TNTL 123 (2007) 22-36

Max Louwerse en Willie van Peer

Waar het over gaat in cijfers

LSA als kwantitatieve benadering in tekst- en literatuurwetenschap

Abstract — The present article argues that the content analysis of literature may profit from computational techniques such as *Latent Semantic Analysis* (LSA). LSA is able to calculate the semantic distance between textual items by locating them in a vast multi-dimensional space. The results show remarkable similarity when compared to psychological data. LSA has not, however, been employed for content analysis in Dutch. We offer two explorative examples, one with Dutch lexical items and one with Dutch literary texts, to demonstrate that LSA also works in Dutch. At the same time, we hope to have demonstrated the usefulness of the technique in answering research questions bearing on literature.

In dit artikel wordt betoogd dat inhoudsanalyses in letterkunde en literatuurwetenschap zouden kunnen profiteren van computationele technieken zoals *Latente Semantische Analyse* (LSA). LSA is in staat de semantische afstand tussen tekstelementen te berekenen op basis van hun locatie in een enorme semantische ruimte. De hierbij verkregen semantische gegevens komen sterk overeen met gegevens die worden verkregen uit psychologische experimenten. LSA is echter nooit gebruikt voor inhoudsanalyse in het Nederlands. Wij geven twee exploratieve voorbeelden, een met Nederlandstalige woorden en een met Nederlandstalige literaire teksten, om aan te tonen dat LSA ook werkt voor het Nederlands en om aan te tonen dat een dergelijke techniek van groot belang is voor de beantwoording van onderzoeksvragen uit de letterkunde en literatuurwetenschap.

In letterkunde en literatuurwetenschap lijkt er een misverstand te bestaan dat kwantitatieve onderzoeksmethoden niet thuis zouden horen in de cultuurwetenschappen. Zo zou in de bestudering van literatuur geen gebruik kunnen worden gemaakt van psychologische experimenten of corpus-analyses. Inhoudsanalyses zouden zijn weggelegd voor getrainde individuen. Het lezen van literatuur zou tenslotte een bijzonder individuele ervaring zijn die niet zo maar gemeten kan worden; literatuur zou structurele kenmerken bezitten die zo verheven zijn, dat banale experimenteer- en computertechnieken deze nooit zouden kunnen analyseren. Wanneer we deze benaderingen (experimenteel lezersonderzoek en computergestuurde tekstanalyses) als 'empirisch' bestempelen, dan is voor de meeste literatuurwetenschappers de literatuurwetenschap *niet* empirisch. Mochten deze opvattingen al niet expliciet voorhanden zijn, impliciet zijn ze overduidelijk aanwezig. Een eenvoudige vergelijking dient ter illustratie. Zo kunnen we nagaan hoe vaak binnen de literatuurwetenschap het trefwoord 'literair' samen voorkomt met het trefwoord 'empirisch'. Wanneer we dit nagaan in de grootste bibliografische literatuurwetenschappelijke database, die van de Modern Language Association of America, is de verhouding van treffers voor de zoekwoorden *literary* en *empirical* ten opzichte van het zoekwoord *literary* 0,28%. Met andere woorden: in literatuurwetenschappelijke publicaties wordt vrijwel nooit gewag gemaakt van een

empirische benadering. Dit kan licht de indruk wekken dat dit ook onvermijdelijk of zelfs noodzakelijk zou zijn. Niets is echter minder waar. Wanneer we PsycINFO (de database van de American Psychological Association) erop naslaan, krijgen we een heel ander beeld te zien: in deze database ligt die verhouding op 4%, 14 keer hoger dan in de database van de MLA. Met andere woorden: de vraag of de bestudering van literatuur met een empirische benadering verenigbaar is, wordt in de disciplines literatuurwetenschap en psychologie zeer verschillend beantwoord. Psychologen schatten die verenigbaarheid aanzienlijk hoger in dan literatuurwetenschappers. De vrijwel volledige afwijzing van de empirische benadering door de literatuurwetenschap heeft dus niets 'natuurlijks', maar is een product van een ideologie die binnen deze wetenschap in stand wordt gehouden. Andere disciplines beschouwen deze afwijzing als ongegrond. We kunnen deze analyse ook voor andere trefwoorden herhalen. Zo ligt de verhouding van de woorden *literary* en *psychological* ten opzichte van *literary* bij MLA op 0,3% en bij PsycINFO op 17%, 57 keer hoger. Natuurlijk is dit allerminst een volledige en betrouwbare analyse, maar het geeft een aardige indruk hoe empirisch letterkunde en literatuurwetenschap zijn.

De anti-empirische opvatting die in de literatuurwetenschappen kan worden geconstateerd heeft geleid tot een decennialang durende methodologische identiteitscrisis. Verscheidene wetenschappers hebben in de loop der tijden koortsachtig getracht de literatuurwetenschap tot wetenschap te verheffen door onderzoek naar literatuur empirisch te maken (zie de uitvoerige bibliografie op http://www.igelweb.org). Organisaties als de Association for Empirical Studies of Literature and Media (IGEL), the Poetics and Linguistics Association (PALA), en de International Association for Empirical Studies of the Arts (IAEA) hebben in dit proces een belangrijke rol gespeeld. Hierdoor lijkt langzaamaan het belang van empirische benaderingen van de literatuurwetenschap duidelijk te zijn geworden. Het is echter verbazingwekkend dat een methodologische identiteitscrisis in de literatuurwetenschap überhaupt kan worden gediagnosticeerd. Immers, het is moeilijk voor te stellen dat in disciplines als natuurkunde of psychologie een fundamentele discussie zou plaats hebben over de waarde van empirisch onderzoek. Bovendien, het feit dat deze discussie plaatsheeft in de literatuurwetenschap toont niet bepaald de kracht aan van het vakgebied. Immers, van de voorstanders van het verbannen van kwantitatieve methoden uit cultuurwetenschappen is nooit duidelijk geworden waarom de regels die zo goed werken voor andere wetenschappelijke disciplines níet zouden werken voor cultuurwetenschappen. Anderzijds, voorstanders van empirische tekst- en literatuurwetenschap hebben veelvuldig aangetoond dat psycholinguïstische en computationeel-linguïstische technieken wel degelijk op de letterkunde kunnen worden toegepast (bijv. Hakemulder 2000; Louwerse 2004; Louwerse & Van Peer 2002; Miall & Kuiken 1994; Van Peer 1986). Als deelgenoten van deze laatste groep is ons gevraagd de angst voor empirie (die er klaarblijkelijk heerst onder een grote groep literatuurwetenschappers) weg te nemen, door een licht te werpen op empirische methoden naar de inhoud van teksten, in het bijzonder literaire.

Dit artikel is als volgt opgebouwd. Allereerst wordt een zeer beknopt overzicht gegeven van de recentste geschiedenis van de inhoudsanalyse. In onze beknoptheid doen we groot onrecht aan de vele andere benaderingen, methoden en tech-

nieken en we verwijzen daarom op voorhand naar betere overzichten, zoals Graesser, Gernsbacher & Goldman (2002), Jurafsky & Martin (2000), Louwerse & Van Peer (2002), Manning & Schütze (1999). Vervolgens spitsen we ons toe op de computationele linguïstische techniek Latente Semantische Analyse (LSA) die semantische relaties berekent tussen tekstelementen. Twee LSA-illustraties voor de Nederlandse taal worden vervolgens gepresenteerd: een die de semantische relaties blootlegt tussen Nederlandse woorden, de ander die deze relaties blootlegt tussen Nederlandse literaire teksten.

1 Ontwikkelingen in de laatste decennia

Het ontstaan van de empirische tekst- en literatuurwetenschap, en in het bijzonder van de empirische inhoudsanalyse, moet worden gezocht aan het begin van de twintigste eeuw in het werk van de volkskunde. Folkloristen zochten in de lijn van de positivistische traditie naar een classificatie van verhalen in de hoop de universaliteit van volksverhalen te kunnen duiden. Thematiek leek het antwoord op zowel het probleem hoe de enorme hoeveelheid volksverhalen geclassificeerd kon worden en of er universele kenmerken aan volksverhalen ten grondslag lagen. In zijn *Morfologie van het toversprookje* (1997/1928) analyseerde Vladimir Propp een corpus van 100 toversprookjes en abstraheerde die constante elementen die in het merendeel van de verhalen voorkomen. Deze elementen, zogenaamde verhaalfuncties, vormden de bouwstenen van het genre sprookje (Louwerse 1997). Propp werd daardoor de grondlegger van de narratologie, en introduceerde door zijn (op de biologie geïnspireerde) classificatiesysteem van verhaalelementen voor het eerst een systematisch-wetenschappelijke benadering in de literatuurwetenschap.

Propps werk is van onschatbare waarde voor de cultuurwetenschappen. Het legde de basis voor uitgebreide corpuslinguïstische analyses (Bremond 1973) en leidde ertoe dat de fundamentele bouwstenen van verhaalteksten systematisch werden blootgelegd en aan systematisch onderzoek werden onderworpen. Door dat onderzoek weten we vandaag dat die bouwstenen ook psycholinguïstische waarde hebben (Rumelhart 1977). Met andere woorden: de verhaalelementen die Propp in zijn tekstanalyses vond corresponderen met categorieën die de lezers van die teksten hanteren bij het verwerken van het gelezene. Maar dit inzicht is het resultaat van empirisch lezersonderzoek (dat grotendeels *buiten* de institutionele literatuurwetenschap heeft plaatsgevonden); zonder dit empirisch onderzoek stond de literatuurwetenschap nog in het stadium waar ze op het einde van de negentiende eeuw stond. Propp stond daarmee in zekere zin aan de wieg van psycholinguïstiek en computationele linguïstiek. De waarde van Propps werk lijkt hier wellicht overschat te worden, maar de referenties naar zijn werk in psycholinguïstiek en corpuslinguïstiek bewijzen het tegendeel. Het Westen, met name na de eerste vertaling in het Engels in 1968, erkende de methodologische waarde van zijn werk. Propps structuralistische opvattingen hadden bijvoorbeeld een grote invloed op tekstwetenschappers (Van Dijk 1972), antropologen (Levi-Strauss 1958) en cognitiewetenschappers (Kintsch 1974). Van Dijk (1972) toonde bijvoorbeeld aan dat de universele regels die gelden voor de zinsstructuur uitgebreid konden worden tot regels die ook aan de structuur van teksten ten grondslag liggen. Een van de eerste ver-

haalgrammatica's werd voorgesteld door Rumelhart (1977), een van de grondleggers van latere connectionistische en neurale netwerk theorieën, die de psychologische waarde ervan in een reeks experimenten bewees (Rumelhart & McClelland 1986). Schank en Abelson (1977) toonden verder aan dat verhalende teksten begrepen kunnen worden door het toepassen van cognitieve schema's en scripts van bepaalde stereotype situaties. Van Dijk en Kintsch (1983) bouwden deze theorieën verder uit. Zij wezen er bijvoorbeeld op dat teksten twee structuurniveaus hebben. Een microstructuur is de locale structuur van de tekst, terwijl een macrostructuur de globale structuur van de tekst representeert. Het model van Van Dijk en Kintsch werd later vertaald in een psychologisch model van tekstverwerking dat grote invloed heeft gehad op de cognitieve psychologie van de jaren tachtig en negentig (zie Gernsbacher 1994 voor een overzicht). Hun model kwam er op neer dat bij het lezen van teksten zinnen werden vertaald in proposities, abstracte betekeniseenheden bestaande uit een predicaat (bijv. een handeling) en een variabele (bijv. uitvoerder van de handeling). Een hiërarchisch netwerk van die proposities vormde een globale representatie van de tekst. Achtergrondinformatie van de taalgebruiker vulde dit netwerk aan met bijzonderheden opdat een coherent situatiemodel ontstond. Kintsch's Construction-Integration Model (1988, 1998) werkte deze theorie verder uit, wat leidde tot uitbreidingen en alternatieven. Graesser, Millis en Zwaan (1997), bijvoorbeeld, beargumenteerden dat lezers naast een propositie- en situatiemodel ook een genre- en communicatiemodel vormen: de lezer verwerkt dus niet slechts de 'inhoud' van een verhaal, maar vormt zich tijdens de verwerking ook een beeld van het type waartoe de tekst behoort en van de intenties van de auteur. In het hedendaagse wetenschappelijke landschap hebben verdere ontwikkelingen deze modellen steeds verder verfijnd, onder meer in het model van Gernsbacher (1990) een Structure Building Framework, in dat van Zwaan en Radvansky (1998) een Event Indexing Model, of dat van Van den Broek et al. (1996) een Landscape Model.

Deze modellen hebben veel overeenkomsten. Belangrijker is dat deze modellen niet uit de lucht kwamen vallen, maar het gevolg waren van de resultaten van talloze empirische onderzoeken waarin gekeken werd hoe lezers teksten lazen en wat ze zich van die teksten herinnerden. In veel gevallen werd daarbij gebruikt gemaakt van eenvoudige teksten (soms hekelend aangeduid als 'Mickey Mouse texts' of 'textoids'). Aanvankelijk had men om die reden dit onderzoek nog als irrelevant voor de literatuurwetenschap kunnen afdoen: literatuur lag een stap verder. Intussen kan dit argument echter niet meer worden gehanteerd: het aantal empirische onderzoeken waarin de verwerking van literatuur middels bestaande literaire teksten wordt onderzocht, is inmiddels uitgegroeid tot een aanzienlijk corpus (Louwerse & Kuiken 2005). Ook de opvatting dat dergelijk onderzoek slechts met enkele proefpersonen is uitgevoerd, en daardoor beperkte waarde heeft, kan niet meer worden verdedigd. Van Peer (2007) bijvoorbeeld laat zien dat empirisch onderzoek naar de theorie van *foregrounding* (Van Peer 1986; Van Peer, Zyngier & Hakemulder, in druk; Zyngier, van Peer & Hakemulder, in druk) met rond 2000 proefpersonen is uitgevoerd. Zwaan (1993) had bovendien aangetoond dat literaire teksten niet zelden als literair begrepen worden, simpelweg omdat de lezer *verwacht* dat ze literair zijn.

Een verklaring voor de populariteit van tekstgrammatica's van de jaren zeventig

en de tekstbegrip-modellen van de jaren tachtig en negentig ligt in de opkomst van de computationele linguïstiek. Met de cognitieve revolutie in de jaren vijftig werden hogere psychologische functies, zoals intelligentie, redeneren, geheugen en besluitvorming belangrijker in de psychologie. Niet toevallig viel die aandacht samen met de opkomst van de computer. Langzaamaan won het besef veld dat computers wellicht mentale functies konden simuleren. Newell en Simons (1972) informatieverwerkingstheorie bijvoorbeeld beschouwde de menselijke hersenen als de hardware die de menselijke geest, de software, bestuurde. Tenslotte vertaalden de hersenen zintuiglijke informatie in een (neurale) code, verwerkten deze code, bewaarden relevante aspecten van de code en waren in staat deze aspecten op te vragen, net als een computer. Een van de meest invloedrijke computationele modellen voor tekstbegrip werd geïntroduceerd in de jaren tachtig en won aan populariteit in de jaren '90. Het werd aanvankelijk gebruikt voor automatische vraagen antwoordsystemen, maar groeide uit tot een techniek die de betekenis van woorden, zinnen, alinea's en teksten kon berekenen. Bovendien had deze computationele methode erg veel weg van menselijk taalbegrip (Landauer en Dumais 1997). Latente Semantische Analyse (LSA; Landauer, McNamara, Kintsch, en Dennis 2006) en haar minder bekende broertje Hyperspace Analogue to Language (HAL; Lund en Burgess 1996) berekenen de semantische afstand tussen tekstelementen door deze in een enorme multi-dimensionele ruimte te plaatsen en de afstand tussen woordvectoren te berekenen. Daarbij is van belang dat woorden en hun context worden gebruikt in de berekening van semantische relaties. Dus woorden met gemeenschappelijke buren (bijv. *koe* en *schaap* hebben gemeenschappelijke buren als *boer*, *weide*, *grazen*) zijn semantisch gerelateerd. LSA analyseert echter niet slechts de buren van sleutelwoorden, maar ook de buren van de buren (van de buren van de buren, etc.) We komen later nog uitvoeriger terug op LSA. Op dit moment is het voldoende te stellen dat er empirisch bewijs is dat LSA menselijk taalbegrip goed kan simuleren.

Recentelijk is echter kritiek geuit: computermodellen mogen dan wellicht interessante resultaten bieden, maar deze komen geenszins in de buurt van menselijk tekstbegrip: taalgebruikers combineren geen woorden zoals LSA dit doet, maar ze onderhandelen met de wereld (Pecher en Zwaan 2005). Zoals Searle (1980) stelde, met een woordenboek alleen kom je er niet als pasgeboren taalgebruiker. In plaats daarvan moet taal worden 'geaard' in de wereld. Deze 'embodiment'-beweging, die hamert op de 'symbol grounding' in het begrijpen van taal, stelt dat we weten wat een *stoel* is, niet omdat deze een semantische relatie heeft met *tafel* en *zitten*, maar omdat we onze ervaringen met stoelen kunnen simuleren. Deze simulaties zijn mogelijk omdat we op stoelen hebben gestaan, met stoelen hebben gegooid, stoelen hebben aangeschoven en op stoelen hebben gezeten. Een ander voorbeeld: als we zeggen dat de beurs is *gestegen*, dan verbinden we die betekenis met onze ervaringen met letterlijk *stijgen*, het tegengestelde van *vallen* – en omdat wij rechtop lopende wezens zijn, vinden wij *vallen* niet leuk, maar *stijgen* integendeel wel. Datzelfde geldt voor functiewoorden als voorzetsels: embodiment maakt dat *boven* en *op* positieve betekenisassociaties oproepen, maar *onder* en *neer* juist negatieve. Op gelijksoortige wijze begrijpen we iets positiefs wanneer we van iemand zeggen dat hij aan de *top* van een bedrijf staat, en iets negatiefs wanneer die iemand aan *lager* wal is geraakt. Op zichzelf is er niets negatiefs aan *laag* en niets positiefs

aan *hoog*, maar onze lichamelijke ervaringen met de verticale dimensie maken die betekenissen tot wat ze zijn. Embodiment constitueert betekenis. Experimenten hebben aangetoond dat er inderdaad veel bewijs is voor embodiment (Pecher en Zwaan 2005). Binnen letterkunde- en literatuurwetenschap heeft deze richting geleid tot het ontstaan van Cognitive Poetics, een richting die 'embodiment' voor het begrijpen van literatuur noodzakelijk acht (Stockwell, 2002, Semino en Culpeper 2002). Hoewel hier het belang van embodiment niet wordt ontkend, en ook niet wordt onderschat, zijn er ook auteurs die er nadrukkelijk op wijzen dat embodiment mogelijk, maar zeker niet altijd noodzakelijk is; zo ondermeer Louwerse (2007), Louwerse, Cai, Hu, Ventura en Jeuniaux (2007). Veel taalbegrip kan namelijk puur symbolisch plaatshebben omdat de symbolische taalstructuren belichaamde structuren hebben gecodeerd. Louwerse (2007) stelt een Symbol Interdependency Hypothesis voor, die stelt dat taalgebruikers zowel symbolische als belichaamde representaties van betekenissen vormen. LSA kan daarbij als model worden gebruikt om die symbolische representaties te op te sporen. Hoewel we ervan overtuigd zijn dat LSA als cognitief model kan worden beschouwd (Landauer en Dumais 1997; Louwerse et al. 2006), speelt een dergelijke overtuiging geen beslissende rol in het onderstaande betoog. In wat volgt richten we ons vooral op LSA als een computationele techniek om betekenis te berekenen in taal, tekst en literatuur.

2 Inhoud meten

2.1 Latente Semantische Analyse

LSA berekent met welke frequentie welke woorden gebruikt worden in welke context. LSA construeert een multi-dimensionele semantische ruimte uit een groot corpus van miljoenen woorden en tienduizenden alinea's. Het doet dat door een enorme matrix samen te stellen waarbij elke cel de frequentie weergeeft van een woord in een alinea. Neem bijvoorbeeld een 'corpus' dat bestaat uit de 'alinea' *lees maar, er staat niet wat er staat*. Een rij uit de matrix representeert bijvoorbeeld deze alinea in de matrix, waarbij de kolommen corresponderen met de woorden *lees*, *maar*, *er*, *staat*, *niet*, *wat*, met respectievelijk de waarden 1, 1, 2, 2, 1, 1, want *lees* en *maar* komen elk 1x voor en *er* en *staat* 2x, *niet* 1x en *wat* 1x. Aangezien een corpus groter is dan zes woorden en niet elk woord uit het corpus in elke alinea voorkomt, ontstaat er dus een enorme matrix met enorm veel lege cellen. Omdat we geïnteresseerd zijn in de zinvolle informatie, wordt deze gigantische matrix gefilterd door middel van een decompositietechniek van singuliere waarden, waarbij het aantal dimensies van de matrix wordt gereduceerd tot ongeveer 300. De 'samenvatting' die nu is ontstaan heeft elk woord en elke alinea vertaald naar een vector in een semantische ruimte. De afstand tussen de vectoren, en daarmee de semantische afstand tussen woorden en alinea's, wordt berekend door de cosinus te nemen tussen de betreffende vectoren. Het reduceren van dimensies tot ongeveer 300 blijkt optimaal te zijn voor het berekenen van semantische relaties: niet te veel (waardoor elk woord een unieke semantische relatie heeft met elk ander woord) en niet te weinig (waardoor elk woord vrijwel dezelfde semantische relatie heeft met een ander woord). Neem ter illustratie de volgende zinnen:

1. De hond rende rondom de bomen in het park.
2. De kat klom in de bomen van het park.
3. De eekhoorn sprong van tak naar tak.

Het is niet moeilijk om te aan te nemen dat *hond* en *kat* in de eerste twee voorbeelden semantisch gerelateerd zijn aan elkaar, omdat ze vrijwel dezelfde context hebben (*bomen* en *park*). In de praktijk werkt LSA het beste met inhoudswoorden, omdat deze het meeste afhankelijk zijn van een semantische context (en dus betekenis hebben). Functiewoorden, zoals *rondom* en *in*, hebben een zeer hoge frequentie en zo'n variërende context dat deze doorgaans minder goed werken, ofschoon er theoretisch geen reden is dat zij niet kunnen werken.

Maar de semantische relatie in LSA is niet beperkt tot de relaties tussen woorden en hun context, zoals in voorbeeld 1) en 2). Het betreft ook de relaties tussen de woorden die de buren zijn van andere woorden. *Kat* en *eekhoorn* in de voorbeelden hierboven hebben geen enkele context met elkaar gemeen, maar de context van de context (bijvoorbeeld de woorden die samengaan met *bomen* en de woorden die samengaan met *takken*) overlappen wel, waardoor *kat* en *eekhoorn* wel een semantische relatie kunnen hebben. LSA berekent dus de afstand van de context (van de context van de context van de context van de context, etc.) van woorden. Het kan hetzelfde doen voor zinnen, alinea's en zelfs hele teksten en kan daarmee de inhoud van woorden, zinnen, alinea's en teksten berekenen.

Deze techniek om statistisch een representatie te geven van kennis blijkt uiterst vruchtbaar te zijn. Landauer en Dumais (1997) evalueerden bijvoorbeeld of LSA zou slagen voor de Test of English as a Foreign Language (TOEFL) die elke buitenlander moet doen om toegelaten te worden aan een Amerikaanse universiteit. In dit examen moet het juiste synoniem bij een woord worden gezocht. Op 80 meerkeuzevragen gaf LSA in 64% van de gevallen het juiste antwoord, even goed als de gemiddelde student die het examen doet. Maar LSA kan meer dan de synoniemen van woorden vinden. Landauer, Foltz en Laham (1998) trainden LSA met tekstboeken psychologie om te zien hoe goed het in staat zou zijn om het juiste antwoord te vinden in meerkeuze-examens die gebruikt worden in colleges. LSA deed het even goed als een gemiddelde student, niet geweldig, maar het wist een examen te halen. In een vervolgstudie toonden Landauer et al. (1998) aan dat LSA beter de inhoud kon beoordelen dan universiteitsdocenten. Bovendien was LSA in staat plagiaat te identificeren, ook in de gevallen dat teksten niet letterlijk over waren geschreven.

LSA wordt ook gebruikt in *Summary Street*, een lees- en schrijfvaardigheidsprogramma dat de kwaliteit beoordeelt van samenvattingen die studenten schrijven (Wade-Stein & Kintsch 2004). Verder dient LSA als model van het langetermijngeheugen van kunstmatig-intelligente docenten, zoals *AutoTutor* en *iSTART*. *AutoTutor* (Graesser et al., 2004; Louwerse, Graesser en Olney, 2002) heeft gesprekken met studenten zoals een menselijke tutor die zou hebben. LSA evalueert wat de student zegt en houdt daarmee de conversatie gaande én beoordeelt de kennis van de student. *iSTART* gebruikt LSA in het onderwijzen van leesstrategieën, waarbij LSA gebruikt wordt in de beoordeling van de antwoorden van studenten (McNamara, Levinstein & Boonthu 2004). LSA wordt ook gebruikt in *Coh-Metrix*, een systeem dat de coherentie van teksten berekent met meer dan

honderd verschillende maatstaven (Graesser, McNamara, Louwerse & Cai 2004). Louwerse (2004) gebruikte LSA om het idiolect en sociolect van literaire schrijvers te evalueren door te kijken naar de coherentie in verschillende literaire teksten. Kintsch (2002) gebruikte LSA voor het identificeren van thema's en subthema's in teksten en zelfs voor het analyseren van de betekenis van metaforen (Kintsch 2000).

Hierboven zijn verschillende voorbeelden gegeven van toepassingen van LSA op de Engelse taal. In de volgende twee secties richten we ons op het Nederlands. Dit laatste is met name van belang, omdat voor zover ons bekend LSA niet tot nauwelijks gebruikt is voor het Nederlands. Uitzonderingen zijn Bestgen, Degand, Spooren (2006) die de techniek gebruikten voor een automatische identificatie van voegwoorden en Van Bruggen, Rusman, Giesbers, & Koper (ingediend) die de mogelijkheden van LSA hebben onderzocht voor onderwijs op het Internet. Voor de Nederlandse voorbeelden die volgen richten we ons zowel op taal- en tekstwetenschap (2.2) en literatuur (2.3).

2.2 Waar het over gaat in woorden

Voor het eerste voorbeeld richten we ons op Nederlandstalige woorden. Een 300-dimensionele LSA-ruimte werd geconstrueerd van teksten die een totaal van 45480 verschillende woorden en 24607 alinea's bevatten (de totale grootte van het document was 11 MB). Dit corpus komt overeen met 3500 bladzijden van een artikel in *TNTL*. Zo'n corpus is relatief klein vergeleken met de Engelstalige corpora waar eerder over gesproken werd. De gebruikte teksten voor dit corpus zijn afkomstig van het Eindhoven-corpus (inl.nl), *de Volkskrant* (volkskrant.nl) en verschillende in het Nederlands vertaalde teksten uit het Gutenberg-corpus (gutenberg.org), waaronder omvangrijke werken van Tolstoy en Jules Verne.

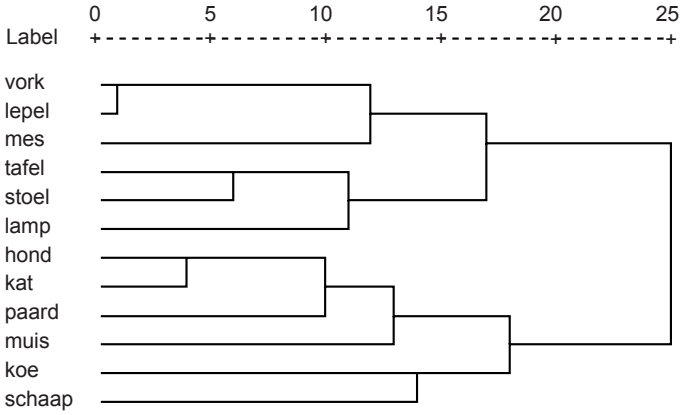
Nadat de semantische ruimte is geconstrueerd kunnen nu de semantische afstanden worden berekend door de cosinus te nemen tussen vectoren. Is deze laag (minimaal -1) dan staan woorden semantisch ver van elkaar, is deze hoog (maximaal 1) dan staan woorden semantisch dicht bij elkaar. In de praktijk komt het nooit voor dat een LSA cosinus waarde van -1 wordt verkregen, aangezien een woord indirect altijd een relatie heeft met een ander woord, simpelweg vanwege het feit dat de twee woorden ergens in het corpus voorkomen en ergens in de semantische ruimte de buren (van de buren van de buren etc.) semantisch overlappen. Een waarde van 1 wordt anderzijds verkregen door de relatie tussen twee identieke woorden, die per definitie altijd in de dezelfde documenten voorkomen.

Ter illustratie gebruiken we hier een twaalftal Nederlandse woorden: *koe*, *paard*, *schaap*, *hond*, *kat*, *muis*, *tafel*, *stoel*, *lamp*, *vork*, *mes* en *lepel* en berekenen we hun betekenis op basis van de cosinuswaarden uit het Nederlandstalige corpus. De semantische relaties tussen die woorden zijn gegeven in Tabel 1. In deze tabel staan relaties die zinnig zijn (een *tafel* is semantisch het sterkst gerelateerd aan *stoel*), maar ook die onzinnig zijn (een *koe* is het sterkst gerelateerd aan *lamp*, maar bijvoorbeeld niet aan *paard*).

Tabel 1 LSA cosinus waarden tussen twaalf woorden.

	koe	paard	schaap	hond	kat	muis	tafel	stoel	lamp	vork	mes	lepel
koe	1,00	-0,05	0,09	0,07	0,13	0,09	0,08	-0,08	0,13	0,02	-0,07	0,13
paard	-0,05	1,00	0,05	0,27	0,09	0,05	-0,12	-0,04	-0,11	0,03	0,01	-0,05
schaap	0,09	0,05	1,00	0,04	0,08	-0,05	-0,01	-0,13	0,00	-0,05	-0,06	0,03
hond	0,07	0,27	0,04	1,00	0,32	0,12	-0,01	-0,07	-0,12	0,06	0,04	-0,06
kat	0,13	0,09	0,08	0,32	1,00	0,20	-0,07	-0,05	0,02	-0,03	-0,02	-0,05
muis	0,09	0,05	-0,05	0,12	0,20	1,00	-0,01	-0,05	0,02	0,01	-0,02	-0,02
tafel	0,08	-0,12	-0,01	-0,01	-0,07	-0,01	1,00	0,28	0,25	0,24	0,14	0,29
stoel	-0,08	-0,04	-0,13	-0,07	-0,05	-0,05	0,28	1,00	0,11	0,02	0,01	0,07
lamp	0,13	-0,11	0,00	-0,12	0,02	0,02	0,25	0,11	1,00	0,00	-0,16	0,18
vork	0,02	0,03	-0,05	0,06	-0,03	0,01	0,24	0,02	0,00	1,00	0,30	0,42
mes	-0,07	0,01	-0,06	0,04	-0,02	-0,02	0,14	0,01	-0,16	0,30	1,00	0,06
lepel	0,13	-0,05	0,03	-0,06	-0,05	-0,02	0,29	0,07	0,18	0,42	0,06	1,00

Een dergelijke tabel mag er dan misschien indrukwekkend uitzien, maar hij zegt niet zo veel. De verhoudingen worden pas duidelijk wanneer de LSA resultaten worden geanalyseerd door middel van een techniek zoals Hierarchical Clustering (Louwerse et al., 2006), waarbij alle relaties tussen alle woorden in beeld kunnen worden gebracht. Deze verhoudingen zijn weergegeven in Figuur 1, waarbij de lengte van de horizontale lijnen de afstand bepalen.



Figuur 1 Hiërarchische clustering van LSA-cosinus-waarden tussen twaalf woorden.

Bestudering van de hiërarchische relaties in Figuur 1 toont onmiddellijk een tweedeling aan tussen dieren en objecten. Bovendien is binnen de laatste groep een tweedeling zichtbaar tussen eetgerei en meubilair. Deze classificatie lijkt volledig overeenkomstig die in de wereld. In de groep dieren lijken vervolgens graasdieren zich te onderscheiden van de andere dieren. *Muis* heeft daarbij een uitzonderlijke positie, maar deze moet waarschijnlijk worden toegeschreven aan de gedeelde rol met zijn electronische naamgenoot. Het volstaat hier echter te zien dat LSA semantische relaties blootlegt die ook in de wereld om ons heen waar te nemen zijn.

Het bovenstaande mag weliswaar een interessante illustratie zijn dat LSA ook voor het Nederlands gebruikt kan worden, het levert echter geen bewijs dat LSA een natuurgetrouwe simulatie biedt. Om een aanzet te geven tot dat bewijs maakten we gebruik van de gegevens die Ruts et al. (2004) rapporteren in een studie waarbij 2100 proefpersonen werd gevraagd woorden neer te schrijven die associeerden met een stimulus. Zo leverden het woord *kabeljauw* 61 keer de associatie *vis* op, *bakker* 57 keer de associatie *brood*, *psycholoog* 5 keer *dokter*, en *bloes* 6 keer *knopen*. Deze frequenties kunnen gebruikt worden om te toetsen in hoeverre LSA tot vergelijkbare resultaten komt. Voor elk van de 425 woordparen bestaande uit stimulus en associatie werd de LSA cosinus berekend op basis van de Nederlandstalige semantische ruimte. De resultaten van LSA correleerden met de experimentele data van Ruts et al. ($r = .12$, $p = .01$, $N = 425$), een mate die niet aan toeval kan worden toegeschreven.

Een groot aantal woorden kwam echter niet voor in de corpora waarop we de LSA-ruimte hadden getraind (bijvoorbeeld woorden als *informaticus* en *eddywally*). Daarom werden vervolgens alle woordparen verwijderd waarvoor geen cosinus kon worden berekend doordat een woord niet in het corpus voorkwam. Voor 201 woorden was dit het geval. Een analyse op basis van de overgebleven woorden gaf opnieuw een significante correlatie aan ($r = .21$, $p < .01$, $N = 224$). Deze resultaten tonen aan dat de resultaten van experimentele data van proefpersonen overeenkomen met die van LSA, ook voor het Nederlands.

2.3 Waar het over gaat in literatuur

Verschillende LSA-voorbeelden zijn eerder gegeven voor taal. De vraag is in hoeverre LSA ook, en in het bijzonder, van toepassing kan zijn op literatuur. We hebben recentelijk aangetoond dat een model als LSA evenzeer haar vruchten afwerpt in literaire analyses (Louwerse en Van Peer, in druk) door voorbeelden te nemen uit Stockwell (2002) en een LSA-analyse te vergelijken met een embodiment analyse, zoals gegeven door Stockwell. In een van de LSA-analyses keken we bijvoorbeeld naar de LSA-waarden tussen woorden als *Chaucer*, *Dante*, *Dickens*, *Faulkner*, *Joyce*, *Shakespeare* en *Woolf*. Een hiërarchische clustering vergelijkbaar met die hierboven gepresenteerd, gaf aan dat *Chaucer* en *Dante* semantisch het dichtst bij elkaar staan, *Shakespeare* en *Dickens* zijn vervolgens het meest gerelateerd aan die eerste twee. *Woolf*, *Joyce* en *Faulkner* staan daar respectievelijk het verst vanaf, opmerkelijk genoeg. LSA weet auteurs dus goed in literaire periodes te plaatsen, louter op basis van de namen van de auteurs, woorden dus die erg vaak samen voorkomen, of meer technisch uitgedrukt: waartussen slechts een kleine semantische afstand bestaat. Daarbij moet worden aangemerkt dat het corpus geen literatuurgeschiedenissen bevatte die een dergelijk resultaat minder opzienbarend zouden maken.

Wat natuurlijk opmerkelijker zou zijn dan de semantische afstanden van de *namen* van auteurs zijn de semantische afstanden van de *teksten* zelf. En voor een Nederlandstalig tijdschrift zou het bovendien aardig zijn Nederlandse literatuur te vergelijken. Dat is precies wat we hebben gedaan in de volgende analyse. De onderzoeksvraag die we daarbij stellen is of LSA in staat is Nederlandse literaire teksten te zinvol categoriseren op basis van (literaire) periodes.

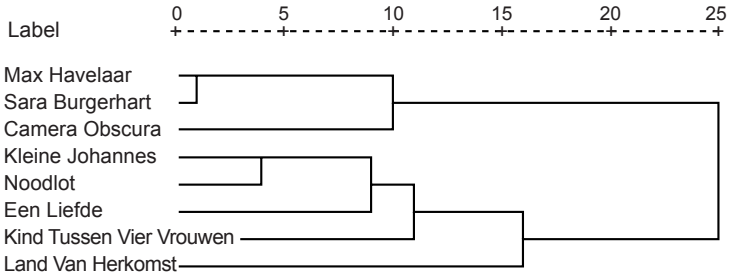
Om deze vraag te beantwoorden werd dezelfde LSA-ruimte gebruikt als die van de vorige sectie waarin we Nederlandstalige woorden vergeleken. Acht Nederlandstalige literaire teksten werden gekozen: Couperus' *Noodlot* (1890), Du Perrons *Het land van herkomst* (1935), Hildebrands *Camera Obscura* (1839), Multatuli's *Max Havelaar* (1860), Van Deyssels *Een liefde* (1887), Van Eedens *Kleine Johannes* (1887), Vestdijks *Kind tussen vier vrouwen* (1933), en Wolff en Dekens *Sara Burgerhart* (1782). Elektronische versies van deze teksten werden verkregen via Project Gutenberg (gutenberg.org) of door het inscannen van de boeken. Vervolgens werden de LSA-cosinus-waarden berekend tussen de volledige inhoud van alle werken. Dus waar in de analyse uit 2.2 de cosinus tussen twee woorden werd berekend, werd in de huidige analyse de cosinus berekend tussen (*alle* woorden van) twee teksten.

Het is daarbij van belang te melden dat het allerminst noodzakelijk is dat de teksten dezelfde woorden bevatten. Immers, LSA brengt relaties van een hogere orde tot stand door niet te kijken naar specifieke woorden, maar de context (van de context van de context etc.) van die woorden. De vergelijking van acht teksten resulteerde in een 8 x 8 matrix die in Tabel 2 is weergegeven. Zoals te zien is, blijkt *Sara Burgerhart* semantisch veel te lijken op de *Max Havelaar* (.56), terwijl de laatste ook veel lijkt op de *Camera Obscura* (.54). Maar *Kind tussen vier vrouwen* toont de meeste semantische verwantschap met *De kleine Johannes* (.45), terwijl *Het land van herkomst* de meeste verwantschap toont met *Camera Obscura* (.34). Evenals de vergelijking tussen individuele woorden, zeggen deze waarden echter niet zoveel. Ze worden betekenisvoller wanneer groepen semantische waarden in verhouding tot elkaar worden beschouwd. Net zoals in de vorige analyse werden de waarden daarom gegroepeerd in een hiërarchische clustering. Het resultaat daarvan is gepresenteerd in Figuur 2.

Tabel 2 Cosinus-waarden van de inhoud van acht Nederlandstalige literaire teksten

	1	2	3	4	5	6	7	8
1. *Camera obscura*	1,00	0,47	0,26	0,34	0,54	0,38	0,41	0,42
2. *De kleine Johannes*	0,47	1,00	0,45	0,23	0,17	0,47	0,04	0,33
3. *Kind tussen vier vrouwen*	0,26	0,45	1,00	0,30	0,01	0,26	0,01	0,32
4. *Het land van herkomst*	0,34	0,23	0,30	1,00	0,12	0,20	0,22	0,28
5. *Max Havelaar*	0,54	0,17	0,01	0,12	1,00	0,13	0,56	0,14
6. *Noodlot*	0,38	0,47	0,26	0,20	0,13	1,00	0,17	0,40
7. *Sara Burgerhart*	0,41	0,04	0,01	0,22	0,56	0,17	1,00	0,15
8. *Een liefde*	0,42	0,33	0,32	0,28	0,14	0,40	0,15	1,00

Twee categorieën kunnen allereerst worden onderscheiden. De eerste groepeert *Max Havelaar*, *Sara Burgerhart*, en *Camera Obscura*, de tweede de overige werken (*De Kleine Johannes*, *Noodlot*, *Een liefde*, *Kind tussen vier vrouwen*, en *Het land van herkomst*). Hoewel we binnen het bestek van dit artikel aarzelen literaire periodes toe te kennen, kan het argument worden gemaakt dat deze tweedeling een scheiding weergeeft tussen Realisme enerzijds en Naturalisme/Modernisme anderzijds (vgl. Schenkeveld-Van der Dussen 1993). Bovendien is het meest Modernistische werk van de reeks van acht, *Het land van herkomst*, het verst gelegen van de overige werken, maar het meest gerelateerd aan *Kind tussen vier vrouwen*, doorgaans ook Modernistisch geïnterpreteerd (Fokkema & Ibsch 1987).



Figuur 2. Hiërarchische clustering van LSA*-cosinus-waarden tussen acht Nederlandstalige literaire teksten.*

Maar ook als we literaire periodes terzijde leggen brengt de groepering een interessant patroon naar voren. Als de acht werken op jaar van publicatie worden geordend, komen patronen tot stand die vergelijkbaar zijn met die in de hiërarchische groepering: (1782-1839-1860), (1887- 1887-1890), (1933), (1935). De mathematische techniek achter LSA staat niet toe dat dit patroon verklaard zou kunnen worden op basis van taalverandering over de jaren heen. LSA is namelijk vrijwel ongevoelig voor specifieke woorden omdat het semantische relaties van een hogere orde berekent. Het is daarnaast de vraag of die taalverandering zo snel zou verlopen en zich uitgerekend in deze literaire werken zou manifesteren Bovendien is 9,02% (SD=0,048) van alle woordtypen (unieke woorden) terug te vinden in alle acht teksten en is het niet zo dat oudere (of nieuwere) teksten meer woorden gemeen hebben dan nieuwere (of oudere). Bijvoorbeeld hebben zowel *Camera Obscura* als *Kind tussen vier vrouwen* 4% van de woorden gemeen met de overige werken en *Sara Burgerhart* en *Noodlot* respectievelijk 11% en 14%. Een andere, interessantere, verklaring moet dus worden gezocht, een die we in dit artikel naar voren hebben gebracht: de relaties tussen de inhoud van verschillende literaire werken kunnen louter op basis van kwantitatieve computertechnieken worden bepaald.

3 Conclusie

Aan het begin van dit artikel stelden we dat het met de empirie in letterkunde en literatuurwetenschap niet al te best gesteld is. Deze identiteitscrisis is vreemd, aangezien ze niet bij andere wetenschappen voorkomt en verbazingwekkend, omdat onderzoek overtuigend heeft aangetoond dat psychologische experimenten en computationele modellen van grote waarde zijn voor de beantwoording van talloze onderzoeksvragen. We hebben vervolgens een kort overzicht gegeven van de geschiedenis van de inhoudsanalyse van de afgelopen decennia, daarbij de lezer waarschuwend dat binnen het bestek van dit artikel dit niet veel meer dan een subjectieve momentopname kan zijn. Voor een uitvoeriger overzicht van thematische analyses zij de lezer verwezen naar Louwerse en Van Peer (2002).

Als succesvolle kwantitatieve methode voor inhoudsanalyse hebben we ons toegespitst op LSA, een statistische techniek die op basis van de context van woorden semantische relaties tussen woorden, zinnen, alinea's en teksten kan berekenen. Voor zover ons bekend zijn er geen studies waarin LSA gebruikt is voor inhoudsanalyses

in het Nederlands. In onze bijdrage hebben we taalkundige en letterkundige voorbeelden gegeven. In het eerste voorbeeld werd LSA vergeleken met woordassociaties die experimenteel zijn verkregen (Ruts et al. 2004). In het tweede hebben we de inhoud vergeleken van acht Nederlandstalige literaire teksten. Beide analyses waren exploratief van aard, maar hun resultaten nodigen uit tot verder onderzoek.

Analyses zoals hier gepresenteerd roepen talloze vragen op die ons van belang lijken te zijn voor letterkunde en literatuurwetenschap. Een selectie van vragen: hoe verhouden de verschillende hoofdstukken binnen een boek zich tot elkaar (bijv. *Het land van herkomst, Max Havelaar*)? Hoe verhouden verschillende werken van auteurs zich tot elkaar (bijv. *Kind tussen vier vrouwen* in verhouding tot de verschillende Anton Wachter-romans? Kunnen genres onderscheiden worden binnen (en tussen) literaire werken (bijv. *Het land van herkomst,* en *Sara Burgerhart*)? Kan de toegankelijkheid van bepaalde literaire werken berekend worden op basis van de toegankelijkheid van een ander literair werk? Kan de kwaliteit van samenvattingen van literaire werken computationeel beoordeeld worden? De kwaliteit van recensies? Kunnen thema's van literaire werken objectief geabstraheerd worden? Hoeveel groepen van literaire werken kunnen worden geconstrueerd? Wat is de intertekstualiteit van bepaalde literaire werken? Hoe verhouden literaire werken zich semantisch tot niet-literaire werken? Antwoorden op deze en tal van andere vragen kunnen worden berekend door middel van statistische technieken zoals we die hier hebben gepresenteerd. Cijfers als begin van een antwoord waar het over gaat in literatuur is waar het onzes inziens ook om zou moeten gaan in letterkunde en literatuurwetenschap.

Bibliografie

Bestgen, Degand, & Spooren 2006 – Y. Bestgen, L. Degand, W. Spooren: 'Toward automatic determination of the semantics of connectives in large newspaper corpora'. In: *Discourse Processes* 41 (2006), 175-193.

Bremond 1973 – C. Bremond: *Logique du récit*. Paris: Seuil, 1973.

Van den Broek et al. 1996 – P. van den Broek, K. Risden, C.R. Fletcher & R. Thurlow: 'A "landscape" view of reading: fluctuating patterns of activation and the construction of a stable memory representation'. In: B.K. Britton & A.C. Graesser: *Models of understanding text*. Mahwah: Lawrence Erlbaum, 1996, 165-187.

Van Bruggen, Rusman, Giesbers & Koper (ingediend) – J.M. van Bruggen, E. Rusman, B. Giesbers & R. Koper: 'Latent Semantic Analysis of small-scale corpora for positioning in learning networks'.

Van Dijk 1972 – T.A.van Dijk: *Some aspects of text grammars. A study in theoretical linguistics and poetics*. The Hague: Mouton, 1972.

Van Dijk & Kintsch 1983 – T.A. van Dijk & W. Kintsch: *Strategies of discourse comprehension*. New York: Academic Press, 1983.

Fokkema & Ibsch 1987 – D. Fokkema & E. Ibsch: *Modernist conjectures. A mainstream in European literature 1910-1940*. London: Hurst, 1987.

Gernsbacher 1990 – M.A. Gernsbacher: *Language comprehension as structure building*. Hillsdale: Erlbaum, 1990.

Gernsbacher 1994 – M.A. Gernsbacher: *Handbook of psycholinguistics*. San Diego, CA: Academic Press, 1994.

Graesser, Gernsbacher & Goldman 2002 – A.C. Graesser, M.A. Gernsbacher & S.J. Goldman (Red.): *Handbook of discourse processes*. Mahwah, NJ: Erlbaum, 2002.

Graesser, Lu, Jackson, Mitchell, Ventura, Olney & Louwerse 2004 – A.C. Graesser, S. Lu, G.T. Jack-

son, H. Mitchell, M. Ventura, A. Olney & M.M. Louwerse: 'AutoTutor: A tutor with dialogue in natural language'. In: *Behavioral Research Methods, Instruments, and Computers* 36 (2004), 180-193.

Graesser, McNamara, Louwerse & Cai 2004 – A.C. Graesser, D. McNamara, M.M. Louwerse & Z. Cai: 'Coh-Metrix: Analysis of text on cohesion and language'. In: *Behavior Research Methods, Instruments, and Computers* 36 (2004), 193-202.

Graesser, Millis & Zwaan 1997 – A.C. Graesser, K.K. Millis & R.A. Zwaan: 'Discourse comprehension'. In: *Annual Review of Psychology* 48 (1997), 163-89.

Gutenberg corpus 2006 – http://www.gutenberg.org. Opgevraagd 20 juni, 2006.

Hakemulder 2000 – F. Hakemulder: *The moral laboratory; Experiments examining the effects of reading literature on social perception and moral self-knowledge.* Amsterdam: Benjamins, 2000.

Jurafsky & Martin 2000 – D. Jurafsky & J.H. Martin: *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, NJ: Prentice Hall, 2000.

Kintsch 1974 – W. Kintsch: *The representation of meaning in memory*. Hillsdale, NJ: Erlbaum, 1974.

Kintsch 1988 – W. Kintsch: 'The role of knowledge in discourse comprehension: a construction-integration model'. In: *Psychological Review* 95 (1988), 163-182.

Kintsch 1998 – W. Kintsch: *Comprehension: A paradigm for cognition*. New York: Cambridge University Press, 1998.

Kintsch 2000 – W. Kintsch: 'Metaphor comprehension: A computational theory'. In: *Psychonomic Bulletin and Review* 7 (2000), 257-266.

Landauer & Dumais 1997 – T.K. Landauer & S.T. Dumais: 'A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge'. In: *Psychological Review* 104 (1997), 211-240.

Landauer, Foltz & Laham 1998 – T.K. Landauer, P.W. Foltz & D. Laham: 'An introduction to latent semantic analysis'. In: *Discourse Processes* 25 (1998), 259-284.

Landauer, McNamara, Dennis & Kintsch – T. Landauer, D. McNamara, S. Dennis & W. Kintsch: *LSA: A road to meaning*. Mahwah, NJ: Erlbaum.

Lévi-Strauss 1958 – C. Lévi-Strauss: *Anthropologie structurale*. Paris, Plon, 1958.

Louwerse, Cai, Hu, Ventura, Jeuniaux 2007 – M.M. Louwerse, Z. Cai, X. Hu, M. Ventura, P. Jeuniaux: 'Cognitively inspired natural-language based knowledge representations: Further explorations of Latent Semantic Analysis'. In: *International Journal of Artificial Intelligence Tools*, 15 (2006), 1021-1040.

Louwerse, Graesser, Olney & Tutoring Research Group 2002 – M.M. Louwerse, A.C. Graesser, A. Olney & Tutoring Research Group: 'Good computational manners: Mixed-initiative dialog in conversational agents'. In: C. Miller (red.): *Etiquette for human-computer work. Papers from the 2002 Fall Symposium, Technical Report FS-02-02*. Menlo Park, CA: AAAI Press, 2002, 71-76.

Louwerse & Kuiken 2005 – M.M. Louwerse en D. Kuiken (red.): *The effects of personal involvement in narrative discourse.* Themanummer *Discourse Processes* 38 (2005).

Louwerse & Van Peer 2002 – M.M. Louwerse & W. Van Peer (red.): *Thematics: Interdisciplinary studies*. Philadelphia, John Benjamins, 2002.

Louwerse 2007 – M.M. Louwerse: 'Iconicity in amodal symbolic representations'. In: T. Landauer, D. McNamara, S. Dennis & W. Kintsch (Red.): *LSA: A road to meaning*. Mahwah, NJ: Erlbaum.

Louwerse & Van Peer, in druk – M.M. Louwerse & W. Van Peer: 'How cognitive is cognitive poetics? The interaction between symbolic and embodied cognition'. In: G. Brone & J. Vandaele (Red.): *Cognitive Poetics*. Berlin, Germany: De Gruyter.

Louwerse 1997 – M.M. Louwerse: 'Inleiding'. In: Propp 1997.

Louwerse 2004 – M.M. Louwerse: 'Semantic variation in idiolect and sociolect: Corpus linguistic evidence from literary texts'. In: *Computers and the Humanities* 38 (2004), 207-221.

Lund & Burgess 1996 – K. Lund & C. Burgess: 'Producing high-dimensional semantic spaces from lexical co-occurrence'. In: *Behavior Research Methods, Instrumentation, and Computers* 28 (1996), 203-208.

Manning & Schütze 1999 – C. Manning & H. Schütze: *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press, 1999.

McNamara, Levinstein & Boonthum 2004 – D.S. McNamara, I.B. Levinstein & C. Boonthum: 'i-START: Interactive strategy trainer for active reading and thinking'. In: *Behavioral Research Methods, Instruments, and Computers* 36 (2004), 222-233.

Miall & Kuiken 1994 – D.S. Miall & D. Kuiken: 'Beyond text theory: Understanding literary response'. In: *Discourse Processes* 17 (1994), 337-352.
Newell & Simon 1972 – A. Newell & H.A. Simon: *Human problem solving*. Englewood Cliffs, NJ: Prentice Hall, 1972.
Pecher & Zwaan 2005 – D. Pecher & R.A. Zwaan (Red.): *Grounding cognition: The role of perception and action in memory, language, and thinking*. New York: Cambridge University Press, 2005.
Van Peer 1986 – W. Van Peer: *Stylistics and psychology; Investigations of foregrounding.* London: Croom Helm, 1986.
Van Peer 2007 – W. Van Peer: 'Introduction. Thematisch nummer over "foregrounding"'. *Language and Literature*.
Van Peer, Zyngier & Hakemulder, in druk – W. Van Peer, S. Zyngier & F. Hakemulder: 'Foregrounding: past, present, future'. In: David Hoover (Ed.): *Prospect and retrospect. Papers from the Poetics and Linguistics Association International Conference, New York, 2004*. Amsterdam: Rodopi.
Propp 1997/1928 – Vladimir Propp, *De morfologie van het toversprookje. Vormleer van een genre* [The Morphology of the folktale. Formal study of a genre; 1928]. Utrecht, Het Spectrum. [transl. M.M. Louwerse], 1997.
Rumelhart & McClelland 1986 – D.E. Rumelhart & J.L. McClelland: *Parallel distributed processing. Explorations in the microstructure of cognition.* Cambridge, MA: MIT Press, 1986.
Rumelhart 1977 – D.E. Rumelhart: 'Understanding and summarizing brief stories'. In: D. LaBerge and S.J. Samuels (Red.): *Basic processes in reading: Perception and comprehension*. Hillsdale, NJ: Erlbaum, 1977, 265-303.
Ruts et al. 2004 – W. Ruts, S. De Deyne, E. Ameel, W. Vanpaemel, T. Verbeemen en G. Storms: *Behavior Research Methods, Instruments & Computers* 36 (2004), 506-515.
Schank & Abelson 1977 – R.C. Schank & R.P. Abelson: *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Hillsdale, NJ: Erlbaum, 1977.
Schenkeveld-Van der Dussen 1993 – M.A. Schenkeveld-Van der Dussen (Red.): *Nederlandse literatuur: Een geschiedenis*. Groningen: Nijhoff, 1993.
Searle 1980 – J.R. Searle: 'Minds, brains, and programs'. In: *Behavioral and Brain Sciences* 3 (1980), 417-57.
Semino & Culpeper 2002 – E. Semino & J. Culpeper: *Cognitive stylistics: Language and cognition in text analysis*. Philadelphia: John Benjamins, 2002.
Stockwell 2002 – P. Stockwell: *Introduction to cognitive poetics*. London: Routledge, 2002.
Wade-Stein & Kintsch 2004 – D. Wade-Stein & W. Kintsch: 'Summary Street: Interactive computer support for writing'. In: *Cognition and Instruction* 22 (2004), 333-362.
Zwaan 1993 – R.A. Zwaan: *Aspects of literary comprehension: A cognitive approach*. Philadelphia: John Benjamins, 1993.
Zwaan & Radvansky 1998 – R.A. Zwaan & G.A. Radvansky: 'Situation models in language comprehension and memory'. In: *Psychological Bulletin* 123 (1998), 162-185.
Zyngier, Van Peer & Hakemulder, in druk – S. Zyngier, W. Van Peer & F. Hakemulder: 'Love in literature. Complexity, foregrounding, and evaluation'. In: *Poetics Today*.

Correspondentie-adres van de auteurs

Max M. Louwerse, Department of Psychology / Institute for Intelligent Systems, University of Memphis, Psychology Building, Memphis, TN 38152, mlouwerse@memphis.edu