Rogier Kraf, Leo Lentz en
Henk Pander Maat

Drie Nederlandse instrumenten voor het automatisch voorspellen van begrijpelijkheid

Een klein consumentenonderzoek

Samenvatting

Tientallen jaren van leesbaarheidsonderzoek hebben laten zien dat het ontwikkelen van een betrouwbaar apparaat voor het meten van leesbaarheid van teksten geen eenvoudige zaak is. Ondersteund door nieuwe taaltechnologie lijkt het leesbaarheidsonderzoek voor het Engels de afgelopen jaren begonnen te zijn aan een tweede leven. Op de markt voor het Nederlands is vandaag de dag een drietal applicaties te vinden. Over de werking van deze instrumenten is echter nog weinig bekend. Om duidelijkheid te krijgen zochten wij contact op met de makers van deze programma's. Daarnaast voerden we een klein experiment uit waarbij we de output van de instrumenten hebben vergeleken met elkaar en met enkele tekstkenmerken waarbij we een hoge correlatie met leesbaarheidsscores verwachtten. Resultaten laten zien dat de instrumenten het vaak niet met elkaar eens zijn en dat er uitgebreider onderzoek nodig is om meer te weten te komen over hun betrouwbaarheid en validiteit.

1. Inleiding

Zou het niet geweldig zijn wanneer de computer aan de hand van een serie tekstkenmerken automatisch een moeilijkheidsscore kan geven aan een tekst?[1] Een halve eeuw lang heeft het onderzoek naar automatische leesbaarheidsvoorspelling, kortweg leesbaarheidsonderzoek, een grote bloei gekend. Al vanaf de begindagen (Vogel & Washburne, 1928) werd dit onderzoek gestuurd door sterke behoeften uit de praktijk. Vooral in het onderwijs was er grote vraag naar een snelle, goedkope en objectieve methode om teksten van een geschikt niveau te selecteren voor leerlingen die sterk van elkaar verschilden in hun leesvaardigheid. Leesbaarheidsformules leken decennialang het ideale antwoord op die vraag: door simpelweg het aantal letters, woorden en zinnen in de tekst te tellen, kon de leesbaarheidsformule vertellen voor welke lezers een tekst geschikt was. Beroemde formules werden ontworpen door onderzoekers zoals Flesch (1948) en Dale en Chall (1948). Deze formules voorzien teksten van Grade levels: een 10-jarige leerling uit klas 5 van het Amerikaanse onderwijssysteem zou een tekst die door de formule gelabeld is als "Grade level 5" goed moeten kunnen begrijpen.

Maar in de jaren '70 en '80 verschenen nogal wat publicaties met fundamentele kritiek op het leesbaarheidsonderzoek (zie voor een overzicht Kraf & Pander Maat, 2009). Zo spelen lezerkenmerken en interacties tussen lezer- en tekstkenmerken geen rol in een leesbaarheidsformule en wordt er qua tekstkenmerken alleen rekening gehouden met een beperkt aantal simpele woord- en zinskenmerken (de formule is blind voor globale tekststructuur en coherentie), kenmerken die bovendien weinig te maken hebben met de werkelijke oorzaken

van leesproblemen in een tekst. Langere zinnen zijn niet altijd moeilijkere zinnen; sterker nog, de suggestie dat men een tekst kan verbeteren door zinnen in tweeën te knippen en voegwoorden te verwijderen is gevaarlijk. Daarnaast is de voorspellende kracht van leesbaarheidsformules systematisch overschat door onderzoekers, doordat altijd is gewerkt met gemiddelde groepsprestaties in plaats van individuele begripsprestaties.

Hoewel het leesbaarheidsonderzoek door deze kritiek gedurende ongeveer een kwart eeuw op een laag pitje is komen te staan, bleven leesbaarheidsformules in de praktijk onverminderd populair, althans in de bakermat van het leesbaarheidsonderzoek, de VS. In Nederland kennen we niet zo'n omvangrijke traditie op het gebied van leesbaarheidsvoorspelling. De enige serieuze leesbaarheidsformule in de 20e eeuw voor het Nederlands is de CLIB formule van het CITO geweest (Staphorsius, 1994). Deze formule is nog altijd onderdeel van het AVI-systeem, waarmee boeken voor kinderen in de basisschoolleeftijd geclassificeerd kunnen worden op moeilijkheid.

De laatste jaren is het onderzoek naar leesbaarheidsvoorspelling weer op gang gekomen, onder andere doordat taaltechnologische vooruitgang ervoor gezorgd heeft dat nieuwe inzichten over tekstverwerking geïmplementeerd kunnen worden en we op die manier in staat zouden moeten zijn "slimmere" tekstkenmerken te bouwen die dichterbij de werkelijke oorzaken van leesproblemen staan. Het meeste van dit nieuwe leesbaarheidsonderzoek vindt plaats voor het Engels (o.a. Collins-Thompson & Callan, 2005; Schwarm & Ostendorf, 2005, Crossley et al., 2006; Heilman et al., 2007), voor het Nederlands zijn Kraf & Pander Maat (2009) en Van Oosten et al. (2010) actief op het gebied van leesbaarheidsonderzoek.

Onafhankelijk van dit onderzoek zijn er de afgelopen jaren enkele leesbaarheidsinstrumenten op de Nederlandse markt gekomen. Ons zijn er drie bekend:

- Texamen van BureauTaal;
- Klinkende Taal van het taaltechnologiebedrijf Gridline;
- De Accessibility Leesniveau Tool van de Stichting Accessibility.

De makers van deze tools stellen "het taalniveau" van een tekst te kunnen vaststellen, maar hoe deze instrumenten dat doen, wordt niet helemaal duidelijk uit de informatie op de bijbehorende websites.

Het doel van dit kleine consumentenonderzoek is de drie instrumenten te vergelijken wat betreft hun werking en de voorspellingen die ze opleveren. Texamen en Gridline combineren de niveauvoorspelling met een pakket aan hulp bij het verbeteren van teksten, maar in het verslag van dit onderzoek naar deze instrumenten gaan we voorbij aan die aanvullende diensten. We benaderden de drie organisaties met drie vragen:

1. Met welke tekstkenmerken werkt het programma?
2. Hoe wordt op basis van die kenmerken een leesbaarheidsniveau gedefinieerd?
3. Welk onderzoek ligt er ten grondslag aan de kenmerkenkeuze en de werking van het programma?

Daarnaast leverden we iedere organisatie een twintigtal teksten van ca. 275 woorden met het verzoek deze teksten te analyseren met het programma. Voor die teksten vroegen we om een bepaling van het leesniveau, met daarbij een overzicht van de scores voor de tekstkenmerken aan de hand waarvan dat niveau is toegekend.

2. Vooraf: de link tussen taalniveaus en de begrijpelijkheid van teksten

Voordat we de antwoorden op onze vragen en de analyses door de instrumenten bespreken, is het goed nader in te gaan op wat deze tools nu precies voorspellen. Met 'taalniveau' wordt regelmatig verwezen naar het Common European Framework of Reference (CEFR) met verschillende niveaus van taalvaardigheid. Dit referentiekader is vastgesteld door de Raad van Europa. Deze schaal is ontwikkeld om de taalvaardigheid van mensen die een vreemde Europese taal leren te kunnen duiden. De CEFR-schaal kent 6 niveaus: A1, A2, B1, B2, C1 en C2. Zo moeten in Nederland inburgeringsplichtigen sinds 2007 een taalniveau van A2 hebben. Op A2-niveau is men als spreker van het Nederlands in staat:

- simpele zinnen en veelgebruikte uitdrukkingen te begrijpen
- korte, simpele gesprekken te voeren over alledaagse onderwerpen zoals familie, winkelen en persoonlijke achtergrond.

Voor het begrip van geschreven taal werden de niveaus oorspronkelijk als volgt gedefinieerd:

Tabel 1: Officiële definities van de taalniveaus volgens het CEFR. Bron: Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR). http://www.coe.int/t/dg4/linguistic/CADRE_EN.asp; geraadpleegd 23-2-2011

C2	Can understand and interpret critically virtually all forms of the written language including abstract, structurally complex, or highly colloquial literary and non-literary writings. Can understand a wide range of long and complex texts, appreciating subtle distinctions of style and implicit as well as explicit meaning
C1	Can understand in detail lengthy, complex texts, whether or not they relate to his/her own area of speciality, provided he/she can reread difficult sections.
B2	Can read with a large degree of independence, adapting style and speed of reading to different texts and purposes, and using appropriate reference sources selectively. Has a broad active reading vocabulary, but may experience some difficulty with low frequency idioms.
B1	Can read straightforward factual texts on subjects related to his/her field and interest with a satisfactory level of comprehension.
A2	Can understand short, simple texts on familiar matters of a concrete type which consists of high frequency everyday or job-related language. / Can understand short, simple texts containing the highest frequency vocabulary, including a proportion of shared international vocabulary items.
A1	Can understand very short, simple texts a single phrase at a time, picking up familiar names, words and basic phrases and rereading as required.

In het wetenschappelijk onderzoek naar de leesvaardigheid van moedertaalsprekers en naar de begrijpelijkheid van teksten spelen de taalniveaus tot dusver geen rol. Het is dan ook niet zo eenvoudig om uit een taalvaardigheidsniveau af te leiden welke kenmerken een tekst moet hebben, wil een lezer met dat niveau de tekst begrijpen. We zouden dan eigenlijk moeten weten hoe het taalniveau van een lezer samenhangt met de componenten die ten grondslag liggen aan leesvaardigheid. Dat zijn bijvoorbeeld decodeervaardigheid (technisch lezen), woordenschat, syntactische vaardigheden en redeneervaardigheid (Macaruso & Shankweiler 2010). Ons is geen onderzoek bekend dat het taalniveau verbindt aan dit soort deelvaardigheden. Daarom is het nog niet goed mogelijk om wetenschappelijk te bepalen welke

tekstkenmerken passen bij lezers met een bepaald niveau van leesvaardigheid. Naast deze indirecte route van taalniveau naar tekstkenmerken (via deelvaardigheden van de lezer) is ook onderzoek denkbaar naar een directe link tussen taalniveaus van lezers en tekstkenmerken die zij 'aan kunnen' wat betreft begrip. Wij kennen zulk onderzoek niet. Ook de makers van de softwareprogramma's hebben ons niet gewezen op het bestaan van zulk onderzoek.

Hoe zien de programma's eruit en waar zijn de niveautoekenningen dan wel op gebaseerd? Wij bespreken hieronder de drie programma's, waarbij we ons baseren op de gegevens die we ontvingen van de makers. Telkens bespreken we de volgende vragen:

- Wat zeggen de makers zelf over het programma?
- Met welke tekstkenmerken werkt het programma?
- Hoe wordt op basis van die kenmerken een leesbaarheidsniveau gedefinieerd? En welk onderzoek ligt aan die definitie ten grondslag?

3. Texamen

- Eigenaar/ontwikkelaar: BureauTaal
- URL: http://www.texamen.nl/

Het gebruiksrecht voor Texamen kost € 1000 per jaar. Daarnaast wordt er € 1 per geanalyseerde tekst betaald (beide bedragen zijn exclusief btw).

Wat zegt de ontwikkelaar zelf over het programma?

´Texamen is een instrument waarmee u het taalniveau van teksten kunt:

- diagnosticeren (Wat is het taalniveau van de tekst?);
- analyseren (Welke elementen in de tekst bepalen het taalniveau?);
- aanpassen (Wat moet ik doen om mijn tekst op het gewenste taalniveau te krijgen?).

Texamen is dus een instrument waarmee u het taalniveau van teksten op een objectieve en efficiënte manier kunt meten.´

Met welke tekstkenmerken werkt het programma?

Drie van de tekstkenmerken die Texamen gebruikt, moeten door de gebruiker zelf worden ingevoerd:

- Staat het belangrijkste vooraan?
- Hoeveel figuurlijke uitdrukkingen komen in de tekst voor?
- Legt de schrijver jargonwoorden uit?

Daarnaast berekent Texamen de waardes voor de volgende kenmerken zelf:

- Lengte van de tekst
- Aantal kopjes
- Gemiddeld aantal zinnen per alinea
- Gemiddeld aantal woorden per zin
- Gemiddeld aantal letters per woord.
- Aantal formele uitdrukkingen

- Aantal formele woorden
- Aantal hoogfrequente woorden
- Aantal laagfrequente woorden
- Aantal passiefconstructies
- Aantal tangconstructies (in een tangconstructie is de afstand tussen delen die bij elkaar horen - zoals onderwerp en persoonsvorm - erg groot)
- Aantal voorzetselgroepen (*met de trein, op de tafel*, ...)
- Aantal nominalisaties (bijv. *werking, ontdekking*, ...)
- Aantal vooropplaatsingen (bijv. *Die auto, ik wil er nooit meer in rijden*, of *In de krant, op tv, op de radio, je komt die man echt overal tegen*.")

We zien dat er naast klassieke kenmerken ook maten zijn gebruikt die iets kunnen zeggen over de grammaticale complexiteit van de tekst (passiefconstructies, tangconstructies, vooropplaatsingen en aantal voorzetselgroepen). Nominalisaties en formele woorden vallen onder de noemer "woordmoeilijkheid".

BureauTaal vermeldt wel dat Texamen deze kenmerken vaak herkent in de tekst aan de hand van heuristieken. Dit houdt in dat er in sommige gevallen geen harde regels zijn gebruikt die met 100% zekerheid scoren, maar dat er ervaringsregels gebruikt zijn die niet altijd (maar hopelijk vaak) tot het goede resultaat leiden. Dat is op zich niet zo verwonderlijk: taaltechnologie is niet perfect en voor bijvoorbeeld het herkennen van formele uitdrukkingen bestaan geen algoritmes.

We leverden elke maker van een programma onder andere vijf teksten uit roddelbladen. Als we het kenmerk "Aantal formele woorden" nader bekijken, dan valt bijvoorbeeld op dat Texamen in die teksten 7 formele woorden terugvindt, terwijl wij er zelf geen hebben kunnen vinden. Helaas gaf de output van Texamen ons alleen aantallen aan; het wees de woorden zelf niet in de tekst aan. We kunnen dus niet nagaan of wij formele woorden gemist hebben, of dat Texamen zogenaamde "false alarms" voor formele woorden produceerde.

Hoe wordt op basis van die kenmerken een leesbaarheidsniveau gedefinieerd? En welk onderzoek ligt aan die definitie ten grondslag?

BureauTaal heeft geen onderzoek gepubliceerd over de totstandkoming van Texamen. Het laat weten dat het programma niveaus heeft leren toekennen aan de hand van input door een onbekend aantal MBO-docenten. Deze MBO-docenten hebben aan 200 teksten een taalniveau toegekend, in termen van de zes Europese taalniveaus. Tekstkenmerken werden gekozen "op basis van intuïtie, onderbuikgevoel en ervaring als tekstschrijvers: zeg maar best practice". Verder moesten de kenmerken uit te drukken zijn in een getal en makkelijk te herkennen zijn aan de hand van heuristieken.

De verschillende tekstkenmerken zijn vervolgens door middel van een neuraal netwerk (een classificatiemethode uit de Kunstmatige Intelligentie) aan deze taalniveaus gekoppeld. Volgens BureauTaal is bij Texamen "de leesbaarheid van een tekst een afgeleide van het taalniveau dat het programma toekent aan een tekst". BureauTaal adviseert gebruikers van Texamen die een groot publiek willen bereiken "om op taalniveau B1 te schrijven".

Het is lastig om op deze basis het programma te beoordelen. Het is onbekend in hoeverre de experts het met elkaar eens waren wat betreft de toekenning van de taalniveaus. Verder weten we niet in hoeverre de experts de begripsprestaties van echte lezers goed voorspellen. We kunnen hierdoor helaas niets zeggen over de prestaties van het programma. Een andere onduidelijkheid betreft de link tussen de CEFR-taalniveaus en de daarbij passende tekstkenmerken; daarover schreven we hierboven al.

Biedt Texamen nieuwe oplossingen voor de problemen met klassieke leesbaarheidsformules? Het biedt een zekere vooruitgang wat betreft het introduceren van meer relevante tekstkenmerken. Wat diagnostiek betreft geeft de interface een grove indicatie van het aantal laagfrequente en formele woorden, de hoeveelheid abstract taalgebruik, het aantal "ingewikkelde zinnen" (waarschijnlijk zinnen met lange tangconstructies) en de hoeveelheid figuurlijk taalgebruik in de tekst. Het lijkt zinniger een tekst te reviseren op deze kenmerken dan op woord- en zinslengte (kenmerken waartoe veel oude leesbaarheidsformules zich beperken). Maar er zijn natuurlijk nog meer factoren die een tekst moeilijker of makkelijker te begrijpen maken. En we weten ook bij de meeste kenmerken niet hoe nauwkeurig de heuristiek is.

Voor de overige kritiekpunten die we hierboven noemden, biedt Texamen geen oplossingen. De interactie tussen de lezer en de tekst blijft afgezien van het taalniveau nog buiten beeld. Zouden alle lezers op een bepaald taalniveau evenveel moeite hebben met dezelfde teksten? Texamen lijkt dit wel te suggereren, omdat alle lezers die op taalniveau B1 of hoger zitten een tekst met het Texamen label B1 zonder problemen zouden moeten kunnen begrijpen. En er wordt dus nog steeds met gemiddelde data voor groepen lezers gewerkt. Omdat daarmee veel variantie wordt weggecijferd, is de voorspellende waarde voor individuele lezers uit die groep B1 veel geringer. Verder krijgen teksten nog steeds één niveau (een gemiddeld niveau dus) toegekend, en komen we weinig te weten over de locatie van ingewikkelde passages in de tekst zelf of over de mate waarin de moeilijkheid binnen de tekst zelf varieert. De ene passage kan immers moeilijker zijn dan de andere.

We weten ten slotte weinig over de 200 teksten die gebruikt zijn bij de totstandkoming van Texamen. Om welk genre gaat het? Er wordt geen voorbehoud gemaakt wat betreft de verschillende typen teksten waarvoor Texamen voorspellingen doet. Idealiter vormen de 200 teksten een goede afspiegeling van de totale diversiteit aan tekstgenres; het is bijvoorbeeld maar de vraag in hoeverre een programma dat getraind is op krantenartikelen in staat is voorspellingen te maken voor juridische teksten.

Op één punt is Texamen een verslechtering ten opzichte van de klassieke leesbaarheidsformules. De oude formules werden ontwikkeld en getoetst aan de hand van leesonderzoek met echte lezers. Bij Texamen waren het experts en aanbieders van leesmateriaal (de opdrachtgever) die bepaalden of een lezer een tekst al dan niet zou moeten begrijpen. We kunnen daarom niet meer spreken van 'begripsvoorspelling'. De relatie tussen tool en het begripsniveau van een lezer is indirecter. Het programma levert een voorspelling van een begripsvoorspelling: de begripsvoorspelling die besloten ligt in het oordeel van de experts die met de 200 teksten gewerkt hebben. Over de waarde van zo'n diagnose voor de schrijfhulp die Bureau Taal biedt, kunnen wij geen uitspraak doen.

- Eigenaar/ontwikkelaar: Gridline
- URL: http://www.klinkendetaal.nl/

De kosten van een licentie op Klinkende Taal zijn afhankelijk van het aantal werknemers van de afnemende instelling: vanaf € 3000 per jaar. De meting met het instrument wordt niet afzonderlijk geleverd, hoewel een eenmalige scan op een hoeveelheid teksten wel tot de mogelijkheden behoort. Gridline ziet het instrument uitdrukkelijk als een hulpmiddel bij het revisieproces. Het instrument is bedoeld als aanvulling op het aanbod van bureaus die schrijftrainingen en coaching verzorgen. Gridline biedt zelf dergelijke trainingen niet aan, afgezien van een introductie in het instrument. Naast de licentiekosten zijn er eenmalige opstartkosten voor installatie en optionele kosten voor hosting en applicatiebeheer. Per geanalyseerde tekst wordt geen extra bedrag gevraagd.

Wat zegt de ontwikkelaar zelf over het programma?

´Klinkende Taal helpt u duidelijke teksten te schrijven. Onze producten controleren uw brieven, brochures en webteksten snel en effectief op leesbaarheid.
Onze Word Plugin meet het taalniveau (A1 t/m C2). Bovendien worden alle moeilijke zinnen, woorden en passages duidelijk voor u aangestreept. Zo kunt u makkelijk en snel het gewenste taalniveau bereiken.
Lange zinnen, moeilijke woorden, passieve constructies? U vindt ze makkelijk, verbetert wat u wilt verbeteren, en komt zo tot een beter leesbare tekst. Zodat er meer tijd overblijft voor de inhoud.´

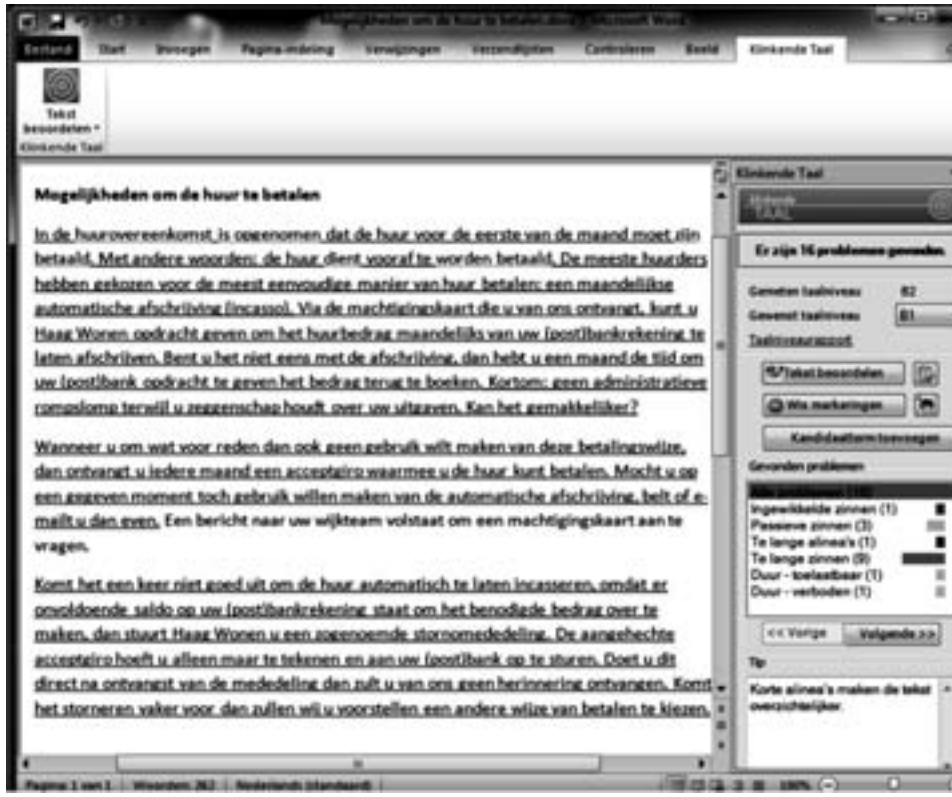
Met welke tekstkenmerken werkt het programma?

Klinkende Taal test het taalniveau op de volgende kenmerken en markeert passages op die punten:
- Dure woorden
- Vaktermen
- Lange zinnen
- Ingewikkelde zinnen
- Passieve zinnen
- Lange alinea's
- Gemiddeld aantal bijzinnen per zin
- Gemiddeld aantal woorden per zin en per alinea

Klinkende Taal brengt bovendien markeringen aan in de tekst voor enkele kenmerken die niet meetellen bij de beoordeling van het taalniveau:
- Opsommingen
- Uitdrukkingen
- Hulpwerkwoorden
- Aanspreekvormen (jij, men, etc.)
- Lange bijzinnen aan het begin van de zin
- Ontbrekende tussenkopjes

- Lexicale samenhang
- Lange woorden
- Dubbele ontkenningen
- Naamwoordstijl
- Tangconstructies
- Abstracte woordkeuze

We weten niet hoe nauwkeurig Klinkende Taal is in het extraheren van deze tekstkenmerken. We ontvingen van slechts 1 van de 20 ingezonden teksten een volledige output met enkele markeringen, en van een andere tekst een onvolledige output. Ter illustratie laten we de eerste tekst hieronder volgen (Figuur 1).



Figuur 1: Screenshot van de analyse met Klinkende Taal.

In deze tekst zijn *huurovereenkomst* en *dient* dure woorden. Maar bijvoorbeeld *incasso, machtigingskaart, (post)bankrekening, acceptgiro, stornomededeling* en *storneren* niet. Mogelijk gaat het hier om vaktermen die de gebruiker zelf als zodanig dient op te geven bij de ontwikkelaar. Hiervoor stelt Klinkende Taal een speciale beheertool beschikbaar.

Wanneer of waarom een zin te lang is, is niet duidelijk. De makers van het instrument zeggen in een toelichting dat de beoordeling afhangt van het soort woorden in de zin: hoe

meer korte woorden (6 letters of minder), hoe langer de toegelaten zinslengte. Dit beloont het omzetten van moeilijke naar makkelijke woorden.

Over de andere 18 teksten werd geen extra informatie verstrekt; we ontvingen alleen het toegekende taalniveau. Het aantal teksten met kenmerkanalyse dat ons toegestuurd werd, is te laag om conclusies aan te verbinden. De twee geleverde voorbeelden illustreren wel dat de tekstkenmerken door Klinkende Taal niet helemaal foutloos geëxtraheerd worden. Het is overigens ook niet reëel dat van taaltechnologie te verwachten. Een kenmerk als "dure woorden" zal gebaseerd zijn op een lijst die door de ontwikkelaars mogelijk samen met gebruikers moet worden opgesteld. Daar zullen altijd woorden op blijven ontbreken.

Hoe wordt op basis van die kenmerken een leesbaarheidsniveau gedefinieerd? En welk onderzoek ligt aan die definitie ten grondslag?

Gridline heeft geen onderzoek gepubliceerd over de totstandkoming van Klinkende Taal, en evenmin over de nauwkeurigheid waarmee het programma taalniveaus voorspelt. We kunnen hierdoor helaas niets zeggen over de prestaties van de tool.

Gridline stelt dat zij weliswaar geen wetenschappelijk onderzoek doen, maar wel ervaring opgedaan hebben met het gebruik van het programma door partners en klanten. Deze gebruikers komen met opmerkingen aan de hand waarvan Gridline vervolgens het programma kan aanpassen. Voor Gridline is dit erg nuttig en waardevol.

Over de toekenning van taalniveaus aan teksten meldt Gridline dat die "in de praktijk heel goed blijkt te werken". Bij de ontwikkeling van het programma is de hulp ingeroepen van communicatie-experts, die ervaring hebben als trainer. Teksten van cursisten werden gebruikt als trainingsmateriaal. Klinkende Taal geeft volgens Gridline in 95% van de gevallen hetzelfde oordeel als de communicatie-expert. De cursisten zouden meer waarde hechten aan het oordeel van Klinkende Taal, dan aan het oordeel van de expert. Gridline erkent dat deze praktijkervaringen geen wetenschappelijke data vormen, maar vindt dat zij een acceptabel startpunt vormen voor het product, aangezien er op dit moment geen alternatief is.

Klinkende Taal is net als Texamen niet gebaseerd op onderzoek bij lezers. In feite is Klinkende Taal een programma dat het oordeel van experts over leesbaarheid probeert te voorspellen. Hierbij richt het programma zich op de vormkenmerken van de tekst. Om die reden kunnen we dan ook niet echt spreken van 'begripsvoorspelling'. Doordat Klinkende Taal is gebaseerd op data die door een expert geleverd zijn, is de relatie tussen voorspelling en het begripsniveau van een lezer indirect. Het programma levert een voorspelling van een begripsvoorspelling: de begripsvoorspelling die besloten ligt in het oordeel van de expert. Over de waarde van dergelijke diagnoses voor de schrijfpraktijk kunnen we in deze analyse geen uitspraak doen. Een verschil met Texamen is de presentatie van de output. Bij Klinkende Taal vindt de gebruiker markeringen in de tekst die wijzen op revisiemogelijkheden. Bij Texamen ontvangt de gebruiker scores op verschillende variabelen, die echter niet direct naar concrete passage verwijzen.

5. Accessibility Leesniveau Tool

- Eigenaar/ontwikkelaar: Stichting Accessibility
- URL: http://www.accessibility.nl/internet/tools/leesniveau_tool

Als enige van de drie besproken instrumenten is de Accessibility Leesniveau Tool voor iedereen gratis online te gebruiken.

Wat zegt de ontwikkelaar zelf over het programma?

´De Accessibility Leesniveau Tool is een programma dat, op basis van een ingevoerde tekst, een indicatie geeft van het niveau van de technische leesbaarheid van de tekst. Buiten de technische leesbaarheid zijn ook de inhoud van de tekst, de structuur van de tekst en het design van de tekst van belang voor de totale leesbaarheid. Deze punten zijn echter niet met dit programma vast te stellen. Het programma geeft dan ook geen enkele garantie over het precieze leesniveau, maar moet puur indicatief gebruikt worden!´

Met welke tekstkenmerken werkt het programma?

Volgens de Stichting Accessibility baseert de Leesniveau Tool zijn oordeel op o.a. de volgende kenmerken:

- de percentages tekstwoorden die voorkomen op 5 lijsten van frequente woorden;
- het aantal woorden per zin;
- het aantal lettergrepen per woord.

In totaal werkt het programma op dit moment met vijf criteria waarbij de woordenlijsten de doorslag geven. De gebruiker geeft verder zelf aan hoeveel namen er in de tekst staan.

Alle kenmerken zijn kenmerken op woordniveau; de zinsopbouw wordt niet gemeten. Deze kenmerken zijn simpel, zonder verdere taaltechnologische middelen, uit de tekst te extraheren. Alle kenmerken werden al gebruikt in de klassieke leesbaarheidsformules uit de vorige eeuw. De kritiek op het gebruik van zulke formules is dus ook op dit instrument van toepassing. Kenmerken als woord- en zinslengte hebben bovendien geen diagnostische waarde: een tekst reviseren door de zinnen in tweeën te knippen en voor kortere woorden te kiezen leidt niet tot een beter leesbare tekst.

Anders dan Texamen en Klinkende Taal, toont de Accessibility Leesniveau Tool de gebruiker geen informatie over de waardes van tekstkenmerken. Stichting Accessibility is bescheiden in wat het programma kan: het gaat om een *indicatie* van het niveau van *technische* leesbaarheid. Over de leesbaarheid van teksten in ruimere zin doet het programma geen uitspraak, omdat o.a. inhoud en structuur niet meegenomen worden. Stichting Accessibility benadrukt in een reactie dat het programma geen garantie geeft over het leesniveau, maar puur indicatief gebruikt dient te worden.

Hoe wordt op basis van die kenmerken een leesbaarheidsniveau gedefinieerd? En welk onderzoek ligt aan die definitie ten grondslag?

Stichting Accessibility meldt dat het programma in overleg met Stichting Lezen en Schrijven is getest door Eenvoudig Communiceren[2]. Zij hebben tientallen teksten per niveau aange-

leverd. De gewichten van de kenmerken zijn vervolgens aangepast om de voorspelling van het programma overeen te laten komen met de niveaus zoals aangegeven door Eenvoudig Communiceren. Verdere details ontbreken. Het is mogelijk dat er bij het ontwikkelen van het programma sprake is van wat ook wel overfitting wordt genoemd: het programma lijkt zo sterk getraind op één specifieke dataset dat ook de ruis in deze dataset door het programma gemodelleerd wordt.

Voor zover bekend is er geen onderzoek verricht bij de ontwikkeling van het programma. Stichting Accessibility meldt dat de lezer geen rol had in het ontwikkelingsproces. Er is ons geen validatie-onderzoek bekend.

De Accessibility Leesniveau Tool lijkt van de drie besproken instrumenten het meest op de klassieke leesbaarheidsformules, omdat het werkt met dezelfde soort tekstkenmerken. Net als Klinkende Taal en Texamen zijn er geen lezers direct betrokken bij het ontwikkelingsproces noch bij de validatie ervan. Wat de overige kritiekpunten op leesbaarheidsformules betreft lijkt de Accessibility Leesniveau Tool evenmin dichterbij een oplossing te zijn gekomen. De ontwikkelaars hebben met hun tool echter weinig pretenties: het resultaat ervan moet gezien worden als een grove schatting van het technisch leesniveau van de tekst.

6. Een vergelijking van de drie leesniveauvoorspellingen

Het is onmogelijk uitspraken te doen over de betrouwbaarheid van de tools wanneer een validatie-onderzoek niet uitgevoerd of niet gepubliceerd is. Een dergelijk validatie-experiment is erg tijdrovend en vraagt bovendien onbeperkte toegang tot de instrumenten. Die toegang is ons niet geboden. Hieronder rapporteren wij over een iets bescheidener analyse.

Wij hebben de drie instrumenten getest door ze in te zetten op een beperkt aantal teksten. Daartoe zijn vier soorten teksten van elk rond de 270 woorden gebruikt:

- vijf teksten afkomstig van een woningbouwvereniging die zijn gericht aan huurders;
- vijf krantenberichten over uiteenlopende binnenlandse onderwerpen;
- vijf fragmenten uit polisvoorwaarden van verzekeringen;
- vijf teksten uit de roddelrubriek van een landelijk dagblad.

De bedoeling was een zekere spreiding in zowel onderwerp als complexiteit in de verzameling te brengen. De leesniveaus van deze teksten zoals bepaald door de drie instrumenten vindt u terug in Tabel 2. Van één tekst ontbrak bij Texamen het resultaat.

Wat laat deze vergelijking zien? Ten eerste komen de taalniveaus A1 en A2 niet voor. Dat is niet zo gek, wanneer we bedenken dat de Europese taalniveaus ontworpen zijn voor mensen die een vreemde taal leren. Op de laagste niveaus is het taalgebruik van deze mensen nog enorm beperkt (zie Tabel 1). Aangenomen dat men kan spreken van een tekst die op een zeker taalniveau geschreven is, dan zullen A1 en A2 nauwelijks worden toegekend aan teksten die mensen dagelijks voorgeschoteld krijgen (zoals kranten, tijdschriften en brieven). Eigenlijk blijven er van de 6 taalniveaus nog 4 over die in aanmerking komen. De Accessibility tool kent ook nog 3 tussenniveaus en kent dus 9 taalniveaus (waarvan er 7 operationeel waren in deze analyse). Dit programma maakt dus meer onderscheid tussen teksten.

Tabel 2: Taalniveauvoorspellingen door Texamen, Klinkende Taal en de Accessibility Leesniveau Tool

Tekst	Texamen	Klinkende Taal	Accessibility Leesniveau Tool
huur1	C1	C1	B2
huur2	C1	C2	C2
huur3	C1	C1	B2
huur4	B2	B2	B1/B2
huur5	C1	C1	B2
krant1	C1	C1	B2
krant2	C1	B2	B2
krant3	C1	C1	B2/C1
krant4	C1	B2	C1
krant5	C1	C1	C1
polis1	C1	C2	C2
polis2	C1	C2	C2
polis3	C1	C2	C2
polis4		B2	B1/B2
Polis5	C1	B2	B2
roddel1	B2	B2	B2
roddel2	B1	B2	B1/B2
roddel3	C1	C1	B2/C1
roddel4	B2	B2	B2
roddel5	B1	B2	B2

Het valt verder op dat bij Texamen 14 van de 19 teksten taalniveau C1 kregen toegewezen. De af en toe vrij pittig ogende polisvoorwaarden worden zo op hetzelfde niveau geplaatst als een roddelrubriektekst over Lady Gaga en de krantenberichten van de binnenlandpagina.

Omdat er slechts vier bruikbare niveaus zijn, gaat het om een behoorlijk grove classificatie. Toch verschillen de instrumenten nog al eens in hun schattingen. Bij de vergelijking hieronder hebben we de tussenniveaus van de Accessibility Tool moeten afronden naar een van de aangrenzende niveaus.

- Wanneer we de tussenniveaus van de Accessibility tool in 2 gevallen gunstig afronden en in 1 geval ongunstig afronden, zijn de drie tools het in 5 van de 19 gevallen met elkaar eens.
- Texamen en Klinkende Taal zijn het in 10 van de 19 gevallen eens.
- Texamen en de Accessibility Tool zijn het in 5 van de 19 gevallen eens, waarbij we de Accessibility score 1 keer gunstig en 1 keer ongunstig afronden.
- Klinkende Taal en de Accessibility Tool zijn het in 13 van de 20 gevallen eens, waarbij we de Accessibility score 3 keer gunstig en 2 keer ongunstig afronden.

We hebben twee eenvoudige vervolganalyses gedaan op de gegevens. Ten eerste hebben we met een rangordetoets gekeken of er verschillen zijn tussen de instrumenten wat betreft de hoogte van de toegekende niveaus. Voor de analyse daarvan hebben we de taalniveaus omgezet naar numerieke waarden (B1 = 3, B2 = 4, C1 = 5, C2 = 6 en de tussenniveaus tussen deze getallen in). Het gaat natuurlijk niet om intervalgegevens maar om ordinale data en daarom is een Wilcoxon Signed Ranks Test uitgevoerd. In Tabel 3 staan desondanks voor het gemak de gemiddelde rangordes.

Tabel 3: Gemiddelde taalniveaus en standaarddeviatie

Tool	**Gemiddelde**	**Standaarddeviatie**
Texamen	4.63	0.684
Klinkende Taal	4.79	0.787
Accessibility Leesniveau Tool	4.48	0.881

De Wilcoxon test laat een significant verschil zien tussen de hoogte van de taalniveaus toegekend door Accessibility en Klinkende Taal: Accessibility kent lagere scores toe ($Z = 2.04$; $p < .05$; zie Tabel 3 voor de gemiddelden).

Ten tweede zijn rangordecorrelaties (Spearmans rho) berekend tussen de tools. Het is immers mogelijk dat, ook al verschillen de toegekende niveaus per tool, elke tool uiteindelijk dezelfde rangorde toekent aan onze verzameling teksten. De correlaties zijn te vinden in Tabel 4. Hoewel ze allemaal significant zijn op 1%-niveau ($p < .01$), blijft er heel wat variantie tussen de instrumenten onverklaard, met name tussen Texamen en de twee andere tools. Ter vergelijking: een correlatie van .64 op intervaldata zou 41% van de variantie verklaren.

Tabel 4: Rangordecorrelaties tussen de taalniveaus toegekend door de drie instrumenten

	Texamen	**Klinkende Taal**	**Accessibility**
Texamen	-	.64	.60
Klinkende Taal		-	.75
Accessibility			-

7. *Het verband tussen de voorspellingen en vier tekstkenmerken*

Ten slotte zijn we nagegaan of de toegekende taalniveaus correleren met een aantal tekstkenmerken waarvan op basis van de literatuur bekend is dat ze begripsprestaties helpen voorspellen, zoals beschreven in Kraf & Pander Maat (2009). Hoewel deze tekstkenmerken niet allemaal onderdeel zijn van de onderzochte systemen, verwachten we op grond van eerder onderzoek wel significante correlaties te vinden met de toegewezen taalniveaus en hebben we duidelijke verwachtingen over de richtingen van deze correlaties (positief/negatief).

7.1 Proportie frequente woorden Deze maat drukt uit hoeveel woorden uit de tekst voorkomen in een lijst meest frequente woorden. Wij gebruikten een frequentielijst die we zelf bouwden aan de hand van een verzameling teksten uit het D-COI corpus. Dit corpus bevat ongeveer 25 miljoen woorden uit o.a. krantenteksten, tijdschriften, Wikipedia en teletekstpagina's. We vormden daaruit gerangordende frequentielijsten. Uit die lijsten hebben we op twee manieren de meest frequente woorden gekozen. We hebben een lijst gemaakt waarop de meest frequente woorden voorkomen die samen 50% van het totaal aantal woorden (woordtokens) in het corpus vormen. Daarbij bleek het te gaan om de 98 meest frequente woordtypen. Dit aantal woordtypen is zo laag, dat we verwachten er weinig onderscheid mee te kunnen maken tussen teksten; we zien er daarom vanaf om deze lijst te gebruiken. De tweede lijst bevat de meest frequente woorden die samen 77% van de woordtokens in het corpus voor hun rekening nemen, waarbij het ging om de 1520 meest frequente woordtypen. Voor deze lijst berekenden we per tekst hoeveel procent van de tekstwoorden gedekt wordt door de lijst; we noemen die variabele de lexicale dekking. We verwachten dat hoe hoger deze dekking is, hoe eenvoudiger de tekst en dus hoe lager het taalniveau zal zijn. In het onderzoek van Staphorsius (1994) was deze lexicale dekking de belangrijkste voorspeller van de begrijpelijkheid van teksten op basisschoolniveau.

7.2 TTR: de type-token ratio, eveneens aanwezig in de CLIB formule Deze wordt berekend door het aantal unieke woorden (types) in de tekst te delen op het totale aantal woorden (tokens). Hoe lager deze waarde, hoe meer verschillende woorden voorkomen in de tekst en des te minder woorden herhaald worden. We verwachten daardoor wederom een negatieve correlatie. TTR zegt dus iets over woordgebruik, maar kan ook gezien worden als een grove maat voor informatiedichtheid.

7.3 De gemiddelde afstand tussen onderwerp en persoonsvorm en tussen lijdend voorwerp en persoonsvorm In het verleden is door Gibson (1998) aangetoond dat deze afstand (ook wel afhankelijkheidslengte genoemd) een grotere invloed heeft op de complexiteit van de zin dan de lengte van de gehele zin. Deze waarde werd automatisch berekend met behulp van de automatische zinsontleder Alpino. Hoewel erg nauwkeurig, maakt deze ontleder altijd nog twee maal zoveel fouten als een menselijke expert: de berekende waarde voor dit kenmerk zou daarom kunnen afwijken van de werkelijke waarde. We verwachten een positieve correlatie tussen de taalniveaus en de waarde van dit kenmerk: langere afstanden maken een tekst complexer en horen dus bij een hoger taalniveau.

Ook voor de gemiddelde afstand tussen lijdend voorwerp en de persoonsvorm in de tekst verwachten we een positieve correlatie tussen kenmerkwaarde en leesniveau.

7.4 Resultaten Tabel 5 geeft informatie over de scores op deze vijf maten van de twintig gebruikte teksten. Hier is dus nog geen vergelijking met de instrumenten aan de orde. De tabel laat zien hoe groot de spreiding is op deze scores in het corpus van die twintig teksten. We zien dat er meer spreiding is tussen de teksten wat betreft de afstand tussen object en persoonsvorm dan wat betreft de afstand tussen subject en persoonsvorm. Dat kan kloppen, omdat in langere zinnen vooral de eerste afstand toeneemt, en niet zozeer de tweede.

Tabel 5: Descriptieve gegevens over vijf tekstkenmerken

	N	Minimum	Maximum	Gemiddelde	Std. Deviatie
Lexicale dekking	20	.56	.68	.62	.038
TTR	20	.44	.64	.55	.059
Afstand subject – pv	20	1.48	4.58	2.86	.95
Afstand object – pv	20	.40	7.89	3.76	2.06

De laatste stap is een vergelijking tussen deze scores en de uitkomsten van de drie instrumenten. We gaan na hoe hoog de correlatie is tussen de scores van de tools en elk van de bovengenoemde maten.

Tabel 6: Rangordecorrelaties tussen de taalniveaus per tool en de vijf tekstkenmerken

	Lexicale dekking	TTR	Subject-pv	Object-pv
Texamen	-.051	-.549*	.379	.466*
Klinkende Taal	-.153	-.406	.426	.542*
Accessibility	-.436	-.241	.460*	.644**

** = $p < .05$; ** = $p < .01$*

Wat betreft woordmoeilijkheid valt op dat Texamen en Klinkende Taal weinig verband laten zien met de proportie frequente woorden in de tekst. Dat is opmerkelijk, omdat in onderzoek woordfrequentie als de meest robuuste predictor van begripsprestaties aangewezen wordt (zie de referenties in Kraf en Pander Maat, 2009). Bovendien hebben beide instrumenten veel aandacht voor het woordniveau. Beslissend is natuurlijk hoe een tool dure woorden of vaktermen definieert, en die definities kennen wij niet. Accessibility is wel enigszins gevoelig voor frequentie. Dat is ook te verwachten, aangezien dit programma hoofdzakelijk met frequentiematen werkt.

Wat betreft informatiedichtheid correleren de scores van Texamen redelijk met de type-token-ratio van de geanalyseerde teksten. Dat is opmerkelijk, omdat de makers daar zelf weinig over zeggen. Klinkende Taal laat een minder duidelijke samenhang zien met deze score.

Wat betreft zinscomplexiteit correleren alle drie de leesniveauvoorspellingen redelijk met de beide afhankelijkheidslengtes, vooral die tussen persoonsvorm en lijdend voorwerp. Waarschijnlijk komt dit doordat alle tools de zinslengte als tekstkenmerk gebruiken.

8. Conclusie

Voor het Nederlands zijn op dit moment drie leesbaarheidsinstrumenten op de markt. Deze plaatsen een tekst op een schaal met de zes taalniveaus uit het Common European Framework. Deze niveau-indeling is ontwikkeld als indicatie van de taalvaardigheid van mensen die een vreemde taal leren. In de verzameling aangeboden teksten kwamen slechts vier van

die niveaus te voorschijn. Blijkbaar zijn de eenvoudige tekstjes uit de roddelrubriek van een landelijk dagblad al te moeilijk voor die laagste niveaus. Maar er is geen onderzoek bekend waar uit blijkt dat er echt een verband is tussen dit Europees raamwerk voor taalvaardigheid en kenmerken van teksten. De makers van de instrumenten hebben ons ook niet gewezen op eigen onderzoek waaruit wel een relatie blijkt tussen begripsprestaties van volwassen Nederlanders enerzijds en de scores van hun tools anderzijds. Er ontbreekt dus een empirische basis voor de claim dat zo'n instrument de leesbaarheid of begrijpelijkheid van een tekst zou kunnen voorspellen.

Voor zover de instrumenten zijn geijkt, is dat gebeurd door de voorspellingen te vergelijken met die van experts. Het blijft daardoor onduidelijk voor welke teksten en wat voor lezers deze tools de begrijpelijkheid voorspellen. We kunnen daarom niets zeggen over hun validiteit.

Dankzij de medewerking van de makers konden we voor twintig teksten in beperkte mate nagaan wat de scores dan wel zouden kunnen betekenen. Voor het schalen van deze twintig teksten werden in de praktijk slechts vier van de zes beschikbare niveaus gebruikt. Een klein experiment waarbij we de uitkomsten van de instrumenten op deze tekstjes met elkaar vergelijken, laat zien dat de instrumenten het regelmatig met elkaar oneens zijn over het taalniveau van teksten. We mogen daarom voorzichtig aannemen dat ook de experts van wie de oordelen in de tools gemodelleerd werden, van mening zouden verschillen over de begrijpelijkheid.

Een vergelijking van de scores met een analyse op een beperkt aantal algemeen aanvaarde factoren die de begrijpelijkheid van teksten beïnvloeden, leidt tot een volgende conclusie. Texamen en Klinkende Taal zijn niet erg gevoelig voor verschillen in woordfrequentie, wat opvalt aangezien dat in veel onderzoek de sterkste voorspeller van begrijpelijkheid is. De correlaties tussen de automatisch toegekende niveaus en de scores op zinsmoeilijkheid en informatiedichtheid zijn hoger.

Voor zowel Bureau Taal als voor Klinkende Taal geldt dat deze instrumenten ingezet worden in een context van training en schrijfhulp. In onze analyse blijft die toepassing buiten beschouwing. Het is goed denkbaar dat cursisten de feedback waarderen die met deze instrumenten geleverd wordt. Het is ook denkbaar dat zij beter gaan schrijven wanneer ze van die feedback goed gebruik maken. Maar wanneer diezelfde cursisten menen dat een tekst met de score B2 door de meeste Nederlanders begrepen zal worden, is dat nog niet zeker. Er is namelijk geen onderzoek dat de Europese taalniveaus definieert in termen van tekstbegrip. De vraag of zo'n definitie überhaupt mogelijk is, staat ook nog open. De grenzen tussen de niveaus worden op dit moment dus gebaseerd op de intuïties van enkele experts. Daadwerkelijk onderzoek naar de relatie tussen tekstkenmerken en tekstbegrip is daarom een absolute noodzaak.

Noten

1 Dit onderzoek is mogelijk gemaakt met steun van het NWO-programma Begrijpelijke Taal (project. Nr. 321-70-050). Dit programma wil bestaande expertise op dit terrein bundelen en vergroten.

2 Eenvoudig Communiceren is een uitgeverij die boeken en kranten uitgeeft speciaal voor slechte lezers.

Bibliografie

Collins-Thompson, K., & Callan, J. (2005). Predicting reading difficulty with statistical language models. *Journal of the American Society for Information Science and Technology, 56*(13), 1448-1462.

Crossley, S.A., Dufty, D.F., McCarthy, P.M., & McNamara, D.S. (2006). Toward a new readability: a mixed model approach. In D.S. McNamara & G. Trafton (Red.), *Proceedings of the 29th annual conference of the Cognitive Science Society*, Austin, Texas: Cognitive Science Society, 197-202.

Dale, E., & Chall, J.S. (1948). A formula for prediciting readability. *Educational Research Bulletin, 27*, 37-54.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*(3), 221-233.

Gibson, E., & Pearlmutter, N. (1998). Constraints on sentence comprehension. *Trends in Cognitive Sciences, 2*(7), 262-268.

Heilman, M.J., Collins-Thompson, K., Callan, J., & Eskenazi, M. (2007). Combining lexical and grammatical features to improve readability measures for first and second language texts. *Proceedings of NAACLHLT* 2007, 460-467.

Kraf, R., & Pander Maat, H. (2009). Leesbaarheidsonderzoek: oude problemen, nieuwe kansen. *Tijdschrift voor Taalbeheersing, 31*(2), 97-123.

Macaruso, P., & Shankweiler, D. (2010). Expanding the simple view of reading in accounting for reading skills in community college students. *Reading Psychology, 31*, 454-471.

Schwarm, S.E., & Ostendorff, M. (2005). Reading level assessment using support vector machines and statistical language models. *Proceedings of ACL* 523-530.

Staphorsius, G. (1994). Leesbaarheid en leesvaardigheid. De ontwikkeling van een domeingericht meetinstrument. Arnhem: Cito.

Van Oosten, P., Tanghe, D., & Hoste, V. (2010). Towards an Improved Methodology for Automated Readability Prediction. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, et al. (Red.), *Proceedings of the seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association, Valletta, Malta.

Vogel, M., & Washburne, C.W. (1928). An objective method of determining garde placement of childrens reading material. *Elementary School Journal*, *28*, 373-381.