Taal op recept: de vereenvoudiging van medische brochures door mensen en AI: Een onderzoek naar taalvariabelen waarmee menselijke redacteuren en ChatGPT pogen medische brochures te vereenvoudigen voor een hogere mate van begrijpelijkheid

Tuin, Lisanne

Taal op recept: de vereenvoudiging van medische brochures door mensen en AI

Een onderzoek naar taalvariabelen waarmee menselijke redacteuren en ChatGPT pogen medische brochures te vereenvoudigen voor een hogere mate van begrijpelijkheid

Masterscriptie

Neerlandistiek (MA): Taalbeheersing van het Nederlands

Universiteit Leiden

Student:	Lisanne Tuin
Studentnummer:	S3386856
Begeleider:	Dr. A. Reuneker
Tweede lezer:	Dr. H. Jansen
Inleverdatum:	16 juni 2025
Aantal ECTS:	20 EC
Aantal woorden:	22361 woorden

Voorwoord

Drie jaar geleden had ik mezelf voorgenomen niet meer te gaan studeren. Ik had een hbo-bachelor behaald en de universitaire bachelor Nederlands afgerond en ik was klaar met het onderwijs. Ik verlangde ernaar het werkveld te verkennen en te ontdekken welke functie daadwerkelijk bij mij zou passen. Dat ik niet meer uit het onderwijs weg zou zijn te slaan, had ik toen ook niet bedacht. Inmiddels werk ik al ruim tweeënhalf jaar met veel plezier als hbo-docent taalbeheersing. Om mijn kennis en kunde verder te verdiepen besloot ik deze master te volgen. Wat ik leer, kan ik vrijwel direct toepassen in mijn dagelijkse praktijk. Wat voor mij erg waardevol en relevant is. Dat geldt eveneens voor het onderzoeksobject van deze scriptie. Mijn interesse gaat uit naar generatieve AI en in het bijzonder hoe AI een rol kan vervullen binnen het schrijfonderwijs. In de praktijk gebruik ik zelf ook generatieve AI als docent: niet alleen als hulpmiddel bij het ontwikkelen van opdrachten, maar ook als sparringpartner bij het vormgeven van lesideeën. Hoe generatieve AI-technologie precies functioneert, is mij echter nog grotendeels onbekend. Juist daarom leek het mij een uitgelezen kans om dit thema te onderzoeken in het kader van mijn afstudeertraject.

Allereerst wil ik mijn scriptiebegeleider, dr. Alex Reuneker, bedanken. De oplossingsgerichte en pragmatische aanpak van de heer Reuneker heeft niet alleen geholpen bij de structuur en helderheid van mijn onderzoek, maar ook bij mijn academische groei. Wat ik tevens erg heb gewaardeerd, is zijn open houding ten opzichte van nieuwe ideeën. In plaats van te sturen op onderzoek binnen al bestaande theoretische kaders, moedigt hij studenten juist aan om buiten de gebaande paden te treden en nieuwe invalshoeken te verkennen. Die vrijheid heb ik als heel stimulerend ervaren. Het gaf me de ruimte om eigen keuzes te maken, theorieën te combineren en zelfstandig een academisch perspectief te ontwikkelen.

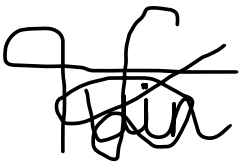
Daarnaast wil ik mijn studiemaatje en vriendin Jennifer Duin bedanken: mijn steun en toeverlaat. Samen hebben we talloze weekenduren studerend doorgebracht, waarin we elkaar bleven aanmoedigen en motiveren tijdens het hele scriptietraject. Met haar gedeelde achtergrond (een bachelor in Nederlandse taal en cultuur) kon zij waardevolle suggesties aandragen op momenten dat ik mijn ideeën wilde aanscherpen of bepaalde keuzes in mijn scriptie wilde onderbouwen. Haar aanwezigheid en frisse blik maakten het schrijfproces niet alleen productiever, maar ook een stuk gezelliger!

Tot slot wil ik graag mijn collega Enno Wiggers bedanken voor zijn bereidwilligheid om mee te denken en mij te begeleiden bij de data-analyse die een belangrijk onderdeel vormt van deze scriptie. Zijn hulp bij het uitvoeren van tientallen statistische toetsen en zijn heldere uitleg hebben mij in staat gesteld de analyses zelfstandig en met meer vertrouwen uit te voeren. Dankzij zijn deskundige toelichting op complexe statistische aspecten kon ik mijn onderzoeksresultaten op een verantwoorde en onderbouwde wijze interpreteren. Daarnaast ben ik andere lieve collega's van Hogeschool Windesheim dankbaar, die me met veel begrip en flexibiliteit hebben ondersteund. Meerdere keren zijn roosters en planningen aangepast om ruimte te maken voor mijn dubbelfunctie als docent en student, iets wat ik enorm waardeer.

Met gepaste trots leg ik deze scriptie voor, waarin ik met veel betrokkenheid en nieuwsgierigheid een relatief nieuw fenomeen heb onderzocht. Ik hoop dat deze nieuwsgierige blik ook doorklinkt in de pagina's die volgen.

Veel leesplezier,

Lisanne Tuin



Inhoudsopgave

Hoofdstuk 1 Inleiding

Begrijpelijke communicatie in de gezondheidszorg is essentieel voor effectieve zorgverlening, benadrukt Lentz (2011) in zijn oratie. Desondanks blijkt uit onderzoek van Kenniscentrum Nivel (2019) dat 7,7 procent van de Nederlandse volwassenen onvoldoende gezondheidsvaardig is; 21,1 procent beperkte gezondheidsvaardigheden heeft en 28,8 procent moeite ervaart met het nemen van beslissingen over hun eigen gezondheid. Een belangrijke oorzaak hiervan is dat veel (medische) informatieve teksten niet goed zijn afgestemd op het kennis- en leesniveau van het grootste deel van de Nederlandse bevolking. Dit werd vastgesteld in het rapport *Begrijpelijke Taal: fundamenten en toepassingen van effectieve communicatie* (NWO, 2010). Hierin staat dat veel teksten te formeel, juridisch en jargonrijk zijn. Successcores op het correct vinden en begrijpen van informatie zijn daarom vaak laag en revisies van teksten leiden lang niet altijd tot verbetering en een verhoogd begrip van de tekst. Het rapport benadrukt dat er geen universeel antwoord is op de vraag wat een tekst begrijpelijk maakt. Meer onderzoek is nodig zodat duidelijk wordt welke factoren daadwerkelijk bijdragen aan een hogere mate van begrijpelijkheid (Lentz, 2011).

Ander recent onderzoek van Oliffe et al. (2019) toont aan dat medische informatiebladen van reumatologen vaak slecht worden begrepen door patiënten. De auteurs laten zien dat het begrip toeneemt wanneer teksten eenvoudiger zijn geformuleerd, kortere zinnen bevatten en visueel worden ondersteund met afbeeldingen en infographics. Een beperkt vermogen om (medische) teksten te begrijpen kan negatieve gevolgen hebben. In de literatuur wordt dit vaak gekoppeld aan het begrip functionele geletterdheid. Hoewel er verschillende definities van functionele geletterdheid bestaan, richten de meesten, onder wie Nutbeam (2008), zich op het vermogen om eenvoudige teksten te lezen, te begrijpen en vervolgens te vertalen naar eenvoudige verklaringen over het dagelijks leven. Functionele geletterdheid is volgens Nutbeam (2008) van belang, omdat deze het mogelijk maakt om volledig deel te nemen aan de samenleving. Functionele geletterdheid stelt mensen beter in staat om dagelijkse gebeurtenissen te begrijpen en er controle over uit te oefenen.

De opkomst van generatieve kunstmatige intelligentie (GenAI) kan mogelijk bijdragen aan het vergroten van functionele geletterdheid. Grote taalmodellen (LLM's) kunnen met minimale input uitgebreide, grammaticaal correcte teksten genereren en bieden nieuwe mogelijkheden voor het herschrijven van bestaande teksten in een vereenvoudigde vorm. Dit

is met name relevant in de medische sector, waar het van groot belang is dat patiënten zelfstandig uitleg, adviezen en instructies kunnen begrijpen en toepassen (Pharos, 2025).

Hoewel GenAI de potentie heeft om teksten automatisch te vereenvoudigen, is het de vraag of dit daadwerkelijk leidt tot een hogere mate van begrijpelijkheid dan herschrijvingen door menselijke redacteuren. Om bij te dragen aan deze discussie en beter inzicht te krijgen in de effectiviteit van AI bij het (her)schrijven van medische brochures, richt dit onderzoek zich op een vergelijking tussen door AI herschreven teksten en herschrijvingen door menselijke redacteuren. Het doel van dit onderzoek is om een inzicht te verkrijgen in de effectiviteit van generatieve AI bij het herschrijven van medische brochures. De begrijpelijkheid van de teksten wordt beoordeeld met de LiNT-formule (Leesbaarheidsinstrument voor Nederlandse Teksten), die dertien objectieve kenmerken van tekstcomplexiteit omvat. De onderzoeksvraag die centraal staat luidt als volgt:

In hoeverre is GenAI in staat medische brochures begrijpelijker te formuleren dan menselijke redacteuren, gemeten op basis van objectieve tekstuele kenmerken?

Om antwoord te geven op deze vraag wordt in hoofdstuk 2 het theoretisch kader uiteengezet, met aandacht voor tekstbegrijpelijkheid, tekstvereenvoudiging, AI-gestuurde herschrijving en de functie van medische brochures. In hoofdstuk 3 wordt de methode toegelicht. Hierin wordt beschreven hoe het corpus is samengesteld, hoe de AI-teksten zijn gegenereerd via twee prompts, en hoe de vier tekstversies worden vergeleken met behulp van de LiNT-formule. Met statistische modellen (CLMM en LMM) worden verschillen in begrijpelijkheid onderzocht. Daarnaast wordt er naast een kwantitatieve analyse in hoofdstuk 4 ook een kwalitatieve analyse uitgevoerd, die verschillen verklaren tussen menselijke herschrijvingen en GenAI. In hoofdstuk 5 volgt de discussie en hoofdstuk 6 sluit af met conclusie door middel van beantwoording van de onderzoeksvraag en aanbevelingen voor vervolgonderzoek en praktijk.

Hoofdstuk 2 Kaders rondom begrijpelijke taal in medische brochures: mens en AI

In dit hoofdstuk wordt het onderzoek ingebed in bestaande wetenschappelijke en maatschappelijke kaders rondom begrijpelijkheid, tekstoptimalisatie en AI-gestuurde herschrijving. Begrijpelijke taal is de afgelopen jaren steeds belangrijker geworden in de communicatie met leken en met name in de medische context. Daarom worden concepten als 'begrijpelijkheid', 'tekstkwaliteit' en 'tekstvereenvoudiging' hier theoretisch onderbouwd en van elkaar onderscheiden. Eveneens wordt ingegaan op de rol van GenAI en op de genrekenmerken van medische brochures. Tot slot wordt aandacht besteed aan het juridische kader dat bepaalt hoe helder en toegankelijk medische informatie moet zijn geformuleerd.

2.1 Menselijke tekstoptimalisatie: schrijven op B1-niveau

In het publieke debat wordt *begrijpelijke taal* vaak vertaald naar de term *B1*. Dit concept is inmiddels breed geaccepteerd en staat hoog op de agenda van veel overheidsinstellingen, verzekeraars, banken en zorginstellingen, die hun communicatie expliciet op dit zogenoemde B1-niveau willen formuleren.[1] Een tekst op B1-niveau is een tekst die uit korte, actieve zinnen bestaat en waarin alleen woorden staan die veel voorkomen in onze taal (Visser, 2009, p. 308). De nuances in terminologie wijzen soms op een verschil in doelgroep: *eenvoudige taal* wordt eerder geassocieerd met communicatie voor laaggeletterden; terwijl *duidelijke taal* een breder publiek beoogt. Over het algemeen wordt aangenomen dat teksten op B1 toegankelijk zijn voor ongeveer tachtig tot negentig procent van de Nederlandse bevolking (Onze taal, z.d.).

Van oorsprong komt B1 uit het Gemeenschappelijke Europees Referentiekader voor Talen (ERK) van de Raad van Europa. Dit kader beschrijft zes niveaus van taalvaardigheid van A1 (beginnend) tot C2 (zeer gevorderd) en verwijst naar het niveau waarop iemand een vreemde taal beheerst. Deze niveaus gelden voor meerdere vaardigheden: lezen, luisteren, schrijven, spreken en gesprekken voeren. In bijlage A (Meijer en Noijons, 2008) is zichtbaar dat mensen met leesniveau B1 in staat zijn om vertrouwde en alledaagse teksten in een vreemde taal te begrijpen, vooral als die teksten voorspelbare woordenschat bevatten en in herkenbare contexten voorkomen, bijvoorbeeld op het werk of in persoonlijke communicatie.

[1] https://digitaaltoegankelijk.nl/direct-duidelijk/; https://www.communicatierijk.nl/vakkennis/rijkswebsites/aanbevolen-richtlijnen/taalniveau-b1

Samenvattend kan gesteld worden dat B1 van oorsprong een niveau van taalvaardigheid aanduidt in een vreemde taal. Tegenwoordig heeft dit niveau een bredere, normatieve functie gekregen: in het Nederlandse taalbeleid en binnen institutionele praktijken waar communicatie een belangrijke rol speelt, fungeert B1 steeds vaker als richtlijn voor begrijpelijke en toegankelijke taal in algemene zin.

De praktische vertaling van wat taalniveau B1 kenmerkt, is grotendeels te vinden in taaladviesboeken. In het boek *Schrijf eens even normaal joh!* pleiten Evers en Verdaasdonk (2020) voor helder en verzorgd taalgebruik dat aansluit bij alledaagse spreektaal. Zij adviseren schrijvers onder andere om jargon en formele beleidstaal te vermijden, om concrete formuleringen te hanteren, zinnen beknopt te houden, kopjes te gebruiken, opsommingen en variatie aan te brengen in zinslengte. Verder wijzen zij op het belang van actieve formuleringen en het vermijden van hulpwerkwoorden als *worden, zullen, kunnen, proberen* en *hopen.* Tegelijkertijd bestaat er scepsis over de haalbaarheid en effectiviteit van B1-niveau in de praktijk. Hendrikx (2012) erkent in zijn boek *Schrijven voor het Beeldscherm* dat de Nederlandse overheid B1 als norm hanteert, maar constateert dat het in de praktijk vaak niet haalbaar is om boodschappen daadwerkelijk op B1-niveau te brengen. Volgens Hendrikx (2012) is het gemiddelde taalniveau in Nederland B2 en zorgt taalgebruik op B2-niveau ervoor dat lezers zich niet als een kind voelen aangesproken. Daarom geeft hij in zijn boek adviezen voor schrijven op B2-niveau:

- Gebruik geen modieus taalgebruik of jargon: teksten moeten helder/begrijpelijk zijn.
- Gebruik concrete woorden en gebruik voorbeelden.
- Kies voor een duidelijke structuur (vraag-antwoord, discussie, interview).
- Schrijf actief en persoonlijk: gebruik geen lijdende vorm, gebruik persoonlijke voornaamwoorden om mensen aan te spreken en schrijf in verzorgde spreektaal.
- Schrijf beknopt: gebruik geen lange zinnen, vermijd zowel tangconstructies als hulpwerkwoorden.
- Maak gebruik van visuele elementen: opsommingen, kopjes en niet-tekstuele symbolen als typografische symbolen (Hendrikx, 2012, p.78 – p.84).

Een alternatief voor het werken met vaste taalniveaus bieden De Geus en Loomans (2023), die in hun boek *Stukken beter schrijven* kiezen voor een lezersgerichte benadering. Zij leggen de nadruk op het afstemmen van de tekst op de kennis, denkbeelden en verwachtingen van de lezer, in plaats van op een vaststaand taalniveau. Toch komen zij met vergelijkbare praktische adviezen als eerdergenoemde auteurs. Op het gebied van stijl adviseren zij schrijvers: schrijf

verzorgde spreektaal; houd de zinslengte beperkt; gebruik actieve taal; zorg voor afwisseling in zinslengte, kies lekentaal en vermijd jargon. Op het gebied van structuur adviseren zij om kopjes, opsommingen en structuuraanduiders te gebruiken.

Uit wetenschappelijk oogpunt is er kritiek op de normatieve opvattingen van begrijpelijkheid die aan het B1-niveau ten grondslag liggen. Jansen (2013, p.56) bekritiseert het groeiende vertrouwen in het schrijven op B1-niveau als dé oplossing voor begrijpelijke communicatie. De auteur wijst erop dat het Europees Referentiekader voor Talen (ERK) is bedoeld om taalvaardigheid in een vreemde taal te meten, niet om de moeilijkheid van moedertaalteksten te bepalen. De vaak genoemde claim dat B1-teksten voor vrijwel alle Nederlanders begrijpelijk zijn, mist volgens hem wetenschappelijke grondslag en de bijbehorende schrijfadviezen zijn volgens hem al lang bekend. Ook tools die het B1-niveau van teksten meten, zijn volgens Jansen (2013) niet transparant en weinig voorspellend.

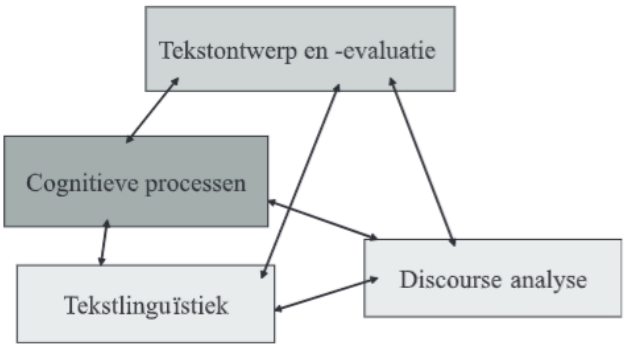
Kortom, hoewel B1-niveau in Nederland breed wordt gedragen als norm, ontbreken zowel wetenschappelijke onderbouwing als praktische effectiviteit voor alle doelgroepen. Dit roept de vraag op wat ‘begrijpelijkheid’ nu precies betekent en hoe dit concept op valide en betrouwbare wijze kan worden gedefinieerd en gemeten. In de volgende paragraaf worden daarom verschillende concepten rondom tekstbegrijpelijkheid systematisch uiteengezet, om de relaties en nuances tussen deze verschillende begrippen te verhelderen en zo een stevige basis te bieden voor verdere analyse.

2.2 Tekstbegrijpelijkheid: conceptuele kaders

Het plaatsen van *begrijpelijkheid* binnen een wetenschappelijke context is een complexe uitdaging. Er lijkt sprake te zijn van conceptuele verwarring tussen termen zoals ‘tekstkwaliteit’, ‘begrijpelijkheid’, ‘tekstbegrip’, ‘simplificatie’ en ‘vereenvoudiging’. Deze begrippen worden vaak door elkaar gebruikt zonder dat ze expliciet gedefinieerd of geoperationaliseerd worden. In deze paragraaf worden deze verschillende concepten systematisch naast elkaar gezet om hun onderlinge relaties en nuances te verduidelijken.

2.2.1 Tekstkwaliteit

Binnen het domein van de taalbeheersing vormt (tekst)kwaliteit een kernbegrip bij het analyseren en verbeteren van communicatie. Sanders (2005) onderscheidt hierin twee fundamentele benaderingen: een verklarende en een evaluatieve. De eerste richt zich op de vraag hoe mensen taal gebruiken in verschillende contexten, en welke regels en patronen daarbij een rol spelen, bijvoorbeeld hoe formuleringen de interpretatie door de lezer beïnvloeden. De tweede benadering betreft de evaluatie van teksten op hun communicatieve effectiviteit: in hoeverre slaagt een tekst erin om zijn doel (aantrekkelijkheid, begrijpelijkheid en/of overtuigingskracht) te bereiken bij een specifieke doelgroep? Wat de taalbeheersing als discipline kenmerkt, is de wisselwerking tussen twee kernbenaderingen: enerzijds het verklaren van taalgebruik in context, anderzijds het evalueren en optimaliseren van communicatie (Sanders, 2005).



Figuur 1 Onderzoeksprogramma taalbeheersing als netwerk van subdisciplinaire projecten

In dit onderzoek worden deze perspectieven met elkaar verbonden. Enerzijds wordt kwantitatief onderzocht of de AI-herschreven medische teksten begrijpelijker zijn dan die van menselijke redacteuren, gemeten aan de hand van objectieve tekstuele kenmerken. Anderzijds wordt via een aanvullende kwalitatieve analyse onderzocht welke linguïstische en structurele keuzes AI doorgaans maakt, en hoe deze zich verhouden tot de herschrijvingsstrategieën van menselijke redacteuren. Daarmee sluit dit onderzoek aan bij zowel het verklarende als het evaluatieve perspectief binnen de taalbeheersing. Tekstkwaliteit vormt hierin een centraal concept, zoals ook door Sanders (2005) wordt benadrukt in zijn beschrijving van de tweeledige kern van de discipline.

Een gangbaar kader voor *tekstkwaliteit* vormt het CCC-model van Renkema (2010, p. 37 - 47). Dit model evalueert teksten aan de hand van drie centrale criteria: het correspondentiecriterium (de mate waarin de tekst is afgestemd op de behoeften, voorkennis en verwachtingen van de lezer); het consistentiecriterium (de mate waarin de keuzes binnen de tekst op het gebied van stijl en structuur consequent worden doorgevoerd) en het correctheidscriterium (de mate waarin de tekst grammaticaal, inhoudelijk en vormtechnisch correct is). Deze criteria worden toegepast op vijf tekstniveaus: tekstsoort, inhoud, opbouw, formulering en presentatie. Hoewel dit model breed toepasbaar is, erkent Renkema (2010,

p.46) zelf de abstracte aard ervan. Het CCC-model biedt een diagnostisch raamwerk, maar werkt niet met concrete, objectieve tekstkenmerken. Hierdoor is het model minder geschikt voor dit onderzoek dat gericht is op meetbare indicatoren van begrijpelijkheid.

Een meer empirisch gefundeerde benadering komt van Langer et al. (2019, p. 192), die op basis van systematisch lezersonderzoek drie centrale dimensies van tekstkwaliteit identificeren: stijl (eenvoudige woorden en zinnen); structuur (logische samenhang); en inhoud (tekstlengte staat in verhouding tot de benodigde informatie). Zij legden in de jaren zeventig daarmee de grondslag voor veel hedendaagse richtlijnen voor begrijpelijk schrijven. Deze dimensies werden toegepast op herschrijvingen van meer dan tweehonderd uiteenlopende teksten, die vervolgens werden beoordeeld door circa 4500 proefpersonen. Uit hun resultaten bleek dat herschreven teksten gemiddeld tot vijftig procent beter werden begrepen dan de oorspronkelijke variant. Daarbij kwam met name naar voren dat een eenvoudige stijl een essentiële voorwaarde bleek voor begrijpelijkheid, oftewel: teksten met een complexe stijl werden systematisch slechter begrepen, ongeacht de kwaliteit van structuur en/of inhoud.

Hoewel het model van Langer et al. (2019) relatief beperkt is in omvang (slechts drie dimensies) biedt het wel een empirisch toetsbaar en tekstintern kader, dat in methodologisch opzicht beter aansluit bij de doelstelling van dit onderzoek dan bredere en lezersafhankelijke modellen zoals het CCC-model van Renkema (2010). Waar Renkema uitgaat van contextuele afstemming, consistentie en correctheid op meerdere tekstniveaus, is het model van Langer et al. meer geschikt voor systematische vergelijking op basis van objectieve stilistische en structurele kenmerken. Deze bevindingen van Langer sluiten ook aan bij de praktische richtlijnen van het B1-schrijven, dat eveneens inzet op stilistische vereenvoudiging om teksten toegankelijker te maken voor een breed publiek.

2.2.2 Tekstbegrip: begrijpelijkheid vanuit lezersperspectief

Tekstbegrip en *begrijpelijkheid* zijn qua definities nauwverwant aan elkaar en worden zowel in de literatuur als in de praktijk vaak door elkaar gebruikt. Beide concepten richten zich op de wisselwerking tussen tekst en lezer. Lentz (2020) erkent dat er verschillende opvattingen zijn rondom *begrijpelijke taal.* De hoogleraar aan de Universiteit Utrecht benadrukt in tegenstelling tot Langer et al. (2019) dat begrijpelijke taal niet louter gedefinieerd kan worden aan de hand van specifieke linguïstische kenmerken, zoals eenvoudig woordgebruik of korte zinnen. In plaats daarvan pleit hij ervoor om begrijpelijkheid te evalueren vanuit het

perspectief van de lezer. Lentz (2020) stelt in zijn definitie centraal dat de ontvanger in staat moet zijn om benodigde informatie makkelijk te vinden, te begrijpen en toe te passen:

> ‘’Een document, website of ander product van communicatie is begrijpelijk als de beoogde gebruikers de informatie die zij nodig hebben kunnen vinden, kunnen begrijpen en kunnen toepassen op hun eigen situatie en aldus de voor hun relevante taak ermee uit kunnen voeren, dankzij de heldere verwoording, de toegankelijke structuur en het (grafisch) ontwerp’’ - (Lentz, 2020).

Deze benadering sluit aan bij *plain language* de vertaling van begrijpelijke taal die in de Verenigde Staten en Engeland wordt gehanteerd. Plain language wordt gekarakteriseerd door een woordkeuze, structuur en vormgeving die aansluiten bij de behoeften en mogelijkheden van de doelgroep. De ontvanger moet in staat zijn om informatie eenvoudig te vinden, te begrijpen en praktisch te benutten (Schriver, 2017).

Dat begrijpelijkheid geen vaststaand kenmerk van een tekst is, maar mede afhangt van de lezer, wordt onderstreept door Van der Bruggen (2020), die onderzoekt wanneer er gesproken kan worden van een begrijpelijke rechterlijke uitspraak. Zij verklaart evenals Lentz (2011; 2020) dat begrijpelijkheid niet alleen een relatief maar ook een gradueel begrip is. Een tekst die voor de ene lezer begrijpelijk is, hoeft dat niet te zijn voor een andere lezer. Van der Bruggen (2020, p. 2028) definieert begrijpelijkheid daarom in termen van de cognitieve belasting die een tekst oplegt aan de lezer en diens benodigde voorkennis: hoe minder inspanning en voorkennis vereist zijn, hoe begrijpelijker een tekst is. Een begrijpelijke tekst stelt de lezer in staat om met minimale moeite tot een dieper tekstbegrip te komen.

Tekstbegrip verwijst naar de mentale representatie die een lezer van een tekst construeert. Vanuit cognitief-linguïstisch perspectief en *discourse processing* (Kintsch, 1998; Singer, 1990) kunnen drie verschillende niveaus van tekstbegrip onderscheiden worden: oppervlakteniveau (grammatica en woordbetekenis), betekenisrepresentatie (een lezer kent betekenis toe aan afzonderlijke zinnen) en situatiemodel (de lezer integreert nieuwe informatie uit de tekst met reeds aanwezige voorkennis). Het situatiemodel geeft weer hoe goed een tekst daadwerkelijk begrepen wordt; het verwijst naar coherente mentale representatie, waarbij een koppeling plaatsvindt met bestaande kennisstructuren van de lezer. Land et al. (2002) geven aan dat het meten van het situatiemodel complex is, omdat iedere tekst uniek is en omdat iedere lezer een eigen kennisstructuur bezit waardoor methoden en

technieken uiteenlopend zijn en vaak intuïtief worden toegepast en er niet altijd duidelijkheid bestaat over wat exact gemeten wordt.

Kortgezegd richt begrijpelijkheid zich op het faciliteren van tekstbegrip: een tekst die begrijpelijk is volgens criteria van Lentz (2020) ondersteunt lezers bij het vormen van adequate mentale representaties. Tekstbegrip wordt in dit onderzoek opgevat als een resultaat of de uitkomst van begrijpelijkheid. Beide begrippen benadrukken dat het proces interactief en contextgevoelig is en worden beïnvloed door eigenschappen van de lezer. De besproken theoretische perspectieven maken duidelijk dat tekstbegrijpelijkheid een meerdimensionaal fenomeen is, dat zowel afhangt van tekstinterne kenmerken (zoals stijl en structuur) als van lezerskenmerken, waaronder voorkennis en cognitieve verwerkingscapaciteit. Waar modellen als dat van Langer et al. (2019) uitgaan van meetbare, objectieve kenmerken die bijdragen aan tekstkwaliteit, benadrukken andere benaderingen, zoals die van Lentz (2020) en Van der Bruggen (2020), juist het belang van begrijpelijkheid als gradueel en situationeel begrip. Begrijpelijkheid is in deze optiek niet alleen het product van formele tekstkenmerken, maar ook van de mate waarin de lezer erin slaagt een coherente mentale representatie te construeren.

2.2.3 Tekstvereenvoudiging

Een andere definitie die verwant is aan de term begrijpelijkheid is *tekstvereenvoudiging,* ook wel bekend onder de noemer *tekstsimplificatie* (letterlijk vertaald vanuit het Engels). Siddharthan (2014, p. 259) en Shardlow (2014, p. 58) definiëren tekstvereenvoudiging als ht proces waarin de taalkundige complexiteit van een tekst automatisch wordt verminderd om deze begrijpelijker en beter leesbaar te maken, waarbij de oorspronkelijke informatie en betekenis behouden blijven. Volgens Siddharthan (2014) en Shardlow (2014) helpt tekstvereenvoudiging lezers met beperkte leesvaardigheden, tweedetaalleerders en kinderen. Tekstvereenvoudiging welke gericht is op zulke specifieke doelgroepen, vragen volgens beide auteurs om een bredere definitie van vereenvoudiging. Binnen de context van dit onderzoek is TS vooral relevant als benadering om medische brochures begrijpelijker te formuleren.

Vanuit historisch perspectief functioneerde automatische tekstvereenvoudiging als een voorverwerkingsstap om taken binnen natuurlijke taalverwerking (*Natural Language Processing,* afgekort NLP) te verbeteren (Althunayyan en Azmi, 2021). Binnen NLP wordt veel onderzoek gedaan naar automatische lexicale en syntactische vereenvoudiging. Naast

oppervlakkige taalkundige bewerkingen (lexicaal en syntactisch), onderscheiden Siddharthan en Shardlow (2014) ook conceptuele vereenvoudiging, elaboratieve aanpassingen en tekstsamenvattingen. Deze bredere benadering (zie bijlage B) benadrukt dat vereenvoudiging niet louter een grammaticaal, maar ook een inhoudelijk en communicatief proces is.

Syntactische vereenvoudiging wordt doorgaans uitgevoerd in drie opeenvolgende stappen: eerst wordt in de analysefase de grammaticale structuur van een zin geclassificeerd en beoordeeld op complexiteit. Vervolgens worden in de transformatiefase herschrijfregels toegepast, zoals het opsplitsen van samengestelde zinnen of het herschikken van bijzinnen. In de afsluitende regeneratiefase wordt de herschreven tekst geoptimaliseerd op het vlak van samenhang en cohesie, zodat het resultaat helder en toegankelijk is voor de beoogde doelgroep. Door deze gefaseerde aanpak draagt syntactische vereenvoudiging bij aan het verbeteren van de begrijpelijkheid van teksten voor doelgroepen met beperkte taalvaardigheid, zonder dat inhoudelijke nauwkeurigheid verloren gaat (Shardlow, 2014). Met name in domeinen waar tekstbegrip van cruciaal belang is, zoals de medische sector, biedt syntactische vereenvoudiging concrete aanknopingspunten voor het herschrijven van complexe teksten ten behoeve van lezers met beperkte taalvaardigheid.

Diverse studies tonen aan dat er een positieve correlatie is tussen tekstvereenvoudiging en leesbegrip. Zo tonen Beck et al. (1991) aan dat het expliciet maken van tekstrelationele verbanden het begrip vergroot. Linderholm et al. (2000) ondervonden soortgelijke overeenkomsten in hun onderzoek; zij zagen dat het begrip bij zwakke lezers aanzienlijk verbeterde wanneer causale verbanden in moeilijke teksten werden herschreven. Daarnaast bleek uit studies van Anderson & Davison (1988) en Irwin (1980) dat de volgorde van informatie belangrijk is voor verschillende lezersgroepen. Zo zijn vooraf geplaatste bijwoordelijke bijzinnen (bijv. "Omdat X...") lastig voor kinderen in groep 5 tot 7, zelfs wanneer de volgorde van vermelding overeenkomt met de causale volgorde (Anderson & Davison, 1988). Deze bevinding sluit aan bij het natuurlijke ontwikkelingspatroon van connectieven in kindertaal (Levy, 2003).

Een relevante toepassing van tekstvereenvoudiging in de medische sector is het onderzoek van Ong et al. (2007), waarin het NLP-gebaseerde systeem *SimText* wordt geëvalueerd. Dit systeem past zowel lexicale als syntactische vereenvoudigingen toe op medische teksten van onder andere de Mayo Clinic. In zeventig procent van de gevallen leidde deze bewerkingen tot een hogere Flesch Reading Ease-score, waarbij vooral het vervangen van moeilijke termen door synoniemen en het opsplitsen van samengestelde zinnen

resulteerde in een lager leesniveau. De studie laat echter ook zien dat niet alle vereenvoudigingsstrategieën even effectief zijn: het toevoegen van definities verhoogde soms juist de complexiteit. Bovendien wijst het onderzoek op de beperkingen van leesbaarheids-formules, die vooral oppervlakkige kenmerken meten en geen rekening houden met voorkennis of coherentie. Ook werd het effect op daadwerkelijk begrip niet getest met menselijke lezers. Toch toont de studie aan dat automatische vereenvoudiging via NLP-technieken in veel gevallen bijdraagt aan betere leesbaarheid van medische teksten, en levert zij een waardevolle empirische onderbouwing voor de relevantie van TS binnen begrijpelijke gezondheidscommunicatie.

Tot slot kan tekstvereenvoudiging eveneens gekoppeld worden aan de begrippen die hiervoor behandeld zijn. Kincaid et al. (1975) leggen uit dat het bepalen van complexiteit van een tekst een essentieel punt is binnen tekstvereenvoudiging, omdat dit bepalend is voor de beslissing of een tekst wel of niet vereenvoudigd dient te worden. Zij verklaren dat een lezer hier de bepalende factor in is, omdat dat wat complex is voor de ene taalgebruiker, niet noodzakelijk complex hoeft te zijn voor de andere. Dit komt overeen met de definities van tekstbegrip en begrijpelijkheid, waarin de mentale representaties van de lezer centraal staan. Ook veel AI-toepassingen (zoals *Large Language Models*) maken gebruik van lexicale en syntactische vereenvoudigingstechnieken, omdat de modellen grotendeels gebaseerd zijn op NLP. Daarmee vormt TS een theoretisch kader waarmee de geobserveerde herschrijvingen systematisch kunnen worden geanalyseerd.

2.3 Kwantitatieve analyse-instrumenten

Naast kwalitatieve benaderingen en theoretische kaders rondom tekstbegrijpelijkheid zijn er ook kwantitatief geautomatiseerde instrumenten ontwikkeld om de leesbaarheid en begrijpelijkheid van Nederlandstalige teksten objectief te voorspellen. Dergelijke tools ondersteunen tekstschrijvers en communicatieprofessionals bij het signaleren van potentieel moeilijke passages en het vergelijken van tekstversies op basis van meetbare kenmerken. In deze paragraaf worden drie veelgebruikte instrumenten besproken: Texamen, Klinkende Taal en de Accessibility Leesniveau Tool. Daarbij wordt kritisch ingegaan op wetenschappelijke onderbouwing en de voorspellende waarde van deze tools. Tot slot wordt aandacht besteed aan de recent ontwikkelde LiNT-formule, die expliciet is gebaseerd op empirisch gemeten tekstbegrip en daarmee antwoord biedt op de beperkingen van eerdere leesbaarheidsmodellen.

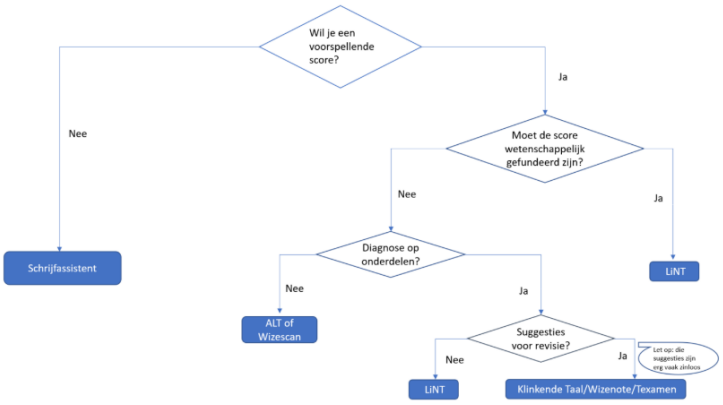
2.3.1 Texamen, Klinkende Taal en Accessibility Leesniveau Tool

Jansen en Boersma (2013) onderzochten drie automatische instrumenten voor het voorspellen van de leesbaarheid van Nederlandse teksten: Texamen, Klinkende Taal en de Accessibility Leesniveau Tool. Eerder onderzoek van Kraf et al. (2011) had al aangetoond dat deze instrumenten onderling sterk kunnen verschillen in hun inschatting van leesniveau. Jansen en Boersma (2013) bouwden hierop voort en gebruikten dezelfde negentien teksten, waaronder fragmenten van een woningbouwvereniging, krantenberichten, polisvoorwaarden en roddelrubrieken. In totaal namen 125 proefpersonen deel aan het onderzoek.

Het eerste doel was om te bepalen in hoeverre de voorspellingen van de drie instrumenten overeenkwamen met het werkelijke tekstbegrip van lezers, gemeten via cloze-tests. De resultaten toonden aan dat Texamen geen significante relatie vertoonde met de cloze-scores, wat suggereert dat dit instrument het begrip niet betrouwbaar voorspelt. Klinkende Taal en de Accessibility Leesniveau Tool lieten wel significante negatieve correlaties zien met de cloze-scores, maar deze verbanden bleken relatief zwak. Het tweede doel was om te onderzoeken welke tekstkenmerken het begrip het beste voorspellen. Een regressieanalyse liet zien dat gemiddelde woordlengte en zinslengte samen zestig procent van de variantie in begripsscores konden verklaren. Andere kenmerken, zoals de proportie frequente woorden, de type-token ratio, en de afstand tussen onderwerp en persoonsvorm, voegden geen verklarende waarde toe.

Jansen en Boersma concluderen dat de onderzochte instrumenten methodologisch tekortschieten als betrouwbare leesbaarheidsindexen. Ze bekritiseren met name het feit dat deze tools zijn ontwikkeld zonder grondige validatie met *echte* lezers, daarmee bedoelen zij: personen uit de werkelijke doelgroep van de tekst, die de tekst lezen zoals deze in de praktijk bedoeld is. Deze tools baseren zich vooral op oppervlakkige kenmerken zoals woord- of zinslengte. Het gebruik van empirische data van echte lezers is essentieel om inzicht te krijgen in hoe teksten daadwerkelijk worden begrepen in de praktijk. Daardoor ontbreekt volgens de auteurs een solide wetenschappelijke onderbouwing.

2.3.2 De LiNT-formule als empirisch onderbouwde maat voor begrijpelijkheid



Figuur 2 Keuzetool analyse-instrument (Lentz, 2021)

De LiNT-formule is een leesbaarheidsinstrument dat in tegenstelling tot veel bestaande algoritmes wel gebaseerd is op metingen met echte lezers. Lentz (2021) beschrijft op de website *Didactiek Nederlands* dat LiNT werkt met wetenschappelijk gefundeerde voorspellingen van tekstbegrip. Dat is belangrijk, omdat veel bestaande modellen alleen gebaseerd zijn op expertinschattingen of geclassificeerde leesniveaus (Collins-Thompson, 2014; Vajjala, 2021). Zulke modellen voorspellen vooral de aannames van experts, niet hoe goed lezers een tekst werkelijk begrijpen. Een beter alternatief is om oordelen van echte lezers te gebruiken (Vor der Brück et al., 2008; De Clercq et al., 2014), maar ook die blijken niet betrouwbaar, omdat lezers hun eigen begrip vaak overschatten. Een tweede probleem is dat moderne modellen vaak moeilijk te doorgronden zijn. Zo gebruiken Martinc, Pollak & Robnik-Šikonja (2021) neurale modellen gebaseerd op woordperplexiteit om tekstcomplexiteit te voorspellen, maar deze modellen geven geen uitleg waarom een tekst als moeilijk wordt ingeschat. Dat maakt ze ongeschikt voor situaties waarin uitleg of verantwoording nodig is.

Om bovenstaande tekortkomingen in leesbaarheidsonderzoek te ondervangen, ontwikkelden Pander Maat et al. (2023) de LiNT-formule: de eerste empirisch gevalideerde leesbaarheidsmaat voor het Nederlands buiten het basisonderwijs. LiNT benadert leesbaarheid als onderdeel van tekstbegrijpelijkheid, waarbij ook opbouw, vormgeving en inhoud een rol spelen. De auteurs hielden bij de ontwikkeling van de formule expliciet rekening met bekende problemen in traditioneel onderzoek, zoals:

- Beperkte en onbetrouwbare voorspellers: veel oudere formules gebruiken kenmerken zoals zins- of woordlengte als maat voor complexiteit, terwijl deze niet altijd direct iets zeggen over tekstbegrip. Zulke oppervlakkige maten kunnen misleidend zijn.
- Gebrekkige nuancering in metingen: vaak wordt slechts één gemiddelde score per tekst gegeven, zonder inzicht in moeilijkheidsverschillen binnen de tekst of specifieke probleemlocaties (zoals een complexe zin of paragraaf).

- Gebrek aan aansluiting bij de lezer en context: veel modellen houden geen rekening met verschillen tussen lezers (zoals voorkennis of taalvaardigheid), of met de invloed van tekstsoort. Ook gebruiken ze soms teksten die extreem van elkaar verschillen, wat leidt tot overschatting van het voorspellend vermogen van de formule.

De LiNT-formule is ontwikkeld op basis van tekstbegripdata van ruim 2700 leerlingen uit de tweede, derde en vierde klas van het voortgezet onderwijs (13–16 jaar). Pander maat et al. (2023) gingen ervan uit dat de formule ook toepasbaar zou zijn op zakelijke teksten voor volwassenen, aangezien uit geletterdheidsonderzoek blijkt dat hun leesvaardigheid vaak vergelijkbaar of zelfs lager is dan die van zestienjarigen (OECD, 2013). Voor het onderzoek werd een corpus van 120 tekstvarianten samengesteld, bestaande uit schoolboek- en voorlichtingsteksten, waaronder gemanipuleerde versies om de invloed van formulering los te koppelen van de inhoud. Het tekstbegrip werd gemeten met de Hybrid Text Comprehension (HyTeC) cloze-methode: een test waarbij vooral minder voorspelbare inhoudswoorden werden weggelaten, terwijl technische termen, eigennamen en getallen werden vermeden. Teksten werden geanalyseerd met T-Scan, dat kenmerken uit negen taalkundige categorieën identificeert zoals woordfrequentie, zinscomplexiteit en lexicale diversiteit. Via stapsgewijze lineaire regressieanalyse werden vier kenmerken geselecteerd die tekstbegrip het best voorspellen: woordfrequentie**,** concreetheid van naamwoorden**,** afhankelijkheidslengte en deelzinslengte**.**

Pander Maat et al. (2023) stellen dat het niet haalbaar is om de LiNT-score direct te koppelen aan leesbaarheidsniveaus voor specifieke onderwijsniveaus of CEFR-niveaus (zie paragraaf 2.1). In plaats daarvan stellen de auteurs voor om te schatten hoeveel Nederlanders waarschijnlijk moeite hebben met het taalgebruik in een tekst. De voorspelde HyTeC cloze-score wordt omgezet naar een LiNT-score (0-100, hoger = moeilijker). De vier niveaus krijgen de volgende betekenis toegekend in de LiNT-formule:

- **Niveau 1** (< 34 LiNT-score): Streefniveau, geschat dat ongeveer **15%** moeite heeft.
- **Niveau 2** (34-46 LiNT-score): 'Next best', geschat dat ongeveer **30%** moeite heeft.
- **Niveau 3** (46-60 LiNT-score): Moeilijk, geschat dat ongeveer **55%** moeite heeft.
- **Niveau 4** (> 60 LiNT-score): Extreem moeilijk, geschat dat ongeveer **82%** moeite heeft.

De LiNT-schaal is geëvalueerd aan de hand van diverse zakelijke tekstgenres. De scores bleken grotendeels overeen te komen met intuïtieve moeilijkheidsgraden: zo werden reisblogs en roddelberichten als makkelijk geclassificeerd en gerechtelijke uitspraken en wetenschappelijke artikelen als moeilijk. Hierdoor is de tool bruikbaar voor tekstprofessionals zoals schrijvers, schrijfcoaches en communicatieadviseurs, bijvoorbeeld om teksten te beoordelen, versies te vergelijken of moeilijkheden op te sporen.

Pander Maat et al. (2023) stellen dat LiNT een belangrijke verbetering is ten opzichte van eerdere formules. Het model is empirisch onderbouwd omdat het is gebaseerd op feitelijke begripsscores, de tool werkt genre-overstijgend en houdt rekening met verschillen tussen lezers. Tegelijkertijd kent de tool beperkingen: aspecten zoals tekststructuur, vormgeving en inhoud blijven buiten beschouwing. Bovendien zijn scores op zins- of alineaniveau nog niet afzonderlijk gevalideerd. Belangrijk om te benoemen is dat een lage score niet automatisch betekent dat een tekst goed begrepen wordt, zo kan een willekeurige lezer niet beschikken over benodigde voorkennis. LiNT is daarom vooral geschikt om moeilijke teksten te signaleren, niet om teksten definitief goed te keuren. De tool is bruikbaar voor verantwoording van tekstkeuzes en mogelijk ook voor probleemopsporing. Verder onderzoek is nodig, vooral naar begrip bij volwassenen, laaggeletterden (Pander Maat et al., 2023).

Dit onderzoek biedt daarmee meerwaarde doordat het beoogt als een van de eersten de koppeling te maken tussen empirisch gevalideerde leesbaarheidsmetingen (zoals de LiNT-formule) en AI-gebaseerde herschrijftechnieken. (Voor de Engelse taal is dit al wel gedaan zie daarvoor paragraaf 2.4). Deze verbinding met de Nederlandse taal is in bestaand onderzoek tot op heden nauwelijks gelegd, terwijl ze juist van belang is voor het evalueren van automatische tekstoptimalisatie op begrijpelijkheid. Hoewel de LiNT-formule in de literatuur uitvoerig wordt gepresenteerd en onderbouwd door de ontwikkelaars zelf, ontbreekt tot op heden een onafhankelijke, kritische bespreking van deze formule door externe onderzoekers. Dit betekent dat de objectiviteit van de LiNT-formule niet volledig kan worden gegarandeerd. Hierdoor is het mogelijk dat bepaalde beperkingen of knelpunten van de formule onderbelicht blijven.

2.4 AI-gestuurde tekstadaptatie

Artificiële intelligentie (AI) speelt een steeds grotere rol in het herschrijven van teksten. Vooral generatieve AI, zoals het taalmodel GPT-4, maakt het mogelijk om automatisch teksten te bewerken of herschrijven op basis van ingevoerde informatie. In sectoren zoals de medische voorlichting, waar duidelijke en begrijpelijke communicatie essentieel is, ontstaat de vraag hoe AI-gegenereerde teksten kunnen bijdragen aan bijvoorbeeld het vereenvoudigen van voorlichtingsteksten. In dit onderzoek is gekozen voor ChatGPT, omdat dit model als enige getraind is als taalmodel, het meest wordt toegepast in de praktijk en het beste scoort volgens verschillende onderzoeken.

2.4.1 ChatGPT als herschrijfhulpmiddel

De opkomst van grote taalmodellen, in het Engels *Large Language Models* (LLMs), zoals ChatGPT, Perplexity, Gemini en talloze andere modellen hebben de manier waarop mensen informatie zoeken en verwerken ingrijpend veranderd. LLMs zijn geavanceerde kunstmatige intelligentiesystemen die in staat zijn om teksten te genereren die nauwelijks te onderscheiden zijn van menselijke communicatie. Door hun vermogen om complexe vragen te beantwoorden en contextueel relevante antwoorden te geven, worden deze modellen steeds vaker ingezet als alternatief voor traditionele zoekmachines, ook binnen de gezondheidszorg (Cao et al., 2024). Cao et al. (2024) benoemen dat LLMs veelbelovend zijn voor medische toepassingen en dat ze potentie hebben om de gezondheidszorg op verschillende manieren te ondersteunen, onder andere bij het beantwoorden van patiëntvragen, het toegankelijker maken van informatie, het beter begeleiden van patiënten en het ondersteunen van zorgverleners. ChatGPT, ontwikkeld door OpenAI, is één van de bekendste voorbeelden van een LLM. Onderzoek van Kung et al. (2023) heeft aangetoond dat het model zelfs in staat is om complexe examens zoals het United States Medical Licensing Examination (USMLE) te halen en diepgaande medische uitleg te geven. Cao et al. (2024) onderzochten de prestaties van drie veelgebruikte LLMs: ChatGPT-3.5, ChatGPT-4.0 en Perplexity bij het beantwoorden van veelgestelde vragen over de aandoening osteoartritis. ChatGPT-4.0 scoorde hierbij het hoogst, gevolgd door ChatGPT-3.5 en Perplexity als laatst. Vooral bij algemene medische kennis waren de antwoorden accuraat en informatief. Bij vragen over behandeling en preventie bleken de antwoorden minder betrouwbaar. De auteurs benadrukken dat, hoewel LLMs veel potentie tonen, het essentieel is om hun antwoorden kritisch te evalueren en voortdurend te werken aan het verbeteren van hun betrouwbaarheid, vooral op het gebied van behandeling en preventie waar medische kennis snel evolueert en misvattingen hardnekkig kunnen zijn (Cao et al. 2024).

Hoewel Cao et al. (2024) laten zien dat ChatGPT-4 inhoudelijk correcte antwoorden kan geven op medische vragen, richt hun onderzoek zich vooral op de juistheid van informatie. Wat met name relevant is voor dit onderzoek is in hoeverre taalmodellen ook geschikt zijn voor het herschrijven van complexe medische teksten tot begrijpelijke voorlichting voor een algemeen breed publiek. Swisher et al. (2024) onderzochten of ChatGPT-4 gebruikt kan worden om de leesbaarheid van patiëntfolders binnen de keel-, neus- en oorheelkunde (KNO) te verbeteren. Ze analyseerden vijf folders van professionele KNO-organisaties die informatie gaven over veelvoorkomende aandoeningen en behandelingen. De oorspronkelijke teksten werden herschreven door ChatGPT, met als opdracht: ‘’Herschrijf dit op het leesniveau van een leerling uit groep 8.’’ Vervolgens werden zowel de originele als de herschreven teksten beoordeeld met verschillende leesbaarheidsformules (zoals Flesch-Kincaid en SMOG) en met een beoordelingsinstrument voor begrijpelijkheid en bruikbaarheid van patiëntinformatie.
De herschreven teksten waren duidelijker en beter leesbaar dan de oorspronkelijke versies. De herschreven teksten waren veel korter (gemiddeld 489 versus 1175 woorden) en bevatten veel minder complexe (polysyllabische: uit meerdere lettergrepen) woorden (40 versus 231). Ze hadden gemiddeld een leesniveau van de brugklas, terwijl de originele teksten op het niveau van de bovenbouw van de middelbare school lagen. Ook scoorden de herschreven teksten hoger op leesbaarheid (91 procent tegenover 74 procent). De mate waarin patiënten concrete acties uit de tekst konden halen (handelingsgerichtheid) was iets beter, maar het verschil was niet statistisch significant. Tegelijkertijd wijzen de onderzoekers op een aantal risico’s:

- ChatGPT reduceerde de tekstlengte met gemiddeld 58%, waardoor mogelijk nuance en detail verloren gaan.
- Soms kiest het model voor informele of minder passende bewoordingen (“Sinusitis can be a drag...”), wat niet altijd wenselijk is in medische communicatie.
- De focus op eenvoud kan ten koste gaan van volledigheid en accuratesse; daarom is review door een arts essentieel.
- Leesbaarheidsformules meten vooral zins- en woordlengte, niet per se daadwerkelijke begrijpelijkheid of inhoudelijke kwaliteit (Swisher et al. 2024).

De bevindingen van Swisher et al. (2024) laten zien dat door ChatGPT herschreven teksten weliswaar beter leesbaar zijn, maar dat ze soms belangrijke nuances missen of formuleringen bevatten die niet goed passen bij medische communicatie. Ze duiden op dieperliggende eigenschappen van het model zelf. Om te begrijpen waarom ChatGPT bepaalde keuzes maakt

in bijvoorbeeld woordgebruik, hoeveelheid informatie en schrijfstijl, is het belangrijk om na te gaan hoe zulke taalmodellen zijn ontwikkeld en getraind.

Last en Sprakel (2024, P. 39) benadrukken dat de kwaliteit van een taalmodel sterk wordt beïnvloed door de data waarop het is getraind. Tijdens de pre-trainingsfase leert het model op basis van grote hoeveelheden tekst patronen, grammatica, wereldkennis en ook eventuele vooroordelen herkennen. Vervolgens wordt het model via *supervised fine-tuning* aangescherpt, waarbij menselijke beoordelaars feedback geven op modeloutput volgens vaste richtlijnen. Tot slot volgt *reinforcement learning from human feedback* (RLHF), waarbij het model leert om menselijke voorkeuren na te bootsen door meerdere antwoordopties te genereren en daaruit de meest geschikte te selecteren.

Een GPT-taalmodel doorloopt verschillende stappen om invoertekst om te zetten in begrijpelijke en contextuele output. Het proces begint met *tokenisatie*, waarbij tekst wordt opgedeeld in kleine eenheden, zoals woorden of woorddelen. Daarna volgt ***vectorisatie***, waarbij deze tokens worden omgezet in getallen (vectoren), zodat het model er wiskundig mee kan werken. In de volgende stap krijgen de vectoren via *woordverankering* betekenis. Dit gebeurt door elk woord een positie te geven in een virtuele ruimte, waarbij de afstand tot andere woorden iets zegt over de betekenis en het gebruik ervan in context. Het model leert deze relaties op basis van patronen in grote hoeveelheden tekst. Daarna verwerkt de *transformerarchitectuur* deze informatie. Dit is een structuur die bestaat uit lagen met aandachtmechanismen en neurale netwerken. Deze analyseren de relaties tussen woorden en houden rekening met de context waarin een woord voorkomt. Wanneer je de vraag stelt: ‘’Hoeveel dagen kent één week?’’ Dan zorgt de transformer ervoor dat het model begrijpt dat ‘dagen’ iets te maken heeft met ‘week’. Na het analyseren genereert het model stap voor stap een reeks vectoren die mogelijke vervolgzinnen voorstellen. Tijdens het genereren van tekst gebruikt het model verschillende *parameters*. De belangrijkste is de *temperatuur*, die bepaalt hoe voorspelbaar of creatief het antwoord is. Bij een lage temperatuur kiest het model voor het meest waarschijnlijke antwoord; bij een hogere temperatuur ontstaat meer variatie. Andere parameters, zoals de *frequency*- en *presence penalty*, zorgen ervoor dat woorden niet te vaak herhaald worden en er meer afwisseling in woordgebruik is. Tot slot zet het model de gegenereerde vectoren via *detokenisatie* om in begrijpelijke woorden en zinnen. Deze output wordt gecontroleerd met *guard rails*: ingebouwde veiligheidsmaatregelen die ervoor zorgen dat het model geen ongepaste of schadelijke inhoud genereert. Zo waarborgt het systeem dat

de gegenereerde antwoorden passen binnen ethische en maatschappelijke normen (Last & Sprakel, 2024, p. 33 – 38).

Last en Sprakel (2024, p. 32) benadrukken dat *Natural Language Processing* (NLP) de basis vormt van taalmodellen; het stelt machines in staat om menselijke taal te begrijpen, te interpreteren en te genereren. Het vakgebied bevindt zich op het snijvlak van informatica en taalkunde. Door NLP te combineren met *transformerarchitectuur* hebben taalmodellen grote vooruitgang geboekt, vooral in hun vermogen om tekst te voorspellen op basis van kansberekening. Transformermodellen werken *autoregressief*: zij genereren tekst woord voor woord, waarbij elk volgend woord gebaseerd is op de volledige reeks eerder gegenereerde woorden. Dit verklaart waarom sommige modellen hun antwoorden zichtbaar opbouwen, één woord tegelijk (Last & Sprakel, 2024, p. 37).

2.4.2 De rol van prompt engineering bij AI-herschrijving

Een prompt is de input (in de vorm van een instructie of vraag) die aan een generatieve AI wordt gegeven om specifieke output te verkrijgen. De kwaliteit van deze prompt bepaalt in sterke mate de relevantie en bruikbaarheid van de gegenereerde respons. *Prompt engineering* verwijst naar de vaardigheid om effectieve prompts te formuleren, waarbij taalvaardigheid, contextbewustzijn en digitale geletterdheid essentieel zijn (Last & Sprakel, 2024, p. 47)

Hoewel stappenplannen voor het schrijven van goede prompts kunnen variëren, volgen ze volgens Last en Sprakel (2024, p. 50 – 51) doorgaans dezelfde kernprincipes, die bijdragen aan gerichte, consistente en contextueel passende output:

1. **Taakbeschrijving:** geef een duidelijke, specifieke instructie of vraag.
2. **Persona-simulatie:** wijs het model een rol of identiteit toe om toon en stijl te sturen.
3. **Stappenstructuur:** geef indien nodig aan welke stappen het model moet volgen.
4. **Context en beperkingen:** specificeer relevante achtergrondinformatie en grenzen.
5. **Doel en doelgroep:** benoem het beoogde doel en voor wie de output bedoeld is.
6. **Outputformat:** geef instructies over de vorm, lengte of stijl van de output.

Het effectief gebruik van een taalmodel verloopt via een interactief en iteratief proces, waarbij gebruikers stapsgewijs toewerken naar betere output. Een gespreksmatige benadering (met een menselijke toon en gerichte feedback) helpt het model om de gewenste respons steeds beter te benaderen. Dit kan onder meer door: herhaalde feedback te geven, vervolgvragen te stellen, om variaties te vragen, aan te geven wat behouden moet blijven, toelichting te vragen

op keuzes, extra informatie toe te voegen en de lengte van de output aan te passen. Deze strategieën versterken het samenspel tussen gebruiker en model en verhogen de kwaliteit van de gegenereerde tekst (Last & Sprakel, 2024).

Een efficiënte strategie om tot het gewenste resultaat te komen bij het gebruik van generatieve AI, is het inzetten van een zogeheten *prompt generator* (Roemmele, 2023). Hierbij werkt de gebruiker iteratief samen met het taalmodel om stap voor stap een optimale prompt te ontwikkelen. In dit proces beoordeelt het model elke versie van de prompt met een score op een schaal van 1 tot 10, vergezeld van inhoudelijke feedback over mogelijke verbeteringen. Wanneer een prompt een score van 8 of hoger behaalt, stelt het model voor om deze prompt daadwerkelijk uit te voeren. Indien dit niet het geval is, genereert het model een verbeterde versie van de prompt. Dit iteratieve mechanisme bevordert doelgerichte communicatie en verhoogt de kwaliteit van de uiteindelijke output (Roemmele, 2023).

Uit het onderzoek van Eid et al. (2023) blijkt dat prompt engineering een cruciale rol speelt in het optimaliseren van de begrijpelijkheid van door large language models (LLM's) gegenereerde teksten. De onderzoekers onderzochten. De onderzoekers wilden weten of teksten die door AI-chatbots, zoals ChatGPT 4.0 en Google Bard, zijn geschreven, makkelijker te begrijpen zijn dan de bestaande brochures van de *American Society of Ophthalmic Plastic and Reconstructive Surgery* (ASOPRS). In de eerste onderzoeksfase werd de chatbots gevraagd om informatieve teksten te genereren zonder nadere specificaties omtrent het gewenste leesniveau. De resultaten toonden aan dat ChatGPT 4.0 in deze setting de minst toegankelijke teksten produceerde, terwijl Google Bard zonder aanvullende instructies al duidelijk leesbaardere output genereerde dan zowel ChatGPT als de originele ASOPRS-teksten. Bij het expliciet instrueren van de AI-systemen om te schrijven op leesniveau groep 8, verbeterde de leesbaarheid van de output significant. In het bijzonder leverde ChatGPT 4.0 onder deze condities de meest toegankelijke teksten op. Op bijna alle leesbaarheidsindices, zoals de *Gunning Fog Index* (10,3), *Flesch-Kincaid Grade Level* (7,8), *Coleman–Liau Index* (10,2), *SMOG Index* (10,2) en *Automated Readability Index* (8,9), behaalde ChatGPT 4.0 met prompt lagere en daarmee betere scores dan de andere opties. Toch bleek dat zelfs met deze aanpassingen het gemiddelde leesniveau nog steeds rond klas 4 havo/vwo lag, terwijl de *American Medical Association* adviseert om patiëntinformatie op leesniveau groep 8 te houden. De onderzoekers benadrukken daarom dat AI-chatbots kunnen helpen om medische informatie begrijpelijker te maken, maar dat de inhoud altijd gecontroleerd moet worden door experts voordat deze aan patiënten wordt aangeboden.

Eveneens onderstrepen Ratnayake en Wang (2023) dat prompt engineering essentieel is voor het optimaliseren van de prestaties van large language models (LLMs). Ze constateren dat standaard promptingstrategieën (zie bijlage C) vaak tekortschieten, met name wanneer de taak complex is of een hoge mate van nuance vereist. Om dit probleem te ondervangen ontwikkelden zij het *PERFECT-framework:* een modulair raamwerk dat gebruikers helpt om prompts stap voor stap op te bouwen. Het framework bestaat uit zeven bouwstenen:

1. **P**urpose – Bepaal het doel van de prompt;
2. **E**lement – Selecteer passende modulaire elementen;
3. **R**ole – Geef aan vanuit welk perspectief het model moet antwoorden;
4. **F**ormat – Definieer de gewenste vorm van het antwoord;
5. **E**xamples – Voeg indien nodig voorbeelden of casussen toe;
6. **C**onditions – Stel voorwaarden of beperkingen voor het antwoord vast;
7. **T**imeframe – Geef een tijdskader aan (bijv. historische of toekomstige context).

Het empirisch onderzoek van Ratnayake en Wang (2023) toont aan dat prompts die dit framework volgen aanzienlijk meer inhoudelijke diepgang en lengte genereren dan conventionele prompts. In een test met GPT-3.5 leidde het gebruik van dit framework tot gemiddeld 2100 woorden meer output. Het framework stimuleert bovendien het model om stapsgewijs te redeneren (de accuratesse steeg van 28% naar 36%) en expliciet verbanden te leggen, wat effectief is voor analytische of probleemoplossende taken. Een belangrijke nuance die uit de experimenten naar voren komt, is echter dat meer detail in de prompt niet altijd leidt tot betere prestaties. In sommige creatieve taken, zoals het schrijven van fictie, bleek een eenvoudige, open prompt langere output op te leveren dan een gespecificeerde prompt. Een andere beperking is dat het succes van een promptingstrategie deels afhankelijk is van het specifieke gedrag van het taalmodel, welke onderhevig kan zijn aan interne updates van de ontwikkelaars (zoals bij GPT-4), waardoor eerder werkende strategieën opeens minder effectief kunnen worden. Daarnaast geven de auteurs te kennen dat zij slechts twee datasets hebben gebruikt, waardoor de resultaten niet gegeneraliseerd kunnen worden (Ratnayake en Wang, 2023).

2.4.3 Beperkingen en risico's generatieve AI

Hoewel taalmodellen zoals ChatGPT veel potentie tonen als hulpmiddel bij het herschrijven van complexe medische teksten naar begrijpelijke patiëntinformatie, brengt het gebruik ervan ook duidelijke beperkingen en risico's met zich mee. Deze hebben zowel betrekking op de technische werking van het model als op de wijze waarop gebruikers, via prompt engineering, sturing geven aan de gegenereerde output.

Een eerste aandachtspunt betreft het reduceren van tekstinhoud. Uit het onderzoek van Swisher et al. (2024) blijkt dat herschreven teksten door ChatGPT gemiddeld 58 procent korter zijn dan de oorspronkelijke patiëntfolders. Hoewel deze reductie gepaard gaat met verbeterde leesbaarheidsscores, kan het ook leiden tot verlies van nuance, medische precisie en contextuele volledigheid. Korte teksten zijn immers niet per definitie begrijpelijker of betrouwbaarder en met name niet in medisch-educatieve contexten waarin volledigheid van informatie van groot belang is.

Daarnaast blijkt dat vereenvoudiging van tekst niet altijd gepaard gaat met passende woordkeuze. Swisher et al. (2024) bespreken bijvoorbeeld een zin uit een informatieve handout over sinusitis waarin staat: "Sinusitis can be a drag, making you feel blocked up and sore." De auteurs geven aan dat het woord *drag* ongeschikt en mogelijk verwarrend is binnen medische communicatie. Bovendien zou de zin, ondanks zijn toegankelijke toon, te veel samenvatten en daarmee belangrijke details verdoezelen. Dit onderstreept dat eenvoud in taalgebruik niet ten koste mag gaan van inhoudelijke helderheid of gepaste formulering.

Een ander relevant punt betreft dat veel van dit onderzoek is uitgevoerd met leesbaarheidsformules zoals Flesch-Kincaid en SMOG. Deze zich richten op zinslengte en woordkeuze, maar zeggen weinig over de daadwerkelijke begrijpelijkheid of medische correctheid van de inhoud. Dit betekent dat een tekst formeel wel eenvoudig kan zijn, maar inhoudelijk tekortschiet of zelfs incorrect kan zijn (Swisher et al. 2024)

Daarbij komt dat de output van een taalmodel sterk afhankelijk is van de input: de zogeheten prompt. Zoals besproken in paragraaf 2.4.2 bepaalt de kwaliteit van de prompt in hoge mate de kwaliteit van de gegenereerde tekst. Dit brengt risico's met zich mee, omdat effectieve prompt engineering specialistische kennis en digitale geletterdheid vereist. Bovendien is het gedrag van het taalmodel onderhevig aan regelmatige updates vanuit de ontwikkelaar (zoals bij GPT-4), wat tot gevolg heeft dat eerder succesvolle promptstrategieën plots minder effectief kunnen zijn (Ratnayake & Wang, 2023).

Verder laten onderzoeken van Eid et al. (2023) zien dat zonder specifieke instructies over het gewenste leesniveau de gegenereerde teksten vaak juist minder toegankelijk zijn dan de originele. Pas na expliciete aanwijzingen om bijvoorbeeld op niveau ‘groep 8’ te schrijven, leverde ChatGPT-4.0 teksten op die significant beter scoorden op diverse leesbaarheidsindices. Toch bleef het resultaat gemiddeld op het niveau van klas 4 havo/vwo, wat hoger is dan het door de American Medical Association aanbevolen niveau voor patiëntcommunicatie (groep 8). Deze bevinding bevestigt dat AI-gegenereerde teksten niet vanzelfsprekend geschikt zijn voor een breed publiek, en dat menselijk toezicht noodzakelijk blijft.

Tot slot onderstrepen Meskó en Topol (2023) een aantal fundamentele risico’s verbonden aan de inzet van LLMs binnen de medische context. Zij waarschuwen voor:

- **Misinformatie**: LLMs kunnen zogenoemde “hallucinaties” genereren: foutieve maar overtuigend klinkende informatie. In een medische context kan dit leiden tot onjuiste behandelingen.
- **Bias en ongelijkheid:** Wanneer de trainingsdata van een LLM bestaande vooroordelen of onevenwichtige trainingsdata bevat, kan dit leiden tot ongelijke behandeling van bijvoorbeeld etnische minderheden.
- **Gebrek aan transparantie**: LLMs zijn complexe ‘black box’-systemen zijn, is het vaak onduidelijk waarom ze tot bepaalde aanbevelingen of output komen. Dit maakt het lastig om de output te controleren of te verantwoorden, doordat de redeneerstructuur van het model niet inzichtelijk is.
- **Verlies van menselijke expertise**: Overmatige afhankelijkheid van AI kan ertoe leiden dat zorgverleners hun klinische vaardigheden verliezen of overschatten wat AI-systemen kunnen.
- **Privacyrisico’s**: De inzet van LLMs vraagt om strikte waarborgen rondom patiëntgegevens, zeker wanneer deze modellen gevoed worden met medische informatie.

Samengevat geldt dat taalmodellen zoals ChatGPT waardevolle ondersteuning kunnen bieden bij het vereenvoudigen van medische teksten, mits deze met de nodige deskundigheid worden ingezet. Prompt engineering en menselijke controle zijn daarbij onmisbaar. Alleen op die manier kan gewaarborgd worden dat AI-gestuurde herschrijving niet alleen begrijpelijk, maar ook betrouwbaar, gepast en ethisch verantwoord is.

2.5 Medische brochures: functie- en genrekenmerken

Om medische brochures verantwoord te analyseren, is het essentieel de begrippen genre en teksttype helder te definiëren. Deze termen worden in de literatuur vaak gebruikt, maar niet altijd eenduidig omschreven.

2.5.1 Genre en teksttype: een theoretische grondslag

Swales (1990) definieert een *genre* als een concept dat gedefinieerd kan worden aan de hand van drie kernideeën. De meest bepalende is de communicatieve intentie. Dit betekent dat het hoofddoel van de communicatie de belangrijkste factor is bij het bepalen van het genre waartoe een tekst behoort. Dit idee wordt eveneens onderschreven door andere auteurs (Berkenkotter & Huckin, 1995; Bhatia, 1993, 2004, 2010; Dudley-Evans, 1986, 1994; Kathpalia, 1992). Ten tweede wordt een genre als genre bestempeld wanneer het wordt geaccepteerd en erkend door experts uit een specifieke discoursegemeenschap, waarin eveneens dit genre is ontwikkeld. Tot slot kan een genre worden omschreven als een tekstsoort die zich onderscheidt door specifieke gemeenschappelijke kenmerken op het gebied van structuur, stijl en inhoud. Hatim en Mason (1990) definiëren een genre als ‘’conventionalised forms of texts which reflect the functions and goals involved in particular social occasions as well as the purposes of the participants in them’’ (p. 69). Deze definitie benadrukt dat genres niet willekeurig zijn, maar gestoeld op conventies die voortkomen uit sociale contexten en gedeelde doelen.

Deze zogenoemde *genreconventies* fungeren als signalen voor de lezer en helpen bij de interpretatie van een tekst (Hurtado, 2002, p. 477). Gamero (1998) en Alcaraz (2000) benadrukken dat herhaling van kenmerkende eigenschappen genres herkenbaar maakt voor zowel schrijver als lezer. Hatim en Mason (1990, p. 69) verbinden genres bovendien aan cognitieve processen: ze zijn gebaseerd op gedeelde kennis en communicatieve intenties. In later werk onderstrepen ze het dynamische karakter van genres, die evolueren over tijd, context en cultuur (Hatim & Mason, 1995, vii). Naast het concept *genre* wordt in de literatuur ook vaak het verwante begrip *teksttype* aangehaald. Hoewel beide begrippen soms door elkaar worden gebruikt, verwijzen ze naar verschillende classificaties van tekst. Waar een genre sterk contextueel en sociaal bepaald is, richt het begrip teksttype zich voornamelijk op de functionele en structurele kenmerken van teksten.

Hatim en Mason (1990, p. 140) bieden een duidelijke definitie van teksttype: ‘’a conceptual framework which enables us to classify texts in terms of communicative intentions serving an overall rhetorical purpose.’’ Zij onderscheiden drie kernelementen:
(a) de communicatieve intentie van de tekst;
(b) de mogelijkheid om teksten systematisch te classificeren;
(c) het overkoepelende retorische doel.

Op basis van het retorische doel kunnen drie type teksten worden onderscheiden: verklarende, betogende en instructieve teksten. De context is hierin bepalend, omdat deze de primaire functie van een tekst vaststelt en als bijgevolg de functie die het teksttype daarmee krijgt.

In later werk beschouwt Hatim (1997) teksttypen niet langer enkel vanuit een brede communicatieve intentie, maar ook vanuit linguïstisch perspectief. Hij gaat ervan uit dat de structuur en textuur (zoals lexicale en grammaticale keuzes) van een tekst specifiek worden aangepast om te reageren op de context waarin de tekst functioneert en op het teksttype dat weergegeven moet worden. Tegenwoordig bekend als register.

Het is waardevol om te benoemen dat de definities *tekstgenre* en *teksttype* door sommige auteurs door elkaar worden gebruikt en deze bezien als synoniemen van elkaar. Desondanks beschouwt de meerderheid van de auteurs het als van elkaar te onderscheiden termen (Ornia, 2016, p.12). Zo wijst Quijada (2008, p.189) erop dat beide concepten niet ver van elkaar af staan; ze delen bepaalde elementen als externe en interne tekstkenmerken, conventies, cultureel erfgoed en de contextuele situaties van gesprekspartners. Volgens Trosborg (1997, p.15) ligt het fundamentele verschil tussen beide concepten in hun theoretische basis. Genres worden ondersteund door externe criteria en niet door linguïstische kenmerken. Teksttypen daarentegen zijn gebaseerd op interne criteria en dus linguïstische aspecten.

2.5.2 Medische brochures als genre

Op basis van de vorige paragraaf kan de conclusie worden getrokken dat medische brochures een tekstgenre zijn. Medische brochures zijn een geconventionaliseerde vorm: ze omvatten de herhaling van linguïstische en structurele elementen, waardoor teksten herkenbaar en voorspelbaar zijn voor de lezer. Dit suggereert dat ze worden geaccepteerd door de discursieve gemeenschap waarin ze worden geproduceerd. Bovendien hebben ze een specifieke functie die de communicatieve intentie van de afzender weerspiegelt, namelijk om een niet-gespecialiseerd publiek te informeren over een medisch probleem.

Volgens Ornia (2016, p. 45) betreft een medische brochure een document van enkele pagina's dat de meest relevante informatie bevat over een bepaald medisch onderwerp en gericht is op een niet-expertpubliek, doorgaans bestaande uit patiënten of hun naasten. Volgens de auteur richt een medische brochure zich specifiek op lezers die niet gespecialiseerd zijn in medische onderwerpen. De lezers zijn meestal patiënten of hun familieleden die eenvoudige toegankelijke informatie en praktische adviezen zoeken over medische onderwerpen, risicosituaties of ziektes. Daartegenover staan de afzenders van medische brochures, die juist wel gespecialiseerd zijn (farmaceutische bedrijven, medici, organisaties gericht op onderzoek, preventie en behandeling van ziekten. Deze combinatie tussen gespecialiseerde afzenders en niet-gespecialiseerde ontvangers maakt medische brochures tot een bijzondere vorm van kennisoverdracht, waarbij complexe medische kennis begrijpelijk en toegankelijk gemaakt moet worden voor een breed publiek.

Volgens Mayor Serrano (2010, p.30) hebben medische brochures expliciete communicatieve functies, namelijk: op een begrijpelijke en onderhoudende manier medische basisinformatie verkondigen; aanbevelingen doen voor het voorkomen van ziekten, risicovolle situaties of hersteltoestanden en te proberen enige invloed uit te oefenen op het gedrag van de ontvanger. Hierin is het vooral van belang dat de informatie op een duidelijke, georganiseerde en beknopte manier wordt gepresenteerd, zodat ook mensen met lage gezondheidsvaardigheden de informatie begrijpen.

Ornia (2016) heeft gepoogd een typologie te bedenken voor het medische veld, met als doel: een duidelijk, algemeen en volledig kader te creëren waarin nieuwe medische genres gemakkelijk geclassificeerd kunnen worden om de teksttypen binnen deze genres te observeren en vergelijken. Voor het ontwerp heeft Ornia de voorstellen van Gamero (1998; 2001), Hurtado (2002) en Muñoz (2002) genomen. Hierin introduceert zij een flexibele benadering waarin tekstgenres niet statisch zijn, maar dynamisch van het ene naar het andere teksttype kunnen verschuiven naarmate hun communicatieve functie verandert. Binnen dit onderzoek is deze typologie toegepast op één representatieve brochure met als doel om erachter te komen wat de functie is van een medische brochure (zie tabel 1).

Functie	Onderwerp	Uitleg functie	Toon
Expositorische functie (verklarend)	Wat is hoofdpijn?	Deze vraag impliceert een neutrale uitleg van hoofdpijn als fenomeen (definitie, kenmerken, oorzaken, enz.). Concepten en feiten worden neutraal gepresenteerd zonder directe evaluatie of instructies.	Algemene communicatie gericht op niet-gespecialiseerde groepen.
Instructieve functie	Kan ik er zelf iets tegen doen? In welke gevallen kan ik beter naar de huisarts gaan?	Deze onderdelen richten zich op het geven van concrete aanwijzingen en instructies om het gedrag van de lezer in de toekomst te sturen of te beïnvloeden.	Algemene communicatie gericht op niet-gespecialiseerde groepen.
Argumentatieve functie	Welke medicijnen worden gebruikt bij hoofdpijn? Wat kan de apotheker voor mij doen?	Argumenten voor waarom bepaalde medicijnen worden aangeraden of juist worden afgeraden. Het doel van de apotheker kan argumentatief zijn als deze vertrouwen wil wekken in deskundigheid en het nut van het raadplegen van een apotheker wil verklaren.	Algemene communicatie gericht op niet-gespecialiseerde groepen.

Tabel 1 Functie medische brochure hoofdpijn

2.6 Juridisch verankering voor begrijpelijke communicatie

De noodzaak voor begrijpelijke communicatie in de gezondheidszorg, zoals beschreven in hoofdstuk 1, heeft de afgelopen jaren niet alleen geleid tot meer aandacht in onderzoek en beleid, maar is inmiddels ook vastgelegd in wet- en regelgeving. Daarmee wordt duidelijk dat begrijpelijke taal in de zorg niet langer een wenselijke, maar een vereiste norm is.
Deze ontwikkeling komt onder andere tot uiting in recente aanpassingen van bestaande wetten, waarin expliciet wordt vastgelegd dat medische informatie begrijpelijk moet zijn voor patiënten. Zo is op 1 januari 2020 de Wet op de geneeskundige behandelingsovereenkomst (WGBO) aangescherpt, met name op het punt van de informatieplicht van zorgverleners. In artikel 448 lid 1 staat:
''De hulpverlener licht de patiënt op duidelijke wijze, en desgevraagd schriftelijk in over het voorgenomen onderzoek en de voorgestelde behandeling en over de ontwikkelingen omtrent het onderzoek, de behandeling en de gezondheidstoestand van de patiënt" (Stb. 2019, 284).

Zorgverleners zijn verplicht om in begrijpelijke taal met patiënten te communiceren, zodat deze weloverwogen toestemming kunnen geven voor een medische behandeling. Dit principe geldt eveneens voor het verstrekken van heldere medische informatie op schrift. Heijmans, et al. (2018, p3). benadrukken het belang voor het aanscherpen van de wet, omdat in Nederland ruim een derde van alle volwassenen over beperkte gezondheidsvaardigheden (*health literacy)* beschikt. Dit betekent dat het voor deze kwetsbare groep een uitdaging is om informatie te begrijpen dan wel op te volgen, met als mogelijk gevolg dat medicatie op onjuiste wijze wordt gebruikt of adviezen niet worden opgevolgd. Volgens Heijmans, et al. (2018, p3) kunnen deze lage gezondheidsvaardigheden aangepakt worden door schriftelijke, mondelinge en digitale gezondheidsinformatie eenvoudiger en toegankelijker te maken. Ter juridische borging hiervan is per 26 mei 2021 de Europese verordening Medical Device Regulation (MDR) in werking getreden. Deze nieuwe Europese wet voor medische hulpmiddelen houdt onder andere in dat patiënten meer duidelijkheid moeten krijgen over medische hulpmiddelen, zodat zij gerichter een passend hulpmiddel kunnen kiezen (Ministerie van Volksgezondheid, Welzijn en Sport, 2024). Tot slot gaat op 28 juni 2025 de European Accessibility Act (EAA) van kracht. Hierin worden organisaties verplicht om digitale informatie op taalniveau B1 aan te bieden, zodat deze begrijpelijk is voor een breed publiek.

De analyse van het theoretisch kader in hoofdstuk 2 laat zien dat het begrip ‘begrijpelijkheid’ in medische brochures complex en meerduidig is: het wordt beïnvloed door tekstinterne kenmerken, lezerskenmerken en de context waarin de tekst functioneert. De bespreking van bestaande leesbaarheidsmodellen maakt duidelijk dat traditionele instrumenten vaak tekortschieten doordat ze vooral oppervlakkige tekstkenmerken meten en zelden empirisch zijn gevalideerd met echte lezers. De LiNT-formule biedt hierin een belangrijke vooruitgang, maar kent eveneens beperkingen, zoals het ontbreken van een onafhankelijke, kritische toetsing en het niet meenemen van alle aspecten van tekstbegrip. Daarnaast maakt de literatuur over AI-gestuurde herschrijving duidelijk dat taalmodellen zoals ChatGPT veel potentie hebben om teksten begrijpelijker te maken. Die potentie is echter sterk afhankelijk van hoe de AI wordt aangestuurd (bijvoorbeeld via de prompt) en of er menselijke controle plaatsvindt. Om te toetsen in hoeverre generatieve AI daadwerkelijk in staat is medische brochures begrijpelijker te formuleren dan menselijke redacteuren, is een systematische, empirische vergelijking noodzakelijk. In het volgende hoofdstuk wordt eerst de methodologische opzet van het onderzoek uiteengezet.

Hoofdstuk 3 Methode

In dit hoofdstuk wordt de methodologische opzet van het onderzoek uiteengezet. Het doel van dit onderzoek is tweeledig: enerzijds wordt onderzocht in hoeverre generatieve AI in staat is om medische brochures begrijpelijker te herformuleren dan menselijke redacteuren. Dit wordt gemeten aan de hand van objectieve linguïstische kenmerken volgens de LiNT-formule. Anderzijds omvat dit onderzoek een verklarende, kwalitatieve analyse van de linguïstische en structurele herschrijfstrategieën die AI enerzijds en menselijke redacteuren anderzijds hanteren. Doel van deze analyse is om systematisch in kaart te brengen waarom specifieke verschillen ontstaan in de herschrijfoutput, en hoe deze verschillen te duiden zijn vanuit theoretische kaders over begrijpelijkheid, herschrijfstrategieën en automatische tekstgeneratie.

Het hoofdstuk begint met een beschrijving van de onderzoeksopzet en de corpusselectie, gevolgd door een methodologische verantwoording van de gehanteerde prompts in de AI-conditie. Vervolgens wordt ingegaan op de toepassing van de LiNT-formule als meetinstrument voor tekstbegrijpelijkheid. Aansluitend worden de toegepaste statistische analysemethoden toegelicht, waaronder een CLMM voor het algemene moeilijkheidsniveau en LMM's voor de dertien afzonderlijke tekstkenmerken. Tot slot wordt de kwalitatieve analysemethode besproken, gericht op het interpreteren van de aard en functie van herschrijfkeuzes in de medische context.

3.1 Onderzoeksopzet

Dit onderzoek is opgezet als een experimentele vergelijkingsstudie, met als doelstelling om vast te stellen in hoeverre generatieve AI in staat is om medische brochures begrijpelijker te herschrijven dan menselijke redacteuren. De centrale onderzoeksvraag luidt:

In hoeverre is GenAI in staat medische brochures begrijpelijker te formuleren dan menselijke redacteuren, gemeten op basis van objectieve tekstuele kenmerken?

Om de onderzoeksvraag te beantwoorden, worden 25 originele medische brochures geanalyseerd, telkens in vier varianten: (1) de oorspronkelijke versie zoals verspreid via fysieke folders, (2) een herschreven versie door menselijke redacteurs (apothekers), en twee AI-gegenereerde versies op basis van verschillende prompts (3 en 4). Elke tekstvariant wordt geanalyseerd met de LiNT-formule, een wetenschappelijk gevalideerd instrument voor het

objectief meten van tekstuele begrijpelijkheid op dertien linguïstische kenmerken (zie paragraaf 2.3.2 en 3.4).

Dit onderzoek hanteert een mixed methods design, waarbij zowel kwantitatieve als kwalitatieve analysetechnieken worden ingezet. De kwantitatieve component richt zich op meetbare verschillen in tekstbegrijpelijkheid via de LiNT-formule. De kwalitatieve component onderzoekt de onderliggende linguïstische herschrijfstrategieën van AI en menselijke redacteuren, met als doel deze te verhelderen binnen bestaande theoretische kaders. De opzet maakt het mogelijk om alle vier tekstversies onderling te vergelijken, zowel op het globale moeilijkheidsniveau als op de afzonderlijke linguïstische dimensies die bijdragen aan tekstbegrip. Deze vergelijkingen worden uitgevoerd binnen een gepaarde onderzoeksdesign, waarbij elke brochure fungeert als eigen controlegeval. Door dit design wordt gecontroleerd voor inhoudelijke verschillen tussen brochures, zodat effecten van herschrijfwijze (mens of AI) zuiver en geïsoleerd geanalyseerd kunnen worden. Naast deze kwantitatieve benadering wordt aanvullend een kwalitatieve analyse uitgevoerd om inzicht te verkrijgen in de onderliggende linguïstische strategieën en herschrijfkeuzes van AI ten opzichte van menselijke redacteuren.

3.2 Corpusselectie

Het corpus van dit onderzoek is een verzameling voorlichtende medische teksten gericht aan patiënten die als leek kunnen worden beschouwd op het gebied van medische kennis. In de praktijk functioneren deze teksten met name als medische brochures. Kenmerkend voor dit corpus is dat de medische brochures:

medisch-inhoudelijk van aard zijn;
een voorlichtend en niet-overtuigend doel dienen;
bedoeld zijn voor een breed publiek inclusief mensen met beperkte taal- of gezondheidsvaardigheden;
vaak voorkomen in de vorm van folders, webteksten of informatiemateriaal in de zorg.

Het corpus van dit onderzoek beslaat 25 voorlichtende teksten die oorspronkelijk afkomstig zijn uit fysieke folders die verkrijgbaar waren in apotheken. Deze folders bevatten informatie voor patiënten over medische aandoeningen zoals klachten en ziektes. In tabel 2 volgt een overzicht van alle geanalyseerde originele teksten in dit onderzoek.[2] Het corpus is

[2] Patiëntenfolders of via: Patientenfolders – Apotheek Het Recept

samengesteld op basis van materiaal afkomstig uit drie fysieke apotheekvestigingen in Nederland (Wormerveer, Amsterdam en Zwolle). Volgens de betrokken apothekers gaat het om verouderde folders: het zijn de laatst overgebleven fysieke exemplaren die toen nog in de publieksrekken lagen. Hoewel sommige folders lokaal nog wel verspreid worden, worden ze in de praktijk niet meer actief gebruikt voor voorlichting. De meeste teksten zijn voor het laatst herzien in 2011 of 2012, met uitzondering van twee folders (over worminfecties en koortslip) die zijn bijgewerkt in 2018.

1. **Acne -** september 2012	**14. Maagklachten –** augustus 2011
2. **Aambeien -** augustus 2011	**15. Menstruatiepijn –** september 2012
3. **Astma en COPD -** april 2011	**16. Osteoporose –** maart 2011
4. **Diabetes mellitus -** september 2012	**17. Reisziekte –** september 2011
5. **Diarree -** april 2013	**18. Reumatische aandoening –** april 2012
6. **Eczeem -** januari 2012	**19. Spier- en gewrichtspijn** – april 2013
7. **Hoest –** maart 2009	**20. Vaginale schimmelinfectie** – mei 2012
8. **Hoofdluis** – december 2012	**21. Verkoudheid en griep –** augustus 2011
9. **Hoofdpijn –** september 2012	**22. Verstopping –** april 2011
10. **Hoofdroos –** april 2013	**23. Voetschimmel –** juli 2012
11. **Hooikoorts en allergie –** juli 2012	**24. Wormen –** augustus 2018
12. **Kanker –** september 2011	**25. Wratten** – maart 2011
13. **Koortslip** – juni 2018	

Tabel 2 Geanalyseerde originele folders op alfabetische volgorde

Om de volledigheid en betrouwbaarheid van het corpus te waarborgen, zijn alle fysiek beschikbare teksten opgenomen die voldeden aan de inhoudelijke focus van dit onderzoek (klachten- en ziekteteksten voor lekenpubliek). Het corpus is aanvullend verrijkt met archiefmateriaal afkomstig van websites van lokale apotheken, voor zover dit inhoudelijk aansloot bij de bestaande folders. Dit heeft geleid tot een corpus van in totaal 25 unieke medische voorlichtingsteksten, die representatief zijn voor laagdrempelige publieksvoorlichting binnen de eerstelijnszorg.

Nike Everaarts, teamleider Patiënteninformatie bij KNMP en tevens apotheker, heeft schriftelijk bevestigd dat de teksten op de website van de Apotheek volledig zijn herschreven door mensen (apothekers). Deze apothekers hebben taaltrainingen ‘Schrijven op B1-niveau’

gevolgd bij een gerenommeerd taalbureau en gebruikten eveneens een tool[3] die het taalniveau heeft gecontroleerd. De klacht- en ziekteteksten zijn volgens Everaarts (2025) een samenstelling van teksten van de website van thuisarts.nl en van apotheek.nl en zijn beide louter door mensen geschreven. De focus van de herschrijving door mensen lag op het toegankelijk maken van medische informatie voor een breed publiek, met specifieke aandacht voor begrijpelijk taalgebruik.

Het corpus bevat de originele versie zoals eerder fysiek verspreid, de aangepaste vorm die online is gepubliceerd en herschreven door mensen en wordt aangevuld met twee extra varianten geschreven door AI. Voor elke brochure worden in totaal vier versies geanalyseerd:

1. **Originele versie**: zoals gepubliceerd door KNMP in een fysieke folder (zie tabel 4).
2. **Menselijke herschreven versie**: door experts (apothekers) herschreven op B1-niveau met het doel de begrijpelijkheid te verbeteren.[4]
3. **AI-versie 1:** gegenereerd door een Large Language Model (ChatGPT) op basis van een uniforme prompt gericht op vereenvoudiging van de tekst.
4. **AI- versie 2:** gegenereerd door een Large Language Model (ChatGPT) op basis van een verbeterde uniforme prompt gericht op vereenvoudiging van de tekst. Zie paragraaf 3.3 voor een verantwoording van deze twee prompts.

Voor de analyse is telkens slechts één deel van de folder geselecteerd, namelijk de eerste paragraaf onder de kop *"Wat is ziekte/klacht X"*. De keuze voor deze specifieke sectie is bewust gemaakt op basis van drie overwegingen. Ten eerste omdat deze paragraaf in begrijpelijke taal moet verhelderen wat kenmerkend is voor een bepaald ziektebeeld en/of klacht voor een patiënt. Ten tweede is deze paragraaf in vrijwel identieke vorm overgenomen op de website van de apotheek, waardoor een directe vergelijking mogelijk is tussen de oorspronkelijke foldertekst en de herschreven digitale versie. Tot slot, is ervoor gekozen om de tekstfragmenten te beperken tot een lengte van circa tweehonderd tot vijfhonderd woorden, conform de aanbevelingen uit de documentatie van LiNT. Kortere teksten zorgen vergroten de betrouwbaarheid van de analyse. Bij langere teksten kunnen verschillen in moeilijkheidsgraad binnen eenzelfde tekst worden afgevlakt door een gemiddelde, waardoor scanresultaten minder representatief zijn.

[3] De tool en de naam van het taalbureau is gedeeld via de e-mail, maar mag vanwege vertrouwelijke aard hier niet worden gedeeld.
[4] Klachten & Ziektes | Apotheek.nl = herschreven teksten door menselijke redacteuren

3.3 Formulering en verantwoording van AI-prompts

Naast de oorspronkelijke teksten zijn ook de herschreven versies door menselijke redacteurs al reeds beschikbaar. Deze zijn door apothekers herschreven op B1-niveau op en gepubliceerd op de website van de apotheek. Deze teksten dienen in dit onderzoek als vergelijkingsmateriaal voor de AI-gegenereerde versies. Voor de AI-output zijn twee verschillende prompts opgesteld. Deze keuze is gemaakt om te onderzoeken in hoeverre de begrijpelijkheid van AI-gegenereerde teksten beïnvloed wordt door de formulering van de gebruikersinstructie. De inzet van twee varianten maakt het mogelijk om de gevoeligheid van generatieve AI voor variatie in promptspecificiteit systematisch in kaart te brengen. Er is bewust gekozen voor een algemeen prompt en een specifieker prompt. Hiermee kan in kaart worden gebracht in hoeverre de outputkwaliteit afhankelijk is van de specificiteit en precisie van de gebruikersinstructie en sluit aan bij eerdere bevindingen in paragraaf 2.4.

Prompt 1: Herschrijf deze tekst zodat de tekst begrijpelijker wordt voor lezers, vereenvoudig de tekst naar B1-niveau (algemeen prompt)

Prompt 1 is zorgvuldig opgebouwd en voldoet grotendeels aan de principes van effectieve promptengineering zoals beschreven in hoofdstuk 2. Ten eerste bevat de formulering een duidelijk doel: het herschrijven van een tekst met als expliciete instructie om de begrijpelijkheid te vergroten. Door te verwijzen naar het B1-niveau van het Europees Referentiekader voor Talen (CEFR) wordt bovendien een concrete beperking opgelegd (negatieve prompt), die richting geeft aan het taalgebruik, de zinsstructuur en de woordenschat die in de output verwacht mag worden. Tot slot veronderstelt de prompt geen specifieke vorm (bijvoorbeeld een opsomming of paragrafen), maar de opdracht "herschrijf deze tekst", wat impliceert dat het model een volledige, herschreven tekst als output moet genereren in doorlopende vorm.

In de prompt wordt bewust B1-niveau benoemd, omdat ook aan menselijke redacteuren in dit geval de apothekers voor een breed lezerspubliek, conform taalniveau B1 (CEFR). Bij de formulering van het prompt is ervan uitgegaan dat ook niet-expertgebruikers, zoals leken zonder diepgaande kennis van generatieve AI, met deze prompt kunnen werken. Daarom wordt er middels de term ‘B1-niveau’ impliciet verwezen naar het Europees Referentiekader voor Talen (CEFR), maar wordt deze niet voluit gespecificeerd in de prompttekst zelf. Deze keuze is bewust gemaakt om de prompt compact te houden en representatief voor hoe gebruikers doorgaans taalniveaus aanduiden. In Prompt 2 is dit

explicieter geoperationaliseerd door meerdere specificaties toe te voegen. De instructie is kort, bondig en vrij van vakjargon, wat aansluit bij hoe mensen in de praktijk vaak op een natuurlijke manier vragen stellen aan een taalmodel. Daarmee weerspiegelt de prompt een realistische interactiewijze, zoals die ook buiten specialistische contexten voorkomt. Juist doordat de formulering dicht bij het spontane taalgebruik van alledaagse gebruikers blijft, vergroot dit de kans op correcte interpretatie en bruikbare output, ook zonder kennis van onderliggende AI-processen.

Prompt 2: Herschrijf deze tekst zodat deze begrijpelijker is voor een breed publiek. Gebruik taal op B1-niveau, vermijd jargon, en zorg voor duidelijke, korte zinnen. Behoud de kernboodschap van de originele tekst.

Prompt 2 is tot stand gekomen met behulp van een prompt generator, zoals beschreven in paragraaf 2.4.2, waarin deze werkwijze wordt gepresenteerd als een efficiënte strategie om gerichte en kwalitatieve output te verkrijgen bij het gebruik van generatieve AI. Als uitgangspunt werd de oorspronkelijke prompt ingevoerd, met de instructie aan de prompt engineer van ChatGPT: *"Improve my prompt: Herschrijf deze tekst zodat de tekst begrijpelijker wordt voor lezers, vereenvoudig de tekst naar B1-niveau."*

De prompt engineer beoordeelde zowel de oorspronkelijke als de gegenereerde prompt aan de hand van een scoresysteem. Uit deze beoordeling bleek dat de tweede prompt, gegenereerd na iteratie en verfijning, duidelijk hoger scoorde op effectiviteit en precisie. Deze bevindingen uit de promptgenerator geven aanleiding om te veronderstellen dat zorgvuldig geoptimaliseerde prompts kunnen bijdragen aan effectievere AI-output. In dit onderzoek worden beide prompts als afzonderlijke varianten meegenomen om te toetsen in hoeverre de mate van promptspecificiteit daadwerkelijk invloed heeft op de begrijpelijkheid van de gegenereerde teksten.

📌 **Beoordeling originele prompt:**

"Herschrijf deze tekst zodat de tekst begrijpelijker wordt voor lezers, vereenvoudig de tekst naar B1-niveau"

Categorie	Beoordeling (1–10)	Toelichting
Duidelijkheid	6	De opdracht is globaal duidelijk, maar het doelpubliek en wat "begrijpelijker" betekent is vaag.
Specifiekheid	5	"Begrijpelijker" en "B1-niveau" zijn niet nader gespecificeerd qua stijl of structuur.
Contextgerichtheid	4	Geen aanwijzingen over het type tekst of wat behouden moet blijven (zoals toon of boodschap).
Resultaatsturing	5	Er is geen expliciete richtlijn voor zinslengte, jargon of opbouw.

▶️ **Eindcijfer: 5/10** – Basisinstructie is aanwezig, maar mist precisie en context.

📌 **Beoordeling verbeterde prompt:**

"Herschrijf deze tekst zodat deze begrijpelijker is voor een breed publiek. Gebruik taal op B1-niveau, vermijd jargon, en zorg voor duidelijke, korte zinnen. Behoud de kernboodschap van de originele tekst."

Categorie	Beoordeling (1–10)	Toelichting
Duidelijkheid	9	De opdracht is helder en goed geformuleerd.
Specifiekheid	8	Er wordt uitgelegd wat met 'begrijpelijk' bedoeld wordt.
Contextgerichtheid	7	Er is basiscontext over doelgroep en toon, al kan die nog specifieker als gewenst.
Resultaatsturing	9	Duidelijke richtlijnen voor taalgebruik en behoud van inhoud.

▶️ **Eindcijfer: 8.5/10** – Een duidelijke, bruikbare prompt die consistentere en betere output zal genereren.

Daarnaast is het relevant om te vermelden dat de prompts zijn ingevoerd in de betaalde versie van ChatGPT, gebaseerd op het GPT-4o-model. Dit is relevant aangezien de prestaties en functionaliteiten van deze versie kunnen afwijken van die van de gratis beschikbare modellen. GPT-4o biedt onder meer een hogere verwerkingscapaciteit, verbeterde contextgevoeligheid en toegang tot geavanceerdere functies, wat van invloed kan zijn op de kwaliteit en consistentie van de gegenereerde output. Deze keuze waarborgt een maximale benutting van de mogelijkheden van generatieve AI binnen dit onderzoek.

3.4 Analysekader: toepassing van de LiNT-formule

Voor de beoordeling van de begrijpelijkheid wordt uitsluitend de LiNT-formule toegepast. Deze meetmethode toetst op dertien linguïstische kenmerken. Hieronder volgt de lijst van deze dertien kenmerken. In bijlage D is in een uitgebreide uitleg opgenomen hoe een kenmerk volgens Pander Maat en Dekker (2016) wordt gemeten en hoe het bijdraagt aan het meten van begrijpelijkheid.

Hoe moeilijk zijn uw woorden?

(hoe lager de score op deze kenmerken, des te kleiner de kans dat lezers moeite zullen ervaren met de tekst):

- Onbekende woorden;
- Abstracte zelfstandige naamwoorden;
- Woorden die nieuw zijn in de tekst;

Hoe moeilijk zijn uw zinnen?

(hoe lager de score op deze kenmerken, des te kleiner de kans dat lezers moeite zullen ervaren met de tekst):

- Lengte van deelzinnen;
- Afhankelijkheidslengtes;
- Bijvoeglijke bepalingen;
- Opsommingen in doorlopende zinnen;
- Bijzinnen;
- Zinslengte;

Hoe persoonlijk is uw tekst?

(hoe hoger de score op deze kenmerken, des te kleiner de kans dat lezers moeite zullen ervaren met de tekst).

- Mensen
- Persoonlijke voornaamwoorden
- Aansprekingen
- Actieve werkwoordsvormen

De LiNT-score van elke tekstversie (origineel, menselijk herschreven, AI prompt I en AI-prompt 2) wordt afzonderlijk berekend. Per kenmerk worden de scores statistisch vergeleken. Ten eerste beoordeelt LiNT de moeilijkheidsgraad van teksten globaal op een schaal van 1 tot 100, waarbij een lagere score wijst op een eenvoudiger tekst en een hogere score op een complexere. Deze moeilijkheidsscore wordt visueel weergegeven met behulp van een schuifbalkgrafiek. Hierdoor is in één oogopslag zichtbaar op welk leesniveau een tekst zich bevindt. Daarnaast biedt LiNT een gedetailleerde uitsplitsing van afzonderlijke tekstkenmerken die bijdragen aan de totale moeilijkheidsgraad. Elk kenmerk (zoals woordfrequentie of abstractheid van zelfstandige naamwoorden) wordt afzonderlijk gekwantificeerd en eveneens visueel gepresenteerd in grafieken. Deze weergave maakt het mogelijk om verschillende tekstvarianten onderling te vergelijken en gericht te analyseren op specifieke aspecten van tekstcomplexiteit.

Om de leesbaarheid van teksten automatisch en betrouwbaar te analyseren met LiNT, is het van belang dat de teksten vooraf goed worden voorbereid. In deze studie zijn de richtlijnen

van LiNT nauwkeurig gevolgd: Voorafgaand aan de analyse zijn de teksten voorbereid volgens de richtlijnen van LiNT:

- Bestandsnamen zijn kort gehouden en spelfouten gecorrigeerd om fouten in de verwerking te voorkomen;
- Kopjes en vaste tekstonderdelen zijn gemarkeerd met ###. Dit voorkomt dat LiNT korte kopjes als losse zinnen ziet, wat de leesbaarheidscijfers zou vertekenen.
- Zinnen zijn consequent afgesloten met leestekens (punt, vraagteken, uitroepteken) om zinsherkenning te waarborgen.
- Opsommingen zijn alleen opgenomen indien zij bestonden uit volledige zinnen met een werkwoord; overige opsommingselementen zijn genegeerd om vertekening van de moeilijkheidsscore te voorkomen.

3.5 Statistische analyse van leesniveaus en tekstkenmerken

In dit onderzoek zijn zowel kwantitatieve als kwalitatieve analysemethoden toegepast. In deze paragraaf wordt de kwantitatieve analysemethode beschreven en in de volgende paragraaf de kwalitatieve analysemethode. De data waarmee is gewerkt is opgenomen in bijlage E.

Om te bepalen of de vier tekstversies (origineel, herschreven door mensen, AI prompt I en AI prompt II) van elkaar verschillen op het gebied van begrijpelijkheid, zijn statistische toetsen uitgevoerd. Daarbij is allereerst gekeken naar het algemene moeilijkheidsniveau van teksten en vervolgens naar de dertien kenmerken van de LiNT-formule. Daarbij worden twee typen uitkomsten onderscheiden:

- **Ordinale uitkomsten:** het algemene moeilijkheidsniveau van de tekst, weergegeven in vier leesniveaus.
- **Schaaluitkomsten:** de afzonderlijke scores op dertien linguïstische kenmerken, weergegeven als continue variabelen.

Bij beide analysetypen wordt gebruikgemaakt van statistische modellen die rekening houden met clustering van data binnen dezelfde brochures (ook wel *nesting* genoemd). Dit wordt statisch gemodelleerd via een *random intercept* per brochure, ook wel aangeduid als een block-effect. Op deze manier wordt gecontroleerd op onderlinge verschillen tussen de brochureonderwerpen, zodat de effecten van de tekstversies zuiverder worden geanalyseerd.

3.5.1 Analyse van algemeen moeilijkheidsniveau (CLMM)

De LiNT-formule kent elke tekst een moeilijkheidsscore toe op een schaal van 0 tot 100. Deze score is vervolgens ingedeeld in vier niveaus van begrijpelijkheid. Deze score wordt in categorieën weergegeven:

- **Niveau 1 (< 34)**: streefniveau (ongeveer 15% van de lezers ervaart moeite)
- **Niveau 2 (34–46)**: next best (±30% ervaart moeite)
- **Niveau 3 (46–60)**: moeilijk (±55% ervaart moeite)
- **Niveau 4 (> 60)**: extreem moeilijk (±82% ervaart moeite)

Omdat deze niveaus ordinaal zijn (ze hebben een volgorde, maar de verschillen zijn niet gelijk verdeeld), is een *Cumulative Link Mixed Model* (CLMM) gebruikt. Dit model maakt het mogelijk om verschillen tussen de vier tekstversies te analyseren, terwijl er tegelijk rekening wordt gehouden met het feit dat dezelfde brochures meerdere keren in het onderzoek voorkomen. De brochures zijn daarom opgenomen als *random effect*.

Om te bepalen tussen welke tekstversies (origineel, menselijk herschreven, AI prompt I en AI prompt II) het verschil in moeilijkheidsniveau precies zit, zijn post-hoc toetsen uitgevoerd. Daarbij worden de tekstversies per brochure onderling vergeleken. Omdat het hier gaat om 25 brochures die elk vier varianten kent, worden er veel vergelijkingen gemaakt. Hoe meer toetsen er worden uitgevoerd, hoe groter de kans op toevallig significante resultaten. Om deze kans te beperken, is gebruikgemaakt van een *Holm-correctie*. De Holm-correctie is een methode die stap voor stap de grenswaarde voor wat als ‘statistisch significant’ geldt aanpast. Dit betekent dat de kans op het onterecht vinden van een verschil kleiner wordt, zonder dat de toets te streng wordt voor echte verschillen. Deze correctie is daarmee betrouwbaar en geschikt voor meerdere vergelijkingen tussen tekstvarianten binnen één dataset.

3.5.2 Analyse van afzonderlijke tekstkenmerken (LMM)

Voor de dertien kenmerken van de LiNT-formule, die elk een continue (schaal)score opleveren, is per kenmerk een *Linear Mixed Model* (LMM) uitgevoerd. Het model is geschikt voor getalsmatige uitkomsten (zoals gemiddelde zinslengte, of het aantal onbekende woorden) en houdt net als CLMM rekening met clustering binnen dezelfde brochures. Het model maakt onderscheid tussen twee soorten factoren:

- Vaste effecten (*fixed effects*): dit zijn de variabelen waarvan het effect wordt onderzocht. In dit onderzoek is dat de tekstversie (origineel, menselijk herschreven, AI prompt I, AI prompt II).
- Willekeurige effecten (*random effects*): dit zijn variabelen die niet centraal staan in het onderzoek, maar wel invloed kunnen hebben. Hier betreft dat de brochures zelf, die qua inhoud en stijl van elkaar kunnen verschillen.

Binnen dit model:

- is de **tekstversie** opgenomen als vaste factor (fixed effect);
- is de **LiNT-score** per kenmerk de afhankelijke variabele;
- en is de **brochure** opgenomen als random intercept (block effect).

Door de brochures als *random intercept* op te nemen, corrigeert het model voor deze verschillen, zodat betrouwbaar vergeleken kan worden wat het effect is van de herschrijfwijze. Na het uitvoeren van het model wordt gekeken of er verschillen zijn tussen de tekstversies. Als dat zo is, worden post-hoc toetsen uitgevoerd. Die vergelijken de versies twee aan twee (bijv. AI 1 met AI2, mens met origineel). Omdat er meerdere vergelijkingen worden gedaan, is de kans groter dat er toevallig een verschil lijkt te zijn. Om dit te voorkomen, wordt een False Discovery Rate (FDR) toegepast. Deze methode helpt om vals-positieve resultaten te beperken, terwijl het model gevoelig blijft voor echte verschillen. De FDR-correctie voorkomt daarmee dat toevallige verschillen als ‘betekenisvol’ worden geïnterpreteerd bij veel vergelijkingen.

3.6 Kwalitatieve vergelijking: AI versus menselijke redacteuren

Deze analyse beoogt niet enkel het identificeren van oppervlakkige verschillen in formulering, maar richt zich op het verklaren van onderliggende keuzes. Daarom wordt in dit onderzoek naast een kwantitatieve analyse ook een aanvullende kwalitatieve analyse uitgevoerd, met als doel om de linguïstische en structurele herschrijfstrategieën van AI en menselijke redacteuren systematisch met elkaar te vergelijken.

De kwalitatieve vergelijking focust op de herschreven paragrafen 'Wat is klacht/ziekte X’ en analyseert op microniveau welke aanpassingen AI doorgaans genereert ten opzichte van menselijke redacteuren. Daarbij wordt onder meer gekeken naar de moeilijkheidsgraad van woorden en zinnen en de mate van persoonlijkheid van de tekst (zie paragraaf 3.4).

Deze verschillen worden vervolgens functioneel geïnterpreteerd aan de hand van bestaande theoretische kaders over linguïstische en structurele herschrijfstrategieën zoals die door menselijke redacteuren en AI worden toegepast bij tekstvereenvoudiging. Doel daarbij is om te analyseren welke impliciete normatieve principes of herschrijfstrategieën generatieve AI lijkt te hanteren bij het optimaliseren van teksten, en in hoeverre deze strategieën overeenkomen met of afwijken van conventionele herschrijfstrategieën door menselijke taalgebruikers. Door deze verklarende invalshoek draagt bij aan een verdiepend inzicht in de wijze waarop AI-gegenereerde herschrijvingen functioneren binnen het bredere veld van begrijpelijkheidsonderzoek. Zo wordt zichtbaar in hoeverre AI begrijpelijker schrijft dan menselijke redacteuren, en welke linguïstische keuzes daaraan ten grondslag liggen.

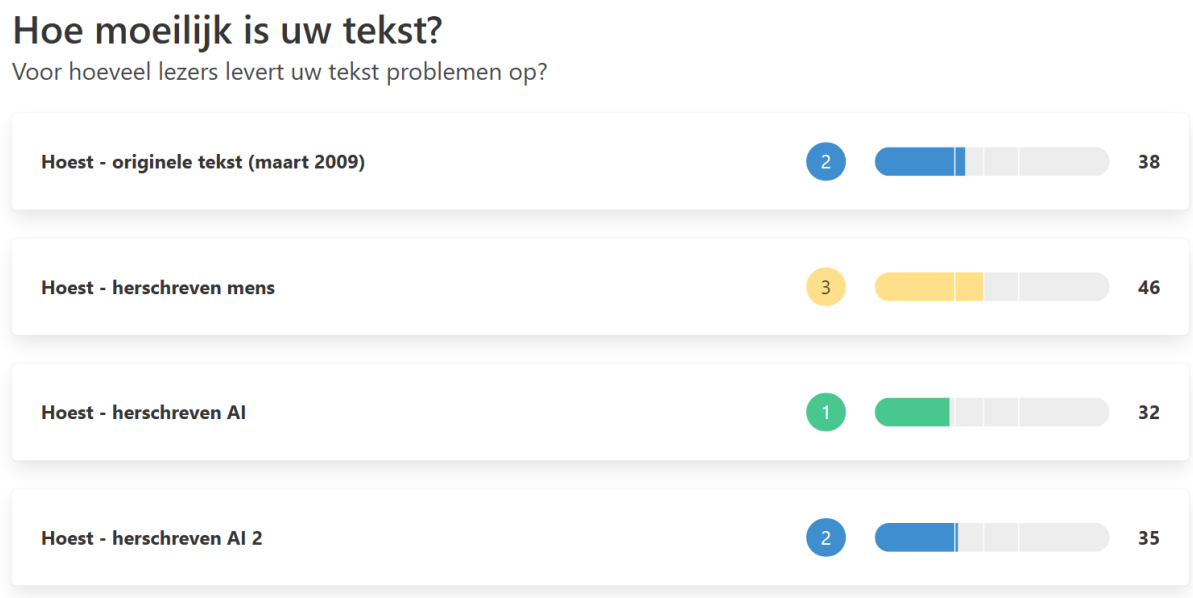
In dit hoofdstuk is toegelicht hoe het onderzoek is opgezet om verschillen in begrijpelijkheid tussen menselijke en AI-herschrijvingen van medische brochures te analyseren. Door vier tekstvarianten per brochure systematisch met elkaar te vergelijken, wordt onderzocht in hoeverre generatieve AI begrijpelijke teksten kan produceren. De LiNT-formule biedt daarbij een objectieve manier om de leesbaarheid van elke tekstversie te meten op dertien taalkenmerken. Daarnaast geeft de aanvullende kwalitatieve analyse meer inzicht in hoe en waarom AI andere keuzes maakt dan menselijke redacteuren. Met deze gecombineerde aanpak wordt zowel gemeten wat er verandert in de tekst, als verklaard uit theoretische inzichten waarom dat mogelijkerwijs gebeurt. In hoofdstuk 4 worden de resultaten van deze analyses gepresenteerd.

Hoofdstuk 4 Resultaten

In dit hoofdstuk worden de resultaten gepresenteerd van de kwantitatieve analyse die is uitgevoerd op de vier tekstversies van de medische brochures: de originele versie, de menselijke herschrijving en de twee AI-herschrijvingen. De analyse is gebaseerd op de LiNT-formule, die inzicht biedt in het algemene moeilijkheidsniveau van een tekst en in dertien afzonderlijke linguïstische kenmerken die samenhangen met begrijpelijkheid. De resultaten worden eerst besproken op globaal niveau (leesniveau), gevolgd door een gedetailleerde vergelijking van de dertien tekstkenmerken. Verschillen tussen de tekstversies worden gerapporteerd op basis van de uitkomsten van het CLMM en de LMM's, inclusief relevante post-hoc toetsen. De laatste paragraaf biedt een aanvullende kwalitatieve analyse van structurele en stilistische verschillen tussen de AI-herschrijvingen en de menselijke herschrijvingen, waarin voorbeelden worden geanalyseerd.

4.1 Begrijpelijkheid op vier leesniveaus (CLMM)

Figuur 3 laat zien in welke mate een medische brochure over hoesten begrijpelijk is voor lezers, aan de hand van vier verschillende tekstversies: een versie herschreven door AI, een versie herschreven door AI met een specifieke prompt (AI II), een versie herschreven door een mens, en de originele brontekst. Met behulp van de LiNT-formule wordt per variant weergegeven hoeveel lezers waarschijnlijk moeite hebben met het begrijpen van de tekst: hoe lager het getal, hoe beter de tekst te begrijpen is voor het grootste deel van het publiek.



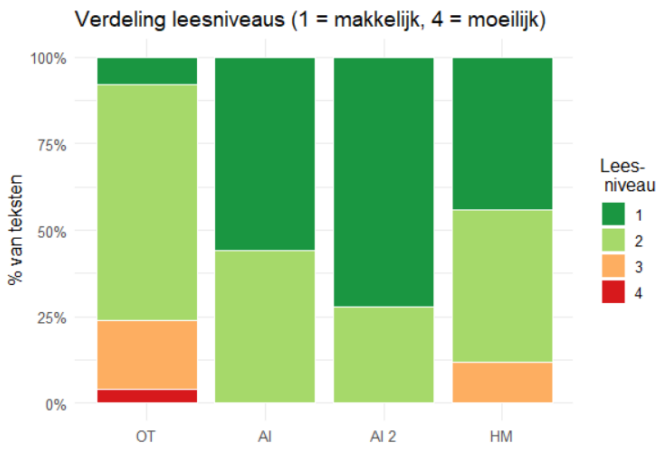
Figuur 3 Overzicht van de begrijpelijkheid een medische brochure in vier varianten gemeten met de LiNT-formule

Om te onderzoeken hoe begrijpelijk de teksten waren, is een statistisch model gebruikt dat de kans berekent dat een tekst op een bepaald *leesniveau* wordt ingedeeld. Hier is bewust voor gekozen. De LiNT-formule levert een leesniveau op dat geordend is in vier categorieën, waarbij elk hoger niveau een moeilijker tekst aanduidt: niveau 1 staat voor teksten die goed leesbaar zijn voor vrijwel iedereen (±15% ervaart moeite); terwijl niveau 4 staat voor teksten die als zeer moeilijk worden ingeschat (±82% ervaart moeite). Dit betekent dat de uitkomstvariabele ordinaal is: de niveaus hebben een natuurlijke volgorde, maar de afstanden tussen de niveaus zijn niet noodzakelijk gelijk. Een directe vergelijking van LiNT-scores zou alleen zinvol zijn als het om een continue schaal zou gaan (bijvoorbeeld een gemiddelde LiNT-score per tekst). In dit onderzoek zijn de uitkomsten echter discrete categorieën: de variabele kan, in tegenstelling tot een continue variabele, een beperkt aantal afzonderlijke waarden aannemen (leesniveaus). Om deze reden is een ordinaal model zoals een CLMM beter geschikt voor deze data. Het voordeel is dat CLMM de volledige spreiding over de niveaus benut, in plaats van de data te reduceren tot één gemiddelde score. Voor dit onderzoek zijn 25 medische voorlichtingsbrochures geselecteerd. Elke brochure is in vier verschillende varianten opgenomen: (1) automatisch herschreven door AI met een algemene prompt (AI I), (2) automatisch herschreven door AI met een specifieke prompt gericht op begrijpelijkheid (AI II), (3) herschreven door een mens, en (4) de originele, niet-herschreven tekst. In totaal zijn er dus 100 tekstversies geanalyseerd (25 brochures × 4 varianten).

Leesniveau	AI- I	AI – II	Herschreven Mens	Originele tekst
Niveau 1	48,5%	**63,1%**	39,5%	**10,9%**
Niveau 2	35,8%	27,6%	33,7%	17,1%
Niveau 3	12,5%	7,1%	19,3%	30,0%
Niveau 4	3,2%	**2,2%**	7,5%	**42,0%**

Tabel 3 Algemeen leesniveau (CLMM)

De gegevens uit tabel 5 zijn in de visualisatie hiernaast overzichtelijk weergegeven, zodat in één oogopslag zichtbaar is hoe de leesniveaus per tekstvariant verdeeld zijn (figuur 4).



Figuur 4 Algemeen leesniveau - LiNT

Uit de analyse blijkt dat de manier waarop een tekst is herschreven, grote invloed heeft op het leesniveau. De LiNT-formule deelt teksten in vier moeilijkheidsniveaus in, waarbij niveau 1 het meest toegankelijk is en niveau 4 het moeilijkst. De originele teksten scoren het laagst op begrijpelijkheid: slechts 10,9 procent van deze teksten wordt als niveau 1 beoordeeld, terwijl maar liefst 42 procent in niveau 4 valt. Dit betekent dat een groot deel van de lezers moeite zal hebben met het begrijpen van de oorspronkelijke brochures.

VOORBEELDZIN UIT BROCHURE SPIER- EN GEWRICHTSPIJN PER CONDITIE:		KORTE TOELICHTING
ORIGINELE TEKST	Door het langdurig of overmatig belasten van spieren, kan er in de spier een ophoping ontstaan van afvalstoffen, vooral melkzuur (denk maar aan het 'verzuren' van de benen bij een sporter).	Lange, complexe zin met jargon, abstracte begrippen, passieve vorm en meerdere bijzinnen.
HERSCHREVEN MENS	Door het langdurig of overmatig belasten van spieren kan er een ophoping ontstaan van afvalstoffen, met name melkzuur.	Korter, minder jargon, iets eenvoudiger opgebouwd, nog steeds abstract en deels passief.
AI – I	Als u uw spieren lang of extra veel gebruikt, hopen afvalstoffen zoals melkzuur zich op in uw spieren.	Kort, actief geschreven en er is een (persoonlijke) aanspreking toegevoegd ‘u’.
AI – II	Spieren kunnen pijn gaan doen als u ze te veel of te lang gebruikt. In de spier hopen dan afvalstoffen op, zoals melkzuur.	Kort, helder, persoonlijk (aanspreking ‘u’), gesplitst in twee eenvoudige zinnen, actief.

Tabel 4 Voorbeeldzinnen brochure spierpijn toegespitst op alle varianten

Wanneer de teksten herschreven worden door mensen neemt de begrijpelijkheid toe. Ongeveer 39,5 procent van de menselijk herschreven teksten wordt ingedeeld als niveau 1 en slechts 7,5 procent als niveau 4. Dit laat zien dat herschrijvingen door menselijke experts al een duidelijke verbetering oplevert ten opzichte van het origineel. De meest toegankelijk geschreven teksten worden bereikt door middel van AI-herschrijvingen en dan met name wanneer er een specifieke en doelgerichte prompt wordt gebruikt, zoals AI-II. In deze conditie wordt 63,1 procent van de teksten beoordeeld als niveau 1. Slechts 2,2 procent als niveau 4. Dat betekent dat teksten die door AI met een specifieke prompt zijn herschreven, de grootste kans hebben om als zeer toegankelijk te worden beoordeeld en de kleinste kans hebben om als zeer moeilijk te worden ingeschat. Ook de herschrijvingen met een wat algemener prompt

(AI-I) scoren goed: 48,5 procent op niveau 1 en 3,2 procent op niveau 4. Dit geeft aan dat AI, zelfs met een algemene prompt, doorgaans beter presteert dan menselijke herschrijvers en daarmee de begrijpelijkheid van teksten aanzienlijk kan vergroten.

Samenvattend laat de analyse zien dat AI-herschrijvingen, vooral met een specifieke prompt, het meest effectief zijn in het verhogen van de begrijpelijkheid van medische brochures. Menselijke herschrijvingen verbeteren de leesbaarheid ten opzichte van het origineel, maar blijven iets achter bij de AI-varianten. De resultaten onderstrepen het belang van duidelijke instructies bij het inzetten van AI voor tekstoptimalisatie.

4.2 Invloed van dertien taalvariabelen op leesniveau (LMM)

In dit onderdeel worden de verschillende tekstversies (origineel, menselijk herschreven en twee AI-varianten) met elkaar vergeleken op dertien kenmerken die samenhangen met begrijpelijkheid. Hiervoor is gebruikgemaakt van een Lineair Mixed Model (LMM). Deze analysemethode is geschikt voor situaties waarin meerdere metingen zijn gedaan binnen één eenheid, in dit geval: vier versies per brochure. Omdat brochures onderling kunnen verschillen in inhoud en stijl, is bij het gebruiken van het model rekening gehouden met deze variatie door elk brochure als een apart ‘blok’ op te nemen (*random intercept*). Zo wordt beter zichtbaar welk effect de herschrijfwijze zelf heeft, los van het onderwerp van de brochure.

Per kenmerk wordt eerst een tabel weergegeven, gevolgd door de uitkomsten van het statistische model. Voor elk kenmerk is onderzocht of de manier van herschrijven invloed heeft op de begrijpelijkheid van de tekst. Het getal achter ‘F’ geeft aan hoe groot het verschil is tussen de tekstversies. De ‘p-waarde’ laat zien of het verschil betrouwbaar is (een p-waarde kleiner dan .05 betekent: het verschil is echt, niet toevallig). De effectgrootte (marginale R^2) geeft aan hoeveel van de verschillen in bijvoorbeeld zinslengte, aantal onbekende woorden of andere kenmerken, daadwerkelijk veroorzaakt worden door een verschil in wijze van herschrijven en dus niet door toevallige verschillen tussen de brochures zelf. Een waarde van 0,01 is klein, 0,06 is middelgroot en 0,14 of hoger is groot. Dit helpt om te bepalen welke kenmerken het meest worden beïnvloed door herschrijven. Indien er sprake is van een significant verschil worden er ter illustratie een voorbeeldzin per variant toegevoegd uit de brochures zodat er een gedetailleerd beeld ontstaat van de invloed van de herschrijving op begrijpelijkheid. Om rekening te houden met het feit dat iedere brochure vier keer is gemeten, is een lineair gemengd model gebruikt (random intercept per brochure) voor ieder kenmerk.

4.2.1 Onbekende woorden

Een tekst is makkelijker te begrijpen als er minder onbekende woorden in staan. Dat betekent hoe lager de score, hoe beter de tekst te begrijpen is.

Beschrijvende cijfers (N = 25 brochures)		
Versie	Gem. onbekende woorden	SD
AI – 1	1,40	0,23
AI – 2	1,30	0,18
Menselijk	1,45	0,30
Origineel (OT)	1,60	0,22

Tabel 5 onbekende woorden

De originele brochuretekst bevat het grootste aantal onbekende woorden. Beide AI-versies en de menselijke herschrijving verminderen dit, maar AI-2 doet dat het sterkst (ongeveer 0,3 onbekende woorden minder dan het origineel en 0,1 minder dan AI-1). De menselijke herschrijving scoort net iets beter dan het origineel, maar niet beter dan AI-1.

De analyse laat zien dat dit kenmerk significant verschilt tussen de vier tekstversies ($F(3, 72) = 12{,}57$, $p < .001$). De tekst die is herschreven door AI met een specifieke prompt gericht op begrijpelijkheid (AI-II) bevat gemiddeld de minste onbekende woorden. Daarna volgen de AI-herschrijving met een algemene prompt (AI-I) en de menselijke herschrijving. De originele tekst bevat juist het hoogste aantal onbekende woorden. De verschillen tussen het origineel en beide AI-versies blijven ook na correctie voor multiple testing significant, wat erop wijst dat deze bevinding betrouwbaar is. De effectgrootte, uitgedrukt als marginaal R^2, wijst op een middelgroot effect: ongeveer 17% van de variatie in het aantal onbekende woorden is toe te schrijven aan de wijze van herschrijven, los van de inhoud van de brochure zelf. De manier waarop een tekst is herschreven heeft een duidelijke invloed op de woordkeuze, en dan met name op het gebruik van minder bekende termen. De tweede AI-bewerking levert daarbij de duidelijkste winst op voor wat betreft begrijpelijkheid.

4.2.2 Abstracte zelfstandige naamwoorden

Teksten met minder abstracte zelfstandige naamwoorden zijn doorgaans beter te begrijpen. Dat betekent hoe lager de score, hoe beter de tekst te begrijpen is.

Beschrijvende cijfers (N = 25 brochures)		
Versie	Gem. abstracte znw	SD
AI – 1	48,43	17,05
AI – 2	48,39	17,62
Menselijk	47,87	18,67
Origineel (OT)	53,16	15,88

Tabel 6 Abstracte zelfstandige naamwoorden

Voor het kenmerk *abstracte zelfstandige naamwoorden* werd geen significant verschil gevonden tussen de vier tekstversies ($F(3, 72) = 2{,}57$, $p = .06$). Hoewel de originele tekst gemiddeld iets meer abstracte termen bevat dan de herschreven versies, zijn de verschillen klein. Na correctie voor multiple testing blijft geen enkel contrast statistisch significant. De effectgrootte, uitgedrukt in marginaal R^2, bedraagt slechts 0,02. Dit betekent dat slechts 2% van de variatie in het gebruik van abstracte zelfstandige naamwoorden verklaard wordt door de wijze van herschrijven. Samengevat heeft de herschrijfstrategie nauwelijks invloed op dit kenmerk. Het gebruik van abstracte zelfstandige naamwoorden lijkt eerder samen te hangen met inhoudelijke kenmerken van de brochure dan met de manier waarop de tekst is aangepast.

4.2.3 Lengte deelzinnen

Langere deelzinnen bemoeilijken doorgaans het lezen. Dat betekent hoe lager de score, hoe beter de tekst te begrijpen is.

Beschrijvende cijfers (N = 25 brochures)		
Versie	Gem. lengte deelzinnen	SD
AI – 1	7,02	0,66
AI – 2	6,82	0,57
Menselijk	7,57	0,91
Origineel (OT)	8,33	1,19

Tabel 7 Lengte deelzinnen

De originele brochuretekst bevat de langste deelzinnen. Alle herschrijvingen verkorten de zinsdelen, maar AI-2 doet dat het sterkst. (gemiddeld ruim 1,5 woord korter dan het origineel en 0,2 woord korter dan AI-1). De menselijke herschrijving verbetert wel ten opzichte van het origineel (-0,76 woord) maar blijft langer dan zowel AI-1 als AI-2.

Uit de analyse blijkt dat er een duidelijk verschil is in de lengte van deelzinnen tussen de verschillende tekstversies ($F(3, 72) = 15{,}42$, $p < .001$). De originele brochures bevatten de langste deelzinnen, terwijl de teksten die met AI-II zijn herschreven juist de kortste zinsdelen laten zien. Dit verschil is statistisch betrouwbaar en blijft overeind na correctie voor het uitvoeren van meerdere toetsen. De effectgrootte is marginaal ($R^2 = 0{,}32$). Dit laat zien dat dit een groot effect is en dit betekent dat bijna een derde van de verschillen in deelzinlengte verklaard kan worden door de manier waarop de tekst is herschreven. Kortgezegd heeft de herschrijfwijze, en met name de inzet van AI met een gerichte prompt, een grote invloed op hoe beknopt of langdradig zinnen zijn. Omdat kortere zinsdelen het lezen vergemakkelijken, draagt de aanpassing van AI- II aantoonbaar het meest bij aan de begrijpelijkheid van de tekst.

4.2.4 Woorden die nieuw zijn

Het gebruik van minder nieuwe woorden verlaagt de cognitieve belasting voor de lezer. Dat betekent hoe lager de score, hoe beter de tekst te begrijpen is.

Beschrijvende cijfers (N = 25 brochures)		
Versie	Gem. woorden die nieuw zijn	SD
AI – 1	87,91	3,67
AI – 2	87,49	3,12
Menselijk	80,77	24,87
Origineel (OT)	90,17	2,62

Tabel 8 Woorden die nieuw zijn

Hoewel het gebruik van minder nieuwe woorden doorgaans bijdraagt aan een lagere cognitieve belasting voor de lezer, werd in dit onderzoek geen significant verschil gevonden tussen de vier tekstversies op dit kenmerk ($F(3, 72) = 2{,}71$, $p = .05$). De originele tekst bevat gemiddeld iets meer nieuwe woorden dan de herschreven versies, maar deze verschillen zijn klein en blijken na correctie voor multiple testing niet statistisch significant. De effectgrootte is marginaal $R^2 = 0{,}07$, wat als klein wordt beschouwd. Samengevat suggereert deze analyse dat het herschrijven van de tekst (of dit nu door AI of een mens gebeurt) weinig invloed heeft op het gebruik van woorden die voor de lezer mogelijk nieuw of ongebruikelijk zijn. Dit kenmerk lijkt dus slechts in beperkte mate gevoelig voor herschrijving.

4.2.5 Afhankelijkheidslengtes

Korte afhankelijkheden in zinnen maken een tekst makkelijker te verwerken. Dat betekent hoe lager de score, hoe beter de tekst te begrijpen is.

Beschrijvende cijfers (N = 25 brochures)		
Versie	Gem. afhankelijkheidslengte	SD
AI – 1	3,65	0,74
AI – 2	3,48	0,54
Menselijk	3,87	1,51
Origineel (OT)	5,31	0,91

Tabel 9 Afhankelijkheidslengtes

De originele brochuretekst bevat veruit de langste afhankelijkheden, wat het lezen bemoeilijkt. Alle herschrijvingen verkorten deze afhankelijkheden. De menselijke herschrijving verbetert wel ten opzichte van het origineel (-1,4 woord), maar blijft langer dan zowel AI-1 als AI-2. AI-2 scoort gemiddeld bijna 1,8 woord korter dan het origineel en is het kortste van alle versies.

In dit onderzoek zijn op dit kenmerk duidelijke en statistisch significante verschillen gevonden tussen de vier tekstversies ($F(3, 72) = 20{,}40$, $p < .001$). De originele teksten bevatten de langste afhankelijkheden, wat de syntactische complexiteit vergroot. Daartegenover staan de AI-herschrijvingen, waarbij vooral AI-II de afhankelijkheden het sterkst verkort. Dit verschil ten opzichte van het origineel blijft ook na correctie voor multiple testing significant, wat de betrouwbaarheid van het effect onderstreept. De effectgrootte bedraagt marginaal $R^2 = 0{,}35$, wat geldt als een groot effect. Dit betekent dat ruim een derde van de variatie in afhankelijkheidslengte verklaard kan worden door de herschrijfwijze – een aanzienlijk aandeel. Deze resultaten laten zien dat herschrijven, en met name herschrijven met een gerichte AI-prompt (AI-II), effectief bijdraagt aan syntactische vereenvoudiging van de tekst, waardoor de tekst begrijpelijker wordt.

4.2.6 Bijvoeglijke bepalingen

Meer bijvoeglijke bepalingen kunnen zinnen verzwaren en daarmee minder leesbaarheid en dus begrijpelijker maken. Dat betekent hoe lager de score, hoe beter de tekst te begrijpen is.

Beschrijvende cijfers (N = 25 brochures)		
Versie	Gem. bijv. bepalingen	SD
AI – 1	0,64	0,18
AI – 2	0,57	0,20
Menselijk	0,82	0,31
Origineel (OT)	0,98	0,39

Tabel 10 Bijvoeglijke bepalingen

De originele teksten bevatten de meeste bijvoeglijke bepalingen. Alle herschrijvingen verlagen dit aantal, maar AI-2 doet dat het sterkst (gemiddeld 0,42 minder dan het origineel en 0,26 minder dan de menselijke versie). De menselijke herschrijving reduceert wel ten opzichte van het origineel (-0,16), maar voegt juist extra bijvoeglijke bepalingen toe in vergelijking met beide AI-versies.

In dit onderzoek blijken de vier tekstversies significant van elkaar te verschillen op dit kenmerk ($F(3, 72) = 15{,}06$, $p < .001$). De originele brochures bevatten het hoogste aantal bijvoeglijke bepalingen, wat de zinnen complexer maakt. In contrast daarmee bevatten de AI-herschrijvingen, en met name AI-II, aanzienlijk minder van deze bepalingen. AI-II blijkt het meest effectief in het terugdringen van overbodige uitbreidingen binnen de zin. De verschillen tussen het origineel en beide AI-versies blijven ook na correctie voor multiple testing significant. De effectgrootte is marginaal $R^2 = 0{,}25$, wat wijst op een middelgroot tot groot effect. Dit houdt in dat ongeveer een kwart van de variatie in het aantal bijvoeglijke bepalingen verklaard kan worden door de wijze van herschrijven. Kortom: het herschrijven van medische brochures, met name via een doelgericht aangestuurde AI, leidt tot bondigere zinnen met minder bijvoeglijke bepalingen, wat ten goede komt van de begrijpelijkheid van de tekst.

4.2.7 Opsommingen in doorlopende zinnen

Opsommingen in doorlopende zinnen kunnen de structuur van een tekst bemoeilijken en daarmee de begrijpelijkheid verminderen Ook hier geldt een lagere score is dus beter voor het vergroten van de begrijpelijkheid van een tekst.

Beschrijvende cijfers (N = 25 brochures)		
Versie	Gem. opsommingen	SD
AI – 1	0,30	0,15
AI – 2	0,26	0,13
Menselijk	0,33	0,22
Origineel (OT)	0,45	0,28

Tabel 11 Opsommingen in doorlopende zinnen

De originele brochuretekst bevat de meeste opsommingen in doorlopende zinnen. Zowel AI-I als AI-2 reduceren dit aantal, waarbij AI-2 de sterkste daling laat zien (gemiddeld 0,19 minder opsommingen dan het origineel).

In dit onderzoek zijn significante verschillen gevonden tussen de vier tekstversies op dit kenmerk ($F(3, 72) = 4{,}63$, $p = .005$). De originele brochures bevatten gemiddeld de meeste opsommingen in één zin, terwijl de AI-herschrijvingen het aantal opsommingen het sterkst reduceren. Het verschil tussen het origineel en beide AI-versies blijft significant na correctie, wat betekent dat deze vermindering betrouwbaar een effect oplevert. De effectgrootte bedraagt marginaal $R^2 = 0{,}11$, wat als een klein effect wordt beschouwd. Dit suggereert dat de herschrijfwijze in beperkte mate, maar toch aantoonbaar, bijdraagt aan het verminderen van opsommingen in doorlopende zinnen. Samenvattend: herschrijving met behulp van AI leidt tot iets compactere zinnen door minder opsommingen in te passen, wat de tekst eenvoudiger en overzichtelijker maakt voor de lezer. Wederom draagt AI-II op dit kenmerk het beste bij aan het begrijpelijker maken van de tekst.

4.2.8 Bijzinnen

Bijzinnen voegen vaak complexiteit toe aan zinnen. Dat betekent hoe lager de score op dit kenmerk, hoe meer het bijdraagt aan een begrijpelijke tekst.

Beschrijvende cijfers (N = 25 brochures)		
Versie	Gem. opsommingen	SD
AI – 1	0,27	0,13
AI – 2	0,26	0,09
Menselijk	0,26	0,16
Origineel (OT)	0,41	0,16

Tabel 12 Bijzinnen

De originele brochuretekst bevat veruit de meeste bijzinnen. Alle herschrijvingen reduceren dit aantal aanzienlijk, maar AI-2 en de menselijke versie doen dit in ongeveer gelijke mate. Er is geen verschil tussen de AI-versies onderling of tussen AI-I en de menselijke herschrijving.

In dit onderzoek zijn significante verschillen gevonden tussen de vier tekstversies op dit kenmerk ($F(3, 72) = 9{,}73$, $p < .001$). Alle herschrijfcondities, zowel de menselijke als beide AI-versies, laten een duidelijke daling zien in het aantal bijzinnen ten opzichte van de originele tekst. Deze reductie is in alle gevallen statistisch significant, ook na correctie voor multiple testing. De effectgrootte bedraagt marginaal $R^2 = 0{,}17$, wat duidt op een middelgroot effect. Dit betekent dat 17% van de variatie in het gebruik van bijzinnen verklaard wordt door de wijze van herschrijven, onafhankelijk van brochure-specifieke verschillen. Kortom zowel mensen als beide AI-prompts slagen erin teksten te schrijven met minder bijzinnen, wat een positieve invloed heeft op de begrijpelijkheid van de tekst, waarbij AI-2 en menselijke redacteurs dit het beste doen.

4.2.9 Zinslengte

Kortere zinnen dragen bij aan een betere leesbaarheid, omdat ze minder cognitieve inspanning vragen van de lezer. Dat betekent hoe lager de score op dit kenmerk, hoe meer het bijdraagt aan een begrijpelijke tekst.

Beschrijvende cijfers (N = 25 brochures)		
Versie	Gem. woorden per zin	SD
AI – 1	9,32	1,23
AI – 2	9,14	0,75
Menselijk	10,00	1,92
Origineel (OT)	12,41	1,83

Tabel 13 Zinslengte

De oorspronkelijke teksten hebben de langste zinnen. Alle herschrijvingen verkorten de zinnen aanzienlijk. AI-2 doet dit het beste (ruim 3 woorden per zin korter dan het origineel). De menselijke herschrijving verbetert wel ten opzichte van het origineel (-2,4 woorden), maar blijft langer dan beide AI-versies.

In dit onderzoek zijn op het kenmerk *zinslengte* duidelijke en statistisch significante verschillen gevonden tussen de vier tekstversies ($F(3, 72) = 29{,}73$, $p < .001$). De originele brochures bevatten gemiddeld de langste zinnen, terwijl de AI-herschreven teksten – met name AI-II – de kortste zinnen opleveren. Het verschil tussen het origineel en beide AI-versies blijft ook na correctie voor multiple testing significant, wat de betrouwbaarheid van dit effect bevestigt. De effectgrootte, uitgedrukt als marginaal R^2, bedraagt 0,43. Dit is een groot effect en betekent dat 43% van de variatie in zinslengte verklaard wordt door de wijze van herschrijving. Deze bevinding onderstreept dat herschrijving, en vooral de inzet van AI met een gerichte prompt, een krachtig middel is om zinnen te verkorten en daarmee de toegankelijkheid van medische teksten substantieel te vergroten.

4.2.10 Mensen

Het expliciet benoemen van ‘mensen’ (in plaats van abstracte verwijzingen) vergroot de toegankelijkheid van de tekst. Een hogere score op dit kenmerk leidt tot een begrijpelijkere tekst.

Beschrijvende cijfers (N = 25 brochures)		
Versie	Gem. benoemde mensen	SD
AI – 1	69,40	29,80
AI – 2	67,92	25,11
Menselijk	71,28	38,30
Origineel (OT)	45,60	26,76

Tabel 14 Mensen

De originele teksten verwijzen het minst naar ‘mensen’. Alle herschreven versies voegen aanzienlijk meer concrete verwijzingen naar ‘mensen’ toe (≈ 22–26 extra per 1.000 woorden). In dit onderzoek zijn significante verschillen gevonden in het aantal verwijzingen naar mensen tussen de vier tekstversies ($F(3, 72) = 5{,}07$, $p = .003$). Alle herschrijfcondities, zowel de menselijke als de twee AI-versies, verhogen het aantal concrete verwijzingen naar mensen ten opzichte van de originele tekst. Deze toename is statistisch significant en blijft overeind na correctie voor multiple testing. De effectgrootte is marginaal $R^2 = 0{,}11$, wat als een klein effect wordt beschouwd. Dit betekent dat de wijze van herschrijving een bescheiden, maar aantoonbare invloed heeft op de mate waarin mensen in de tekst worden benoemd. Samenvattend laat deze analyse zien dat iedere herschrijving bijdraagt aan een persoonlijkere

schrijfstijl. Hoewel het effect relatief klein is, vergroot het benoemen van mensen in de tekst wel degelijk de betrokkenheid en begrijpelijkheid voor de lezer.

4.2.11 Persoonlijke voornaamwoorden

Het gebruik van persoonlijke voornaamwoorden maakt een tekst directer en lezersvriendelijker. Een hogere score is gunstiger voor de begrijpelijkheid van teksten.

Beschrijvende cijfers (N = 25 brochures)		
Versie	Gem. persoonlijke vnw.	SD
AI – 1	52,00	28,74
AI – 2	50,64	21,75
Menselijk	61,24	39,36
Origineel (OT)	29,68	24,23

Tabel 15 Persoonlijke voornaamwoorden

De originele brochureteksten bevatten veruit de minste persoonlijke voornaamwoorden. Alle herschreven versies voegen er aanzienlijk meer toe (≈ 21–32 extra), waardoor de tekst toegankelijker en directer wordt. Er zijn geen significante verschillen tussen de AI-versies en de menselijke herschrijving; het belangrijkste is dat men afwijkt van het origineel door de persoonlijke toon te versterken.

In dit onderzoek zijn significante verschillen gevonden in het gebruik van persoonlijke voornaamwoorden tussen de vier tekstversies ($F(3, 72) = 6{,}34$, $p = .001$). Alle herschrijfcondities, inclusief beide AI-versies en de menselijke herschrijving, laten een duidelijke toename zien in het gebruik van persoonlijke voornaamwoorden ten opzichte van de originele tekst. Deze verschillen blijven statistisch significant, ook na correctie voor multiple testing. De effectgrootte bedraagt marginaal $R^2 = 0{,}14$, wat net als een groot effect wordt geïnterpreteerd. Dit betekent dat 14% van de variatie in het gebruik van persoonlijke voornaamwoorden verklaard wordt door de wijze van herschrijving. Zowel mensen als beide AI-versies slagen erin meer persoonlijke voornaamwoorden toe te voegen.

4.2.12 Aansprekingen

Directe aansprekingen verhogen de toegankelijkheid van een tekst doordat ze de lezer actief betrekken bij de inhoud. Een hogere score is daarom gunstiger voor de begrijpelijkheid.

Beschrijvende cijfers (N = 25 brochures)		
Versie	Gem. aansprekingen	SD
AI – 1	40,04	29,31
AI – 2	37,12	22,52
Menselijk	49,04	32,79
Origineel (OT)	18,20	21,78

Tabel 16 Aansprekingen

De originele teksten bevatten veruit de minste aansprekingen. Alle herschrijvingen voegen er substantieel meer toe (≈ 19–31 extra), waarbij de menselijke versie de hoogste frequentie bereikt. De twee AI-versies verschillen echter niet significant onderling.

In dit onderzoek zijn significante verschillen gevonden in het gebruik van *aansprekingen* tussen de vier tekstversies ($F(3, 72) = 7{,}53$, $p < .001$). Alle herschrijfcondities, zowel AI-gestuurd als menselijk, leiden tot een duidelijke toename van directe aansprekingen ten opzichte van de originele tekst. Deze verschillen blijven statistisch significant, ook na correctie voor multiple testing. De effectgrootte bedraagt marginaal $R^2 = 0{,}15$, wat geïnterpreteerd wordt als een middelgroot effect. Dit betekent dat de wijze van herschrijving in aanzienlijke mate bijdraagt aan het aantal directe aanspreekvormen in een tekst. Samengevat laat deze bevinding zien dat iedere herschrijving (AI-I, AI-2 en mensen) de tekst persoonlijker en actiever maakt. Door de lezer direct aan te spreken, wordt de inhoud concreter en begrijpelijker, wat vooral bij voorlichtende teksten zoals medische brochures van groot belang is.

4.2.13 Actieve werkwoordsvormen

Actieve zinnen zijn over het algemeen eenvoudiger te begrijpen dan passieve constructies. Een hogere score draagt bij aan de begrijpelijkheid van de tekst.

Beschrijvende cijfers (N = 25 brochures)		
Versie	Gem. aansprekingen	SD
AI – 1	96,34	3,31
AI – 2	97,41	3,14
Menselijk	92,18	20,53
Origineel (OT)	92,93	5,56

Tabel 17 Actieve werkwoordsvormen

In dit onderzoek is echter geen significant verschil gevonden in het gebruik van actieve werkwoordsvormen tussen de vier tekstversies ($F(3, 72) = 1{,}55$, $p = .21$). Hoewel AI-II descriptief iets meer actieve formuleringen bevat, zijn de verschillen tussen de versies klein en statistisch niet betrouwbaar. Na correctie blijven er geen significante contrasten over. De effectgrootte, uitgedrukt als marginaal R^2, bedraagt 0,04. Dit is een verwaarloosbaar effect en wijst erop dat de wijze van herschrijving slechts in zeer beperkte mate invloed heeft op het gebruik van actieve versus passieve zinnen. Kortom: in tegenstelling tot veel andere kenmerken lijkt de zinsstructuur in termen van actief of passief taalgebruik relatief stabiel te blijven, ongeacht of de tekst herschreven is door een mens of door AI.

4.2.14 Conclusie kwantitatieve analyse

De kwantitatieve resultaten uit dit onderzoek laten duidelijke patronen zien in de mate waarin de verschillende herschrijfwijzen bijdragen aan de begrijpelijkheid van medische brochures. Op basis van het algemene leesniveau blijkt dat AI-herschrijvingen – en in het bijzonder de versie met een specifieke prompt gericht op begrijpelijkheid (AI-II) – het meest effectief zijn. Van de teksten die met AI-II zijn herschreven, wordt 63,1% ingeschaald op het meest toegankelijke leesniveau (niveau 1), tegenover 48,5% bij AI-I, 39,5% bij de menselijke herschrijvingen en slechts 10,9% bij de originele tekstversies. Dit wijst op een duidelijke winst in globale begrijpelijkheid wanneer AI wordt aangestuurd met een gerichte prompt.

Ook op de afzonderlijke linguïstische kenmerken van de LiNT-formule presteert AI-II consequent het best. De teksten herschreven door AI-II bevatten gemiddeld het minste aantal onbekende woorden, de kortste deelzinnen en afhankelijkheidslengtes, het geringste aantal bijvoeglijke bepalingen, en de meest beknopte zinnen. Daarnaast laten zij een sterke reductie zien in het aantal opsommingen in doorlopende zinnen en bijzinnen. Tegelijkertijd verhogen zij het aantal verwijzingen naar mensen en het gebruik van persoonlijke voornaamwoorden. Voor deze kenmerken zijn de verschillen tussen AI-II en de originele versies statistisch significant, met effectgroottes die vaak in de categorie middelgroot tot groot vallen.

De AI-versie met een algemene prompt (AI-I) presteert eveneens beter dan de originele en menselijke herschrijvingen, maar minder uitgesproken dan AI-II. Het verschil tussen AI-I en AI-II is vooral zichtbaar op structurele kenmerken zoals deelzinslengte, afhankelijkheidslengte en het aantal bijvoeglijke bepalingen, waar AI-II telkens meer vereenvoudiging realiseert.

Menselijke herschrijvingen onderscheiden zich met name op het gebied van directheid. Op het kenmerk *aansprekingen* – bijvoorbeeld formuleringen als “u moet” of “let op” – scoren zij het hoogst. Ook het aantal verwijzingen naar mensen ligt relatief hoog bij de menselijke herschrijvingen. Op syntactisch en lexicaal niveau blijven zij echter achter bij beide AI-versies: zinnen zijn langer, onbekende woorden komen vaker voor, en de reductie van grammaticale complexiteit is minder consistent.

Een opvallende gemene deler is dat alle herschrijfwijzen – zowel door AI als door mensen – bijdragen aan een persoonlijker toon. Zowel het gebruik van persoonlijke voornaamwoorden als het expliciet benoemen van mensen neemt toe in de herschreven versies, wat de toegankelijkheid van de tekst versterkt. Hoewel de onderlinge verschillen hier klein zijn, is het contrast met de originele brochures duidelijk.

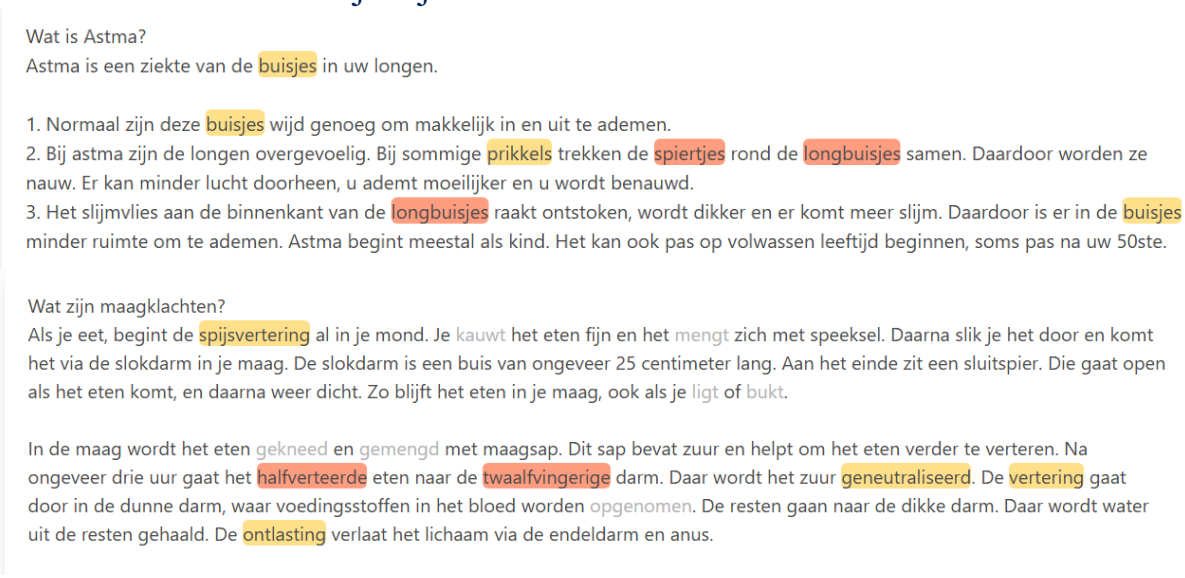
Voor enkele kenmerken zijn de verschillen niet significant: abstracte zelfstandige naamwoorden, nieuwe woorden en actieve werkwoordsvormen. Op deze punten lijken zowel AI als menselijke herschrijvers weinig invloed te hebben, en wordt de uitkomst vermoedelijk sterker bepaald door de inhoudelijke aard van de tekst dan door de herschrijfwijze.

4.3 Kwalitatieve duiding van herschrijfgedrag tussen mens en AI

Voor de kwalitatieve analyse wordt aangesloten bij de beoordelingswijze van de LiNT-formule, die tekstbegrijpelijkheid benadert vanuit drie centrale dimensies: de moeilijkheid van de gebruikte woorden, de moeilijkheid van de zinnen en de mate van persoonlijkheid in de tekst. Op deze manier wordt achterhaald in hoeverre de herschreven medische brochures, zowel door AI als door menselijke redacteuren, daadwerkelijk voldoen aan de criteria van de LiNT-formule. Daarbij wordt systematisch nagegaan of de linguïstische kenmerken waarop LiNT beoordeelt, ook zichtbaar zijn in de concrete herschrijfkeuzes van beide partijen. Met andere woorden: wordt het beoogde effect van vereenvoudiging, zoals gemeten met LiNT, ook daadwerkelijk gerealiseerd in de praktijk van het herschrijven? Om deze vraag te beantwoorden wordt niet alleen gekeken naar de aanwezigheid of afwezigheid van specifieke taalkenmerken, maar ook naar mogelijke verklaringen voor eventuele discrepanties tussen de gemeten LiNT-score en de feitelijke herschrijfstrategieën. Zo wordt geanalyseerd welke linguïstische en structurele keuzes AI en menselijke redacteuren maken, en in hoeverre deze overeenkomen met de normatieve uitgangspunten van de LiNT-formule. Ook wordt onderzocht of specifieke stijlkeuzes samenhangen met de manier waarop AI en mensen

tekstoptimalisatie benaderen. Deze verklarende benadering biedt niet alleen inzicht in de mate van overeenstemming met de LiNT-criteria, maar draagt ook bij aan de inhoudelijke validatie ervan. De analyse maakt zichtbaar hoe moeilijkheidsgraad niet alleen meetbaar is, maar ook taalkundig verklaarbaar en functioneel te duiden valt binnen de context van medische communicatie.

4.3.1 Hoe moeilijk zijn de woorden?


Wat is Astma?
Astma is een ziekte van de buisjes in uw longen.

1. Normaal zijn deze buisjes wijd genoeg om makkelijk in en uit te ademen.
2. Bij astma zijn de longen overgevoelig. Bij sommige prikkels trekken de spiertjes rond de longbuisjes samen. Daardoor worden ze nauw. Er kan minder lucht doorheen, u ademt moeilijker en u wordt benauwd.
3. Het slijmvlies aan de binnenkant van de longbuisjes raakt ontstoken, wordt dikker en er komt meer slijm. Daardoor is er in de buisjes minder ruimte om te ademen. Astma begint meestal als kind. Het kan ook pas op volwassen leeftijd beginnen, soms pas na uw 50ste.

Wat zijn maagklachten?
Als je eet, begint de spijsvertering al in je mond. Je kauwt het eten fijn en het mengt zich met speeksel. Daarna slik je het door en komt het via de slokdarm in je maag. De slokdarm is een buis van ongeveer 25 centimeter lang. Aan het einde zit een sluitspier. Die gaat open als het eten komt, en daarna weer dicht. Zo blijft het eten in je maag, ook als je ligt of bukt.

In de maag wordt het eten gekneed en gemengd met maagsap. Dit sap bevat zuur en helpt om het eten verder te verteren. Na ongeveer drie uur gaat het halfverteerde eten naar de twaalfvingerige darm. Daar wordt het zuur geneutraliseerd. De vertering gaat door in de dunne darm, waar voedingsstoffen in het bloed worden opgenomen. De resten gaan naar de dikke darm. Daar wordt water uit de resten gehaald. De ontlasting verlaat het lichaam via de endeldarm en anus.

Figuur 5 voorbeeld onbekende woorden

In deze studie is het kenmerk ‘onbekende woorden’ vastgesteld op basis van woordfrequentie. Het LiNT-instrument beschouwt woorden als ‘onbekend’ wanneer ze zelden voorkomen in het dagelijkse taalgebruik. In medische brochures gaat het hierbij vooral om vaktermen en inhoudelijk specifieke begrippen, zoals *twaalfvingerige darm*, *buisjes* of *vertering*. Deze termen zijn vaak moeilijk te vervangen door eenvoudigere alternatieven, zonder dat dit ten koste gaat van de medische nauwkeurigheid. Zowel AI als menselijke herschrijvers gaan bewust en zorgvuldig om met dit soort complexe woorden. Wanneer vervanging door een eenvoudiger woord niet mogelijk is, kiezen beide herschrijvers er voor om het moeilijke begrip uit te leggen of toe te lichten in de context. Zo blijft de oorspronkelijke term behouden, maar wordt deze wel begrijpelijker gemaakt. Bijvoorbeeld: de term *twaalfvingerige darm* wordt niet altijd vervangen, maar krijgt in beide herschrijvingen vaak een korte uitleg mee, zoals *het eerste deel van de dunne darm*. Dit draagt bij aan de toegankelijkheid van de tekst, ook als moeilijke woorden behouden blijven. Tegelijkertijd blijkt dat het gebruik van de LiNT-formule op dit punt zijn beperkingen kent. De formule beoordeelt alleen hoe vaak een

woord voorkomt in alledaagse taal, maar houdt geen rekening met de context waarin het woord wordt gebruikt of met de mate waarin het wordt uitgelegd. Medische termen komen nu eenmaal weinig voor in algemene taal, maar kunnen toch begrijpelijk zijn voor een lekenpubliek, wanneer ze worden toegelicht. Samengevat blijkt AI-II het meest consequent in het vermijden van onbekende woorden, wat suggereert dat deze herschrijfvariant het sterkst inzet op toegankelijkheid van het taalgebruik. *Abstracte zelfstandige naamwoorden* en *woorden die nieuw zijn* werden niet significant verklaard en worden daarom buiten beschouwing gelaten.

4.3.2 Hoe moeilijk zijn de zinnen?

VOORBEELDZIN LENGTE DEELZINNEN

ORIGINELE TEKST	Die gaat open als een klepje; en daarna weer dicht, zodat het voedsel in de maag blijft (normaal gesproken ook als u ligt of bukt). Dan kneden de spieren van de maag het voedsel en mengen dit met zuur maagsap, waardoor de vertering op gang komt. In de dikke darm worden water en zout aan de resten onttrokken, die dan, ingedikt, als ontlasting het lichaam verlaten via de endeldarm en de anus.
HERSCHREVEN MENS	Bij maagklachten kunt u last hebben van pijn in uw buik, brandend maagzuur en opboeren. De klachten kunnen uit de maag zelf komen. Maar ook uit de slokdarm. Bijna iedereen heeft wel eens last van brandend maagzuur. Of van eten of drinken dat weer omhoog komt.
AI – I	Na ongeveer drie uur gaat het halfverteerde eten via de maaguitgang naar de twaalfvingerige darm. Daar wordt het maagzuur geneutraliseerd en gaat de vertering verder. Voedingsstoffen komen via de dunne darm in het bloed. In de dikke darm worden water en zout uit de resten gehaald. Wat overblijft, wordt poep en verlaat het lichaam via de endeldarm en anus.
AI – II	Na ongeveer drie uur gaat het halfverteerde eten naar de twaalfvingerige darm. Daar wordt het zuur geneutraliseerd. De vertering gaat door in de dunne darm, waar voedingsstoffen in het bloed worden opgenomen. De resten gaan naar de dikke darm. Daar wordt water uit de resten gehaald. De ontlasting verlaat het lichaam via de endeldarm en anus.

Lengte van deelzinnen wordt gedefinieerd als het aantal woorden tussen komma's, voegwoorden en andere scheidende elementen binnen samengestelde zinnen. Wat opvalt is dat de originele tekst nog relatief wat lange deelzinnen bevat. Ter illustratie de zin: 'Dan kneden de spieren van de maag het voedsel en mengen dit met zuur maagsap, waardoor de vertering op gang komt'. Het eerste deel van de zin bevat de hoofdzin, gevolgd door een betrekkelijke bijzin *waardoor* (…) *op gang komt*. De hoofdzin zelf bestaat uit twee samengestelde werkwoordelijke gezegdes, gescheiden door *en*, wat de deelzinnen relatief lang maakt volgens de LiNT-formule. De laatste zin bevat een passieve constructie *worden water en zout aan de resten onttrokken*, gevolgd door een betrekkelijke bijzin *die dan, ingedikt, als ontlasting het lichaam verlaten via de endeldarm en de anus*. De deelzin *die dan, ingedikt, als ontlasting het lichaam verlaten via de endeldarm en de anus* is lang en bevat veel informatie. Wat niet bijdraagt aan de begrijpelijkheid van de tekst.

Menselijke redacteuren lossen dit op door losstaand zinsgedeelte te hanteren. De bijzin wordt dan losgeplaatst van de hoofdzin. In het voorbeeld hierboven is dat te zien aan de geelgemarkeerde zin: 'De klachten kunnen uit de maag zelf komen. Maar ook uit de slokdarm.' De bijzin wordt in een nieuwe zin geplaatst, terwijl deze hoort bij de hoofdzin. Syntactisch gezien is dit onvolledig, omdat bijzinnen niet op zichzelf kunnen staan. Dit draagt niet bij aan het vergroten van de begrijpelijkheid van de tekst. Ondanks het feit dat mensen er wel in slagen zinnen te verkorten

AI-I plaatst vooral losse zinnen achter elkaar. De tekst bevat weinig expliciete verbindings- of redengevende signaalwoorden (zoals omdat, daardoor, vervolgens en ten slotte). Zo is er een tijdmarker namelijk 'na ongeveer drie uur' en verwijst 'daar' terug naar de twaalfvingerige darm. De chronologische opbouw zorgt enigszins voor samenhang in de tekst, maar echte verbanden tussen zinnen blijven impliciet. Daardoor moeten lezers zelf de samenhang tussen zinnen reconstrueren, wat vooral voor taalzwakkere lezers een extra belasting vormt. Daarnaast begint elke nieuwe zin met een ander onderwerp, zoals *het halfverteerde eten*, *het maagzuur*, *voedingsstoffen*, *de dikke darm* of *het lichaam*. Die continue verschuiving van onderwerp maakt de tekst moeilijker navolgbaar.

AI-II plaatst eveneens losse zinnen naast elkaar. Echter, de structuur is eenvoudiger, de zinnen zijn korter en de relaties worden tussen de stappen iets explicieter gemaakt. Ook het verhaal is iets logischer te volgen, omdat niet iedere zin met een nieuw onderwerp begint. AI II slaagt er ook na de kwalitatieve analyse het beste in om zinnen begrijpelijker te herschrijven. Ook hier geldt dat wat extra signaalwoorden de lezer nog wat beter hadden

kunnen ondersteunen in het tekstbegrip.

VOORBEELDZIN OPSOMMINGEN, BIJZINNEN, AFHANKELIJKHEIDSLENGTES

ORIGINELE TEKST	De oorzaak is een besmetting met een schimmel die zich graag nestelt in de vochtige huid. Daarom wordt voetschimmel ook wel 'zwemmerseczeem' genoemd. Maar het is geen eczeem. Wel raken veel mensen met de schimmel besmet in zwembaden en in gemeenschappelijke douches, bijvoorbeeld bij gymzalen. Daar komen vochtige blote voeten immers veel voor. Maar ook zweetvoeten in slecht ventilerende sportschoenen vallen ten prooi aan schimmel.
HERSCHREVEN MENS	Voetschimmel is een veelvoorkomende en hardnekkige kwaal. De oorzaak is een besmetting met een schimmel die zich graag nestelt in de vochtige huid. Besmetting vindt vaak plaats: ### in zwembaden en gemeenschappelijke douches. ### bij zweetvoeten in slecht ventilerende sportschoenen. ### als de zuurgraad van de huid wordt verstoord doordat. ### de voeten te vaak worden gewassen met zeep. ### de zeep waarmee de voeten worden gewassen niet goed wordt weggespoeld.
AI – I	Voetschimmel ontstaat door een schimmel. Deze schimmel groeit graag op een vochtige huid. Daarom wordt voetschimmel ook wel 'zwemmerseczeem' genoemd. Maar het is geen eczeem. Veel mensen raken besmet in zwembaden of in douches bij sportzalen. Daar lopen veel mensen op blote, natte voeten. Ook als je vaak zweetvoeten hebt in slecht ventilerende sportschoenen, kun je voetschimmel krijgen. Een andere oorzaak kan zijn dat je je voeten te vaak met zeep wast, of de zeep niet goed afspoelt. Daardoor raakt de natuurlijke beschermlaag van de huid beschadigd. Op een droge huid met een gezonde zuurgraad krijgen schimmels minder kans.
AI – II	Voetschimmel ontstaat door een schimmel. Die groeit graag op een warme en vochtige huid. Daarom komt voetschimmel vaak voor bij mensen die veel op blote voeten lopen in zwembaden of douches. Ook mensen die veel zweten in dichte schoenen, zoals sportschoenen, kunnen voetschimmel krijgen.

De originele tekst bevat meerdere *opsommingen die in doorlopende zinnen* zijn verwerkt, zoals: ‘Wel raken veel mensen met de schimmel besmet in zwembaden en in gemeenschappelijke douches, bijvoorbeeld bij gymzalen. Daar komen vochtige blote voeten immers veel voor. Maar ook zweetvoeten in slecht ventilerende sportschoenen vallen ten prooi aan schimmel.’ Dit type verhoogt de cognitieve belasting van lezers en maakt de structuur minder duidelijk. Vooral omdat de originele tekst losstaande zinsgedeelte bevat, waardoor de bijzin bij de hoofdzin hoort. De opsomming wordt onderscheiden door het woordje ‘maar’ waardoor het klinkt als een tegenstelling, wat de opsomming nog meer bemoeilijkt. Dit komt ook tot uiting in de *afhankelijkheidslengte*. In deze zinnen staan onderwerp en persoonsvorm vaak ver uit elkaar, doordat er veel extra informatie tussen wordt geplaatst, bijvoorbeeld bij ‘*gymzalen’* of ‘*vochtige blote voeten immers veel voor*.’ Dit vergroot de afhankelijkheidslengte en maakt het voor de lezer moeilijker om de hoofdstructuur van de zin te volgen. Eveneens zijn de bijzinnen vaak ingebed in langere zinnen, wat de syntactische complexiteit verhoogt en het begrip bemoeilijkt, bijvoorbeeld: ‘(…) *die zich graag nestelt in de vochtige huid’*

Menselijke redacteurs kiezen voor een duidelijk, visueel gescheiden opsomming. Deze zijn niet meegenomen in de kwantitatieve analyse, maar kunnen wel meegenomen worden in de kwalitatieve analyse. Zichtbaar is ten eerste dat de opsomming niet goed loopt: ‘*Besmetting vindt vaak plaats: ### de zeep waarmee de voeten worden gewassen niet goed wordt weggespoeld.’* Opsommingen die visueel zichtbaar worden gemaakt, dragen bij aan het vergroten van de begrijpelijkheid, mits de opsomming klopt. Opvallend is dat zoals ook bleek uit de kwantitatieve analyse mensen geneigd zijn extra bijvoeglijke bepalingen toe te voegen: *hardnekkige* kwaal, *graag* (bijwoord) nestelt in een *vochtige* huid. Veel van deze bepalingen maken zinnen compacter en daarmee moeilijker leesbaar, omdat ze veel informatie in weinig woorden stoppen, volgens Pander Maat en Dekker (2016). Echter, kan het in een medische context ook voor betekenisverfijning zorgen. Ze maken het mogelijk om nuances en medische precisie over te brengen zonder overmatige redundantie of herhaling, bijvoorbeeld: ‘Een pijnlijke, ontstoken keel is vaak het eerste symptoom van keelontsteking.’ Het maakt de zin korter en geeft snel een duidelijk beeld. Ondanks dat ze in medische context functioneel kunnen zijn, blijven bijvoeglijke bepalingen voor patiënten met een lage taalvaardigheid moeilijk, vooral omdat er veel informatie in weinig woorden zit.

De AI-I versie gebruikt vooral doorlopende zinnen, maar verdeelt de informatie over meer losse zinnen: ‘*Veel mensen raken besmet in zwembaden of in douches bij sportzalen.*

Daar lopen veel mensen op blote, natte voeten. Ook als je vaak zweetvoeten hebt in slecht ventilerende sportschoenen, kun je voetschimmel krijgen. Een andere oorzaak kan zijn dat je je voeten te vaak met zeep wast, of de zeep niet goed afspoelt.' Hier worden de oorzaken grotendeels opgesplitst in aparte zinnen, maar soms blijven er nog korte opsommingen in doorlopende zinnen over '*dat je je voeten te vaak met zeep wast, of de zeep niet goed afspoelt*.' In vergelijking met het origineel is het aantal opsommingen in doorlopende zinnen verminderd, maar ze zijn niet geheel verdwenen. Wat ten goede komt aan de begrijpelijkheid van teksten. Over het algemeen zijn ook de afhankelijkheidslengtes korter dan in het origineel: *'Veel mensen raken besmet in zwembaden of in douches bij sportzalen. Daar lopen veel mensen op blote, natte voeten.'* Verder gebruikt AI-I nog steeds bijzinnen, maar zijn de zinnen al een stuk korter. De bijzinnen zijn minder complex dan in het origineel en vaak causaal of conditioneel (als/dat).

De AI-II versie kiest voor de meest eenvoudige structuur: '*Daarom komt voetschimmel vaak voor bij mensen die veel op blote voeten lopen in zwembaden of douches. Ook mensen die veel zweten in dichte schoenen, zoals sportschoenen, kunnen voetschimmel krijgen.*' Hier worden de verschillende oorzaken volledig gescheiden in korte, enkelvoudige zinnen. Er zijn geen opsommingen in doorlopende zinnen meer aanwezig; elke oorzaak wordt apart en helder benoemd. Dit sluit het beste aan bij de LiNT-criteria voor begrijpelijkheid en minimaliseert de cognitieve belasting door opsommingen. Dat geldt ook voor afhankelijkheidslengtes. Elke zin bevat één hoofdgedachte en weinig tot geen ingebedde bijzinnen of tussenzinnen. Dit resulteert in de laagste scores op afhankelijkheidslengte en sluit aan bij de kwantitatieve analyse. De bijzinnen die AI-II gebruikt zijn beperkt tot eenvoudige betrekkelijke bijzinnen ('*die veel op blote voeten lopen*', '*die veel zweten*'), zonder verdere inbedding. Het aantal bijzinnen is minimaal en de structuur is zeer overzichtelijk, wat wederom bijdraagt aan de begrijpelijkheid.

4.2.3 Hoe persoonlijk zijn de teksten?

VOORBEELDZIN OPSOMMINGEN, BIJZINNEN, AFHANKELIJKHEIDSLENGTES

ORIGINELE TEKST	Als degene in de auto die over de hobbelige weg rijdt, niet leest maar naar de horizon kijkt, komen de signalen van ogen en evenwichtsorgaan met elkaar overeen. Hij heeft dan geen last van reisziekte. Elke manier van reizen kan reisziekte veroorzaken: te land, ter zee en in de lucht. Het komt het meeste voor bij kinderen, met name in de leeftijd tussen twee en tien jaar. Met het ouder worden nemen de klachten meestal af. Sommige mensen blijven er echter hun leven lang last van houden.
HERSCHREVEN MENS	Met reisziekte bedoelen we wagenziekte, zeeziekte en luchtziekte. 1. U wordt dan ziek tijdens een autorit, op een boot of tijdens een vliegreis. 2. In de bus, tram, trein of lift kunt u er ook last van hebben. Om reisziekte te krijgen hoeft u niet altijd te reizen. U kunt er ook last van hebben in draaimolens, op schommels of in achtbanen.
AI – I	Soms moet je achteruit rijden in een bus of maak je een oversteek met een boot. Reisziekte ontstaat als je evenwichtsorgaan en je ogen verschillende signalen naar je hersenen sturen. Stel je voor: iemand leest een boek in een auto die over een hobbelige weg rijdt. De ogen zien alleen het boek en geven aan de hersenen door dat alles rustig is. Maar het evenwichtsorgaan merkt dat de auto beweegt. De hersenen krijgen daardoor tegenstrijdige berichten. Daardoor kun je misselijk worden, gaan zweten, bleek zien en soms zelfs overgeven.
AI – II	Sommige mensen worden misselijk tijdens het reizen. Bijvoorbeeld in een bus die achteruit rijdt of op een boot die heen en weer schommelt. Reisziekte komt doordat je hersenen verschillende signalen krijgen. Stel: je leest een boek in een rijdende auto. Je ogen zien de letters en denken dat je stil zit. Maar je evenwichtsorgaan, dat beweging voelt, merkt dat je rijdt.

Bij de originele tekst is de tekst overwegend onpersoonlijk: '*degene in de auto*', '*sommige mensen*'. Persoonlijke voornaamwoorden ontbreken volledig en de tekst gebruikt vooral onpersoonlijke constructies: '*Hij heeft dan*'. De originele tekst is afstandelijk en beschrijvend, met weinig tot geen persoonlijke betrokkenheid of directe aanspreking van de lezer. Dit sluit

aan bij de bevindingen uit de kwantitatieve analyse, waarin originele teksten het laagst scoren op deze kenmerken.

Menselijke redacteuren spreken in dit voorbeeld mensen aan met u. De tekst wordt directer en persoonlijker doordat er ook vanuit wij-perspectief wordt gestart. De tekst bevat expliciete aansprekingen: '*U wordt dan ziek, u kunt last hebben van.*'

AI-I en II benoemen '*sommige mensen'*, maar spreken in beide gevallen de lezer niet direct aan. De tekst gebruikt vooral algemene formuleringen, zonder mensen direct aan te spreken: '*Je leest een boek.*' AI-I maakt de tekst persoonlijker door het gebruik van '*je*' en door de lezer in een voorbeeldsituatie te plaatsen. De directe betrokkenheid is echter minder expliciet dan bij de menselijke herschrijving. AI-II maakt de tekst iets persoonlijker door het gebruik van '*je*' en door situaties te schetsen, maar blijft grotendeels algemeen en minder direct dan de menselijke herschrijving of AI-I.

4.2.3 Conclusie kwalitatieve analyse

De kwalitatieve analyse van de herschrijfstrategieën laat zien dat zowel AI als menselijke redacteuren op verschillende manieren bijdragen aan de begrijpelijkheid van medische brochures. AI, en dan vooral in combinatie met een specifieke prompt, blijkt bijzonder effectief in het reduceren van structurele complexiteit: zinnen worden korter, afhankelijkheden en bijvoeglijke bepalingen nemen af, en het aantal onbekende woorden wordt geminimaliseerd. Dit sluit nauw aan bij de theoretische inzichten uit hoofdstuk 2 over tekstvereenvoudiging, waarbij syntactische en lexicale aanpassingen centraal staan om de cognitieve belasting voor de lezer te verlagen. De specifieke prompt zorgt ervoor dat AI expliciet wordt aangestuurd op het vermijden van jargon, het hanteren van korte zinnen en het behouden van de kernboodschap, wat resulteert in een output die beter aansluit bij de B1-norm en de plain language-principes en wordt bevestigd door Ratnayake & Wang, 2023.

Menselijke redacteuren onderscheiden zich vooral door hun lezersgerichte benadering: zij voegen vaker directe aansprekingen en persoonlijke voornaamwoorden toe, waardoor de tekst persoonlijker en toegankelijker wordt. Opvallend is dat beide AI-prompts, maar vooral de specifieke prompt, deze strategieën steeds beter weten te imiteren. Echter, AI spreekt mensen nog niet direct aan.

Tegelijkertijd wordt duidelijk dat AI, ondanks de sterke vereenvoudiging, soms nuance en contextuele precisie verliest. AI verliest soms nuance en contextuele precisie bij het vereenvoudigen van medische teksten, ondanks de duidelijke winst in begrijpelijkheid en

leesbaarheid. Dit blijkt uit zowel de literatuur als uit de resultaten van het onderzoek in hoofdstuk 4. Zo toont het onderzoek van Swisher et al. (2024) aan dat door ChatGPT herschreven patiëntfolders gemiddeld 58% korter zijn dan het origineel, wat weliswaar leidt tot een lager leesniveau, maar ook tot het verlies van belangrijke details en nuance. De onderzoekers signaleren dat deze verkorte teksten soms essentiële informatie missen of informele bewoordingen bevatten die niet altijd passend zijn in een medische context. Dit kan ertoe leiden dat de tekst wel eenvoudiger is, maar minder volledig of accuraat, waardoor de patiënt mogelijk cruciale informatie mist.

Deze bevindingen sluiten aan bij de theoretische kanttekeningen uit hoofdstuk 2: effectieve tekstvereenvoudiging vraagt niet alleen om het reduceren van linguïstische complexiteit (zoals kortere zinnen, minder bijzinnen en eenvoudiger woordgebruik), maar ook om het behouden van inhoudelijke volledigheid en accuratesse. Siddharthan (2014) en Shardlow (2014) benadrukken in hun definities van tekstvereenvoudiging dat het doel is om de tekst begrijpelijker te maken zonder dat de oorspronkelijke informatie en betekenis verloren gaan. Ook Meskó en Topol (2023) waarschuwen voor het risico dat AI-systemen ‘hallucinaties’ of foutieve, maar overtuigend klinkende informatie genereren, en dat door het weglaten van nuance of context onjuiste of onvolledige medische adviezen kunnen ontstaan1.

De resultaten van de kwalitatieve analyse in hoofdstuk 4 onderstrepen deze risico’s. AI-gegenereerde teksten (vooral met een specifieke prompt) zijn weliswaar structureel eenvoudiger en bevatten minder onbekende woorden, maar laten soms inhoudelijke details of contextuele precisie achterwege die menselijke redacteuren juist wel behouden. Dit bevestigt dat menselijke controle noodzakelijk blijft om te waarborgen dat vereenvoudigde medische teksten niet alleen begrijpelijk, maar ook volledig en accuraat zijn

Hoofdstuk 5 Discussie

5.1 Interpretatie van de resultaten

De resultaten van dit onderzoek tonen aan dat zowel AI- als menselijke herschrijvingen van medische brochures significant begrijpelijker zijn dan de oorspronkelijke bronteksten, gemeten aan de hand van de LiNT-score. Opvallend is dat de AI-herschrijvingen niet onderdoen voor de versies van menselijke redacteuren met een training in B1-schrijven. Dit suggereert dat generatieve AI-toepassingen, zoals ChatGPT, in staat zijn om medische teksten op een vergelijkbare wijze te vereenvoudigen als menselijke experts. Tegelijkertijd wijzen de verschillen in type ingrepen (lexicaal versus structureel) op uiteenlopende strategieën van vereenvoudiging. Mensen herstructureren de tekst expliciet en spreken de lezer aan, terwijl AI werkt met beknopte herformuleringen en synoniemen. Dat verschil valt logisch te verklaren: de menselijke redacteuren in dit onderzoek zijn inhoudelijk expert en kiezen daardoor eerder vaktaal of complexere termen, terwijl generatieve AI is gebaseerd op natural language processing-modellen die geneigd zijn tot het selecteren van frequentere, toegankelijkere of simplistische synoniemen.

Hoewel deze kwantitatieve resultaten veelbelovend zijn, is het van belang te onderkennen dat de LiNT-score zich primair richt op formeel-structurele kenmerken van begrijpelijkheid (zoals zinslengte, tangconstructies en nominalisaties). Begrijpelijkheid in communicatief-functionele blijft daarmee grotendeels buiten beschouwing. Dit onderstreept het belang van aanvullend lezersonderzoek, waarin de daadwerkelijke interpretatie door patiënten centraal staat.

5.2 Reflectie op het gebruik van AI in herschrijvingen

Het feit dat de AI-herschrijvingen in veel gevallen even begrijpelijk zijn als die van menselijke redacteuren roept vragen op over de rol en waarde van menselijke expertise in het herschrijfproces. Waar AI efficiënt en consistent werkt, blijkt uit de kwalitatieve analyse dat menselijke herschrijvers meer contextbewust zijn, nuance en retorische sensitiviteit toepassen. Mensen hebben het vermogen om taalgebruik af te stemmen op het doel, het publiek en de context van een tekst. Tegelijkertijd wijst de grote overlap in LiNT-score en algemene begrijpelijkheid op het potentieel van AI als ondersteunende tool. Een hybride model waarin AI als eerste herschrijver fungeert, gevolgd door een menselijke eindredacteur, lijkt op basis van deze resultaten niet effectief en sluit aan bij eerdere theorieën (Meskó en Topol, 2023).

5.3 Verantwoording methode

De onderzoeksopzet vertoont een redelijke mate van validiteit en betrouwbaarheid. Binnen de context van dit onderzoek is met name de *constructvaliditeit* relevant: meten de gehanteerde instrumenten daadwerkelijk het construct ‘begrijpelijkheid’? De toepassing van het LiNT-model als meetinstrument is in dit kader verdedigbaar, aangezien dit model beoogt de cognitieve belasting van lezers in kaart te brengen. Niettemin blijft het een indirecte benadering: LiNT biedt een indicatie van potentiële begrijpelijkheid, maar toetst deze niet aan daadwerkelijke leeservaringen van eindgebruikers.

Daarnaast is de *interne validiteit* voldoende geborgd doordat het onderzoek gebruikmaakt van mixed models die controleren voor elke brochure-specifieke variatie. Hierdoor kunnen verschillen in uitkomsten met een redelijke mate van zekerheid worden toegeschreven aan de herschrijfconditie (AI I, AI II, menselijk, origineel).

Wel moet worden opgemerkt dat de *externe validiteit* beperkingen kent. Het onderzoek richt zich uitsluitend op informatieve medische brochures, en op de eerste paragrafen per folder. Daarbij zijn enkel tekstinterne aspecten gemeten; elementen zoals lay-out, visuele ondersteuning of daadwerkelijke lezersrespons zijn niet meegenomen. Daarnaast kunnen technische aspecten van de AI-omgeving invloed hebben gehad. De output kan per IP-adres verschillen, omdat het model deels willekeurige keuzes maakt tijdens het genereren van tekst. Verder is in dit onderzoek gebruikgemaakt van de betaalsde GPT-4o-variant van ChatGPT, waarbij teksten gegenereerd zijn binnen een aparte projectmap. Dit zou van invloed kunnen zijn op de consistentie en kwaliteit van de gegenereerde output.

5.4 Theoretische en maatschappelijke implicaties

De resultaten van dit onderzoek laten zien dat AI bruikbaar kan zijn binnen begrijpelijkheidsonderzoek en tekstoptimalisatie, zolang het gebruik ervan kritisch begeleid wordt. Daarmee levert dit onderzoek een bijdrage aan het lopende debat over de betekenis van *begrijpelijkheid*: het gaat daarbij niet alleen om meetbare kenmerken zoals zinslengte of woordgebruik, maar ook om de wisselwerking tussen de tekst, de lezer en de context waarin de tekst gelezen wordt (Jansen, 2013). Voor de praktijk (de gezondheidszorg) bieden de bevindingen kansen. Organisaties die heldere communicatie belangrijk vinden maar beperkt zijn in tijd of middelen, zouden AI kunnen inzetten als hulpmiddel. Tegelijkertijd laten de resultaten zien dat het onverstandig is AI zonder toezicht te gebruiken: menselijke controle blijft nodig om nuance, helderheid en doelgerichtheid te waarborgen.

Hoofdstuk 6 Conclusie

Deze scriptie onderzocht in hoeverre herschreven medische brochures die door AI zijn gegenereerd verschillen van herschrijvingen door menselijke redacteuren, gemeten op basis van objectieve kenmerken van begrijpelijkheid. En dan met name of AI beter in staat is om teksten begrijpelijker te formuleren.

In hoeverre is GenAI in staat medische brochures begrijpelijker te formuleren dan menselijke redacteuren, gemeten op basis van objectieve tekstuele kenmerken?

Uit de resultaten blijkt dat beide typen herschrijvingen significant begrijpelijker zijn dan de oorspronkelijke teksten. Het antwoord op de onderzoeksvraag is tweeledig. Enerzijds laten de AI-II-herschrijvingen de laagste scores zien op de meeste formele leesbaarheidskenmerken, wat wijst op een hogere potentiële begrijpelijkheid. Hoewel de verschillen met menselijke herschrijvingen niet altijd statistisch significant zijn, presteert AI-II het best op tekstinterne kenmerken zoals zinslengte, onbekende woorden of afhankelijkheidslengtes. Daarmee kan geconcludeerd worden dat volgens de LiNT-formule AI formeel gezien beter begrijpelijkere teksten produceert. Anderzijds blijft AI communicatief gezien achter bij menselijke herschrijvers. Waar AI efficiënt en consistent werkt, blijkt uit de kwalitatieve analyse dat menselijke herschrijvers meer contextbewust zijn, nuance en retorische sensitiviteit toepassen. Mensen hebben het vermogen om taalgebruik af te stemmen op het doel, het publiek en de context van een tekst.

De bevindingen ondersteunen de inzet van AI als herschrijfhulpmiddel, met name in situaties waar snelheid, consistentie en toegankelijkheid prioriteit hebben. Tegelijkertijd onderstrepen zij het belang van menselijke controle en redactionele beoordeling. Het onderzoek laat zien dat AI en menselijke expertise het beste samen kunnen worden ingezet, waarbij ze elkaar aanvullen in plaats van vervangen.

Voor het vakgebied taalbeheersing biedt dit onderzoek aanleiding tot reflectie op de conceptualisering van begrijpelijkheid: het reduceren van dit begrip tot louter formele kenmerken (zoals meetbaar via een formule als LiNT) is ontoereikend. Begrijpelijkheid is meer dan alleen wat je met formules kunt meten; het hangt ook af van de context en van wie de tekst leest. Dat vraagt om een aanpak die zowel objectieve metingen als inhoudelijke en situationele factoren meeneemt.

Vervolgonderzoek zou zich kunnen richten op het testen van deze resultaten bij echte lezers, bijvoorbeeld door patiënten met verschillende achtergronden medische teksten te laten beoordelen. Daarbij is het belangrijk om niet alleen te kijken of ze de tekst begrijpen, maar ook hoe ze zich erbij voelen, of ze de informatie vertrouwen, hoe ze erop reageren en of ze de instructies goed interpreteren. Ook is het aan te raden om AI-systemen te beoordelen op meer dan alleen begrijpelijkheid, bijvoorbeeld: of de stijl past bij het genre, of de informatie consistent is, en of de toon geschikt is voor de doelgroep. Daarnaast laat dit onderzoek zien dat het goed is om na te denken over hoe we leesformules zoals LiNT gebruiken in complexe tekstsoorten zoals medische brochures. Het zou waardevol zijn om nieuwe meetinstrumenten te ontwikkelen die beter passen bij dit soort teksten, bijvoorbeeld door zowel genrespecifieke eisen als communicatieve doelen mee te nemen in de beoordeling.

Literatuurlijst

Alcaraz Varó, E. (2000). *El inglés profesional y académico*. Alianza.

Althunayyan, S., & Azmi, A. (2021). Automated text simplification. *ACM Computing Surveys, 54*(2), 1–36. https://doi.org/10.1145/3442695

Anderson, R. C., & Davison, A. (1988). Conceptual and empirical bases of readability formulas. In A. Davison & G. M. Green (Eds.), *Linguistic complexity and text comprehension: Readability issues reconsidered*. Lawrence Erlbaum Associates.

Beck, I. L., McKeown, M. G., Sinatra, G. M., & Loxterman, J. A. (1991). Revising social studies text from a text-processing perspective: Evidence of improved comprehensibility. *Reading Research Quarterly, 26*(3), 251–276. https://doi.org/10.2307/747763

Berkenkotter, C. A., & Huckin, T. N. (1995). *Genre knowledge in disciplinary communication: Cognition, culture, power*. Lawrence Erlbaum Associates.

Bhatia, V. K. (1993). *Analysing genre: Language use in professional settings*. Longman.

Bhatia, V. K. (2004). *Worlds of written discourse: A genre-based view*. Continuum.

Bhatia, V. K. (2010). Interdiscursivity in professional communication. *Discourse & Communication*, *4*(1), 32–50. https://doi.org/10.1177/1750481309351208

Brück, T. vor der, Hartrumpf, S., & Helbig, H. (2008). A readability checker with supervised learning using deep indicators. *Informatica, 32*, 429–435.

Cao, M., Wang, Q., Zhang, X., Lang, Z., Qiu, J., Yung, P. S., & Ong, M. T. (2024c). Large language models' performances regarding common patient questions about osteoarthritis: A comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Perplexity. *Journal Of Sport And Health Science/Journal Of Sport And Health Science*, 101016. https://doi.org/10.1016/j.jshs.2024.101016

Collins-Thompson, K. (2014). Computational assessment of text readability: A survey of current and future research. *ITL - International Journal of Applied Linguistics, 165*(2), 97–135. https://doi.org/10.1075/itl.165.2.01col

De Clercq, O., Hoste, V., Desmet, B., Van Oosten, P., De Cock, M., & Macken, L. (2014). Using the crowd for readability prediction. *Natural Language Engineering, 20*(3), 293–325. https://doi.org/10.1017/S1351324912000344

Dudley-Evans, A. (1986). Genre analysis: An investigation of the introduction and discussion sections of M.Sc. dissertations. In M. Coulthard (Ed.), *Talking about text* (pp. 128–145). University of Birmingham.

Dudley-Evans, A. (1994). Genre analysis: An approach for text analysis for ESP. In M. Coulthard (Ed.), *Advances in written text analysis* (pp. 219–228). Routledge.

Eid, K., Eid, A., Wang, D., Raiker, R. S., Chen, S., & Nguyen, J. (2023). Optimizing ophthalmology patient education via chatbot-generated materials: Readability analysis of AI-generated patient education materials and the American Society of Ophthalmic Plastic and Reconstructive Surgery patient brochures. *Ophthalmic Plastic and Reconstructive Surgery, 40*(2), 212–216. https://doi.org/10.1097/iop.0000000000002549

Gamero Pérez, S. (1998). *La traducción de textos técnicos (alemán-español): Géneros y subgéneros*. Universitat Autònoma de Barcelona.

Gamero Pérez, S. (2001). *La traducción de textos técnicos*. Ariel.

De Geus, J., & Loomans, N. (2020). *Stukken beter schrijven: zo wordt schrijven makkelijker en leuker!* Uitgeverij Thema.

Hatim, B., & Mason, I. (1990). *Discourse and the translator*. Longman.

Hatim, B., & Mason, I. (1995). *Teoría de la traducción: una aproximación al discurso*. Ariel.

Hatim, B., & Mason, I. (1997). *The translator as communicator*. Routledge.

Heijmans, M., Brabers, A., & Rademakers, J. (2018). *Health literacy in Nederland*. Nivel. Geraadpleegd op 19 februari 2025, van https://www.nivel.nl/sites/default/files/bestanden/Gezondheidsvaardigheden_in_Nederland.pdf\

Hendrikx, W. (2012). *Schrijven voor het beeldscherm: Scoren met tekst op internet, intranet en in sociale media* (6e, herz. dr.). Academic Service.

Hurtado Albir, A. (2002). *Traducción y traductología: Introducción a la traductología*. Cátedra.

Irwin, J. (1980). The effects of explicitness and clause order on the comprehension of reversible causal relationships. *Reading Research Quarterly, 15*(4), 477–488. https://doi.org/10.2307/747275

Jansen, C. (2013). Taalniveau B1: de nieuwste kleren van de keizer. *Onze Taal*, 82(2), 56–57.

Jansen, C., & Boersma, N. (2013). Meten is weten? Over de waarde van de leesbaarheidsvoorspellingen van drie geautomatiseerde Nederlandse meetinstrumenten. *Tijdschrift Voor Taalbeheersing*, *35*(1), 47–62. https://doi.org/10.5117/tvt2013.1.jans

Kathpalia, S. S. (1992). *A genre analysis of promotional texts* (Doctoral dissertation, National University of Singapore).

Kincaid, J. P., Fishburne, R. P., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new*

readability formulas (automated readability index, fog count and Flesch reading ease formula) for Navy enlisted personnel. U.S. Naval Technical Training Command.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge University Press.

Kraf, R., Lentz, L., & Maat, H. P. (2011). Drie Nederlandse instrumenten voor het automatisch voorspellen van begrijpelijkheid - Een klein consumentenonderzoek. *Tijdschrift Voor Taalbeheersing*, *33*(3), 249–265. https://doi.org/10.5117/tvt2011.3.drie411

Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., & Tseng, V. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health*, *2*(2), e0000198. https://doi.org/10.1371/journal.pdig.0000198

Land, J., Sanders, T., Lentz, L., & Van den Bergh, H. (2002). Coherentie en identificatie in studieboeken: Een empirisch onderzoek naar tekstbegrip en tekstwaardering op het vmbo. *Tijdschrift voor Taalbeheersing, 4*(4), 281–302.

Langer, I., Schulz von Thun, F., & Tausch, R. (2019). *Sich verständlich ausdrücken* (11e dr.). Ernst Reinhardt Verlag.

Last, B., & Sprakel, T. (2023). *Chatten met Napoleon: Werken met generatieve AI in het onderwijs*. Boom.

Lentz, L. (2011). *Let op: begrip verplicht! Begrijpelijkheid als norm in de wet* [oratie]. Geraadpleegd op 8 juni 2025, van: https://uu.nl/sites/default/files/gw_lentz_leo_oratie_definitief.pdf

Lentz, L. (2020). Begrijpelijke taal: Wat is dat? *Didactiek Nederlands – Handboek*. Geraadpleegd op 9 april 2025, van https://didactieknederlands.nl/handboek/2020/09/begrijpelijke-taal-wat-is-dat/

Lentz, L. (2021, 1 oktober). *Keuzehulp bij tools voor begrijpelijke teksten*. Didactiek Nederlands. Geraadpleegd op 21 mei 2025, van https://didactieknederlands.nl/2021/10/keuzehulp/

Levy, E. T. (2003). The roots of coherence in discourse. *Human Development, 46*(1), 69–88. https://doi.org/10.1159/000070367

Linderholm, T., Everson, M. G., Van Den Broek, P., Mischinski, M., Crittenden, A., & Samuels, J. (2000). Effects of Causal Text Revisions on More- and Less-Skilled Readers' Comprehension of Easy and Difficult Texts. *Cognition And Instruction*, *18*(4), 525–556. https://doi.org/10.1207/s1532690xci1804_4

Martinc, M., Pollak, S., & Robnik-Šikonja, M. (2021). Supervised and unsupervised neural

approaches to text readability. *Computational Linguistics, 47*(1), 141–179. https://doi.org/10.1162/coli_a_00384

Mayor Serrano, M. B. (2010). Revisión y corrección de textos destinados a los pacientes... y algo más. *Panace@, 11*(31), 29–36.

Meijer, D., & Noijons, J. (2008). Gemeenschappelijk Europees Referentiekader voor Moderne Vreemde Talen - Taalunie: Common European Framework of Reference for Languages: Learning, Teaching, Assessment. In *Nederlandse Taalunie*. Geraadpleegd op 13 mei 2025, van https://taalunie.org/publicaties/34/gemeenschappelijk-europees-referentiekader-voor-moderne-vreemde-talen

Meskó, B., & Topol, E. J. (2023). The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *Npj Digital Medicine*, *6*(1). https://doi.org/10.1038/s41746-023-00873-0

Ministerie van Volksgezondheid, Welzijn en Sport. (2024, 31 december). *Wetgeving medische hulpmiddelen*. Geraadpleegd op 17 februari 2025, van https://www.rijksoverheid.nl/onderwerpen/medische-hulpmiddelen/nieuwe-wetgeving-medische-hulpmiddelen

Muñoz Torres, C. A. (2002). Tipología textual y análisis para la traducción. Una tipología de géneros médicos. In *Translating science. Proceedings 2nd International Conference on Specialized Translation (28 febrero–2 marzo)*. Institut Universitari de Lingüística Aplicada.

Nutbeam, D. (2008). The evolving concept of health literacy. *Social Science & Medicine*, *67*(12), 2072–2078. https://doi.org/10.1016/j.socscimed.2008.09.050

OECD. (2013). *OECD skills outlook 2013: First results from the Survey of Adult Skills*. OECD Publishing. https://doi.org/10.1787/9789264204256-en

Oliffe, M., Thompson, E., Johnston, J., Freeman, D., Bagga, H., & Wong, P. K. K. (2019). Assessing the readability and patient comprehension of rheumatology medicine information sheets: a cross-sectional Health Literacy Study. *BMJ Open*, *9*(2), e024582. https://doi.org/10.1136/bmjopen-2018-024582

Ong, E., Damay, J., Lojico, G., Lu, K., & Tarantan, D. (2008). Simplifying text in medical literature. *Journal Of Research in Science Computing And Engineering*, *4*(1). https://doi.org/10.3860/jrsce.v4i1.441

Onze Taal. (z.d.). *Wat betekent B1-niveau?* Geraadpleegd op 16 mei 2025, van https://onzetaal.nl/taalloket/b1-niveau

Ornia, G. F. (2016). *Medical brochure as a textual genre*. Cambridge Scholars Publishing.

Pharos. (2025, 17 maart). *Begrijpelijke voorlichtingsmaterialen & beeldverhalen*. Geraadpleegd op 2 mei 2025, van https://www.pharos.nl/thema/begrijpelijke-voorlichtingsmaterialen-en-beeldverhalen/

Quijada Diez, C. (2008). *Estudio traductológico del texto médico: Propuesta de un modelo de análisis y aplicación a un corpus textual (alemán-español)*. University of Salamanca.

Ratnayake, H., & Wang, C. (2023). A Prompting Framework to Enhance Language Model Output. In *Lecture notes in computer science* (pp. 66–81). https://doi.org/10.1007/978-981-99-8391-9_6

Renkema, J. (2010). *Schrijfwijzer* (4e dr.). Boom.

Roemmele, B. (2023, 15 maart). *This is the revised version 2.0 ChatGPT SuperPrompt that writes, rates and runs your prompt. It will automatically generate a prompt for you on any subject.* [Tweet]. Twitter. Geraadpleegd op 22 mei 2025, van https://twitter.com/BrianRoemmele/status/1647442652047212544?s=20

Sanders, T. (2005). Taal werkt – en hoe! Over taalgebruik, taalverwerking en taalbeheersing. *Tijdschrift voor Taalbeheersing, 27*(1), 4–22. https://www.dbnl.org/tekst/_tij009200501_01/_tij009200501_01_0004.php

Schriver, K. A. (2017). Plain language in the US gains momentum: 1940–2015. *IEEE Transactions on Professional Communication, 60*(4), 343–383. https://doi.org/10.1109/TPC.2017.2768498

Shardlow, M. (2014). Out in the open: Finding and categorising errors in the lexical simplification pipeline. In *The 9th International Conference on Language Resources and Evaluation (LREC 2014)* (pp. 1583–1590).

Siddharthan, A. (2014). A survey of research on text simplification. *ITL - International Journal Of Applied Linguistics*, *165*(2), 259–298. https://doi.org/10.1075/itl.165.2.06sid

Singer, M. (1990). *Psychology of language: An introduction to sentence and discourse processes*. Lawrence Erlbaum Associates.

Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge University Press.

Swisher, A. R., Wu, A. W., Liu, G. C., Lee, M. K., Carle, T. R., & Tang, D. M. (2024). Enhancing Health Literacy: Evaluating the Readability of Patient Handouts Revised by ChatGPT's Large Language Model. *Otolaryngology*, *171*(6), 1751–1757. https://doi.org/10.1002/ohn.927

Taalcentrum-VU. (2025, 29 januari). *De European Accessibility Act: Hoe zorg jij voor betere*

toegankelijkheid? Geraadpleegd op 17 februari 2025, van https://www.taalcentrum-vu.nl/actueel/nieuws/de-european-accessibility-act-en-digitale-toegankelijkheid

Trosborg, A. (1997). Register, genre and text type. In A. Trosborg (Ed.), *Text typology and translation* (pp. 3–24). John Benjamins.

Vajjala, S. (2021). Trends, limitations and open challenges in automatic readability assessment research. *arXiv*. https://arxiv.org/abs/2105.00973

Van der Bruggen, G. (2020). Klare taal in uitspraken: Meer dan stijl alleen. *Nederlands juristenblad*, *2020*(28), 1808-2036. Article 28. https://www.navigator.nl/document/id864b5b847db7470082cb769d16a2afc3/nederlands-juristenblad-klare-taal-in-uitspraken

Visser, W. (2009). Het moet. Het kan ook. Maar we gaan het niet doen. *Trema*, (7), 305–310.

Wet WGBO 2018. (2025, 17 februari). *Boek 7, Titeldeel 7, Afdeling 5, Artikel 448*. Geraadpleegd op 20 februari 2025, van https://wetten.overheid.nl/BWBR0005290/2018-02-01/0/Boek7/Titeldeel7/Afdeling5/Artikel448/informatie

Bijlage A Gemeenschappelijke referentieniveaus zelfbeoordeling

B1				
Luisteren	**Lezen**	**Interactie**	**Productie**	**Schrijven**
Ik kan de hoofdpunten begrijpen wanneer in duidelijk uitgesproken standaardtaal wordt gesproken over vertrouwde zaken die ik regelmatig tegenkom op mijn werk, school, vrije tijd enz. Ik kan de hoofdpunten van veel radio of tv-programma's over actuele zaken of over onderwerpen van persoonlijk of beroepsmatig belang begrijpen, wanneer er betrekkelijk langzaam en duidelijk gesproken wordt.	Ik kan teksten begrijpen die hoofdzakelijk bestaan uit hoogfrequente, alledaagse of aan mijn werk gerelateerde taal. Ik kan de beschrijving van gebeurtenissen, gevoelens en wensen in persoonlijke brieven begrijpen.	Ik kan de meeste situaties aan die zich kunnen voordoen tijdens een reis in een gebied waar de betreffende taal wordt gesproken. Ik kan onvoorbereid deelnemen aan een gesprek over onderwerpen die vertrouwd zijn, of mijn persoonlijke belangstelling hebben of die betrekking hebben op het dagelijks leven (bijvoorbeeld familie, hobby's, werk, reizen en actuele gebeurtenissen).	Ik kan uitingen op een simpele manier aan elkaar verbinden, zodat ik ervaringen en gebeurtenissen, mijn dromen, verwachtingen en ambities kan beschrijven. Ik kan in het kort redenen en verklaringen geven voor mijn meningen en plannen. Ik kan een verhaal vertellen, of de plot van een boek of film weergeven en mijn reacties beschrijven.	Ik kan eenvoudige samenhangende tekst schrijven over onderwerpen die vertrouwd of van persoonlijk belang zijn. Ik kan persoonlijke brieven schrijven waarin ik mijn ervaringen en indrukken beschrijf.

Tabel 18 Meijer en Noijons in Nederlandse Taalunie (2008)

Bijlage B Overzicht van vereenvoudigingstechnieken

Type vereenvoudiging	Beschrijving	Typische bewerkingen
Lexicale vereenvoudiging (LS)	Het vervangen van complexe of minder frequente woorden door eenvoudigere synoniemen, met behoud van betekenis.	• Identificatie van moeilijke woorden • Contextgevoelige vervanging • Rangorde op basis van eenvoud • Behoud van grammaticale consistentie
Syntactische vereenvoudiging (SS)	Het herstructureren van grammaticaal complexe zinnen om de leesbaarheid te vergroten.	• Opsplitsing zinnen • Herschikken van bijzinnen • Weglaten van bijzinnen of reductie zinsdelen
Conceptuele vereenvoudiging	Inhoudelijk herformuleren van abstracte of specialistische ideeën zodat ze begrijpelijker worden voor leken.	• Generalisering • Gebruik van analogieën of concrete voorbeelden • Verheldering van terminologie
Elaboratieve vereenvoudiging	Toevoegen van redundantie en expliciete signaalwoorden om interpretatie te vergemakkelijken.	• Herhaling van kerninformatie • Explicitering van verbanden • Gebruik van verbindingswoorden
Samenvatting (reduction)	Beknopter maken van teksten door irrelevante of complexe informatie te verwijderen.	• Weglaten van details of bijzaken • Focus op hoofdlijn • Compressie van informatie in kernzinnen

Tabel 19 Vereenvoudigingstechnieken (Siddharthan en Shardlow, 2014)

Bijlage C Promptingstrategieën

Type	Definitie	Voordelen	Nadelen	Toepassingssituaties
Zero-shot prompting	Een enkele prompt wordt gegeven zonder enige context of voorbeelden, en het model wordt geacht een nauwkeurig en relevant antwoord te geven (Kojima Brown et al., 2022).	Snel en eenvoudig; geen voorbeelden nodig	Minder nauwkeurig, kan moeite hebben met complexe taken	Eenvoudige vragen of taken, algemene kennisvragen
Few-shot prompting	Het taalmodel krijgt enkele voorbeelden van gewenste input-outputparen voordat de hoofdvraag wordt gesteld.	Verbeterde nauwkeurigheid, helpt het model de taak te begrijpen	Vereist relevante voorbeelden, verhoogde insteltijd	Taken die context vereisen, nieuwe taken aanleren
Chain-of-thought prompting	Complexe vragen worden opgedeeld in eenvoudigere subvraagstukken en het model wordt achtereenvolgens geprompt om elk subvraagstuk te beantwoorden (Wei Brown et al., 2023).	Vereenvoudigt complexe vragen, genereert samenhangender antwoorden	Vereist handmatige opsplitsing van vragen, kan het grotere geheel missen	Complexe vragen, taken in meerdere stappen
Zero-shot chain of thought	Combineert de concepten zero-shot prompting en chain-of-thought prompting. Het model krijgt een reeks verwante prompts zonder voorbeelden en moet samenhangende antwoorden geven.	Combineert zero-shot en chain-of-thought, test begrip van het model	Minder nauwkeurig, geen voorbeelden om het model te sturen	Sequentiële vragen zonder context, evalueren van modelcapaciteiten
Self-consistency	Een vraag wordt meerdere keren met kleine variaties aan het model gesteld, evenals voorbeeldantwoorden, om de consistentie van de responsen te evalueren.	Evalueert modelbetrouwbaarheid, onthult vooroordelen	Meerdere queries nodig, verhoogt mogelijk niet de nauwkeurigheid	Beoordelen van begrip van het model, identificeren van biases
Generate knowledge prompting	Het model wordt gevraagd nieuwe kennis of inzichten te genereren op basis van zijn bestaande kennis.	Stimuleert creatief denken, ideegeneratie	Antwoorden kunnen onpraktisch zijn, afwijken van bekende feiten	Brainstormen, innovatieve denkprocessen
Prompt Generation	Het model wordt gevraagd om zelf een prompt te genereren op basis van een bepaald thema of onderwerp.	Stimuleert creativiteit, verkent nieuwe ideeën	Door het model gegenereerde prompts kunnen vaag zijn of de kern missen	Brainstormen, discussiethema's, creatieve schrijfopdrachten
Model laten optreden als specifieke persoon	Het model wordt gevraagd om vragen te beantwoorden of inhoud te genereren in de stijl of het perspectief van een specifieke persoon.	Persoonlijke responsen, vangt unieke perspectieven	Nabootsing mogelijk onnauwkeurig, risico op misinformatie	Nabootsen van schrijfstijlen, perspectieven vangen, historische of fictionele scenario's

Tabel 20 Promptingstrategieën (Ratnayake & Wang, 2023)

Bijlage D Operationele definitie van kenmerken LiNT-formule

In deze bijlage wordt uitgelegd hoe welk kenmerk in de LiNT-formule gemeten wordt en hoe dit helpt om de begrijpelijkheid van een tekst te beoordelen. De uitleg is afkomstig van Pander Maat en Dekker (2016) in andere gevallen is de tekst overgenomen in eigen woorden van de website van LiNT.

Hoe moeilijk zijn uw woorden?

1. **Onbekende woorden:** Om in te schatten hoe bekend woorden zijn, maken Pander Maat en Dekker (2016) gebruik van woordfrequentie: hoe vaak een woord voorkomt in alledaagse taal. Zij baseerden zich voor deze formule op het Subtlex-corpus, een grote verzameling Engelse ondertitels (44 miljoen woorden), waarbij alleen inhoudswoorden (zelfstandige naamwoorden, werkwoorden, bijvoeglijke naamwoorden en bijwoorden) worden meegenomen. De frequentie van een woord wordt uitgedrukt op een logaritmische schaal (log10) van het aantal keren dat het woord voorkomt per miljard woorden. Bijvoorbeeld: een woord met een frequentie van 4 komt ongeveer 10.000 keer voor per miljard woorden, en een frequentie van 5 betekent dat het woord 100.000 keer voorkomt. Een verschil van 1 op deze schaal betekent dus dat een woord gemiddeld tien keer zo vaak voorkomt, wat een groot verschil in bekendheid impliceert. Om de betrouwbaarheid te vergroten hebben Pander Maat en Dekker (2016) drie correcties toegepast: eigennamen zijn uitgesloten, bij samenstellingen wordt alleen het basiswoord geteld (duwboot wordt gelezen als ‘boot’). Tot slot worden cultuurgebonden woorden die niet in het corpus staan genegeerd via een vaste stoplijst. Dat wil zeggen dat LiNT ze overslaat bij zijn woordfrequentie-berekening.

2. **Abstracte zelfstandige naamwoorden:** Het tweede kenmerk in de LiNT-formule is abstractheid. Abstracte teksten zijn doorgaans moeilijker te begrijpen dan concrete teksten. LiNT meet daarom welk percentage van de zelfstandige naamwoorden in een tekst als abstract wordt beschouwd. Concrete woorden verwijzen naar direct waarneembare zaken, zoals mensen (bijvoorbeeld: voetballer), voorwerpen (koffiepot) of gebeurtenissen (aai). Abstracte woorden verwijzen naar ideeën of begrippen, zoals *democratie* of *beleidsvoorstel*. Sommige woorden kunnen beide betekenissen hebben; deze worden uit voorzorg als abstract geteld. De analyse richt zich op zelfstandige naamwoorden, omdat die het duidelijkst verschil maken in abstractheid tussen teksten.

3. **Woorden die nieuw zijn in de tekst:** In moeilijkere teksten komen minder herhalingen voor; ze behandelen meer onderwerpen of gebruiken telkens nieuwe woorden om naar dezelfde zaken te verwijzen. Daarom hebben Pander Maat en Dekker (2016) dit kenmerk opgenomen in de LiNT-formule, zodat wordt gemeten welk percentage van de inhoudswoorden niet eerder voorkomt in de voorgaande vijftig woorden. De analyse richt zich daarbij op zelfstandige naamwoorden, werkwoorden, namen en voornaamwoorden. Daarnaast worden woordlemma's gebruikt om rekening te houden met verschillende werkwoordstijden en vormen van hetzelfde woord (koopt en verkocht), deze tellen als één. Een zogenoemd 'venster' van 50 woorden schuift over de tekst en bekijkt per nieuw woord of dat woord al eens eerder voor kwam in de tekst. Woorden die niet in dat venster zijn genoemd, tellen als 'nieuw'.

Hoe moeilijk zijn uw zinnen?

4. **Lengte van deelzinnen:** Een deelzin bevat één vervoegd werkwoord. Een zin kan uit meerdere deelzinnen bestaan, bijvoorbeeld hoofdzinnen en bijzinnen.
   Niet de lengte van de hele zin, maar de lengte van de afzonderlijke deelzinnen zegt iets over de moeilijkheidsgraad. Hoe meer woorden per deelzin, hoe compacter de informatie en hoe complexer de zinstructuur (Pander Maat en Dekker, 2016). LiNT vergelijkt deze lengte met teksten met korte deelzinnen (zoals vmbo-schoolboeken) en met lange (zoals wetenschappelijke artikelen).

5. **Afhankelijkheidslengtes:** Afhankelijkheidslengte gaat over de afstand tussen twee woorden die bij elkaar horen, zoals een lidwoord en een zelfstandig naamwoord (bijvoorbeeld: de man) of een onderwerp en het werkwoord (bijvoorbeeld: Peter schrijft). Een tangconstructie (wanneer de afstand tussen deze twee woorden groot is) bemoeilijkt het lezen van een tekst. LiNT meet per zin de langste afhankelijkheid (in aantal tussenliggende woorden) en berekent daar het gemiddelde van. Lange afhankelijkheden, vooral boven de 10 woorden, kunnen de leesbaarheid sterk verminderen. De scores worden vergeleken met eenvoudige teksten (zoals vmbo-schoolboeken) en moeilijke teksten (zoals wetenschappelijke artikelen).

6. **Bijvoeglijke bepalingen**: Dit kenmerk gaat over woordgroepen die extra informatie geven over een zelfstandig naamwoord. Veel van deze bepalingen maken zinnen compacter en daarmee moeilijker leesbaar, omdat ze veel informatie in weinig woorden stoppen. Daarom hebben Pander Maat en Dekker (2016) bijvoeglijke bepalingen opgenomen in de LiNT-formule. LiNT telt het aantal bijvoeglijke bepalingen per deelzin (één per vervoegd werkwoord) en vergelijkt dat met eenvoudige teksten (zoals roddelberichten) en complexe teksten (zoals wetenschappelijke artikelen).

7. **Opsommingen:** LiNT meet daarnaast ook het aantal opsommingen binnen deelzinnen, omdat Pander Maat en Dekker (2016) beargumenteren dat deze bijdragen aan de informatiedichtheid van een tekst. Vooral doorlopende opsommingen (zoals A, B en C) maken zinnen compacter en daarmee moeilijker leesbaar. Opsommingen in bulletvorm zijn vaak duidelijker voor de lezer. LiNT telt het aantal opsommingen per deelzin en vergelijkt dit met eenvoudige teksten (zoals roddelberichten) en complexe teksten (zoals wetenschappelijke artikelen).

8. **Bijzinnen:** LiNT meet daarnaast het gemiddeld aantal bijzinnen per zin, omdat veel bijzinnen een tekst moeilijker maken. Eén of twee bijzinnen zijn meestal geen probleem, maar bij vier of meer kan de lezer de draad kwijtraken — vooral als de hoofdzin wordt uitgesteld (Pander Maat en Dekker, 2016). LiNT telt vier soorten bijzinnen:
    - Betrekkelijke bijzinnen (bijv. de man die daar loopt),
    - Bijwoordelijke bijzinnen (bijv. omdat ik moe was),
    - Complementszinnen (bijv. ik weet wat hij bedoelt),
    - Infinitiefzinnen (bijv. om te helpen).

    De score wordt vergeleken met teksten met weinig bijzinnen (vmbo-schoolboeken) en veel bijzinnen (wetenschappelijke artikelen).

9. **Zinslengte:** LiNT berekent ook de **gemiddelde zinslengte** in woorden. Pander Maat en Dekker (2016) verklaren echter dat zinslengte minder relevante graadmeter is dan bovengenoemde zinsbouwkenmerken. Een zin wordt door LiNT herkend aan leestekens zoals een punt, puntkomma, vraagteken of uitroepteken. Dubbele punten en haakjes tellen volgens LiNT niet als zinsgrens.

Hoe persoonlijk is uw tekst?

10. **Mensen:** LiNT kijkt tevens naar hoe vaak in de tekst verwijzingen naar mensen voorkomen (per 1000 woorden). Teksten met meer persoonlijke verwijzingen zijn doorgaans makkelijker te lezen. LiNT telt: persoonlijke en bezittelijke voornaamwoorden (zoals jij, jouw), persoonsnamen (zoals Klaas) en zelfstandige naamwoorden die mensen aanduiden (zoals voetballer). De score wordt vergeleken met persoonlijke teksten (zoals roddelberichten) en onpersoonlijke teksten (zoals verkiezingsprogramma's).

11. **Persoonlijke voornaamwoorden:** LiNT telt het aantal persoonlijke en bezittelijke voornaamwoorden per 1000 woorden (zoals ik, jij, wij, mijn, jouw). Een hoger gebruik wijst op een persoonlijkere en vaak beter leesbare tekst. De score wordt vergeleken met persoonlijke tekstsoorten (zoals roddelberichten) en onpersoonlijke (zoals wetenschappelijke artikelen).

12. **Aansprekingen:** LiNT meet hoe vaak de lezer rechtstreeks wordt aangesproken via voornaamwoorden in de tweede persoon (zoals jij, u, jouw, uw), per 1000 woorden. Zulke aanspreekvormen maken een tekst persoonlijker en vaak toegankelijker.
De score wordt vergeleken met tekstsoorten die veel gebruik maken van directe aanspreking (zoals medische adviezen) en die dat nauwelijks doen (zoals wetenschappelijke artikelen).

13. **Actieve werkwoordsvormen**: LiNT meet welk percentage van de vervoegde werkwoorden in de **actieve vorm** staat. Actieve zinnen (bijvoorbeeld: u neemt dit middel in) zijn duidelijker en persoonlijker dan passieve zinnen (bijvoorbeeld: dit middel wordt ingenomen), waarin de handelende persoon vaak ontbreekt. Een tekst met minder dan 90% actieve vormen kan onpersoonlijk of afstandelijk overkomen. De score wordt vergeleken met actieve teksten (zoals roddelberichten) en passieve teksten (zoals wetenschappelijke artikelen).

Bijlage E Data kwantitatieve analyse

Microsoft Excel
97-2003 Worksheet

Om het bestand te openen, klik op rechtermuisknop → object → openen.