NT2-onderwijs, ordeningscriteria met behulp van half-automatische tekstanalyse

Renkema, Jan

Published in:
Levende Talen

Publication date:
1989

NT2-onderwijs: Ordeningscriteria met behulp van half-automatische tekstanalyse*)

J. Renkema

De tekstwetenschap houdt zich weinig bezig met het NT2-onderwijs. Mijn enige ervaring op dit terrein bestaat uit een tweejarig docentschap bij de Volksuniversiteit in Amsterdam, zo'n vijftien jaar geleden. Ik behandelde toen met een groep volwassenen van zeer verschillende nationaliteiten het boek van Shetter. (Ik wist nog weinig en er was nauwelijks lesmateriaal.) Er waren wel problemen: Moet je authentieke teksten behandelen? Zo ja, moet je die dan aanpassen? Hoe bepaal je of een tekst niet te moeilijk is? Ik ging af op eigen intuïtie, en als ik aarzelde over de moeilijkheidsgraad van een tekst vroeg ik advies aan een meer ervaren collega.

Een van de belangrijkste onderzoeksvragen in de tekstwetenschap is de vraag naar de relatie tussen vorm en functie. Een concretisering van deze vraag is: Wat is de relatie tussen woordkeus en zinsbouw (de vorm) en de begrijpelijkheid voor bepaalde groepen lezers (de functie)? voor het beantwoorden van deze vraag is het noodzakelijk dat teksten snel en betrouwbaar kunnen worden geanalyseerd. In dit artikel lever ik een discussiebijdrage aan de problematiek over begrijpelijkheid (1, Ordeningscriteria) en geef ik informatie over een systeem waarmee teksten op woordniveau kunnen worden geanalyseerd met behulp van de computer (2, Half-automatische tekstanalyse). Tot slot wordt in paragraaf 3 kort aangegeven voor welke typen vragen dit systeem kan worden gebruikt.

1. Ordeningscriteria

1.1. De notie 'moeilijkheidsgraad'

In NT2-literatuur over de geschiktheid van al dan niet authentieke teksten wordt vaak gesproken over 'moeilijkheidsgraad'. Een analyse van het gebruik van deze term roept echter vragen op. Deze vragen gelden trouwens niet alleen het NT2-onderzoek, maar in dit kader beperk ik me hiertoe. Als referentiepunt neem ik het artikel van Hulstijn en Vedder (*LT* 442). Deze tekst vormde de basis voor het symposium. In paragraaf 7.1 'Syntactische complexiteit' staan de volgende uitspraken.

(1) Niet-continue zinsdelen kunnen het zinsbegrip bemoeilijken. Een voorbeeld hiervan is de uiteenplaatsing van finiet hulpwerkwoorden en infinitief of deelwoord (de zogenaamde 'tangconstructie').

(2) Het aantal argumenten per predikaat kan de moeilijkheid beïnvloeden.

(3) Verwijzende woorden (bijv. voornaamwoorden) zijn lastiger als ze vooruit dan wanneer ze achteruit verwijzen.

Opvallend in de eerste twee uitspraken is het modale karakter: 'kunnen bemoeilijken'; 'kan de moeilijkheid beïnvloeden'. De tweede uitspraak suggereert: hoe meer argumenten, des te moeilijker. Maar dit is lang niet altijd waar. Wanneer in een bepaalde context een noodzakelijk argument niet genoemd wordt, bijvoorbeeld 'Jan slaat', krijgen we ook een moeilijk interpreteerbare zin. De derde uitspraak over verwijzingen naar voren en naar achteren lijkt te algemeen. Vergelijk de volgende zinnen. Zin (4a) met een vooruitwijzend voornaamwoord is niet lastiger dan (4b) waar 'ik' terugwijst

(4a) Ik zal dat wel doen, zei Piet.

(4b) Piet zei: ik zal dat wel doen.

Opmerkingen over de moeilijkheidsgraad van teksteigenschappen zijn dikwijls te algemeen. Zodra een bepaalde eigenschap van een tekst als oorzaak wordt aangeduid voor de moeilijkheidsgraad, wordt de suggestie gewekt dat het vermijden van die teksteigenschap de begrijpelijkheid verhoogt. Men gaat dan voorbij aan het feit dat die eigenschap ook nog andere functies kan hebben. Aan de hand van de drie citaten zullen in de volgende paragraaf voorbeelden gegeven worden van het verschijsel dat één teksteigenschap meer dan één functie kan hebben.

1.2. Eén eigenschap, meer functies

Voor de stelling dat een tangcontructie moeilijker leesbaar is dan een onttangde versie zijn in de psycholinguïstische literatuur verschillende argumenten aangedragen. Neem de volgende zinsparen.

(5a) Het Cito deelde desgevraagd mee dat, hoewel de score alleen een indicatie is voor een eventuele vervolgopleiding, de toets steeds meer de status krijgt van het belangrijkste basis-examen.

(5b) Het Cito deelde desgevraagd mee dat de toets steeds meer de status krijgt van het belangrijkste basis-examen, hoewel de score alleen een indicatie is voor een eventuele vervolgopleiding.

(6a) De langdurige echte minima verkeren op dit moment in een onmenselijke positie aldus een door medewerkers van de Leidse Universiteit geschreven rapport.

(6b) De langdurige echte minima verkeren op dit moment in een onmenselijke postitie aldus een rapport dat door medewerker van de Leidse Universiteit is geschreven.

In (5a) staat een zogenoemde bijzin-tang. Na het woordje 'dat' wordt er een bijzin tussengevoegd. Zo'n constructie is moeilijker dan die in (5b) omdat de lezer een nieuwe procedure moet starten,

en tegelijkertijd het begin van de eerste bijzin moet vasthouden. Een soortgelijk verschijnsel zien we in de lidwoord-naamwoord tangcontructie in (6a). Tussen 'een' en 'rapport' staan woorden die verwerkt moeten worden, terwijl de lezer het woordje 'een' ook moet vasthouden. Maar betekenen de b-versies nu hetzelfde als de a-versie? Nee, er is een belangrijk verschil. Er komt meer accent op het gedeelte dat buiten de tang wordt geplaatst. Of andersom geredeneerd: zodra een gedeelte in een tangcontructie staat, lijkt het alsof de schrijver aan dit deel minder gewicht wil toekennen. Dit valt goed te zien aan de volgende voorbeelden.

(7a) Een huisarts in Apeldoorn heeft gisteren na buurtdemonstraties zijn praktijk gesloten. Hij werd verdacht van ongewenste intimiteiten.

(7b) Een huisarts in Apeldoorn die verdacht werd van ongewenste intimiteiten, heeft gisten na buurtdemonstraties zijn praktijk gesloten.

(7c) Een van ongewenste intimiteiten verdachte huisarts in Apeldoorn heeft gisteren na buurtdemonstraties zijn praktijk gesloten.

in (7a) krijgen de 'intimiteiten' door de aparte zin een veel zwaarder accent dan in de volgende zinnen. Door de rangschikking in (7b) en (7c) krijgt deze mededeling minder gewicht. Ondersteuning van deze intuïtie vinden we in enkele experimenten (zie Renkema 1989) die aantoonden dat typfouten in het middendeel van een bepaalde tangcontructie minder opvallen dan in de onttangde versie. Maar diezelfde experimenten gaven ook aan dat het voor het tekstbegrip bij bepaalde groepen lezers weinig uitmaakte of de 'getangde' of de 'onttangde' versie wordt gebruikt. Met andere woorden: het is lang niet altijd duidelijk of het vermijden van een tangcontructie een gemakkelijker tekst oplevert. Bovendien wordt bij het herschrijven van een tangcontructie ook de 'attentiewaarde' veranderd.

Om ruimte te besparen behandel ik de uitspraken over 'aantal argumenten per predikaat' en het gebruik van verwijswoorden aan de hand van dezelfde voorbeelden. Ook deze verschijnselen kunnen niet alleen in termen van moeilijkheidsgraad worden behandeld. Want de factor 'voorkennis' kan altijd roet in het eten gooien. Neem de volgende zinsparen.

(8a) Auke kocht een boek. Zijn moeder was jarig.

(8b) Auke kocht een boek. Zijn moeder was blij dat hij zijn zakgeld niet versnoepte.

(9a) Maaike vertelde alles aan Brechtje. Zij kon haar mond niet houden.

(9b) Maaike vertelde alles aan Brechtje. Zij kon haar mond niet houden en verraadde Maaike's geheim.

In (8a) wordt niet meegedeeld dat Auke het boek voor zijn moeder kocht. Toch zullen de meeste lezers dit zonder moeite uit de tekst kunnen afleiden. De argumentrelaties zijn hier duidelijk door onze voorkennis. Deze voorkennis speelt zo'n belangrijke rol, dat geen enkele lezer in (8b) op de gedachte komt dat Auke een boek voor zijn moeder koopt. Verwijswoorden kunnen problemen opleveren wanneer er twee mogelijke antecedenten zijn, zoals in (9a). De interpretatieregel luidt dat het verwijswoord in eerste instantie terugslaat op het onderwerp van de vorgie zin. Maar in (9b) is deze interpretatieregel door inhoudelijke aanwijzingen niet van toepassing. Kortom, opmerkingen over mogelijke problemen bij verwijswoorden kunnen nooit los van de context worden gezien.

Naar aanleiding van deze drie teksteigenschappen (tangcontructies, argumentrelaties, verwijswoorden) zal duidelijk geworden zijn dat het bepalen van de moeilijkheidsgraad van een tekst een hachelijke ondernemening is. Ordeningscriteria voor leesteksten moeten dan ook met de grootst mogelijke voorzichtigheid worden gehanteerd.

Dit lijkt een heel sombere cunclusie. Maar vanuit de tekstwetenschap is er nog een andere invalshoek mogelijk, namelijk die van de kwantitatieve tekstanalyse. Uiteraard is deze vorm van analyse niet zaligmakend. Wel kunnen op deze manier teksteigenschappen worden gedetecteerd, waarmee NT2-docenten hun voordeel kunnen doen. Vooral als zo'n analyse kan worden uitgevoerd met behulp van de computer. In het vervolg van dit artikel zal eerst iets gezegd worden over kwantitatieve analyse, met als voorbeeld de type-token ratio. Daarna zal informatie worden gegeven over een analysesysteem waaraan gewerkt wordt aan de Tilburgse Letterenfaculteit.

2. Half-automatische tekstanalyse

2.1 Kwantitatieve analyse

Een bekend voorbeeld van telwerk in tekstanalyse is de type-token ratio. Dit is de verhouding tussen het aantal woorden ineen tekst (de 'types') en het totaal aantal woorden in een tekst (de 'tokens'). Deze maat voor woordenrijkom wordt ook gezien als indicator voor tekstmoeilijkheid. Hoe hoger het aantal verschillende woorden per bijvoorbeeld honderd woorden, des te moeilijker is de tekst. Toch is ook deze uitspraak niet zonder meer juist. Voor zes genres teksten zijn tellingen verricht (zie Uit den Boogaart 1975 en Renkema 1983). De gegevens staan in de volgende tabel.

(10)

Genres	types	tokens	ttr
Dagbladen	11884	48242	0,25
Opiniebladen	11161	48242	0,23
Pop.wet.proza	10557	48242	0,22
Romans/novellen	9801	48242	0,20
Gezinsbladen	10935	48242	0,23
Overheidstaal	7930	48242	0,16

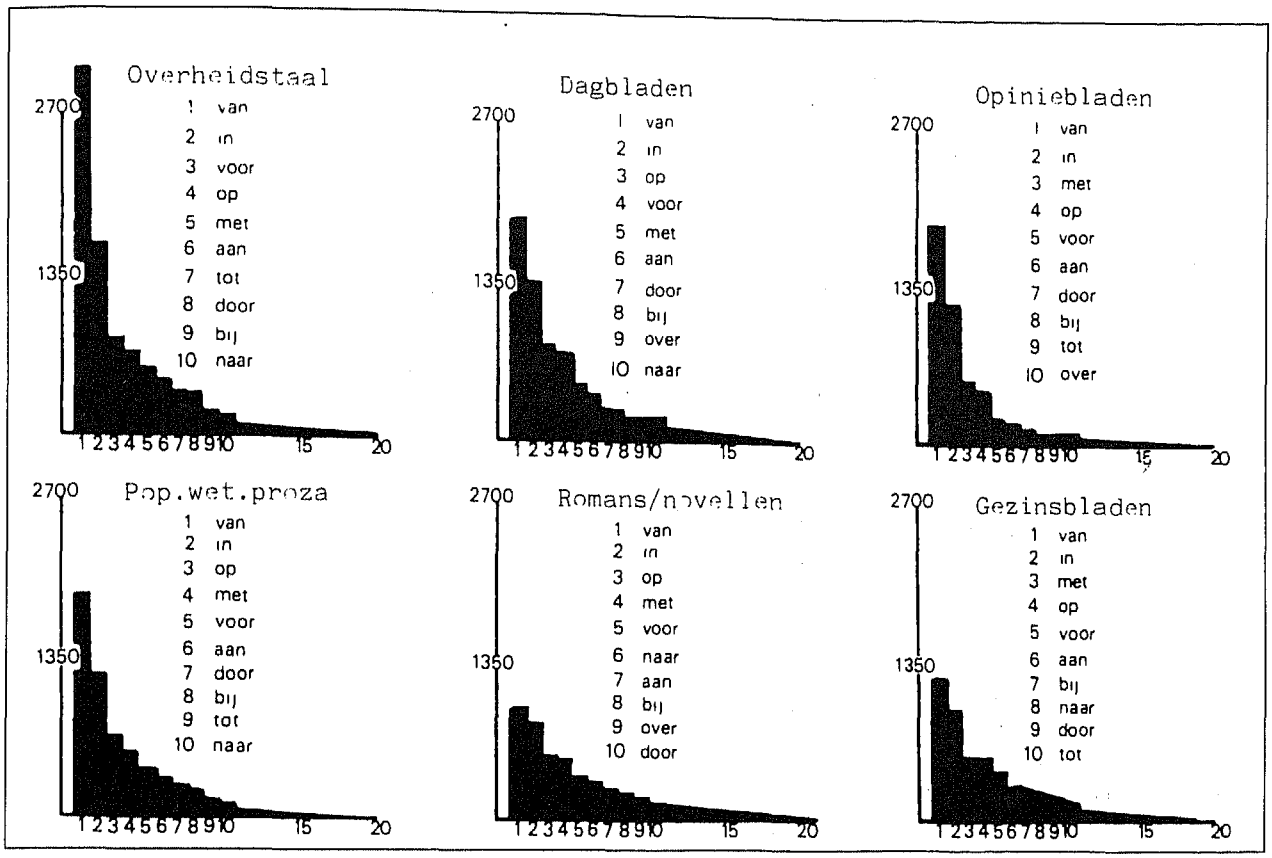
Uit deze tabel zou geconcludeerd moeten worden dat overheidstaal het gemakkelijkst is. Het bevat het laagste aantal verschillende woorden. Toch strookt deze conclusie niet met de gangbare oordelen over de moeilijkheidsgraad van dit type taalgebruik. In bovenstaande tabel ging het om de type-token ratio ongeacht de woord soort. In vervolgonderzoek (Renkema 1984) is ook de type-token ratio voor verschillende woordsoorten gemeten. In de volgende tabel staan de gegevens over bijvoeglijke naamwoorden voor twee genres.

(11) Bijvoeglijke naamwoorden

	types	token	ttr
Opiniebladen	1073	2875	0,37
Overheidstaal	676	2570	0,26

Het verschil is duidelijk, maar het is zeer de vraag of hiermee iets gezegd wordt over een eventueel verschil in moeilijkheidsgraad. Men zou kunnen beargumenteren dat een tekst moeilijker is, naarmate hij saaier is, en dat juist bijvoeglijke naamwoorden zorgen voor levendigheid. Maar voor dit argument is nog geen bewijs geleverd. Soortgelijke gegevens zijn verzameld over voorzetsels. Maar ook hier zijn niet direct conclusies over moelijkheidsgraad te trekken. Wel bracht deze analyse andere gegevens aan het licht, bijvoorbeeld de tien frequentste voorzetsels in de hierboven vermelde zes genres.

(12) De tien frequentste voorzetsels in zes genres.



Een opvallend verschil tussen 'overheidstaal' en de andere genres is de hoge frequentie van het voorzetsel 'van', 2690, tegenover bijvoorbeeld 872 in romans/novellen. Het hoge aantal voorzetsels in 'overheidstaal' hangt samen met het nominale karakter van dit type taalgebruik, waarin veel nominale constituenten aan elkaar worden gekoppeld. (Voor deze koppeling zijn voorzetsels nodig.) De analysevoorbeelden tot nu toe konden worden gepresenteerd op basis van corpora teksten die met de hand zijn gecodeerd. Wanneer een NT2-docent soortgelijke gegevens wil hebben over nieuwe teksten dan is het veel te tijdrovend om dit monnikenwerk te verrichten. Waarschijnlijk kan dan een analysesyteem van nut zijn dat op dit moment in Tilburg wordt ontwikkeld. In de volgende paragraaf zal dit systeem beknopt worden uitgelegd. In de slotparagraaf wordt aangegeven welke typen vragen met dit systeem zijn te beantwoorden.

2.2 Het Hata-systeem

Het Half-automatische analysesysteem is ontwikkeld om tijrovend en onnauwkeurig handwerk te besparen. Het doel van Hata is een zodanige programmatuur te ontwerpen dat de computer, met beperkte interactieve ingrepen, achter elk woord in de tekst een code zet met informatie over de woordsoort. Het codeersysteem is gebaseerd op Uit den Boogaart 1975. Dit betekent dat het aantal onderscheidingsmogelijkheden zeer rijk is. Een aardig voorbeeld is het woord 'waar', dat acht coderingen kan krijgen.

(13) waar (000) voor je geld krijgen. zelfst.naamw.
een waar (100) genoegen. bijv. naamw.
Ik waar (241) door Europa, zei het spook. pv. intrans. ww
Waar (250) ben je? vragend vnw.
de plaats waar (530) ik woon. betr. vnw
Waar (550) ga je heen? vragend vnmw bijw.
Een plaats waar (560) je heen kunt. betr vnmw bijw
Waar (710) wij ons best doen, mag jij je er niet van afmaken. ondersch. voegw,

Met behulp van de nu ontwikkelde programmatuur (zie Kempff, Van Opstal en Renkema 1987) kan de computer aan elk woord in een willekeurige tekst de mogelijke coderingen toekennen, bijvoorbeeld:

(14) mijn rechter hand
000 000 000
330 100

De code 000 staat voor 'zelfstandig naamwoord'; de code 100 voor 'bijvoeglijk naamwoord' en 330 voor 'bijvoeglijk gebruikt bezittelijke voornaamwoord'. Via een reeks van contextregels besluit de computer tot een selectie uit de mogelijke codes. Deze regels zijn gebaseerd op overgangswaarschijnlijkheden, bijvoorbeeld 'geen twee zelfstandige naamwoorden zonder leestekens direct na elkaar' of 'geen bijvoeglijk naamwoord na een zelfstandig naamwoord'. Aan het systeem van contextregels wordt nu nog gewerkt. (Het zal overigens duidelijk zijn dat dit systeem nooit geheel waterdicht kan zijn.)

Het volgende tekstje geeft een indruk van wat Hata wel en niet kan.

(15) Passende arbeid

Werklozen moeten zoeken naar werk. Dat is een voorwaarde voor een uitkering. "Passende arbeid" moeten ze aksepteren. Passende arbeid is "geschikt werk". Het arbeidsbureau zegt: "Dit is passende arbeid voor u". Dan moet je dat werk dus aksepteren. Er is weinig werk. Dus er is niet veel keus. Je moet alles aksepteren. Bijna alles. De Staatssekretaris van Sociale Zaken en Werkgelegenheid zegt namelijk: "Mensen met een verzorgende taak hoeven geen nachtwerk te aksepteren". Vrouwen met kleine kinderen bijvoorbeeld hoeven dus niet 's nachts te werken. Mensen boven de 45 jaar hoeven ook geen nachtwerk te aksepteren. Dat is

geen passende arbeid voor hen. Zij mogen nachtwerk weigeren en houden toch hun uitkering.
Deze tekst werd in gevoerd en kwam met de volgende coderingen uit het programma.

(16) Passende TD3 arbeid N .P
Werklozen ANm moeten 230 (274) zoeken VI naar *710* (600) werk N .P Dat 360 is *272*(273) een *440* (450) voorwaarde *Vev* (N) voor 710 een *440* (450) uitkering N . P "R Passende TD3 arbeid N " R moeten 274 ze 300 aksepteren **Vmt** (VI) .P Passende TD3 arbeid N is *272* (273) " R geschikt A werk N " R . P Het 300 arbeidsbureau N zegt V3e : L " R Dit 360 is 273 passende TD3 arbeid Vle (N) voor *710* (600) u *300* (303) " R . P Dan !M moet 272 je 440 dat *360* (370) werk Vle dus **!M** (5MP) aksepteren Vmt. P Er 5ER is 273 weinig 430 werk N . P dus **!V** (5MP) er 5ER is 273 niet 5ni veel 450 keus N . P Je 300 moet 272 alles 440 aksepteren VI . P Bijna 5HN alles 440 . P De 370 Staatssekretaris N van 600 Sociale A3 Zaken Nm en 700 Werkgelegenheid N zegt V3e namelijk 5V : L " R Mensen Nm met 600 een 450 verzorgende TD3 taak N hoeven **Vmt** (274) geen 450 nachtwerk N te 650 aksepteren VI " R .P Vrouwen Nm met 600 kleine A3 kinderen Nm bijvoorbeeld 5V hoeven **Vmt** (274) dus **!V** (5MP) niet *Vle* (5ni) ' R s 372 nachts N2 te 650 werken VI . P Mensen Nm boven 600 de 370 45 460 jaar N hoeven **Vmt** (274) ook 5MP geen 450 nachtwerk N te 650 sksepteren VI . P Dat 360 is 273 geen 450 passende TD3 arbeid N voor 600 hen 303 . P Zij 300 mogen 274 nachtwerk N weigeren **Vmt** (VI) en 700 houden Vmt toch 5V hun 330 uitkering N . P

Een proefversie van deze tekst van 115 woorden bevatte een twintigtal fouten, de codes in cursief en vet. In een enkel geval gaf het programma niet de mogelijk juiste codes. Een voorbeeld is 'voorwaarde' dat als werkwoord werd gecodeerd, zoals 'bewaarde' en 'ontwaarde'. Zo'n fout kan worden verbeterd door het lexicon aan te passen. De meeste fouten betroffen een verkeerde geselecteerde code, bijvoorbeeld 'een' dat zelfstandig (440) of bijvoeglijk (450) gebruikt kan worden. Deze fouten kunnen worden hersteld door contextregels aan te passen. Na enkele verbeteringen bleven nog een tiental fouten staan (de vetgedrukte codes). Ook deze kunnen op dezelfde wijze worden verbeterd. Voor bijvoorbeeld 'hoeven' moet het lexicon worden aangepast, omdat bij dit woord de code hulpwerkwoord ontbreekt. Bij 'moeten' moet de contextregel worden aangepast om in dit zinsverband de juiste code te krijgen. Belangrijker is het volgende. Verreweg de meeste fouten hebben te maken met het toekennen van het tweede of derde cijfer. Dus de hoofdcategorie (werkwoord, naamwoord enz.) is juist gecodeerd. Door de opbouw van het programma kunnen fouten worden hersteld door aanvullingen op het lexicon of het herformuleren van contextregels. Dit betekent uiteraard niet dat het programma foutloos werkt. De tekstanalist zal altijd correcties moeten aanbrangen in dit *half*-automatisch systeem. Hoofdzaak is dat dit analysesysteem veel tijdrovend werk bespaart, wanneer men van een tekst gegevens wil hebben op woordniveau.

3. Analyse van teksteigenschappen

In het eerste deel van dit artikel is kritiek geleverd op ordeningscriteria. Wat levert het Hata-systeem dat in het tweede deel is geïtroduceerd nu op? Met dit systeem kan een NT2-docent sneller en nauwkeuriger dan tot nu toe mogelijk was bepalen of willekeurig welke tekst bepaalde eigenschappen vertoont. Wanneer een docent teksten wil hebben met veel verschillende voorzetsels, of pronominale verwijzingen of veel hulpwerkwoorden enz. enz. dan is dit analysesysteem een gemakkelijk hulpmiddel. Zodra een vraag naar teksteigenschappen in termen van woordsoorten is te formuleren, kan dit systeem hulp bieden. Het gaat dus om vragen als: Hoeveel onderschikkende voegwoorden bevat deze tekst, en welke zijn dat? Welke verbogen adjectieven staan er in deze tekst? Met welke teksten kan het gebruik van 'er' geoefend worden? Op deze manier kunnen teksten geordend worden op grond van woordsoortcriteria. Hiermee is uiteraard geen uitspraak gedaan over de moeilijkheidsgraad. Maar wanneer veel teksten volgens Hata zijn geanalyseerd, komen er wel meer gegevens beschikbaar waarop uitspraken over teksteigenschappen kunnen worden gebaseerd.

*) Ter gelegenheid van de viering van "25 jaar NT2-onderwijs aan de Vrije Universiteit" werd mij verzocht een tekstwetenschappelijke bijdrage te leveren voor mensen uit het veld. Dit artikel is een bewerking van de lezing op het symposium

Literatuur

Hulstijn, J.H., en I. Vedder (1989), 'Ordeningscriteria voor leesteksten in het tweede-taalonderwijs', in: Levende Taal.
Kempff, H., T. van Opstal en J. Renkema (1987) 'Tekstanalyse per computer; het lexicon', in *Gramma* (3), p. 169-190.
Renkema, J. (1983), 'Onderzoek naar woordenrijkdom; taalstatistische analyse via een microcomputer', in: *Tijdschrift voor Taalbeheersing* (4) p. 275-289.
Renkema, J. (1984), 'On functional and computational LSP-analysis. Officialese as an example', in: Pugh, A.K., and Ulijm, J.M. (eds), *Reading for Professional Purposes: Studies in Native and Foreign Languages*. London: Heinemann Educational, p. 109-119.
Renkema, J. (1989), 'Tangcontructies, experimenteel onderzoek naar leesbaarheid en attentiewaarde', te verschijnen.
Shetter, W.Z. (1973), *Introduction to Dutch. A Practical Grammar*. The Hague: Martinus Nijhoff.
Uit den Boogaart, P.C. (red), (1975), *Woordfrequenties in Geschreven en Gesproken Nederlands*. Utrecht: Bohn, Scheltema & Holkema.

Jan Renkema (1948) doceert tekstwetenschap aan de Katholieke Universiteit Brabant, enis tevens buitengewoon hoogleraar aan de afdeling Cultuurwetenschappen van de Open Universiteit. Adres: Van Beverwijkstraat 5, 5571 BR Bergeyk.

Erratum

In de Bibliografie van het artikel 'Grammatica als instrument voor schrijven' van W.I.M. van Calcar (*Levende Talen* 442, juni 1989) is een gedeelte van de tekst weggevallen op p. 394, waardoor onjuiste informatie wordt gegeven. In plaats van de als tweede genoemde publikatie van Van Calcar moet gelezen worden:
Calcar, W.I.M. van, Grammatica als deel van taalbeschouwing, in: *Levende Talen* 340 (1979), 203-217.
Calcar, W.I.M. van, *Een nieuwe grammatica. Voor taalbeschouwing en taalbeheersing*, Leuven/Amersfooert, 1983: Acco.