, C., and Matthews, D. (2008). Stored word sequences in language learning: The effect of familiarity on children's repetition of four-word combinations. *Psychological Science*, **19:3**, pp. 241–248.
- Daelemans, W., Zavrel, J., Van der Sloot, K., en Van den Bosch, A. (2010). TiMBL: Tilburg memory based learner, version 6.3, reference guide. Technical Report ILK 10-01, ILK Research Group, Tilburg University.
- Dijkstra, T., en Kempen G. (1984). *Taal in uitvoering*. Groningen: Wolters-Noordhoff.
- The European Parliament and the Council (2001). Regulation (EC) No 1049/2001 of the European Parliament and of the Council regarding public access to European Parliament, Council and Commission documents. *Official Journal of the European Union*, 145:43—48, May 2001.
- Goldberg, A.E. (2006). Constructions at work: *The nature of generalization in language*. Oxford University Press.
- Karsdorp, F., Van Kranenburg, P., Meder, T., Trieschnigg, D., en Van den Bosch, A. (2012). In search of an appropriate abstraction level for motif annotations. In *Proceedings of the 2012 Computational Models of Narrative Workshop*, pp. 22–26, Istanboel, Turkije, 2012
- Koehn, P., F.-J. Och, and D. Marcu (2003). Statistical phrase-based translation. In *Proceedings of HLT-NAACL* 2003, pp. 48–54, Edmonton, Canada.
- J. Odijk. Het Nederlands in het Digitale Tijdperk – The Dutch Language in the Digital Age. META-NET White Paper Series. Springer, 2012. Available online at http://www.meta-net.eu/whitepapers
- Oostdijk, N., Goedertier, W., Van Eynde, F., Boves, L., Martens, J.P., Moortgat, M., and Baayen. H. (2002). Experiences from the Spoken Dutch Corpus project. In M. González Rodriguez and C. Paz Suárez Araujo, editors, *Proceedings of the Third International Conference on Language Resources and Evaluation*, pp. 340–347.
- Paul, H. (1920). *Prinzipien der Sprachgeschichte*. Halle: Max Niemeyer. Vijfde druk.
- Reynaert, M., Oostdijk, N., De Clercq, O., Van den Heuvel, H., and De Jong, F. (2010). Balancing SoNaR: IPR versus processing issues in a 500-million word written Dutch reference corpus. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation* (LREC'10), Valletta, Malta, May 2010. European Language Resources Association (ELRA).
- Sinninghe, J.R.W. (1933). *Noord-Brabantsch sagenboek*. Zutphen: Thieme.
- Van den Bosch, A. (1997). Learning to pronounce written words: A study in inductive language learning. Ph.D. thesis, Universiteit Maastricht.
- Van den Bosch, A. (2011). Effects of context and recency in scaled word completion. Computational Linguistics in the Netherlands Journal, **1**:79–94
- Van den Bosch, A., Lendvai, P., Van Erp, M., Hunt, S., Van der Meij, M., and Dekker, R. (2009). Weaving a new fabric of natural history. *Interdisciplinary Science Review*, 34:2–3, pp. 206–23.
- Van den Hoven, M., Van den Bosch, A., and Zervanou, K. (2010). Beyond reported history: Strikes that never happened. In S. Daránvi and P. Lendvai (Eds.), *Proceedings of the First International AMICUS Workshop on Automated Motif Discovery in Cultural Heritage and Scientific Communication Texts*, Vienna, Austria, pp. 20-28.
- Van Pamelen, F. (2012). *De zin van de ommezijde*. De Fontein, 2012.