Computerondersteund beoordelen van essays

Computerondersteund beoordelen van essays

Document

Identificatie	OTEC 2002-1 COMPUTERONDERSTEUND BEOORDELEN VAN ESSAYS1
U-nummer	
Status	Definitief
Soort document	SJB
Auteur(s)	Bruggen, Jan van
Datum afdruk	
Opgeslagen	

Goedkeuring

Acroniem	Handtekening	Datum

Wijzigingshistorie			
Versie	**Acroniem**	**Datum**	**Wijziging**
0.1			

Distributie		
Versie	**Datum**	**Naam**
0.1		

Onderwijstechnologisch expertisecentrum OTEC
Open Universiteit Nederland

Computerondersteund beoordelen van essays

OTEC 2002/1

Colofon

Titel:	Computerondersteund beoordelen van essays
Auteurs:	Jan van Bruggen
Projectondersteuning:	Mieke Haemers
Uitgifte:	OTEC
Datum druk:	8 januari 2002

Onderwijstechnologisch expertisecentrum (OTEC)
Open Universiteit Nederland

Computerondersteund beoordelen van essays

Inhoudsopgave

Inleiding en begripsbepalingen

In dit rapport bezien we de stand van zaken op het gebied van computerondersteund beoordelen van essays – antwoorden op essayvragen. We houden ons, met andere woorden, bezig met de vraag hoe de computer de docent kan ondersteunen bij het scoren en beoordelen van antwoorden op essayvragen. Computerondersteuning bij scoring en beoordeling kan zich beperken tot eenvoudige faciliteiten zoals spreadsheets waarmee scores toegekend kunnen worden, totaalscores kunnen worden bepaald, enzovoort. Dat is niet de vorm van computerondersteuning waarover we het hier hebben. In dit rapport gaat het om de vraag in hoeverre de computer kan worden ingezet om de essays te scoren en te beoordelen.

Open vragen worden onderscheiden in korte antwoordvragen en essayvragen. Bij een korte antwoordvraag vult de student één of enkele begrippen in (aanvulvragen, invulvragen, Cloze procedures). We houden ons in dit rapport niet bezig met korte antwoordvragen, maar alleen met essayvragen. De antwoorden op deze vragen, *essays*, vergen antwoorden van enkele zinnen tot enkele pagina's tekst. Essentieel bij dit type vragen is dat er geen 'goed' / 'fout' oordeel aan kan worden verbonden, maar dat het gaat om *gradaties* van kwaliteit. De volgende open vraag komt uit een deel van de GMAT dat tegenwoordig mede door de computer wordt beoordeeld:

> It is unrealistic to expect individual nations to make, independently, the sacrifices necessary to conserve energy. International leadership and worldwide cooperation are essential if we expect to protect the worlds energy resources for future generations..
>
> Discuss the extent to which you agree or disagree with the opinion stated above. Support your views with reasons and/or examples from your own experience, observations, or reading.

Er kan over gediscussieerd worden of essayvragen het middel bij uitstek zijn om inzicht te toetsen: objectieve toetsen zijn daartoe zeker ook in staat. Een argument voor het gebruik van essayvragen dat o.i. meer hout snijdt, is dat veel problemen geen voor ieder aanvaarde analyse en oplossing kennen, maar dat het er vooral op aan komt om een *goed beargumenteerde* analyse en oplossing te formuleren (Voss, Wiley & Sandak, 1999; Voss, 1991; Voss, Blais, Means, Greene & Ahwesh, 1989). Een tweede argument voor essayvragen is de grotere authenticiteit: het vervaardigen van een heldere tekst met een goed onderbouwd standpunt ligt dichter bij de praktijk dan de beantwoording van meerkeuzevragen.

Beoordeling van essays komt neer op *classificatie* van het essay op een of meerdere aspecten. Classificatie kan holistisch zijn, dat wil zeggen dat één totaaloordeel over de kwaliteit van het essay wordt uitgebracht (bijvoorbeeld 'onvoldoende', 'voldoende', of 'goed') of er kan worden beoordeeld op deelaspecten (zoals inhoud, argumentatie, stijl). De Gruijter (1999) spreekt hier van globale en analytische beoordeling. Volgens Foltz, Laham en Landauer (1999) kan een essay op drie eigenschappen worden beoordeeld: de correctheid en volledigheid van de conceptuele kennis, de geldigheid van de argumentatie en schrijfstijl - hoe vloeiend, elegant en begrijpelijk het essay is geschreven.

Aan het gebruik van essayvragen kleven nadelen: in de eerste plaats is beoordeling van essays kostbaar, omdat normaliter meerdere deskundige beoordelaars ingezet moeten worden. In de tweede plaats is de beoordeling vaak niet erg betrouwbaar: zelfs deskundige

beoordelaars blijken het vaak maar matig met elkaar eens te zijn en het vergt extra training en gedetailleerde scoringsvoorschriften om de betrouwbaarheid te verhogen. In de derde plaats is de beoordeling van essays tijdrovend. Gegeven de kosten en tijd die de beoordeling vergt, is het niet verwonderlijk dat essayvragen vooral worden gebruikt in een toetsingssituatie waarin een summatief oordeel wordt uitgebracht (een tentamen). Formatief gebruik, waarbij de docent(en) essays beoordelen en inhoudelijke feedback geven op grond waarvan de student een nieuwe versie vervaardigt, vergt doorgaans meer tijd en geld dan beschikbaar zijn. Juist hier kan (snelle) computerondersteunde beoordeling van essays interessante nieuwe mogelijkheden bieden. Daarbij is het dan wel gewenst dat de terugkoppeling naar de student zich niet beperkt tot een mededeling over de beoordeling, maar aanwijzingen bevat aan de hand waarvan de student kan werken aan een verbeterd essay.

Hoewel we ons richten op systemen waarbij de computer een beoordeling van essays oplevert, spreken we toch liever van computer*ondersteund* beoordelen dan van computergebaseerd oordelen of van beoordeling door de computer. In de eerste plaats wordt de beoordeling in de praktijk zelden alleen aan de computer overgelaten. In de CODAS nakijkhulp bijvoorbeeld levert de computer een classificatie van antwoorden die door de docent wordt gecontroleerd en zonodig gecorrigeerd, waarna de docent met het systeem interactief en iteratief de overige beoordelingen vastlegt. Het E-rater systeem wordt ingezet als tweede beoordelaar bij een standaardtest, maar de menselijke beoordelaar blijft nodig om de resultaten te controleren. In de tweede plaats vergen beoordelingen door de computer niet onaanzienlijke voorbewerkingen: de docent dient de computer voorbeelden van eerder beoordeelde essays te leveren; bij andere systemen dient de computer ‘te leren’ van ander materiaal over het domein. In de derde plaats vergt beoordeling door de computer nog de nodige verfijning in de parameters die de computer hanteert om te komen tot een instelling waarbij verdere beoordeling aan de computer kan worden overgelaten

In het rapport beschrijven we enkele systemen en technieken voor het beoordelen van essays met de computer. Buiten beschouwing blijven zaken als: invoermogelijkheden; spellingcorrectie of technieken om spelfouten te negeren; opslag van (versies van) de antwoorden; opslag van de analyseresultaten en terugkoppeling.

Of de hier behandelde systemen bruikbaar zijn voor de OUNL hangt uiteraard af van de eisen die de OUNL aan dergelijke systemen wil stellen. Aanvankelijk zijn de eisen die aan de beoordeling van antwoorden in vrije tekst zijn gesteld, vooral gericht geweest op beoordeling en het geven van terugkoppeling in interactieve settings (zie bijlage I). In dit rapport bezien we ook gebruiksmogelijkheden in niet-interactieve settings.

Typering van benaderingen

Het *ideale* systeem voor de beoordeling van inhoud, argumentatie en schrijfstijl vergt geavanceerde technieken voor natuurlijke taalverwerking (NLP – Natural Language Processing) waarmee de tekst eerst grammaticaal en syntactisch wordt ontleed. De semantiek van de herkende tekstfragmenten moet worden vastgelegd en grotere tekstgehelen moeten worden gecombineerd tot, bijvoorbeeld, retorische structuren om de opbouw van de argumentatie weer te geven en te beoordelen. Bij dat alles dient het ideale systeem te beschikken over domeinkennis om de conceptuele kennis van de student te kunnen beoordelen en over linguïstische kennis om een oordeel te kunnen geven over de schrijfstijl van de student. Tenslotte, zou het ideale systeem in staat zijn om de student gerichte terugkoppeling te geven op de beoordeelde aspecten. Een dergelijk ideaal systeem bestaat niet en het valt vooralsnog ook niet te verwachten.

In de praktijk van de computergebaseerde beoordeling van teksten wordt wel gebruik gemaakt van technieken uit de natuurlijke taalverwerking, bijvoorbeeld om de structuren in het antwoorden te ontdekken, maar van een diepe semantische verwerking is geen sprake. We kunnen twee aanpakken onderscheiden die worden gebruikt om te beoordelen essays te classificeren: vergelijking en voorspelling.

Bij *vergelijking* worden teksten vergeleken met al beoordeelde voorbeelden of een normantwoord ('gouden standaard'). Een assumptie die daarbij blijkt te werken is dat goede en zwakke antwoorden zich typeren door hun woordkeus. **CODAS** (Combrink-Kuiters, Elffers, de Mulder & van Noortwijk, 2000) maakt daarvan gebruik en berekent, op basis van het woordgebruik, 'similariteiten' met al beoordeelde voorbeelden en ordent op die manier de teksten. Ook de **Intelligent Essay Assessor** (Landauer, Laham & Foltz, 2000) is indirect gebaseerd op woordgebruik. Middels latente semantische analyse (LSA) wordt een onderliggende semantische structuur geïnduceerd van de woorden en de contexten waarin ze worden gebruikt. Teksten en voorbeelden worden op deze geïnduceerde structuur geprojecteerd en vergeleken. CODAS en op LSA gebaseerde systemen werken alleen op basis van de woorden in de tekst: andere linguïstische informatie, al was het maar de volgorde van woorden, wordt niet gebruikt. Men spreekt wel van 'bag of words' aanpakken.

Bij *voorspelling* wordt geprobeerd om de beoordeling van essays door experts te voorspellen op basis van een reeks tekstkenmerken, zoals woordfrequentie, interpunctie, lengte van de tekst, syntactische en grammaticale karakteristieken, enz. **PEG** (Page, 1994) en **E-rater** (Burstein, Frase & Ginther, 1996; Burstein, Kukich, Wolff, Lu & Chodorow, 1998a) stellen op basis van een groot aantal predictoren enerzijds en een groot aantal beoordeelde essays, anderzijds een regressievergelijking op, die de beoordeling van de expert zo goed mogelijk voorspelt. De gevonden regressievergelijking wordt gebruikt om nieuwe essays te beoordelen.

Vergelijking en voorspelling kunnen worden gecombineerd: een deel van de E-rater beoordeling is bijvoorbeeld gebaseerd op het vergelijken van woorden in de te beoordelen tekst met al beoordeelde voorbeelden. Niettemin bepaalt de basis van de systemen waarover het best kan worden geoordeeld. Woord-gebaseerde systemen zijn vooral geschikt voor beoordeling van de *inhoud* van een tekst. Stijlkenmerken kunnen ze niet, of maar gedeeltelijk (vocabulaire, coherentie) beoordelen. Systemen waarin ook linguïstische aspecten van de tekst worden gemodelleerd, kunnen ook de stijl van de tekst proberen te beoordelen.

Ongeacht de aanpak die wordt gevolgd, moeten de systemen 'leren' hoe de teksten in het domein worden beoordeeld en daartoe is voorbeeldmateriaal nodig. Bij systemen als CODAS en vooral bij PEG en E-rater bestaat het voorbeeldmateriaal uit tientallen tot honderden al beoordeelde essays waarmee het systeem wordt getraind. Systemen die gebruik maken van latente semantische analyse (LSA) gebruiken teksten over het domein als trainingsmateriaal. Alle systemen vergen dat de tekst van de student voldoende informatie moet bevatten. Ze zijn niet geschikt om korte antwoorden te beoordelen. Om die reden is het met de huidige stand van de techniek eenvoudiger om langere essays te beoordelen dan korte (enkele zinnen, alinea). Alle systemen veronderstellen verder dat de student probeert een zinnige tekst te formuleren.

Overeenstemming met menselijke beoordelaars

Het criterium waaraan uiteindelijk de kwaliteit van de systemen wordt afgemeten, is doorgaans de overeenstemming met menselijke beoordelaars en hoe die overeenstemming zich verhoudt tot de mate waarin menselijke beoordelaars het onderling eens zijn. Nu is overeenstemming een wat rekkelijk begrip. In de literatuur over computergebaseerde beoordeling van essays worden verschillende maten gebruikt. We behandelen alleen de meest gebruikte en gaan geheel voorbij aan maten voor de betrouwbaarheid van de beoordelingen.

Een veel gebruikte maat is het *overeenstemmingspercentage* tussen beoordelaars. Daarbij kan een strenge en een milde norm worden gehanteerd. Bij een strenge norm telt alleen een perfecte overeenkomende score mee. Bij een milde norm is sprake van overeenstemming als de oordelen minder dan een bepaald maximum van elkaar verschillen. Een voorbeeld van toepassing van een milde norm zit in een deel van de GMAT, een test waarvoor het te bespreken E-rater systeem momenteel als een van de beoordelaars wordt gebruikt. De score van dit deel wordt uitgedrukt in een waardering lopend van 1 tot en met 6 met stappen van een half punt. Indien de oordelen van de beoordelaars een punt of minder verschillen, wordt dat gezien als overeenstemming.

Een belangrijk nadeel van het overeenstemmingspercentage is dat het een geflatteerde uitkomst oplevert: alleen al door puur willekeurig te beoordelen ontstaat een toevallige overeenkomst waarvoor moet worden gecorrigeerd. Kappa (Cohen, 1960) is een maat die is gecorrigeerd voor toevallige overeenkomsten. Kappa wordt veel gebruikt om de beoordelaarsbetrouwbaarheid te bepalen. De waarde van Kappa kan variëren van −1 tot +1 waarbij een waarde tussen de .60 en .80 geldt als substantiële overeenstemming (Landis & Koch, 1977).

Het overeenstemmingspercentage kent een tweede nadeel: er wordt geen gebruik gemaakt van een ordening in het classificatiesysteem (‘zeer slecht’, ‘slecht’, ‘onvoldoende’ enz.) en die ordening zou mee moeten wegen in de bepaling van de overeenstemming. Immers, als de ene beoordelaar iets als ‘onvoldoende’ en de ander als ‘voldoende’ beoordeelt, dan liggen hun oordelen dichter bij elkaar dan wanneer de één een ‘onvoldoende’ toekent en de ander een ‘zeer goed’. Er bestaat een gewogen Kappa die in dit soort situaties kan worden gebruikt (Cohen, 1968) maar vaker gebruikt men de correlatiemaat van Spearman (ρ) die kan variëren van −1 tot +1, of de Pearson correlatiecoëfficiënt (r). Een correlatiemaat geeft aan in welke mate twee variabelen (hier de oordelen) in dezelfde mate en richting fluctueren. Een correlatiecoëfficiënt is vrij ongevoelig voor systematische verschillen tussen beoordelaars: als beoordelaar 1 systematisch een punt lager scoort dan beoordelaar 2 dan heeft dat vrijwel geen invloed op de correlatiecoëfficiënt.

Beschrijving van systemen

In dit rapport beperken we ons tot systemen waarover meerdere publicaties zijn verschenen en waarin behaalde resultaten worden gerapporteerd[1]. In de literatuur worden weliswaar andere benaderingen geschetst en soms worden ook resultaten gerapporteerd, maar het

[1] Recent is nogal aan de weg getimmerd door Intellimetric (www.intellimetric.com) dat claimt een zeer goed scorend systeem te hebben ontwikkeld dat in verschillende tests wordt geëvalueerd. Er zijn ons echter geen gepubliceerde resultaten bekend. De PEG researchgroep meldt dat er wordt getest met de Intellimetric engine in het kader van een elektronisch portfolio project.

blijft bij incidentele vermeldingen en om die reden laten we ze hier buiten beschouwing. We proberen de systemen kort te karakteriseren en verwijzen naar de literatuur voor uitvoeriger beschrijvingen van de onderliggende techniek en details van de resultaten.

Project Essay Grader

Het oudste voorbeeld van computergebaseerde beoordeling van teksten is Project Essay Grader (PEG) dat in de jaren zestig is gestart en nog steeds loopt (Page, 1994). Het uitgangspunt van PEG is dat de beoordeling door de computer zo veel mogelijk identiek moet zijn aan de beoordeling door menselijke experts. Het gaat daarbij dan niet zozeer om de inhoud als wel om 'the assessment of general writing ability' (Shermis, Mzumara, Olson & Harrington, 2001, p. 248). PEG maakt daarvoor gebruik van een systeem waarin de oordelen van de experts worden voorspeld op basis van een reeks 'proxes', dat wil zeggen oppervlakkige tekstkenmerken, waarmee intrinsieke kwaliteiten van een tekst worden benaderd. Een voorbeeld: de helderheid van een tekst over een algemeen onderwerp kan worden benaderd door te kijken naar woordlengte en woordfrequentie in de taal. Hoe groter de woordlengte en hoe zeldzamer de woorden, des te onduidelijker is de tekst (is de aanname in dit voorbeeld).
PEG veronderstelt de aanwezigheid van een ruim aantal door experts beoordeelde essays. Er wordt een regressieanalyse uitgevoerd waarbij de tekstkenmerken (proxes) van de beoordeelde essays worden gebruikt als predictoren om het expertoordeel te voorspellen. Op basis van een goed voorspellende combinatie van predictoren wordt een regressievergelijking opgesteld waarmee PEG de overige essays beoordeelt.

Vanaf de eerste onderzoeken met PEG is de ervaring opgedaan dat het PEG-oordeel hoog correleert met de oordelen van menselijke beoordelaars en dat deze correlatie vergelijkbaar of zelfs hoger is dan de onderlinge correlaties van beoordelaars. In een aantal heranalyses van data uit een NAEP studie werd geprobeerd om de gemiddelde beoordeling van 599 essays door zes menselijke beoordelaars te voorspellen (Page, 1994). De multiple correlatie tussen PEG's voorspelling en het gemiddelde oordeel bedroeg 0,88 hetgeen, zeker gezien de betrouwbaarheid van de menselijke oordelen, bijzonder hoog is. Bij het gebruik van regressie-analyse loopt men het risico dat het gevonden model (de regressievergelijking) alleen een goede fit geeft voor de onderzochte cases. Om te testen of sprake was van een statistisch artefact werden door Page drie kruisvalidaties uitgevoerd waarin telkens 400 random getrokken cases gebruikt werden om een regressievergelijking met 24 variabelen op te stellen en 199 cases om de voorspelling te toetsen. De multiple correlatiecoëfficiënten lagen dichtbij 0,85.

Als menselijke beoordelaars het onderling onvoldoende eens zijn, wordt het lastig om hun oordelen te voorspellen. Page (1995) maakt melding van een test waarbij PEG werd uitgeprobeerd op een verzameling van 1314 essays die alle door twee beoordelaars waren gescoord. Van deze essays werden er nog eens 600 voorzien van twee nieuwe oordelen. De helft van deze laatste groep werd gebruikt als onderzoeksgroep, dat wil zeggen PEG voorspelde de scores van de beoordelaars, nadat een regresievergelijking was opgesteld met behulp van de andere essays. Opvallend was dat PEG hogere correlaties met de afzonderlijke beoordelaars liet zien (r = .716 – .768) dan de beoordelaars onderling behaalden (r= .550 – 748) (Page & Petersen, 2001). Een zelfde resultaat wordt in recent onderzoek, waarin de predictoren meer zijn verfijnd, nog eens gerapporteerd (Shermis et al., 2001).

In de literatuur worden overwegend holistische beoordelingen door PEG gerapporteerd. Gebruikmakend van regressietechnieken moet PEG echter in staat zijn om specifieke aspecten van kwaliteit te beoordelen (indien er althans voorbeeldscores van menselijke beoordelaars zijn), maar – zoals gezegd - daarover is nog nauwelijks gepubliceerd.

PEG's sterke punt zijn de hoge correlaties met menselijke beoordelaars die uit het regressiemodel rollen. Helaas zijn de publicaties over PEG erg vaag over de gebruikte predictoren in de regressievergelijking (zo ook Chung & O'Neil, 1997). Er wordt wel vermeld dat in de latere versies van PEG niet alleen directe oppervlaktekenmerken worden verwerkt, maar dat ook gebruik wordt gemaakt van parsers en taggers om maten te vinden voor grammaticale correctheid en vocabulaire, maar hoe dat verloopt en om welke maten het precies gaat, blijft onduidelijk. Uit de literatuur over PEG is niet duidelijk geworden hoe kort of lang de te beoordelen essays mogen of moeten zijn. Gezien het aantal predictoren lijkt het ondenkbaar dat het hier om korte antwoorden kan gaan.

PEG kent een aantal voorwaarden die de gebruiksmogelijkheden binnen de OUNL sterk lijken in te perken. Bedenk daarbij dat de opstelling van een regressie-vergelijking *per vraag* dient te gebeuren (en bij voorkeur gebruik maakt van verschillende studenten en beoordelaars). In het onderzoek naar PEG worden grote aantallen (500 tot 1000) beoordeelde essays gebruikt om de regressievergelijking op te stellen. Chung en O'Neill (1997) wijzen er op dat het verband tussen de scores en de tekstkenmerken, zeker bij deze grote aantallen, een statistisch artefact kan zijn (zie echter Page (1994) waarin deze kritiek wordt weersproken).

Voor de OUNL is eerder de vraag hoe gering het aantal voorbeelden (al beoordeelde essays) mag zijn om PEG toch nog stabiel te laten voorspellen. In de literatuur vinden we daarover geen harde uitspraken. Gering zal het aantal echter niet zijn: nemen we vijf beoordeelde essays per voorspeller als zeer lage ondergrens (vijftien geldt als aanbevolen) en 40 predictoren in de regressieanalyse, dan komen we op een ondergrens van 200 en een bovengrens van 600 beoordeelde essays om het systeem te trainen. De beoordelingen van de essays moeten bovendien voldoende variatie vertonen. Dit zijn kwantitatieve eisen waaraan eenvoudig kan worden voldaan als de essayvragen onderdeel uitmaken van een standaardtest (als de GMAT) die op grote schaal wordt afgenomen. Voor de OUNL vergt het dat de essayvragen onderdeel uitmaken van een toetsing waaraan jaarlijks door veel studenten wordt deelgenomen en waarschijnlijk ook dat dezelfde essayvragen gedurende meerdere jaren worden gebruikt.

De inzetbaarheid van PEG in interactieve situaties lijkt niet erg hoog. Hoewel recent melding is gemaakt van een via het web toegankelijk PEG-systeem (Shermis et al., 2001) gaat het daarbij vooral om on-line scoring. De feedback vanuit PEG is weinig informatief.

CODAS

CODAS is ontwikkeld aan de Erasmus Universiteit Rotterdam en inmiddels commercieel beschikbaar gemaakt onder de naam Nationale Nakijkcentrale (http://www.edu-actief.nl/nakijkcentrale/index_nakijkcentrale.htm). De oorsprong van CODAS is gelegen in onderzoek gericht op het achterhalen van relevante juridische documenten uit een grote database. Een gebruiker kan een juridisch concept omschrijven door aan te geven welke, al bekende, documenten zeker wel (en eventueel ook welke zeker niet) achterhaald moeten worden. Het programma bepaalt nu de mate waarin andere documenten in de database overeenkomen met deze voorbeelden. CODAS is, met andere woorden, een systeem dat werkt met een vergelijkingsaanpak.

De crux zit in nu in de wijze waarop de overeenstemming tussen documenten wordt bepaald. Van Noortwijk en De Mulder (1997) geven een reeks karakteristieken die *zouden* kunnen worden gebruikt voor het bepalen van overeenstemming, uiteenlopend van de gebruikte lettertypen, het aantal woorden, het aantal verschillende woorden tot en met de beoogde doelgroep en het onderwerp van het document. Deze reeks stelt de computer voor steeds

grotere problemen en de auteurs kiezen voor een aanpak waarin overeenstemming wordt bepaald op basis van gezamenlijk voorkomende (en ontbrekende) *woorden* en *documentfrequenties* (hoe vaak komt een woord voor in verschillende documenten).

CODAS bestaat uit twee modules: een fraudecheck-module en een nakijk-module. De fraudecheck-module berekent de similariteit tussen essays op basis van de volgende 'hits':

- woorden die twee essays[2] gezamenlijk hebben. Daarbij hebben woorden die verder weinig voorkomen in andere essays een zwaarder gewicht.
- Woorden die in beide essays *niet* voorkomen, maar juist wel in andere essays. Hier hebben veelvoorkomende woorden een hoger gewicht.

Twee essays die sterk overeenkomen, komen in aanmerking voor een handmatige controle op fraude. We gaan hier verder niet in op de fraude-check module, hoewel het een module is die voor de OUNL zeker interessant zal kunnen worden.

In de tweede module, de nakijkmodule, werkt de docent samen met CODAS toe naar een beoordeling van de essays. CODAS ordent de essays door ieder essay te vergelijken met de andere essays en een gemiddelde similariteitsscore toe te kennen. De tekst moet daartoe een voldoende lengte hebben (minstens 300 woorden). De docent markeert een aantal essays als voorbeelden van goede en onvoldoende antwoorden, waarna een herberekening van de volgorde plaats kan vinden. De docent toetst deze volgorde door essays handmatig te beoordelen en eventueel markeringen aan te passen en te vragen om herberekeningen. Gaandeweg wordt op deze manier de lijst gesorteerd en kunnen de grenzen voor de uiteindelijke beoordeling worden gesteld. De onderzoekers melden dat bij een beoordelingsschema met vier categorieën volstaan kan worden met het handmatig nakijken van 20% tot 40% van de essays. Dit zijn dan vooral essays die erg hoog of erg laag scoren (die worden standaard met de hand nagekeken) en essays die nagekeken worden om de overgang (ijkpunten) tussen de beoordelingscategorieën te markeren. Er wordt geclaimd dat CODAS al een tijdwinst oplevert indien meer dan 80 essays moeten worden nagekeken.

CODAS is op diverse aspecten geëvalueerd (Combrink-Kuiters et al., 2000), maar we concentreren ons hier op de deelonderzoeken naar overeenstemming met menselijke oordelen en beoordelaarsbetrouwbaarheid (een aspect dat in het onderzoek naar de andere systemen vrijwel wordt verwaarloosd). Zo is gekeken naar herbeoordeling, naar verschillen tussen beoordelen met en zonder CODAS en naar verschillen tussen beoordelaars. Een deel van de gerapporteerde gegevens toont bevredigende kappa's (rond de .67) en hoge correlaties (rho=.97) maar andere resultaten zijn veel minder rooskleurig. De onderzoekers maken aannemelijk dat het laatste resultaat vooral terug te voeren is op verschillen tussen beoordelaars. Overigens is dit welhaast een standaardbevinding in het onderzoek naar deze systemen: de verschillen met een beoordelaar zijn niet of nauwelijks groter dan verschillen tussen beoordelaars onderling.

CODAS is minder dan andere systemen afhankelijk van hoge correlaties met menselijke beoordelaars, omdat het systeem meer is opgezet als een nakijk*hulp*. Mocht de beoordelaar niet gelukkig zijn met de CODAS-beoordeling, dan kan steeds verder worden verfijnd door meer exemplaren handmatig te bekijken en deze voor het systeem te markeren.

[2] Wij gebruiken overal de term essay, waar in CODAS de termen document en werkstuk worden gebruikt. Als nakijkhulp kan CODAS worden gebruikt voor essays die ook andere inhoud dan tekst alleen bevatten.

Hiermee is ook aangegeven wat de gebruiksmogelijkheden en beperkingen van CODAS zijn voor de OUNL. CODAS kan gebruikt worden als hulp bij het beoordelen van grote aantallen essayvragen (zo'n honderd) die als tekstbestanden zijn aangeleverd. Interactief gebruik van CODAS waarbij het systeem de student terugkoppeling geeft, is niet mogelijk.

E-rater

E-rater is ontwikkeld door ETS (Educational Testing Service) een grote non-profit organisatie die zich bezighoudt met ontwikkeling en afname van tests en toetsen in het onderwijs. Per jaar neemt ETS meer dan 11 miljoen tests en toetsen af. Het onderzoek naar E-rater heeft gebruik gemaakt van gegevens van de Test of Written English en vooral van een deel van de Graduate Management Admission Test (GMAT). Dit is een toelatingstest voor MBA-opleidingen die jaarlijks ruim 200.000 keer wordt afgenomen. De tekst bevat een vast onderdeel met essayvragen (zie bijlage 2 voor enkele voorbeelden). De antwoorden worden beoordeeld op criteria als kwaliteit van ideeën, de organisatie, opbouw en uitdrukkingswijze van de ideeën; de relevante argumenten en voorbeelden en het taalgebruik. Momenteel worden de antwoorden op deze vragen beoordeeld door een menselijke beoordelaar en E-rater die de tweede menselijke beoordelaar heeft vervangen.

E-rater probeert op basis van een reeks tekstkenmerken de oordelen van menselijke beoordelaars te *voorspellen* (bij de GMAT zijn dat oordelen die lopen van 1 tot 6 met een interval van een half punt). In dat opzicht is het systeem goed te vergelijken met PEG. Er is echter een belangrijk verschil in de wijze waarop tekstkenmerken worden bepaald. Bij PEG worden *oppervlakkige* tekstkenmerken gebruikt ('proxes') als indicatoren voor onderliggende kaarakteristieken, terwijl het E-rater onderzoek zich vooral heeft gericht op de vraag hoe de criteria van de GMAT vertaald kunnen worden naar *diepere* tekstkenmerken (Burstein et al., 1998b; Burstein et al., 1998a) die door een computer kunnen worden herkend. Kwaliteit van taalgebruik kan bijvoorbeeld worden geoperationaliseerd als syntactische variatie, maar dat vergt dat een computer de tekst kan parsen. In het E-rater onderzoek werd en wordt dan ook druk geëxperimenteerd met diverse technieken voor Natuurlijke Taalverwerking. Een deel van die technieken, zoals lexicale semantische analyse (Burstein, Wolff & Lu, 1999) lijkt weer verlaten en er komt een beeld naar voren waarin E-rater drie modulen bevat die ieder specifieke karakteristieken van een tekst analyseren: een syntactische module, een discours module en een topic analyse module (Burstein & Marcu, 2000):

- De *syntactische module* breekt de tekst op met behulp van een parser. Deze module herkent diverse werkwoordsvormen en zinsconstructies op basis waarvan de syntactische variatie van de tekst kan worden bepaald. Men kijkt daartoe bijvoorbeeld naar het (relatieve) aantal eenvoudige, samengestelde en complexe zinnen, het gebruik van modale hulpwerkwoorden, ondergeschikte zinsdelen, gebruik van infinitieven e.d.

- De *discours module* poogt zogenaamde 'argument units' op te sporen: onderdelen van de tekst waarin sprake is van argumentatie. E-rater maakt daarvoor gebruik van de syntactische informatie en van cue-woorden of frasen die worden gebruikt als markers voor retorische predikaten als CONTRAST, SAMENVATTING en ELABORATIE. Voorbeelden van dergelijke cue-woorden zijn 'samenvattend', 'concluderend', 'anderzijds' e.d. Er is onderzoek gaande, gebaseerd op het werk van Marcu (2000) om het systeem de retorische structuur in de hele tekst te laten herkennen, maar daarvan zijn ons geen publicaties bekend.

- De *topical content analysis module* kijkt naar het woordgebruik in het antwoord en vergelijkt die met al beoordeelde antwoorden in ieder van de zes scorecategorieën. Er worden twee maten voor inhoudelijke overeenkomsten bepaald: een EssayContent maat

en een ArgContent maat. Bij de EssayContent maat wordt het antwoord, op basis van de daarin gebruikte woorden, vergeleken met de al beoordeelde antwoorden in de zes beoordelingscategorieën. De toegekende score is die van de categorie waarmee de hoogste overenstemming is gevonden. De ArgContent maat wordt op een vergelijkbare manier bepaald, maar nu worden de argumenten (en niet het hele antwoord) vergeleken (Burstein et al., 1998a). Dit deel van E-rater maakt gebruik van een 'bag of words' aanpak die wordt gevolgd in CODAS en in de LSA gebaseerde systemen zoals de Intelligent Essay Assessor.

E-rater kan in totaal een 67 predictoren uit de tekst halen en proberen daarmee menselijke oordelen te voorspellen. Burstein et al (1998a) rapporteren gedetailleerde gegevens van een van de validatiestudies met E-rater. In totaal werden de antwoorden op 13 GMAT en 2 TWE vragen geanalyseerd. Voor iedere vraag werden 280 beoordeelde antwoorden gebruikt om via stapsgewijze regressie een model te bepalen dat de score van de beoordelaars voorspelde (280 is overigens o.i. aan de krappe kant gezien het aantal predictoren). Vervolgens werd berekend in welke mate de oordelen van de beoordelaars en E-rater overeenstemden. Het aantal testcases uit de GMAT varieerde daarbij van 517 tot 915. De correlaties tussen E-rater en menselijke beoordelaars liepen uiteen van .79 tot .87 en dat is hoog als in aanmerking wordt genomen dat de correlatie tussen de (zeer goed getrainde) beoordelaars tussen de .82 en .89 lag.

E-rater ontwikkelde voor ieder antwoord een nieuw scoringsmodel, maar bepaalde predictoren kwamen toch zeer regelmatig voor en – gezien de inhoud en de beoordelingsvoorschriften van de GMAT niet verwonderlijk – dat waren vooral retorische eigenschappen (ArgContent, totaal aantal woorden die een argumentatie opbouwen (retorische vragen, woorden die op bewijs duiden e.d.), EssayContent en vervolgens enkele syntactische eigenschappen.

Een van de aardigste onderzoeken die met E-rater zijn uitgevoerd is dat van Powers et al. (2001) waarin is geprobeerd teksten te vervaardigen die het systeem tot te hoge of te lage scores verleiden. Uit het onderzoek komt naar voren dat het eerste beter lukt dan het laatste. Herhaling van alinea's (eventueel met gebruik van synoniemen), gebruik van woorden die argumentatie suggereren, variatie in syntactische structuren lijken tactieken die E-rater tot hogere scores kunnen brengen. Voor een deel lijken de tactieken ondervangen te kunnen worden door verbetering in het systeem, aan de andere kant heb ik de indruk dat E-rater wel erg gretig afgaat op markers voor argumentatie zonder dat het systeem in staat is om de inhoud van die argumentatie goed te beoordelen.

E-rater heeft een aantal duidelijke sterke punten (zie ook Chung & O'Neill, 1997): er worden hoge overeenstemmings-percentages en correlaties gevonden met menselijke beoordelaars; er is een duidelijk verband gelegd tussen beoordelingscriteria van de test (GMAT vooral) enerzijds en de predictoren die in de tekst worden opgespeurd anderzijds. E-rater deelt met PEG de aandacht voor de schrijfstijl en de argumentatie in het antwoord. Het systeem lijkt daarbij meer dan PEG in staat om ook de inhoudelijke kwaliteiten van de tekst te beoordelen. De gebruikte analysemethode (stapsgewijze regressieanalyse) is bekend en publiek toegankelijk.

E-rater kent ook nadelen: het systeem kan niet beoordelen op basis van een 'gouden standaard': het moet voor iedere vraag getraind worden met al beoordeelde antwoorden. Gezien het aantal predictoren moeten dat er al snel minstens 300 zijn (bij geringere aantallen loopt men een serieus te nemen risico dat het scoringsmodel alleen een statistisch artefact is). Om min of meer dezelfde reden kunnen de antwoorden niet kort zijn. Hoewel het

niet expliciet wordt genoemd, lijkt een ondergrens van 200 woorden redelijk. E-rater is daarmee o.i. een systeem dat vooral zijn waarde kan bewijzen als er jaarlijks zeer veel vragen moeten worden beantwoord die in standaardtests voorkomen (en dat is uiteraard precies de situatie waarvoor ETS zich zag geplaatst).

Hoewel E-rater vooral is ontwikkeld om het hoofd te bieden aan een situatie waarin jaarlijks honderdduizenden essays moeten worden beoordeeld, is sinds kort een versie van E-rater beschikbaar die vooral bedoeld is om oefening in het vervaardigen van essays te bieden. De student kan via het web een essay ter beoordeling aanbieden en ontvangt binnen enkele minuten feedback. Deze feedback is nog erg beperkt en bestaat in feite uit een standaardtekst die hoort bij de categorie waarin het antwoord wordt geclassificeerd. Enige maanden geleden is aangekondigd dat E-rater zal worden geïntegreerd in Question Mark Perception, een veelgebruikt systeem voor het samenstellen en afnemen van interactieve toetsen.

E-rater vergt wat de toepasbaarheid voor de OUNL betreft kwantiteiten die vergelijkbaar zijn met PEG. Daarnaast zal het nodige werk verricht moeten worden om de natuurlijke taaltechnieken van E-rater op Nederlandse teksten los te kunnen laten. De interactieve mogelijkheden, vooral de feedback, van E-rater zijn nog beperkt, maar gezien de inhoud van de tekstkenmerken die E-rater achterhaalt, lijkt verdere verbetering hier haalbaar.

Latente Semantische Analyse

Volgens Foltz, Laham en Landauer (1999) kan beoordeling van een essay op drie criteria plaats vinden: correctheid en volledigheid van de conceptuele kennis, de geldigheid van de argumentatie en de mate waarin vloeiend, elegant en begrijpelijk wordt geschreven. PEG concentreert zich vooral op het derde criterium en E-rater vooral op het tweede en derde. Latente Semantische Analyse (LSA) is vooral gericht op de conceptuele inhoud van het essay. Van de LSA gebaseerde systemen is vooral de Intelligent Essay Assessor, die als webdienst beschikbaar is, bekend geworden, maar LSA is gebruikt in een reeks onderzoeken die voor ons relevant zijn en op die onderzoeken concentreren we ons hier.

LSA is een computationale theorie over de wijze waarop begrippen worden geleerd. Landauer en Dumais (1997) hebben betoogd dat LSA de onderliggende semantiek van begrippen formeert op een wijze zoals een parallel werkend brein dat zou kunnen doen. De theorie heeft geleid tot divers en soms zeer uiteenlopend onderzoek. Zo heeft men LSA getraind met 2000 pagina's Engelse tekst en vervolgens het systeem (met goed gevolg) een synoniementest laten maken. Een ander voorbeeld: het systeem is getraind middels een inleidende tekst op het gebied van de psychologie en kon daarna een voldoende scoren op een multiple choice test over de tekst (Foltz, Laham & Landauer, 1999). Recente toepassingen van LSA op onderwijsgebied zijn gerapporteerd in een themanummer van Interactive Learning Environments (2000, Vol 8, Issue 8).

De aanpak van LSA (vergelijking op basis van 'bag of words') lijkt op die van CODAS en de topical content module van E-rater. In LSA wordt echter niet gewerkt met directe overeenkomsten en verschillen maar met achterliggende, *latente* factoren. Latente Semantische Analyse bouwt een semantisch model van het domein op, gebaseerd op het gebruik van woorden in contexten (zie bijlage 3 voor een beknopte uitleg van de techniek achter LSA). Om dat model te kunnen vormen, moet een LSA systeem worden getraind met domeinkennis. LSA werkt dus nooit alleen met antwoorden op essayvragen, maar heeft achtergrondmateriaal uit het domein nodig om zijn model te kunnen vormen.

Er zijn diverse manieren ontwikkeld om met LSA essays te beoordelen. In een 'holistische aanpak' worden essays vergeleken met eerder beoordeelde (dat wil zeggen gecategoriseerde) essays of met een 'gouden standaard' zoals een modelantwoord dat een docent opstelt. Er is eveneens ervaring opgedaan met een 'componentenbenadering' waarin essays worden vergeleken met delen van de oorspronkelijke tekst; of waarin delen van de essays worden vergeleken met een verzameling deelonderwerpen.

Holistische beoordeling met LSA

Er is in het LSA-onderzoek ervaring opgedaan met diverse maten voor inhoudelijke kwaliteit. Het best vergelijkbaar met E-rater en PEG zijn enkele maten die Foltz, Gilliam en Kendall (2000) gebruikten. Zij lieten LSA een essay vergelijken met andere, al beoordeelde, essays en de vijf meest overeenkomende essays selecteren. Het te beoordelen essay kreeg vervolgens het gemiddelde oordeel van de vijf gevonden essays toegekend. Deze maat correleerde .85 met de oorspronkelijke gemiddelde beoordeling van het essay. In combinatie met een maat voor de hoeveelheid domeinspecifieke informatie in het essay werd een correlatie van .89 met de oorspronkelijke beoordeling gevonden.

LSA kan een essay vergelijken met andere, beoordeelde, essays en zo tot een beoordeling komen. Er kan echter ook worden vergeleken met een 'gouden standaard' (modelantwoord) of met passages die als relevant zijn gemarkeerd in het materiaal waarmee LSA wordt getraind. Foltz (1996) liet vier beoordelaars 24 essays beoordelen op (a) de informatie die was verwerkt uit 21 bronteksten en (b) de kwaliteit van het essay. De correlaties tussen beoordelaars liepen uiteen van .38, .58 tot .77 (een minder ervaren beoordelaar viel wat buiten de boot). De beoordelaars selecteerden bovendien de tien belangrijkste zinnen uit de bronteksten (het expertmodel). LSA werd getraind met behulp van de bronteksten en teksten afkomstig uit naslagwerken. Er werden twee methoden gebruikt om LSA te laten beoordelen. Bij de eerste methode (tekst overlap) werden de zinnen in het essay vergeleken met die van de bronteksten (volgens Foltz een maat voor plagiaat of letterlijk onthouden). Bij de tweede methode werd vergeleken met de tien zinnen die de experts het meest belangrijk vonden (het expertmodel). De LSA beoordeling op basis van tekstoverlap correleerde van .32 tot .55 met de ervaren beoordelaars. De correlatie van de LSA beoordeling op basis van het expert model met de menselijke beoordelaars liep uiteen van .38 tot .63. Foltz, Gilliam en Kendall (2000), die diverse holistische en componentmaten uitprobeerden, vonden in een vergelijkbaar onderzoek een correlatie van .71 tussen beoordeling op basis van het expertmodel en gemiddelde menselijke beoordeling.

Landauer, Laham, Rehder en Schreiner (1997) rapporteren twee experimenten waarin soortgelijk onderzoek werd gedaan met meer proefpersonen. In het eerste experiment vervaardigden 94 studenten essays van ongeveer 250 woorden; ze maakten tevens een toets met korte antwoordvragen. De beoordelingen van de menselijke beoordelaars correleerden onderling 0,77, even hoog als de correlatie tussen LSA en het gemiddelde van de beoordelaars. Opvallend is dat LSA de toetsscores *beter* voorspelde dan de menselijke beoordelaars (r=0,81 tegen r=0,70. Ook hier werd onderzocht of gebruik kon worden gemaakt van een 'gouden standaard', in dit geval een experttekst uit een inleidend biologieboek. De correlatie tussen LSA en de toetsscore was nu 0,71. In een tweede experiment werd de correlatie tussen LSA en menselijke beoordelaars onderzocht in een grotere steekproef. Nu vervaardigden 273 studenten, die een inleiding Psychologie volgden een kort essay (de lengte wordt niet gerapporteerd, de beschikbare tijd was tien minuten) over één van de drie mogelijke onderwerpen. Twee beoordelaars scoorden de essays, waarbij bleek dat ze bij één onderwerp maar weinig overeenstemming hadden (r=0,19). Over alle essays samengenomen correleerden de oordelen van de beoordelaars 0,65. De correlatie tussen LSA en het gemiddeld oordeel was 0,64.

Er is binnen LSA veel ervaring opgedaan met diverse maten om te komen tot een holistische beoordeling. Als LSA gebruik kan maken van vergelijkbare informatie als waarover PEG en E-rater de beschikking hebben, een voldoend aantal beoordeelde essays, boekt het systeem vergelijkbare resultaten. Interessant is dat LSA ook met minder informatie - een 'gouden standaard' of als relevant gemarkeerd materiaal – een heel eind kan komen.

Componentenbeoordeling met LSA

Componentenbeoordeling met LSA komt vooral neer op het opsporen van behandelde onderwerpen in een essay (vgl de topic analysis module van E-rater). Foltz, Gilliam en Kendall (2000) lieten zeven subtopics benoemen die in essays afgehandeld zouden moeten worden en lieten voor ieder subtopic een tot drie zinnen in het voorbeeldmateriaal markeren. Deze voorbeeldzinnen werden vervolgens vergeleken met de zinnen in de essays. Zodra de overeenkomst boven een bepaalde drempelwaarde kwam werd het topic als afgehandeld beoordeeld. Het totaal aantal afgehandelde topics correleerde .78 met de holistische beoordeling van het essay. Een vergelijking van de LSA score per topic met de menselijke beoordeling per topic liet een minder florissant beeld zien. De correlaties liepen uiteen van .09 tot .71. De auteurs merken op dat de lage correlaties vooral optraden bij subtopics waar maar een voorbeeldzin was gemarkeerd.

In een vervolgexperiment bouwden Foltz, Gilliam en Kendall (2000) een webgebaseerde analytische beoordelaar die naast een holistisch oordeel specifieke feedback (vragen en opmerkingen) kon geven op ontbrekende topics. Het systeem werd uitgetest met een take-home opdracht. Studenten bleken met het systeem de tekst gemiddeld drie keer te reviseren en voor een eerste versie meer tijd uit te trekken dan ze gewoonlijk deden. Hun oordeel over het systeem was zeer positief, misschien ook wel door de hoge beoordelingen die het systeem voor de definitieve essays toekende. Deze LSA-beoordelingen werden echter door een onafhankelijke beoordeling bevestigd. Er is nog veel werk nodig, aldus de schrijvers, om uit te zoeken welke feedback het meest gunstig is voor de studenten. Daar valt o.i. aan toe te voegen dat uit dit onderzoek niet overtuigend is gebleken dat LSA in staat is om alle topics in voldoende mate te identificeren en dat ook niet duidelijk is geworden welke minimale training het systeem daarvoor nodig heeft.

Al met al is de componentenbeoordeling met LSA, zoals ook viel te verwachten, veel minder robuust dan de holistische beoordeling. Dat is vooral in situaties waarin inhoudelijke feedback moet kunnen worden gegeven van belang. Hier gelden waarschijnlijk ook zekere minimumnormen waaraan moet worden voldaan en één zin als referentiemateriaal voldoet daar kennelijk niet aan.

Samenvattingen beoordelen met LSA

Een wat andere toepassing van LSA is de beoordeling van samenvattingen. Kintsch et al. (2000) bespreken de ontwikkeling van Summary Street, een systeem waarmee brugklassers kunnen oefenen in het maken van samenvattingen (van zo'n 75 tot 300 woorden). We gaan hier voorbij aan de ervaringen die zijn opgedaan met de ontwikkeling van het interface, de rijkdom van de feedback en de integratie in het onderwijs. Er werd bij de ontwikkeling van het systeem geen gebruik gemaakt van een gouden standaard of eerder beoordeelde samenvattingen. Docenten markeerden in de door de leerlingen te bestuderen teksten de topics die in de samenvatting terug moesten komen. De samenvatting van de leerling werd door LSA vergeleken met de door de docenten genoemde topics. Zodra de overeenkomst tussen de samenvatting en de topics beneden een empirisch bepaalde drempelwaarde kwam, werd feedback gegenereerd. Daartoe werd in eerste instantie de overeenkomst bepaald

tussen iedere afzonderlijke zin van de samenvatting met de samenvatting als geheel. Indien deze overeenkomst beneden een drempelwaarde bleef, werd de zin als irrelevant voor de samenvatting bestempeld. De feedback suggereerde dan om de zin te schrappen. Vervolgens werd de overeenstemming tussen de afzonderlijke zinnen van de samenvatting onder de loep genomen. Indien de overeenstemming tussen twee zinnen *boven* een empirisch bepaalde drempel lag, werd in de feedback gesuggereerd om de zinnen op redundantie te controleren en ze weg te gooien of samen te voegen.

De resultaten van deze onderzoeken leveren een wat gemengd beeld op. Enerzijds wordt gerapporteerd dat vooral de maten die zijn gebaseerd op vergelijking met zinnen weinig stabiel zijn (dit zijn de maten waarop een belangrijk deel van de feedback is gebaseerd, vergelijk ook (Foltz, Gilliam & Kendall, 2000) JBR). Anderzijds blijkt de beoordeling door LSA opmerkelijk overeen te stemmen met de oordelen door experts. Zo correleerde bij een onderzoek met 50 samenvattingen de LSA analyse van de overeenkomst tussen de samenvatting en de brontekst 0,64 met de beoordelingen door de docenten. Bij een tweede test werd onderzocht in hoeverre de zinnen (119) van de samenvatting konden worden verbonden aan topics in de bronteksten. De menselijke beoordelaars vertoonden hier een grote overeenstemming (r=0,92), LSA correleerde hier 0,84 met de beoordelaars. Bij sommige teksten bleek de overeenkomst met menselijke beoordelaars zeer laag uit te pakken en kende LSA soms beduidend hogere scores toe. Volgens de auteurs werd dit veroorzaakt doordat bij de instelling van de parameters voor de LSA-analyse te weinig rekening was gehouden met de verschillen in moeilijkheid van de onderwerpen.

Chung en O'Neill (1997) geven de volgende sterke punten van LSA: de scoring door LSA is vergelijkbaar met die van menselijke beoordelaars. LSA kan zowel relatief als absoluut scoren, dat wil zeggen zowel ten opzichte van beoordeelde essays als ten opzichte van referentieteksten: een gouden standaard of gemarkeerde delen in een tekst. LSA is een relatief open systeem – de gebruikte technieken zijn goed beschreven en over het algemeen in het publieke domein. Daar staan zwakke punten tegenover: LSA vergt veel rekencapaciteit, omdat er veel materiaal moet worden ingevoerd en uitvoerige matrixbewerkingen op het materiaal noodzakelijk zijn.

Wat LSA voor de OUNL interessant kan maken, zijn de meer bescheiden eisen aan de kwantiteiten: LSA vergt geen honderden beoordeelde essays. Daar staat tegenover dat LSA moet worden getraind op domeinmateriaal (en al beschikt de OUNL daar over, dan nog zal dit materiaal moeten worden voorbewerkt). LSA biedt ook voor meer interactieve toepassingen interessante mogelijkheden, al is het onderzoek daarna nog voor een belangrijk deel gaande, maar op zich is dat ook een voordeel: er wordt nog volop gewerkt aan en gerapporteerd over LSA-toepassingen.

Conclusies en aanbevelingen

De stelling dat de computer antwoorden op essayvragen kan beoordelen, zou tien jaren geleden wellicht vooral tot hilariteit of meewarigheid hebben geleid. Momenteel is computergebaseerde beoordeling een feit: het E-rater systeem is een van de beoordelaars van de GMAT, een test die jaarlijks enkele honderdduizend keren wordt afgenomen. In Nederland bestaat een 'nakijkcentrale' waar CODAS als nakijkhulp wordt gebruikt en in de VS is de Intelligent Essay Assessor een webdienst. Bovendien blijken deze systemen het niet veel beter of slechter te doen dan menselijke beoordelaars en ze zouden, in principe, kunnen leiden tot een aanzienlijke besparing van tijd en kosten. Daar komt bij dat sommige systemen (CODAS en LSA) in staat zijn om gevallen van plagiaat op te sporen.

Dat het daarmee nog niet allemaal botertje tot de boom is, blijkt wanneer we wat meer gedetailleerd kijken naar de volgende punten:

- het belang dat wordt gehecht aan het oordeel van de computer
- het soort antwoorden dat kan worden beoordeeld
- de aspecten waarop kan worden beoordeeld
- de feedback die kan worden gegeven
- de noodzakelijke aanpassingen, voor- en nabewerkingen
- de kosten en opbrengsten van het gebruik van deze systemen.

Het belang van het oordeel

Er is ons geen situatie bekend waarin de computer een summatieve evaluatie opstelt over de vorderingen van de student. Het meest vergaande scenario is dat van E-rater die als tweede beoordelaar wordt ingezet bij de GMAT. De vraag of de computer kan worden ingezet als de enige of doorslaggevende beoordelaar van examen- of tentamenopgaven wordt (vooralsnog) ontkennend beantwoord. De reden daarvoor is overigens niet dat men twijfelt aan het oordeel dat de computer over het algemeen velt, maar aan het onvermogen van de computer om de briljante antwoorden te onderkennen.

Er is sprake van een andere situatie als de computer wordt ingezet om formatieve evaluaties uit te voeren. De Intelligent Essay Assessor, de web versie van E-rater en Summary Street zijn er vooral op gericht dat studenten/leerlingen kunnen werken aan een essay of samenvatting en op basis van de feedback van het systeem een verbeterde versie kunnen opleveren. Vooral hier lijkt een niet onaanzienlijke tijdswinst te realiseren (ceteris paribus, zie onder).

Het soort antwoord dat wordt beoordeeld

De systemen die we hier hebben besproken, zijn vooral uitgetest op wat langere antwoorden en dat is ook precies waar ze het sterkst in zijn. Alle systemen halen indicatoren voor onderliggende, te beoordelen factoren, uit de tekst en hoe korter de tekst, hoe kleiner de kans dat ze daarin slagen. Men is niet al te duidelijk over de ondergrenzen van de te beoordelen teksten, maar we gaan er van uit dat een ondergrens van 100 woorden niet irreëel is. Daarmee zijn overigens belangrijke beperkingen aangegeven: de systemen zijn niet in staat om op gedetailleerd niveau inhoudelijke terugkoppeling te geven, omdat ze doorgaans niet beschikken over voldoende informatie uit het antwoord of het materiaal waarop ze getraind zijn.

Aspecten waarop wordt beoordeeld

Het is zeker bij essays niet ongebruikelijk om naast de inhoud ook de schrijfstijl te beoordelen: is het taalgebruik helder en gevarieerd? Is de argumentatie goed opgebouwd en onderbouwd? De GMAT is een voorbeeld van een test waarop zowel inhoud als stijl worden beoordeeld. Van de hier besproken systemen is alleen E-rater er op gericht om beide aspecten te kunnen beoordelen. PEG concentreert zich op stijl, CODAS en LSA beperken zich vooral tot de inhoud.

Feedback door de computer

De toepassing van de systemen is aanvankelijk gericht geweest op het geven van een holistisch oordeel over een essay en op dat oordeel kon wel even worden gewacht. Met het beschikbaar komen van meer computerkracht groeit de vraag of de systemen in staat zijn om in interactieve settings zoals een elektronische leeromgeving, direct zinvolle feedback te geven. Het gaat hier dus om twee aspecten: de snelheid waarmee terugkoppeling kan worden gegeven en de inhoud van de terugkoppeling.

Snelheid lijkt momenteel niet een echt probleem: zowel E-rater als LSA worden momenteel als webservices aangeboden en in de literatuur worden responstijden genoemd die uiteen lopen van enkele seconden tot enkele minuten. Gezien de aard van de taak, lijken dit acceptabele responstijden.

De nauwkeurigheid en zinvolheid van de terugkoppeling lijken voor meer problemen te zorgen. Alle systemen zijn in staat een antwoord te classificeren en ze kunnen dan ook (in principe) een vaste terugkoppeling geven die is gerelateerd aan de categorie waarin het antwoord wordt gescoord. Dat kan het makkelijkst bij een holistische beoordeling en begint problemen op te leveren bij een analytische beoordeling, omdat het oordeel soms moet worden geveld op basis van weinig materiaal (zowel in de voorbeelden als in de antwoorden). De systemen die hier het meest perspectief bieden zijn E-rater en vooral LSA-gebaseerde systemen. De ontwikkeling van meer interactieve toepassingen is van recente datum en het lijkt aannemelijk dat het enige tijd vergt voordat de mogelijkheden en beperkingen van de technieken duidelijker zijn geworden.

Noodzakelijke voor- en nabewerking

Alle systemen vergen diverse vormen van voorbewerking. CODAS en E–rater veronderstellen dat een niet onaanzienlijk aantal beoordeelde essays aanwezig is. De noodzakelijke voorbewerking van nieuwe essays lijkt door CODAS zelf te worden uitgevoerd. Bij de andere systemen (E-rater, LSA, PEG) lijkt de voorbewerking voorbehouden aan de makers van het systeem, maar laten we er hier van uitgaan dat de technologie op een of andere manier beschikbaar komt. De vraag is nu wat er moet worden gedaan om die technologie voor de OUNL inzetbaar te maken.

Aanpassingen binnen de systemen

CODAS lijkt zonder verdere aanpassingen in staat te zijn om de teksten van onze studenten te beoordelen. E-rater maakt gebruik van NLP technieken die geschikt moeten worden gemaakt voor het Nederlands. Dit vergt aanpassing of vervanging van de parser en de daarbij gebruikte regels en lexicons. Een lijst met weg te filteren woorden zal moeten worden vervangen door een Nederlandstalige lijst. De feedback zal moeten worden aangepast. LSA maakt geen gebruik van NLP technieken, wel zal een lijst met weg te filteren woorden moeten worden vervangen door een Nederlandstalige. De Engelstalige terugkoppeling, zoals in de Summary Street implementatie, zal moeten worden aangepast.

Voorbewerken van materiaal

Alle systemen vergen dat er voorbewerkt materiaal is in de vorm van eerder beoordeelde essays en/of bronteksten over het domein. PEG en E-rater vergen een flink aantal essays dat eerder is beoordeeld en dat voldoende spreiding in de beoordeling bevat. Bij CODAS zal de beoordelaar tijdens het werken met het systeem nog een deel van de werkstukken

handmatig moeten beoordelen. LSA kan essays vergelijken met een gouden standaard, of met topics die door een docent zijn gemarkeerd in de bronteksten. LSA moet echter worden getraind in een domein en dat vergt selectie, voorbewerking en invoer van materiaal.

Uiteraard vergen alle systemen dat de te beoordelen essays in elektronische vorm beschikbaar zijn, waarbij eerder moet worden gedacht aan een ASCII-bestand dan een fraai opgemaakt tekstverwerkersresultaat. Handgeschreven tentamens vergen dus een aanzienlijke voorbewerking wil men er met deze systemen iets mee kunnen doen.

Nabewerking

Alle systemen vergen dat een nabewerking (en soms herberekening) wordt uitgevoerd. In CODAS kan de docent interactief de werkstukken verder ordenen en de ijkpunten markeren waarna CODAS de scores verder toekent. In E-rater moet de beoordeling worden geijkt door een serie regressieanalyses uit te voeren om een beoordelingsvergelijking te maken. Het LSA-onderzoek, vooral dat met Summary Street, laat zien, dat na een eerste analyse drempelwaarden moeten worden bepaald die aangeven of er sprake is van voldoende of onvoldoende overeenkomst op basis waarvan feedback kan worden gegenereerd.

Kosten en opbrengsten

De kosten die zijn verbonden aan het gebruik van computerondersteund beoordelen zijn niet erg duidelijk. Er wordt door ETS en de uitbater van Intelligent Essay Assessor wel geschermd met de lage kosten per beoordeeld essay, maar het is volstrekt onduidelijk of in die kosten ook de noodzakelijke voor- en nabewerking zijn begrepen. Dat lijkt voor OUNL materiaal zelfs onaannemelijk, gezien de aanpassingen die nodig zullen zijn om Nederlandstalige teksten te kunnen verwerken. De lage kosten per beoordeeld essay moeten ook in hun context worden bezien: ETS beoordeelt jaarlijks honderdduizenden open vragen die deel uit maken van standaardtests. Nadere studie van de kosten is voor de OUNL dan ook zeker aan te bevelen. Aan de andere kant kan het gebruik van computerondersteund beoordelen ook tot kostenbesparingen leiden: het nakijken en scoren gaat sneller (CODAS), men bespaart een beoordelaar uit (E-rater), of laat het misschien zelfs helemaal aan de computer over als het niet om formele beoordelingen (tentamens) gaat. Ook hier geldt dat moet worden onderzocht of deze besparing bij de OUNL kan worden gerealiseerd.

Een andere opbrengst zit vooral in het interactief gebruik van de systemen: er kan een omgeving worden aangeboden waarin de student kan oefenen en op basis van de ontvangen terugkoppeling kan werken aan een beter product: een werkwijze die momenteel, zoal niet onmogelijk, dan toch wel erg tijdrovend en kostbaar is.

Aanbevelingen

PEG en E-rater (en naar het schijnt ook het systeem van Intellimetrics) proberen de scores van beoordelaars te voorspellen op basis van een reeks tekstkenmerken. Vooral E-rater gebruikt daarvoor diverse NLP-technieken die niet zondermeer voor Nederlandstalige antwoorden vallen te gebruiken. Zowel bij PEG als E-rater moet voor iedere essayvraag een regressiemodel opgesteld worden, waarvoor al beoordeelde antwoorden beschikbaar dienen te zijn. Een voorzichtige rekensom met 5 waarnemingen per predictor en 40 predictoren in de regressieanalyse leert dat per vraag minstens 200 beoordeelde voorbeelden nodig zijn. Om het model vervolgens te kunnen gebruiken, moeten de vragen stabiel zijn. Gezien de

hoeveelheid voorbeeldmateriaal en de NLP voorbewerking van de antwoorden lijken PEG en E-rater voor de OUNL geen systemen die in aanmerking komen.

CODAS werkt met een ‘bag of words’ aanpak – geen NLP, geen woordvolgorde, alleen de inhoudswoorden in de antwoorden worden gebruikt. Bij CODAS kan de docent interactief antwoorden markeren als goed resp. slecht. Het systeem classificeert de documenten in termen van ‘similariteit’ met deze markers. Men claimt dat CODAS al vanaf 80 werkstukken een flinke tijdwinst oplevert. De nakijkhulp van CODAS lijkt voor de OUNL de moeite waard, maar is niet geschikt voor meer interactieve toepassingen. Een proef met de nakijkhulp is zinnig.

Bij LSA-gebaseerde systemen induceert het programma op basis van materiaal over het domein (tekstboeken, artikelen) een semantische ruimte met gereduceerde dimensies. Antwoorden of samenvattingen kunnen vervolgens worden vergeleken met beoordeelde voorbeelden of met een modeltekst. LSA biedt, meer dan CODA, perspectief op interactief gebruik met informatieve feedback voor de student.

Op grond van de state of the art luidt het advies om (a) te experimenteren met CODAS als nakijkhulp en (b) LSA-technologie nader te onderzoeken op meer interactieve mogelijkheden voor de student. Voor beide systemen geldt dat we de hoeveelheid research die in de ontwikkeling van systemen is gestoken niet in kunnen halen. Samenwerking met de makers van de systemen is daarom wenselijk.

Literatuur

Burstein, J., Frase, L. T., & Ginther, A. (1996). Technologies for language assessment. *Annual Review of Applied Linguistics, 16, 240-260*.

Burstein, J., Kukich, K., Wolff, S., Lu, C., & Chodorow, M. (1998a). Computer analysis of essays. *NCME symposium on automated scoring*, 1998a.

Burstein, J., Kukich, K., Wolff, S., Lu, C., Chodorow, M., Braden-Harder, L., & Harris, M. D. (1998b). Automated scoring using a hybrid feature identification technique. *Proceedings of the Annual Meeting of the Association of Compuational Linguistics*, Montreal, Canada.

Burstein, J., & Marcu, D. (2000). Benefits of modularity in an automated essay scoring system. *Proceedings of the workshop on using toolsets and architectures to build NLP systems. 18th International Conference on Computational Linguistics*, Luxembourg.

Burstein, J., Wolff, S., & Lu, C. (1999). Using lexical semantic techniques to classify free-responses. In N. Ide and J. Veronis (Series Eds.) & E. Viegas (Ed.), *Text, speech and language technology series: Vol. 10. The depth and breadth of semantic lexicons* (pp. 227-246). Kluwer Academic Press.

Chung, G. K. W. K., & O'Neil, H. F. (1997). *Methodological approaches to online scoring of essays* (CRESST report 461). Los Angeles: National Center for Research on Evaluation, Standards and Student Testing (CRESST).

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20, 37-46*.

Cohen, J. (1968). Weighted kappa: nominal scale agreement with provisions for scales disagreement of partial credit. *Psychological Bulletin, 70, 212-220*.

Combrink-Kuiters, C. J. M., Elffers, H., de Mulder, R. V., & van Noortwijk, C. (2000). *Computer-ondersteund nakijken van open vragen*. Meppel: Edu'Actief.

de Gruijter, D. N. M. (1999). De scoring van open vragen. In G. Heijnen & S. Meeder (Eds.), *Toetsen en ICT in het hoger onderwijs: stand van zaken en trends in Nederland* (pp. 43-52). Utrecht: Stichting SURF.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41, 391-407*.

Foltz, P. W. (1996). Latent semantic analysis for text-based research. *Behavior Research Methods, Instruments & Computers, 28, 197-202*.

Foltz, P. W., Gilliam, S., & Kendall, S. (2000). Supporting content-based feedback in on-line writing evaluation with LSA. *Interactive Learning Environments, 8, 111-127*.

Foltz, P. W., Laham, D., & Landauer, T. (1999). Automated essay scoring: applications to educational technology. *Edmedia 99.*

Kintsch, E., Steinhart, D., Stahl, G., LSA research group, Matthews, C., & Lamb, R. (2000). Developing summarization skills through the use of LSA-based feedback. *Interactive Learning Environments, 8, 87-109*.

Landauer, T. K., Laham, D., Rehder, B., & Schreiner, M. E. (1997). How well can passage meaning be derived without using word order? A comparison of latent semantic analysis and humans. In M. G. Shafto & P. Langley (Eds.), *Proceedings of the 19th Annual Conference of the Cognitive Science Society* (pp. 412-417). Mahwah, NJ: Lawrence Erlbaum Associates.

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

Landauer, T., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25, 259-284*.

Landauer, T., Laham, D., & Foltz, P. W. (2000). The intelligent essay assessor. *IEEE Intelligent Systems, 15, 27-31*.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33, 159-174*.

Marcu, D. (2000). *The theory and practice of discourse parsing and summarization*. Cambridge, MA: Bradford / MIT Press.

Page, E. B. (1994). New computer grading of student prose, using modern concepts and software. *Journal of Experimental Education, 62, 127-142*.

Page, E. B. (1995). Computer grading of essays: a different kind of testing? *Address for APA Annual Meeting*, New York, 13-8-1995.

Page, E. B., & Petersen, N. S. (2001). The computer moves into essay grading: updating the ancient test. *Phi Delta Kappan : a journal for the promotion of leadership in education, 76, 561-565*.

Powers, D. E., Burstein, J., Chodorow, M., Fowles, M. E., & Kukich, K. (2001). *Stumping E-rater: challenging the validity of automated essay scoring* (ETS Research Report 01-03). Princeton,NJ: Educational Testing Service.

Rehder, B., Schreiner, M. E., Wolfe, M. B., Laham, D., Landauer, T. K., & Kintsch, W. (1998). Using latent semantic analysis to assess knowledge: some technical considerations. *Discourse Processes, 25, 337-354*.

Shermis, M. D., Mzumara, H. R., Olson, J., & Harrington, S. (2001). On-line grading of student essays: PEG goes on the World Wide Web. *Assessment & Evaluation in Higher Education, 26, 247-259*.

van Noortwijk, K., & de Mulder, R. V. (1997) *The similarity of text documents.* The Journal of Information, Law and Technology. http://elj.warwick.ac.uk/jilt/artininf/97_2noor/, Accessed: 08-10-2001.

Voss, J. F., Wiley, J., & Sandak, R. (1999). Reasoning in the construction of argumentative texts. In J. Andriessen & P. Coirier (Eds.), *Foundations of argumentative text processing* (pp. 29-41). Amsterdam: University of Amsterdam Press.

Voss, J. F. (1991). Informal reasoning and international relations. In J. F. Voss & D. N. Perkins (Eds.), *Informal reasoning and education* (pp. 37-58). Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Voss, J. F., Blais, J., Means, M. L., Greene, T. R., & Ahwesh, E. (1989). Informal reasoning and subject matter knowledge in the solving of economics problems by naive and novice individuals. In L. B. Resnick (Ed.), *Knowing, learning, and instruction: Essays in honor of Robert Glaser* (pp. 217-249). Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Wiemer-Hastings, P., & Zipitria, I. Rules for syntax, vectors for semantic. *Proceedings of the Twenty-third Annual Conference of the Cognitive Science Society.* Mahwah,NJ: Lawrence Erlbaum Associates.

Bijlage 1: Requirements patroonvergelijking bij antwoorden op open vragen

1. Invoer

   De student voert een antwoord in in een open tekstveld, volgend op de geformuleerde vraag.
   Nadat het antwoord is ingevoerd wordt nagegaan in hoeverre vooraf bepaalde patronen/constructies in het gegeven antwoord worden herkend.
   Het moet mogelijk zijn dat een student een eerder ingevoerd antwoord kan verbeteren en dat het nieuwe antwoord opnieuw wordt onderworpen aan patroonvergelijking.

2. Hints
   Het moet mogelijk zijn dat studenten door middel van het raadplegen van hints aanwijzingen krijgen voor het beantwoorden van de vraag.

3. Feedback
   Feedback vindt plaats op basis van het al dan niet aantreffen van één of meer constructies in een antwoord. Dit betekent dat afhankelijk van bepaalde patrooncombinaties specifieke feedback moet kunnen worden gegeven.
   Feedback bij een als 'voldoende/goed' beoordeeld antwoord omvat het modelantwoord.
   Na twee onvoldoende pogingen krijgt de student het modelantwoord ter beschikking.

4. Resultaat opslaan en tonen
   Alle pogingen worden opgeslagen in het dossier van de student. Het laatste resultaat wordt 'vastgehouden', c.q. getoond op het moment dat de student betreffende vraag weer oproept.
   Zowel student als docent kunnen de gegeven open antwoorden bekijken via de dossiergegevens.

5. Verwerking
   Het moet mogelijk zijn om één, of een combinatie van patronen te herkennen, en bij elke combinatie specifieke feedback te formuleren. Stel dat vooraf drie patronen worden vastgesteld die in het antwoord van de student dienen voor te komen, wil het antwoord als 'goed' worden beoordeeld, en daarnaast twee patronen worden geformuleerd waarvan studenten deze wellicht maar ten onrechte in hun antwoord zullen verwerken, moet het mogelijk zijn op alle combinaties van patronen feedback te formuleren.

   Daarbij is het wenselijk dat tot op zekere hoogte een tolerantie wordt ingebouwd met betrekking tot spelfouten. Een mogelijkheid is om in de terugkoppeling te werken met een percentage 'gevonden overeenkomst'.

Bijlage 2: Voorbeelden van argument en issue questions in de GMAT

Analysis of Argument Questions for the GMAT

The following appeared in the health section of a magazine on trends and lifestyles.

.People who use the artificial sweetener aspartame are better off consuming sugar, since aspartame can actually contribute to weight gain rather than weight loss. For example, high levels of aspartame have been shown to trigger a craving for food by depleting the brain of a chemical that registers satiety, or the sense of being full. Furthermore, studies suggest that sugars, if consumed after at least 45 minutes of continuous exercise, actually enhance the body.s ability to burn fat. Consequently, those who drink aspartame-sweetened juices after exercise will also lose this calorie-burning benefit. Thus it appears that people consuming aspartame rather than sugar are unlikely to achieve their dietary goals..

Discuss how well reasoned . . . *etc*.

The following appeared in the editorial section of a corporate newsletter.

The common notion that workers are generally apathetic about management issues is false, or at least outdated: a recently published survey indicates that 79 percent of the nearly 1,200 workers who responded to survey questionnaires expressed a high level of interest in the topics of corporate restructuring and redesign of benefits programs..

Discuss how well reasoned . . . *etc*.

Analysis of Issue Questions for the GMAT

In some countries, television and radio programs are carefully censored for offensive language and behavior. In other countries, there is little or no censorship.

In your view, to what extent should government or any other group be able to censor television or radio programs? Explain, giving relevant reasons and/or examples to support your position.

It is unrealistic to expect individual nations to make, independently, the sacrifices necessary to conserve energy. International leadership and worldwide cooperation are essential if we expect to protect the world.s energy resources for future generations..

Discuss the extent to which you agree or disagree with the opinion stated above. Support your views with reasons and/or examples from your own experience, observations, or reading.

Corporations and other businesses should try to eliminate the many ranks and salary grades that classify employees according to their experience and expertise. A flat organizational structure is more likely to encourage collegiality and cooperation among employees.

Discuss the extent to which you agree or disagree with the opinion stated above. Support your views with reasons and/or examples from your own experience, observations, or reading.

Bijlage 3: De techniek van latente semantische analyse

De Intelligent Essay Assessor maakt gebruikt van Latente Semantische Analyse, een techniek die oorspronkelijk is ontworpen ten behoeve van het automatisch indexeren en vinden van documenten in databases. De moeilijkheid daar is niet om documenten te vinden die de opgegeven zoektermen bevatten, maar om documenten te vinden die over het bedoelde onderwerp gaan, met andere woorden de latente semantiek moet worden gevonden.
Het voert te ver om hier in te gaan op de details van LSA, we geven slechts een oppervlakkige schets. Uitvoeriger beschrijvingen van LSA zijn elders te vinden (Deerwester, Dumais, Furnas, Landauer & Harshman, 1990; Landauer, Foltz & Laham, 1998).

Technisch gezien is LSA een generalisatie van factoranalyse. Bij factoranalyse wordt uitgegaan van een symmetrische correlatiematrix die wordt gesplitst in het product van drie matrices $F \Lambda F'$ waar F de matrix met orthogonale eigenvectoren en Λ de diagonaalmatrix met de eigenwaarden is. In een factoranalyse worden de kleinere eigenwaarden op nul gezet. Het resultaat is een vermindering van het aantal dimensies.

In LSA wordt iets soortgelijks gedaan maar niet op een symmetrische matrix, maar op een Term x Document matrix. Term en document moeten hier abstract worden opgevat: een term kan een enkel begrip zijn, maar ook uitvoerige passages bevatten. Een document is de container waarin Termen zijn opgenomen: een alinea, een hoofdstuk, een document enz. In de cellen van de tabel zijn gewichten opgenomen die het belang van de term in het document weergeven. In het meest eenvoudige geval is dat een 0 of een 1 (dat wil zeggen komt niet voor / komt wel voor).

Tabel 1: Term x document matrix

	D1	D2	D3	D4	D5
T1	g_{11}	g_{12}			
T2					
T3					
T4					
T5					

Bij een LSA wordt de Term*Document-matrix gesplitst in het product van drie matrices waarvan de eerste de eigenvectoren van de Termen bevat, de tweede weer een diagonaalmatrix is en de derde de eigenvectoren voor de Documenten bevat. Net als bij factoranalyse vindt vervolgens reductie van dimensies plaats. Overigens gaat het er bij LSA niet om om een zo spaarzaam mogelijk model te krijgen. In de literatuur worden Term* Document Matrices van 1000 bij 6000 genoemd die worden gereduceerd tot een 100 bij 300 matrix. De gereduceerde matrix bevat de latente semantische factoren waarmee de oorspronkelijke matrix kan worden benaderd. Ieder woord en ieder document krijgt 'factorscores' in de nieuwe matrix.
Een standaardmanier om twee documenten met elkaar te vergelijken is de berekening van het inproduct van hun vectoren, gecorrigeerd voor de lengte van de vectoren. Dit product is gelijk aan de cosinus van de hoek die de twee vectoren met elkaar hebben: als deze cosinus 0 is, vallen de vectoren samen. Op analoge wijze kan worden bepaald hoezeer twee termen overeenkomen. Details over mogelijke andere maten zijn elders uitvoerig besproken (Rehder et al., 1998).
Bij de beoordeling van essayvragen wordt de LSA gevoed door diverse teksten over het onderwerp (bijvoorbeeld een cursusboek) en door een reeks voorbeelden van de te

beoordelen teksten. Dit zijn dan in principe weer te voren beoordeelde essays. Nieuw te beoordelen essays worden vervolgens vergeleken met deze voorbeelden.
LSA maakt geen gebruik van linguïstische analyse; syntax is niet van belang: er wordt alleen geanalyseerd op woorden (ongeacht de volgorde waarin ze staan) en contexten. Er is enige evidentie dat het LSA-mechanisme bij korte antwoorden baat kan hebben als syntactische informatie wordt toegevoegd (Wiemer-Hastings & Zipitria,).
LSA is niet in de eerste plaats ingezet voor het beoordelen van teksten, integendeel. Landauer en Dumais (1997) hebben LSA als een proces beschreven waarmee kennis kan worden geïnduceerd. Zij presenteren onderzoeksresultaten op divers gebied (leren van synoniemen en tegenstellingen, priming effecten, leren van vertalingen, enz) om deze stelling te onderbouwen.