Vertelduivels geteld

Een verkennend computationeel onderzoek naar focalisatie

Jurrian Kooiman (0548790)

Masterscriptie Nederlandse literatuur en cultuur (UU)

Begeleider: dr. Sven Vitse

Tweede lezer: dr. Folgert Karsdorp

30 juni 2023

Samenvatting

In deze scriptie wordt het narratologische concept focalisatie in hedendaagse Nederlandse literatuur verkennend onderzocht op systematische en kwantitatieve wijze. Door middel van een analyse op een gedetailleerd tekstniveau en voor grote delen van een tekst wordt getracht een completer beeld van focalisatie te bewerkstelligen. Daarnaast zou het automatisch herkennen van focalisatie het mogelijk maken om de diachrone ontwikkeling van het concept te bestuderen. De hoofdvraag in dit onderzoek is: 'Hoe kan focalisatie computationeel onderzocht worden?'

Hiertoe is focalisatie beschreven in vier aspecten: positie van het focaliserende subject, waarneembaarheid van het gefocaliseerde object, aantal objecten en gedetailleerdheid van de waarneming. Een corpus van 1000 romanfragmenten is vervolgens met de hand geannoteerd voor deze aspecten. Enerzijds diende het annoteren als *close reading* met eigen onderzoeksresultaten, anderzijds is op basis van de annotaties een computermodel getraind op het herkennen van de verschillende aspecten van focalisatie. Vervolgens zijn verschillende manieren vergeleken om de fragmenten numeriek te representeren en zijn logistische regressie en het Nederlandstalige BERT-model ingezet voor de training.

Uit de resultaten van de systematische lezing is gebleken dat interne waarnemers bijzonder vaak voorkomen, terwijl de gedachten van personages niet zo sterk aanwezig zijn als zou worden verwacht van interne waarnemers. Ook worden gefocaliseerde objecten in hoge mate met veel detaillering waargenomen. Gedetailleerdheid blijkt daarbij in hoge mate vatbaar voor interpretatie. Ambiguïteit is daarnaast eerder regel dan uitzondering wanneer focalisatie systematisch wordt onderzocht. Bovendien is de schaal van analyseren een belangrijke factor; wordt focalisatie bestudeerd op hoofdstukniveau, dan levert dat andere resultaten op dan voor fragmenten van 150 woorden.

Het automatisch herkennen van onder meer interne focalisators komt dicht in de buurt van de menselijke oordelen. Het aspect van waarneembaarheid leent zich bovendien goed voor toepassing op onbekende teksten. Een aantal aspecten scoort echter minder goed, waarmee onder meer de complexiteit van lees- en begripsprocessen wordt bevestigd. Gezien de verkennende aard van dit onderzoek leveren de modellen al met al een positief resultaat op. Bovendien maakt dit onderzoek inzichtelijk hoezeer de menselijke hand voortdurend aanwezig is in computationeel onderzoek en dat de schijn van objectiviteit van computers vaak onterecht is.

Inhoudsopgave

1. Inleiding

‘Elke gids heeft zijn gebreken en kwaliteiten, zijn goddelijke kracht en zijn duivelse zwakten. Dat betekent dat we de traditionele verteltheorie verwerpen noch verafgoden’ (p. 18), schrijven Luc Herman en Bart Vervaeck met een knipoog in het woord vooraf in *Vertelduivels* (2009a). Waar binnen de structuralistische theorie als narratologische benadering vaak is geprobeerd een ‘supertheorie’ op te stellen, richten zij zich op het overzicht van ‘vele gidsen die je kunt volgen bij het omgaan met verhalen.’ (p. 18) Deze scriptie wil daar een extra gids aan toevoegen door de narratologische theorie aan te vullen met een computationele benadering. Het betreft echter een verkenning, we weten nog niet waar deze gids ons zal brengen. De relevantie en waarde van dit onderzoek moeten dan ook met name gezocht worden in het experiment, dat wellicht de weg vrijmaakt voor een nieuwe vertelduivel binnen de narratologie.

1.1 Onderzoeksvraag

In deze scriptie wordt focalisatie in hedendaagse Nederlandse literatuur onderzocht op systematische en kwantitatieve wijze. Met ‘systematisch’ bedoel ik dat focalisatie voor grote delen van een tekst wordt geanalyseerd, ‘kwantitatief’ behelst de stap om focalisatie te formaliseren door het concept te vatten in een beperkt aantal aspecten. Op basis van het annotatieschema dat deze stap oplevert, kan focalisatie met de hand worden geannoteerd. Door focalisatie als het ware te beschrijven door middel van annotaties, kan vervolgens worden onderzocht of het kwantificeren van focalisatie standhoudt zodra een computermodel wordt getraind om de aspecten van focalisatie te herkennen. De annotaties dienen namelijk twee doelen: ze fungeren als trainingsmateriaal voor het computermodel en ze genereren resultaten die op zichzelf al het bestuderen waard zijn.

Dit onderzoek heeft dus enerzijds methodologisch het doel om een narratologisch concept, focalisatie, automatisch te herkennen. Anderzijds hoopt dit onderzoek door zijn methode een vernieuwende blik te werpen op focalisatie om zodoende meer te weten te komen over het functioneren van focalisatie. Daarom vraag ik me in deze scriptie af: ‘Hoe kan focalisatie computationeel onderzocht worden?’ Het woord ‘computationeel’ impliceert dat er een stap van formaliseren heeft plaatsgevonden van het concept die een kwantitatief onderzoek mogelijk maakt. Zodoende luidt de eerste deelvraag: ‘In hoeverre kan focalisatie worden geformaliseerd ten behoeve van digitaal onderzoek?’ Focalisatie als narratologisch fenomeen wordt daarnaast onderzocht in de tweede deelvraag, waar de methodologische achtergrond in

doorklinkt: ‘Welke toevoeging aan de narratologie biedt het systematisch onderzoeken van focalisatie?’

Ter beantwoording van beide deelvragen onderscheid ik op basis van de bestaande narratologische literatuur vier aspecten van focalisatie. Deze aspecten vormen samen een beschrijving van focalisatie en kunnen worden uitgedrukt in een cijfer, waarmee de stap van formaliseren wordt gezet. Vervolgens worden deze aspecten met de hand geannoteerd voor vier Nederlandstalige romans. Dit handmatig labelen van focalisatie levert enerzijds op zichzelf al onderzoeksresultaten op over het systematisch onderzoeken van focalisatie. Anderzijds heeft de annotatietaak een methodologisch doel. Op basis van de annotaties wordt namelijk een computermodel getraind op het herkennen van de verschillende aspecten van focalisatie. De mate waarin een computermodel hiertoe in staat is, beantwoordt vervolgens de verkennende vraag of focalisatie digitaal kan worden onderzocht. Dit zou, indien mogelijk, veel nieuwe, relevante onderzoeksmogelijkheden voor de narratologie opleveren. Daarnaast leert het computermodel ons als menselijke lezer met een andere blik naar focalisatie kijken. Het gaat dan om patronen in de fouten die het model maakt bij het herkennen van focalisatie die ons nieuwe informatie geven over ons eigen leesproces, zoals bij teksten waarin de focalisatie ambigu is en de menselijke blik wellicht beperkt is.

1.2 Relevantie van focalisatie

Focalisatie gaat over de relatie tussen wie of wat er waarneemt in een verhaal en dat wat wordt waargenomen (Bal 2009). De waarnemende instantie komt in sommige verhalen overeen met de verteller, maar in veel verhalen vallen de verteller en de focalisator niet (helemaal) samen. Denk bijvoorbeeld aan een verteller die als het ware buiten een verhaal staat of ‘boven het verhaal zweeft’ en gedurende een verhaal steeds wisselt tussen verschillende waarnemende personages. En wanneer een ik-verteller terugblikt op gebeurtenissen uit zijn of haar jeugd is soms onduidelijk wie er waarneemt, de hedendaagse ‘ik’ of de ‘ik’ in het verleden? Het benoemen van de focalisatie geeft inzicht in hoe de informatie die de lezer krijgt gekleurd is. De blik die de lezer wordt gegeven op een bepaalde situatie is afhankelijk van wie of wat er waarneemt. Focalisatie biedt daarom een belangrijke benadering van verhalen om bijvoorbeeld ideologie en representatie te onderzoeken. Zo bespreekt Maaike Meijer in *In tekst gevat* (1996) een fragment uit Willem Elsschots gedicht ‘De klacht van de oude’ om die blik te tonen. De mannelijke ik-verteller geeft daarin een ‘opsomming van levenloze, verhandelbare dingen (kleren, een huis)’ (p. 13) om daar vervolgens zijn vrouw aan toe te voegen. Meijer laat daarbij

zien dat de vrouw onder de blik van de man wordt gemaakt tot levenloos ding dat verkocht kan worden. Het feit dat de vrouw in dit voorbeeld zelf niet waarneemt maar slechts wordt waargenomen draagt bij aan het eenzijdige beeld dat de lezer krijgt van de vrouw. De relevantie van focalisatie wordt door Meijer dan ook als volgt samengevat: 'Omdat de distributie van focalisatie tevens bepaalt wie de macht heeft in het verhaal (wie ziet en/of spreekt, en wie wordt gezien en besproken?) is het vaststellen van de focalisatie onmisbaar in de analyse van de wijze waarop een tekst sekse representeert.' (p. 12) In het verlengde hiervan kunnen we stellen dat focalisatie onmisbaar is voor veel letterkundige analyses. Of nu de betrouwbaarheid van de aangeboden informatie, representatie of ideologische dimensies van een verhaal worden bestudeerd; vaststellen 'wie de macht heeft in het verhaal' is relevant in talloze vertellingen en focalisatie biedt daartoe een belangrijk narratologisch handvat.

1.3 Relevantie van kwantitatieve benaderingen van focalisatie

1.3.1 Focalisatie kwantitatief en systematisch geanalyseerd

In dit onderzoek benader ik focalisatie op kwantitatieve en systematische wijze. De relevantie hiervan zal ik laten zien door een spanning bloot te leggen in het bestaande onderzoek naar focalisatie. Ik geef een voorbeeld uit een narratologische analyse van Luc Herman en Bart Vervaeck (2009b). In hun artikel over *The Echo Maker* (2006) van Richard Powers schrijven ze het volgende: 'The short fifth part of the novel starts with external focalization, then moves to Karin as focalizer and finally to Weber. Mark is left out.' (p. 416-417) De focalisator lijkt dus per paragraaf vast te staan. Als we het deel waarin Karin focaliseert nader bestuderen, blijken er echter zo nu en dan zinnen of zinsdelen op te duiken waarin de waarneming niet direct aan Karin kan worden toegeschreven, bijvoorbeeld het volgende citaat: 'Mark looks at her, bewildered. Some possibility lifts him up. His own loss means nothing. The accident gives him this. "Ask her," he begs. Afraid to suggest even this little.' (p. 446) In de eerste zin is het Karin die Mark focaliseert, vanaf de tweede zin wordt dit echter minder zeker; is het Karin die invult wat Mark voelt of springt de focalisatie wellicht vlug naar Mark? Van de daaropvolgende conclusie in de derde zin en relativering in de vierde zin is ook onzeker van wie die zijn, de focalisator kan niet eenduidig worden toegewezen. Wanneer Mark echter begint te spreken en daarna een gedachte volgt, herhaalt zich een patroon dat veel voorkomt in deze paragraaf: Karin hoort Mark spreken en reageert daar in gedachten op. De focalisatie lijkt dan dus weer volledig bij Karin te liggen. Passages als deze waarin de focalisator ineens instabiel lijkt, komen een

aantal keer voor in deze paragraaf en ze onderbreken de delen waarin door Karin wordt gefocaliseerd.

Deze onzekerheid over de focalisator ligt in lijn met de vaststelling van Herman en Vervaeck over het verdwijnen van centra van focalisatie in *The Echo Maker*: 'It is a novel that seems to give up characters as the centers of focalization. In the mind of the reader, the network of mindreading becomes an almost independent set of connections' (p. 419) Er zit echter een spanning tussen deze uitspraak en die waarin Herman en Vervaeck wel degelijk een focaliserende instantie aanwijzen, zoals de aanduiding dat de paragraaf uit het vijfde deel door Karin wordt gefocaliseerd. Dit heeft te maken met de schaal waarop focalisatie wordt bestudeerd; bekijken we de tekst op paragraafniveau dan lijkt de focalisatie eenduidig, zoomen we in op enkele zinnen dan is dit beeld ineens minder homogeen. Een narratologische analyse is dan ook afhankelijk van de schaal en afstand van de analyse ten opzichte van de tekst. Overigens schrijven Herman en Vervaeck in *Vertelduivels* (2009a) ook over 'de vraag naar de begrenzing van de eenheden die je onderzoekt' (p. 10). In deze scriptie wil ik dan ook onderzoeken wat er gebeurt wanneer we focalisatie op een gedetailleerder tekstniveau bestuderen. Ik kom hier in het methodehoofdstuk op terug.

Daarnaast vraag ik me af in hoeverre een narratologische analyse een volledig beeld geeft van focalisatie in een besproken tekst. Gewoonlijk wordt een tekst geanalyseerd aan de hand van representatieve tekstfragmenten die focalisatie eenduidig weergeven, terwijl bovenstaand citaat uit *The Echo Maker* laat zien dat focalisatie in sommige fragmenten wellicht helemaal niet zo eenduidig te benoemen valt. We weten eigenlijk niet hoe focalisatie eruit ziet als niet alleen de eenduidige fragmenten maar alle fragmenten worden gebruikt als uitgangspunt voor een analyse. Ik veronderstel daarom dat we een completer beeld zouden krijgen van focalisatie als de analyse wordt uitgevoerd op een gedetailleerder tekstniveau en als dat wordt uitgevoerd voor grotere delen van de tekst. Om de bestaande narratologische traditie van focalisatie aan te vullen voer ik daarom in deze scriptie een verkennend onderzoek uit dat focalisatie systematisch (ter aanvulling op de exemplarische *close reading*) en voor een lager tekstniveau (beneden paragraafniveau) bestudeert.

Dit onderzoek positioneert zich dan ook nadrukkelijk *naast* bestaande narratologische analyses. Door focalisatie systematisch te bestuderen kan in dit onderzoek de vraag worden gesteld hoe focalisatie er door de hele tekst heen uitziet. Dit zou me in staat stellen om focalisatie te onderzoeken zonder dit te verbinden met een inhoudelijke interpretatie van de tekst zelf. Daar ben ik in dit onderzoek dan ook benieuwd naar. Uit de analyse van Herman en Vervaeck blijkt bijvoorbeeld dat externe focalisatie minder vaak voorkomt dan interne

focalisatie, maar de extern gefocaliseerde passages blijken vervolgens sleutelpassages te zijn voor hun narratologische analyse. Heeft externe focalisatie vaker deze functie of leent *The Echo Maker* zich hierom juist goed voor een narratologische analyse? We kunnen dit eigenlijk niet weten zonder focalisatie doorheen de gehele tekst te bekijken en te vergelijken met focalisatie in andere teksten, want mogelijk zijn de teksten die narratologisch worden geanalyseerd niet representatief voor focalisatie in een bepaalde periode. Een systematische en kwantitatieve benadering is wellicht beter in staat om lokale narratologische verschijnselen breder te duiden.

1.3.2 Diachroon narratologisch onderzoek

Een belangrijke motivatie om focalisatie computationeel te onderzoeken bestaat eruit dat het de deur kan openen voor diachroon onderzoek naar focalisatie en – in het verlengde hiervan – andere concepten uit de narratologie. Diachrone narratologie richt zich op de beschrijving en analyse van de geschiedenis van narratieve vormen in de literatuur (De Jong 2014). Het richt zich dus op hoe narratologische concepten door de tijd heen op verschillende manieren zijn gebruikt in onder meer romans.

Een referentiepunt voor de diachrone benadering van narratologie is het artikel 'The Diachronization of Narratology' (2003) van Monika Fludernik. Ook zij geeft aan dat, ondanks de vele nieuwe benaderingen die de narratologie door tijd heen heeft zien passeren (denk bijvoorbeeld aan cognitieve benaderingen of de analyse van wat heet 'nieuwe media'), er weinig aandacht is voor de 'history of narrative forms and functions' (p. 331). Dit zou volgens haar een groot aantal nieuwe, relevante onderzoeksvragen kunnen opleveren, waaronder het verkrijgen van inzichten in de historische ontwikkeling van de relatie tussen auteur en verteller, of in hoeverre narratieve functies zoals het aanspreken van de lezer door de eeuwen heen constant is gebleven of juist een andere functie heeft gekregen. Deze ontwikkelingen, waarbij variaties doorheen bijvoorbeeld genre en tijd kunnen worden onderzocht voor een functie of concept, noemt ze het vergelijken van continuïteiten en discontinuïteiten van vorm en structuur (p. 333). In het verlengde hiervan liggen uiteraard vragen over culturele evolutie; welke vormen en functies die in een bepaalde tijd populair zijn 'overleven' en worden twee eeuwen later nog gebruikt en welke overleven in aangepaste vorm? Hetzelfde geldt voor basisbegrippen uit de verhaalanalyse, zoals focalisatie. Wanneer werden welke categorieën van een concept bijvoorbeeld voor het eerst gebruikt en op welk punt werd een andere categorie dominant? Fludernik noemt in dit verband ook het begrip 'refunctionalization' voor de ontwikkeling waarbij verteltechnieken een andere functie krijgen binnen de structurering van een verhaal (p. 334). Ze geeft in haar artikel het voorbeeld van de analepsis, dat aanvankelijk werd gebruikt

om terug te keren naar een eerdere scène, later de functie kreeg van een narratieve pauze zodat de verteller meer de ruimte kreeg en tot slot is verworden tot een meer ironische functie met ruimte voor metacommentaar door de verteller.

Diachrone narratologie lijkt aldus een geschikt middel om een brug te slaan tussen de narratologie en computationeel onderzoek. De beschrijving van de ontwikkeling van literaire vormen herbergt immers al een kwantitatieve component, die zich mogelijk op grotere schaal laat toepassen met behulp van computationele technieken. Als eerste stap in de richting van een synthese zal ik de onderzoekstraditie van de digital humanities en 'cultural analytics' introduceren.

1.4 Digital humanities en cultural analytics

De tweede traditie waar mijn onderzoek, naast de narratologie, bij aansluit is die van de digital humanities. Hoewel verwante begrippen als 'distant reading' of 'digital literay studies' net andere aspecten van het onderzoek benadrukken, gebruik ik ze in dit onderzoek redelijk inwisselbaar. In de kern hebben alle drie gemeen dat het gaat om geesteswetenschappelijk of letterkundig onderzoek met een kwantitatieve dimensie dat wordt uitgevoerd op grote(re) schaal, vaak met behulp van een computer. Tegelijkertijd ga ik hiermee voorbij aan de onderlinge verschillen in ontstaansgeschiedenis. Zo wijst Ted Underwood (2017) op het feit dat distant reading niet per definitie iets met technologie te maken heeft, maar vooral draait om 'the practice of framing historical inquiry as an experiment' (p. 2), waarmee hij het raakvlak tussen letterkunde en sociale wetenschappen benadrukt. Vanuit die invalshoek hebben studies onder de noemer 'distant reading' gemeen dat ze literaire trends over langere periodes onderzoeken, ondersteund door tekstueel bewijsmateriaal. Daarmee gaat de geschiedenis van distant reading volgens Underwood verder terug dan de toevoeging van computers aan de onderzoekspraktijk van letterkundigen. Doordat een term als 'digital humanities' deze toevoeging wel impliceert, wordt volgens Underwood de experimentele methode van het onderzoek gereduceerd tot een dialoog tussen geesteswetenschappers en machines.

In het onderzoek waarin cultuur en data samenkomen is het bovendien belangrijk om de term 'cultural analytics' te noemen. Onderzoek onder deze noemer concentreert zich voornamelijk in *Journal of Cultural Analytics*, waarvan Andrew Piper de hoofdredacteur is. Het uitgangspunt van deze tak van onderzoek, zo schrijft Piper in het eerste artikel van het tijdschrift, is de wederzijdse invloed van computationeel en cultureel onderzoek. Zo stelt computationeel onderzoek vragen als: 'What counts as evidence? What is the relationship

between theory and practice?' (Piper 2016, p. 2) Tegelijkertijd bevraagt cultureel onderzoek het universalisme en de veronderstelde neutraliteit van computationele toepassingen: 'Putting culture into computation cautions us to remember where we are when we think we know something.' (p. 2) Onderdeel hiervan is volgens Piper dat bijvoorbeeld verschillende computationele modellen worden vergeleken 'showing the extent to which different algorithms generate different kinds of meaning' (p. 8) Dit is een interessante opmerking in het licht van de invloed die schaal heeft op de uitkomsten van narratologische analyses, zoals we eerder in deze inleiding zagen. Hier kom ik dan ook nog op terug in het vervolg van deze scriptie.

Een voorbeeld van een computationeel onderzoek dat zich positioneert in de hoek van cultural analytics en dat tevens voortborduurt op de narratologie is het proefschrift van Roel Smeets, *Character Constellations* (2021). Aan de hand van semi-automatisch gegenereerde personagenetwerken onderzoekt Smeets de representatie van verschillende sociale groepen in Nederlandstalige literatuur. Interessant in het licht van dit onderzoek is dat hij zijn digitale methode inspireert op inzichten uit de narratologie. Zo legt hij in het hoofdstuk over centraliteit, de mate waarin een personage centraal staat in een personagenetwerk, de verbinding met focalisatie. Het punt van vergelijking is daarbij dat zowel een personagenetwerk als focalisatie geschikt is als middel om machtsverhoudingen in een vertelling te analyseren. Hoewel ik in deze scriptie een andere methode hanteer, slaat *Character Constellations* een brug tussen narratologie en cultural analytics en dient het daarin als voorbeeld voor mijn scriptie.

1.5 Overzicht van de inhoud

In het vervolg van dit onderzoek werk ik stap voor stap toe naar een antwoord op de vragen die ik aan het begin van de inleiding heb geformuleerd. In het theoretisch kader (hoofdstuk 2) maak ik duidelijk op welke onderzoeken ik voortborduur. Als eerste schets ik de achtergrond van focalisatie om vervolgens op basis van de bestaande narratologische literatuur een aantal aspecten van focalisatie te destilleren. Ten tweede plaats ik mijn streven om focalisatie kwantitatief te bestuderen binnen bestaande debatten rondom modelleren in de geesteswetenschappen. In het omvangrijke derde hoofdstuk beschrijf ik de methode van dit onderzoek. Daarin zal ik in gaan op het proces om een letterkundig concept, via kwantificerende stappen, meetbaar te maken. Vervolgens leg ik uit hoe de computationele stappen van mijn scriptie eruit zien. Hoewel enige technische passages hierbij onvermijdelijk zijn, richt ik me in dit onderzoek op de letterkundige lezer. Dit houdt in dat ik ervoor kies me vaak te beperken tot een conceptuele uitleg van bijvoorbeeld de *machine learning*-technieken die ik toepas zonder

daarbij in te gaan op de wiskundige achtergrond ervan. Ook onderdeel van het methodehoofdstuk zijn de corpusbeschrijving, het opstellen van een annotatieschema en het berekenen van de inter-beoordelaarsbetrouwbaarheid. Tot slot bespreek ik de classificatiemodellen en evaluatiecriteria die ik hanteer in dit onderzoek. In hoofdstuk 4 presenteer ik vervolgens de resultaten van mijn onderzoek. Dit omvat twee delen: eerst interpreteer ik de resultaten op basis van de annotatietaak, daarna analyseer en evalueer ik de resultaten van het classificatiemodel. Als laatste vat ik mijn bevindingen samen in de conclusie (hoofdstuk 5), die ik afsluit met een discussie waarin ik reflecteer op de beperkingen van mijn onderzoek en een aantal voorstellen doe voor vervolgonderzoek.

2. Theoretisch kader

Dit hoofdstuk bestaat uit twee delen die ieder voor een pijler van deze scriptie staan. In de eerste paragraaf richt ik me op het begrip focalisatie, dat ik achtereenvolgens definieer en formaliseer door een viertal aspecten te selecteren. Daarna zal ik in de tweede paragraaf enkele reflecties op modelleren in de letterkunde bespreken om vervolgens dit onderzoek te positioneren binnen het geschetste kader. Ik kies ervoor om eerst een letterkundig en narratologisch kader te schetsen en pas daarna mijn blik te richten op het modelleren. Dit heeft als reden dat ik eerst het letterkundige onderzoeksobject helder wil definiëren alvorens over te gaan op theorie over de methode, die in principe in dienst staat van het object. Tegelijkertijd staat in dit onderzoek de methodologische component dusdanig centraal dat die onderdeel hoort te zijn van het theoretisch kader.

2.1 Focalisatie

Allereerst zal ik focalisatie als concept beschrijven door de theoretische achtergrond te laten zien. Vervolgens licht ik vier aspecten uit die in dit onderzoek focalisatie zullen representeren. Na een definitie te hebben gegeven van focalisatie selecteer ik dus een aantal relevante aspecten van het concept. Dit dient als voorbereidend werk voor de volgende stap van dit onderzoek, namelijk het operationaliseren van focalisatie, waarmee ik het methodehoofdstuk zal beginnen (hoofdstuk 3.1).

2.1.1 Definitie

Het concept focalisatie kent enerzijds consensus wat betreft definitie en gebruik, anderzijds is er twijfel over de ruimte voor interpretatie en zijn sommige categoriseringen niet zo eenduidig als ze op het eerste gezicht lijken. Aangezien focalisatie een decenniaoude wetenschappelijke traditie kent die nauwkeurig staat beschreven in de verschillende handboeken over narratologie, lijken deze handboeken mij geschikt ter inkadering van het begrip focalisatie. In wat volgt schets ik de consensus die bestaat met betrekking tot de definitie en werking van focalisatie. Het eerste deel van deze paragraaf beoogt voort te bouwen op de bestaande literatuur en zodoende mijn eigen interpretatie vooralsnog op de achtergrond te houden. In het tweede deel zal ik een selectie maken van vier aspecten waarmee ik focalisatie denk te kunnen formaliseren. In het tweede deel voeg ik daarom mijn eigen interpretatie toe, hoewel ik ook hier nog voortbouw op aspecten van focalisatie zoals die zijn beschreven in de theorie.

In haar boek *An Introduction to Narratology* (2009) plaatst Monika Fludernik focalisatie binnen het grotere classificatieschema van narratieven dat Gérard Genette (1980) opstelde. Het concept komt dan ook voort uit het structuralisme, dat met onder meer focalisatie een blijvende bijdrage heeft geleverd aan de narratologie. In het schema van Genette wordt focalisatie gebruikt in de betekenis van 'perspectief', waarmee hij een onderscheid maakt tussen de 'reflector' en de verteller, oftewel tussen de vragen 'wie ziet er?' en 'wie spreekt er?' (Fludernik 2009, p. 38 en p. 98; Rimmon-Kenan 2005, p. 72-73). Manfred Jahn vat deze tweedeling in zijn hoofdstuk over focalisatie in de *Cambridge Companion to narrative* (2007) als volgt samen: '*Narration* is the telling of a story in a way that simultaneously respects the needs and enlists the co-operation of its audience; *focalization* is the submission of (potentially limitless) narrative information to a perspectival filter' (p. 94). Focalisatie geldt in dit citaat dus als een filter dat de informatie in een vertelling kleurt. De metafoor van een filter maakt bovendien de geconstrueerdheid van focalisatie inzichtelijk; geen enkel verhaal is volledig gelijktijdig verteld, want verteller en waarnemer vallen nooit exact samen. In de praktijk gaat de lezer echter mee in deze constructie en voelt het alsof een personage het verhaal alsnog gelijktijdig kan waarnemen. In het handboek voor verhaalanalyse *Vertelduivels* (2009a) bouwen Luc Herman en Bart Vervaeck voort op het onderscheid tussen zien en spreken en richten ze zich in hun omschrijving van focalisatie vervolgens op de verschillende aspecten ervan. Focalisatie is dan 'een verhouding tussen een "object" dat waargenomen wordt en een "subject" dat waarneemt' (p. 75). Hierbij bepaalt het focaliserende subject wat de lezer aangeboden wordt. Ze gebruiken de term 'waarnemen' in plaats van 'kijken' of 'zien', omdat focalisatie het gebruik van meerdere zintuigen kan behelzen, waaronder ook denken en beoordelen.

Deze uitbreiding van Herman en Vervaeck is nadrukkelijk beïnvloed door de herziening van Genettes opvatting van focalisatie door Mieke Bal. Zij onderscheidt namelijk de waarnemende instantie en het waargenomen object, ook wel gebruikt als 'focalisator' en 'gefocaliseerd object' (Bal 2009, p. 147; Herman & Vervaeck 2009a, p. 75; Rimmon-Kenan 2005, p. 75). Deze tweedeling stelt impliciet de vragen 'wie neemt waar?' en 'wie wordt waargenomen?' over focalisatie. Doordat deze lezing van het concept de nadruk legt op focalisatie als relatie, pleit Bal ervoor beide polen van de relatie apart te bestuderen (p. 149). Als de focalisator samenvalt met een personage, zal dit de perceptie van de lezer beïnvloeden, aangezien deze dan wordt uitgenodigd om mee te gaan in waarnemingen van het focaliserende personage. De benadering van focalisatie als verhouding tussen waarnemer en waargenomene maakt inzichtelijk dat de lezer wordt gestuurd door de waarnemende instantie. Ook merkt Bal op dat de relatie van waarneming tussen focalisator en gefocaliseerde altijd aanwezig is in een

vertelling: ‘This slanted, or why not say the word, subjective nature of story-telling is inevitable’ (p. 145). De mogelijkheid om enerzijds verteller en waarnemer gescheiden te bestuderen en anderzijds focalisatie als relatie te benaderen zijn volgens Bal dan ook redenen voor het hanteren van het woord ‘focalisatie’ in plaats van bijvoorbeeld *point of view* en *(narrative) perspective*. Dit begrip is dan ook het gebruikelijke concept dat wordt gehanteerd wanneer visies of waarnemingen worden bestudeerd.

Genette (1980) richtte zich in zijn indeling nog enkel op de positie van de focaliserende instantie. Hij onderscheidde drie categorieën van de focalisator: zero focalisatie, interne focalisatie en externe focalisatie. Zero focalisatie slaat op de situatie waarin geen enkel personage focaliseert, namelijk als er een auctoriële verteller optreedt; interne focalisatie vindt plaats wanneer het perspectief van een personage dominant is op het intradiëgetische niveau; wanneer personages worden gefocaliseerd van buitenaf (extern gefocaliseerd object), heet dit externe focalisatie (Fludernik 2009, p. 38 en p. 102). Mieke Bal heeft erop gewezen dat deze indeling onlogisch is omdat zero focalisatie gesitueerd is op het extradiëgetische niveau van de vertelling, terwijl interne focalisatie plaatsvindt op het intradiëgtische niveau en externe focalisatie kan voorkomen op beide niveaus. Daarnaast zijn de termen ‘intern’ en ‘extern’ in de betekenis van ‘van binnenuit/buitenaf’ verwarrend, aangezien interne focalisatie van een personage – de lezer kan de gedachten van dat personage waarnemen – impliceert dat andere personages extern worden gefocaliseerd (Fludernik 2009, p. 38). Hiertoe stelt ze een andere invulling van interne en externe focalisatie voor, zero focalisatie neemt ze hierin niet mee.

Onder *interne focalisatie* verstaat Bal een focalisator die optreedt als actor binnen het vertelde, de focalisatie ligt dan dus bij een personage binnen de fabula. Vervolgens is er sprake van *externe focalisatie* wanneer de focalisator buiten het vertelde staat en dus niet direct is gebonden aan een personage (Bal 2009, p. 152). Het gaat hier dus nadrukkelijk niet om de vraag in hoeverre de binnenwereld van het gefocaliseerde object wordt weergegeven (zoals bij Genette), maar om de *positie* van de focalisator. Herman en Vervaeck wijzen erop dat interne en externe focalisatie geen absolute begrippen zijn en dat de indeling afhankelijk is van bijvoorbeeld het vertelniveau van een tekst en de mate van onderscheid tussen personage en verteller (Herman & Vervaeck 2009a, p. 77). Deze ambiguïteit in het onderscheiden van interne en externe focalisator wordt aanschouwelijk gemaakt door onder meer Shlomith Rimmon-Kenan (2005) en De Coux (2012) die beiden voorbeelden laten zien waarin een ik-verteller vanuit het heden terugblikt op ervaringen uit zijn jeugd. Er wordt dan steeds afwisselend gefocaliseerd door het (externe) ‘ik-nu’ en (interne) ‘ik-toen’ waarbij ambiguïteit in of zelfs

samensmelting van de waarneming kan optreden. Deze en soortgelijke waarnemingen noem ik in navolging van De Coux 'consonant' (p. 67).

Een andere belangrijke toevoeging is de notie dat interne en externe focalisatie onafhankelijk zijn van de gekozen persoonsvorm van de vertelling (Herman & Vervaeck 2009a, p. 77-78). Zo maakt het voor de focalisatie niet uit of een verhaal wordt verteld vanuit de eerste of derde persoon (Bal 2009, p. 151-152 en p. 161; Rimmon-Kenan 2005, p. 74-75); beide kunnen zowel intern als extern zijn gefocaliseerd. Een verhaal kan dus worden waargenomen door meerdere focaliserende subjecten, terwijl de verteller constant blijft. In sommige gevallen kunnen verteller en focalisator (ogenschijnlijk) samenvallen als een ik-verteller gelijktijdig vertelt – Herman en Vervaeck (2009a) noemen dit in navolging van Dorrit Cohn het 'consonante bewustzijnsverhaal' (p. 77)[1] – maar het is voor het bestuderen van de waarneming in een verhaal van belang deze twee actoren in principe los van elkaar te benaderen.

Een laatste onderscheid dat door Mieke Bal wordt gemaakt, gaat erover of het gefocaliseerde object waarneembaar dan wel niet-waarneembaar is (Bal 2009, p. 156; Fludernik 2009, p. 38). Dit houdt in dat voor een interne focalisator zijn of haar eigen zelf een waarneembaar gefocaliseerd object is, terwijl in de wereld om hem of haar heen slechts zichtbare objecten kunnen worden gefocaliseerd (Bal 2009, p. 156; Fludernik 2009, p. 38). Met andere woorden, objecten die buiten de focalisator zelf liggen, die 'echt' zijn, gelden als waarneembaar. Objecten die slechts waarneembaar zijn voor de focalisator en daarmee niet voor andere personages, zoals dromen en gedachten, gelden als niet-waarneembaar. In het geval van een externe focalisator wordt de grens tussen waarneembaar en niet-waarneembaar complexer: sommige externe focalisators zullen 'in het hoofd' van meerdere personages kunnen kijken, terwijl dit voor andere focalisators is beperkt tot één personage (bijvoorbeeld wanneer een ik-waarnemer terugblikt op zijn jongere ik).

Daarnaast merkt Bal op dat de combinatie van focalisator en gefocaliseerd object doorheen een tekst constant kan zijn of kan variëren (p. 153). Dit sluit aan bij de notie van *degree of persistence* ('standvastigheid') door Shlomith Rimmon-Kenan (2005), die hiermee doelt op de mate waarin focalisatie kan veranderen. Ze maakt daarbij een indeling tussen vaste focalisatie, variabele focalisatie voor twee soorten waarnemingen en meervoudige focalisatie voor meerdere waarnemingen (p. 78). Ze bouwt hiermee voort op Genette die deze verdeling

[1] Deze consonantie, waarbij verteller en waarnemer lijken samen te vallen, is een ander soort consonantie dan waar De Coux (2012) over schrijft. Zij beschrijft namelijk dubbelzinnigheid bij het bepalen van twee verschillende waarnemers.

maakte voor de focalisator (Herman & Vervaeck 2009a, p. 78). Volgens Rimmon-Kenan geldt dit echter voor zowel de focalisator als het gefocaliseerde object.

Om deze passage, waarin ik heb geprobeerd de consensus rondom het concept van focalisatie te schetsen, samen te vatten wil ik teruggrijpen op drie vragen die Mieke Bal (2009) opstelt op pagina 153 en die relevant zijn met betrekking tot focalisatie:

1. What does the character focalize: what is it aimed at?
2. How does it do this: with what attitude does it view things?
3. Who focalizes it: whose focalized object is it?

Hoewel alle drie de vragen grofweg refereren aan respectievelijk gefocaliseerd object, de relatie tussen focalisator en gefocaliseerd object, en de focalisator, overlappen ze in hun beschrijving van deze drie factoren. Zo vertelt de manier waarop een object wordt neergezet door een focalisator evengoed iets over de focalisator zelf. Het woord 'character' in de eerste vraag impliceert wellicht dat het hier louter gaat om interne (*character bound*) focalisatie, maar ik vat deze formulering in de bredere betekenis van 'focaliserende entiteit' op die zowel intern als extern kan zijn. In de volgende paragraaf, waarin ik vier aspecten selecteer, dienen deze drie samenvattende vragen als leidraad bij de gekozen aspecten.

2.1.2 Selecteren en formaliseren

Nu het concept focalisatie van een aantal kanten is belicht, kunnen we overgaan op het selecteren en formaliseren. Hoe kan focalisatie, dat zelfs door een getrainde lezer niet altijd eenduidig kan worden geanalyseerd, worden gevangen in enkele variabelen? Voor deze paragraaf is het vooralsnog voldoende om te weten dat ik in dit onderzoek geïnteresseerd ben in de vraag *of* ik een model kan samenstellen dat focalisatie kan classificeren. Dit impliceert dat ik niet zozeer op zoek ben naar het beste of meest complete model. In plaats daarvan zoek ik aspecten die redelijkerwijs focalisatie beschrijven zonder daarbij te streven naar een uitputtende beschrijving. Deze aspecten kunnen worden opgevat als variabelen aan de hand waarvan de focalisatie in tekstpassages kan worden beschreven. Met andere woorden, in deze paragraaf formaliseer ik focalisatie door enkele variabelen te selecteren en voor elke variabele enkele categorieën te onderscheiden. In de afsluitende paragraaf van dit hoofdstuk, waarin ik een theoretisch kader schets rondom het modelleren van een letterkundig concept, zal ik nader ingaan op mijn overwegingen om een concept te representeren door middel van een niet-uitputtende beschrijving.

De variabelen die zullen gelden als dimensies van focalisatie moeten dus uitdrukking geven aan de vragen van Mieke Bal. In wat volgt stel ik daarom vier aspecten op die hieraan

beantwoorden. In deze passage is dan ook in hoge mate mijn eigen hand zichtbaar. Bij de keuze voor de categorieën die ik gebruik maak ik namelijk een afweging tussen enerzijds redelijkerwijs voortbouwen op de bestaande literatuur en anderzijds intersubjectiviteit nastreven voor de annotatietaak. Bij het annoteren wordt namelijk gepoogd een aspect op zo'n manier te beschrijven dat dit aspect door verschillende annotatoren op een soortgelijke manier wordt beoordeeld. Het doel hierbij is om intersubjectiviteit tot stand te brengen, hetgeen een indicatie is van hoe goed de beschrijvingen van aspecten van focalisatie werken. Hier kom ik op terug in hoofdstuk 3.4, waarin ik het annotatieschema opstel. Onderstaande beschrijving van focalisatie aan de hand van vier dimensies van het concept dient als voorwerk voor het annotatieschema. Ik bespreek daarom de verschillende aspecten waarbij ik steeds duidelijk maak wat het aspect behelst, wat de narratologische achtergrond ervan is en uit welke categorieën het aspect bestaat.

Aspect 1: Positie van de focalisator

Het eerste aspect gaat over de meest 'klassieke' vraag binnen de relatie tussen waarnemer en waargenomene, namelijk de positie van de focalisator ten opzichte van de verhaalwereld. Herman en Vervaeck (2009a) wijzen op het continuüm dat loopt van 'totale zekerheid over wat wordt meegedeeld' tot 'totale twijfel' (p. 158). Ze voegen daaraan toe dat David Herman de structuralistische focalisatietypes op dit continuüm plaatst. Daarbij wordt bijvoorbeeld externe focalisatie geïnterpreteerd als een signaal van zekerheid, omdat het een afstand tot de lezer impliceert. Interne focalisatie zit meer aan de kant van de twijfel, aangezien bij een vaste interne focalisator 'de houdingen van meningen meestal in één mogelijke wereld [liggen] verankerd' (p. 159). Dit mogelijke gevoel van onzekerheid wordt versterkt bij meervoudige interne focalisatie. Ze schrijven dat David Herman de 'hypothetische focalisatie' het dichtst bij de pool van twijfel plaatst. Hiermee wordt het type waarnemer bedoeld waarvan het bestaan – als waarnemer – onzeker is, zoals onbezielde zaken (gebouwen, boeken). Volgens Herman en Vervaeck kan een twijfelachtige waarneming door de lezer juist ook worden herkend als een 'zeer conventionele manier waarop een verteller elementen aandraagt zonder de waarde van die elementen in twijfel te trekken.' (p. 159) Ter illustratie noemen ze de waarnemingen van bijvoorbeeld een dubbelganger die veeleer bevestigend dan ondermijnend werken met betrekking tot de ervaringen van de hoofdfiguur.

Hoewel de benadering van focalisatietypes als graden van onzekerheid weer nieuwe vragen oproept, lijkt de idee van externe en interne focalisatie als continuüm me interessant. Bovendien sluit dit aan bij de notie van Jan Christoph Meister & Jörg Schönert (2009) van

descriptive scalars voor de analyse van processen die een vertelling tot stand brengen. In de praktijk zal het gebruikelijke onderscheid tussen interne en externe focalisatie echter ook voor een continuüm leidend zijn. Daarom voeg ik deze twee categorieën hoe dan ook toe aan dit aspect van de positie van de focalisator. Bovendien bouw ik hiermee voort op de consensus die ik hierboven schetste. Daarnaast laten bovenstaande alinea over twijfel en het commentaar van Herman & Vervaeck (2009a), maar ook de al beschreven ambiguïteit bij het indelen van externe en interne focalisatie (De Coux 2012; Rimmon-Kenan 2005) zien dat een extra tussencategorie van consonantie een goede aanvulling is. Hierbij gebruik ik 'consonantie' in de betekenis van De Coux (2012), die het begrip invult als ambiguïteit tussen twee verschillende waarnemers. In haar artikel bespreekt De Coux slechts voorbeelden met waarnemingen van de ik-figuur op twee verschillende tijdstippen (toen en nu). Mijns inziens kunnen consonantie en ambiguïteit in principe plaatsvinden tussen eender welke twee waarnemers, dus ook wanneer de waarnemingen van bijvoorbeeld een heterodiëgetische verteller en die van een personage samenvallen.

De drie categorieën voor de positie van de focalisator worden duidelijker in de volgende zinnen. Wanneer een zin luidt: 'De zon brandt op zijn huid.', dan is dat in de meeste gevallen een interne waarneming van de 'hij' die de zon voelt branden. Als daarop volgt: 'Nooit zal hij de zon nog zo voelen branden op zijn huid.', wordt duidelijk dat de waarnemer waarschijnlijk vanuit het nu terugblikt op het verleden. De tweede zin is dan extern gefocaliseerd. Een waarneming is consonant in de tweede zin als er bijvoorbeeld staat: 'De zon brandt op zijn huid. Zijn pet ligt nog in de tent.' Het is niet direct duidelijk of de tweede zin wordt gefocaliseerd door de interne waarnemer van de eerste zin of dat er een externe waarneming (eventueel de verteller) aan te pas komt. Daarom zou deze tweede zin kunnen doorgaan voor ambigu.

Dit aspect waarin de positie van de focalisator wordt beschreven door middel van drie categorieën volgt enerzijds de narratologische theorie. Anderzijds kent het door zijn eenvoud een aantal mogelijke tekortkomingen. De belangrijkste bestaat uit het feit ik op basis van deze categorisering geen onderscheid kan maken tussen verschillende centra van waarneming. Zodra ik bijvoorbeeld weet dat er intern wordt gefocaliseerd, kan ik op basis van deze indeling dus niets zeggen over welke entiteit focaliseert en of de waarneming wisselt tussen twee interne focalisators. Voor deze scriptie acht ik het echter gerechtvaardigd om me te beperken tot het versimpelde weergave van de positie van de focalisator die me alsnog in staat stelt om te onderzoeken of focalisatie automatisch kan worden herkend. Daarnaast zal de keuze voor de fragmentgrootte van grote invloed zijn tijdens de annotatietaak, zoals ik in de inleiding stelde

naar aanleiding van de analyse van *The Echo Maker* door Herman en Vervaeck. Deze keuze licht ik daarom toe in de corpusbeschrijving (hoofdstuk 3.3).

Aspect 2: Waarneembaarheid van het gefocaliseerde object

Het tweede aspect gaat over de vraag of het gefocaliseerde object waarneembaar of niet-waarneembaar is. Hiermee ken ik een eerste eigenschap toe aan het object. Dit aspect beoogt mogelijke ongelijkheid in informatie en representatie tussen verschillende entiteiten in het verhaal inzichtelijk te maken. In navolging van Mieke Bal (2009), die de vraag van waarneembaarheid voorstelt, is dit aspect binair. Daarmee ben ik in staat om bijvoorbeeld dialogen te onderscheiden van gedachten.

Het belang van dit onderscheid tussen waarneembaar en niet-waarneembaar bestaat eruit dat het inzicht geeft in hoe personages worden gerepresenteerd. De ongelijkheid in informatie die de lezer krijgt over verschillende personages draagt bij aan de manier waarop de lezer ze beoordeelt. Dit is bijvoorbeeld het geval in de volgende passage in *De maaneter* (1980) van Hannes Meinkema. De hoofdpersoon uit in haar gedachten haar onbegrip over haar moeder die gedurende haar jeugd de doodsoorzaak van de vader van de hoofdpersoon heeft verzwegen: ‘Zelfmoord. Waarom heeft moeder dit nooit verteld? En nu pas, in al die jaren haar geheim. En jou stevig opgevoed, steviger dan Saartje’ (p. 126), hierna volgt nog een alinea vol gedachten van de hoofdpersoon over deze onthulling. Pas aan het einde wordt een kort zinnetje toegevoegd van haar moeder, waarna de gedachten van de hoofdpersoon weer overheersen: ‘Ik kon het niet zeggen, zegt ze, voordat ik wist dat het goed met je ging. Dit is typisch moeder, je wordt heel driftig: heb je er niet recht op, geweten te hebben wat er met je vader is gebeurd?’ (p. 126). De hoofdpersoon zet deze gedachtestroom ook na dit citaat nog voort, waarbij ze onder meer generaliseert over het gedrag van haar moeder tijdens haar opvoeding. In deze passage kan de moeder zich niet verdedigen en als lezer krijgen we slechts waarneembare informatie over haar, namelijk de woorden die ze uitspreekt tegen haar dochter. Dit staat in schril contrast met de niet-waarneembare informatie die we krijgen via de gedachten van de hoofdpersoon. De informatie-ongelijkheid in deze passage is dan ook van grote invloed op het beeld dat de lezer vormt van beide personages. Het aspect van waarneembaarheid maakt dus inzichtelijk in welke situaties representatie en informatie mogelijk van invloed zijn op het verhaal.

Aspect 3: Mate van standvastigheid

Dit derde aspect moet een bepaalde mate van dynamiek in de vertelling beschrijven. Rimmon-Kenan (2005) en Herman & Vervaeck (2009a) spreken over respectievelijk *degree of*

persistence (p. 78) en standvastigheid (p. 78), waarmee ze beide bedoelen in hoeverre een verhaal wordt weergegeven vanuit de waarneming van één of juiste meerdere instanties. Rimmon-Kenan voegt hieraan toe dat dit geldt voor zowel de focalisator als voor het gefocaliseerde object. In dat laatste geval is de vraag dus in hoeverre in een verhaal of fragment één of meerdere objecten worden waargenomen.

Voor zowel subject als object geldt de volgende verdeling, zoals beschreven door onder meer Herman & Vervaeck: vaste focalisatie voor één waarnemende/waargenomen instantie, variabele focalisatie voor twee personages die elkaar afwisselen en meervoudige focalisatie indien er meer dan twee centra van waarneming zijn. Deze indeling wordt bij Rimmon-Kenan algauw wat complex, aangezien het gaat over zowel subject als object, en de standvastigheid van beide staat los van elkaar. Ik vat dit aspect dan ook niet op als soorten van focalisatie die precies moeten worden benoemd (waarom krijgen bijvoorbeeld twee centra van waarneming een aparte categorie en geldt dit niet voor drie of vier?). Ik wil daarom standvastigheid operationaliseren als een aspect dat dynamiek aangeeft en daarmee aanvullend werkt voor mate van detail (van hoeveel objecten wordt er detail uitgedrukt?) die hieronder wordt beschreven. Ik richt me dan ook op de standvastigheid van het gefocaliseerde object, omdat het in combinatie met de mate van detail mogelijk een goede beschrijving kan geven van het object. In de beschrijving van het annotatieschema ga ik nader in op de afbakening van de verschillende categorieën van dit aspect en op de vraag waar bijvoorbeeld de grens ligt tussen één of meer waargenomen objecten. Met het oog op het feit dat ik zal werken met fragmenten in plaats van gehele boeken lijkt me de standvastigheid van de focalisator voor dit onderzoek minder betekenisvol. Ik veronderstel namelijk dat de focalisator vaak op hoofdstukniveau zal wisselen, terwijl gefocaliseerde objecten ook op een lager tekstniveau vaker wisselen.

Aspect 4: Mate van detail van het gefocaliseerde object

Het laatste aspect gaat over de mate van detail van het gefocaliseerde object. David Herman pleit voor een bredere opvatting van focalisatie, omdat de menselijke waarneming wordt bepaald door de wereld waarin het lichaam is verankerd. Daarom gaat hij op zoek gaan naar de ‘processes and sub-processes involved in conceptualization’ (Herman 2009, p. 130), waarbij hij een aantal nieuwe parameters voorstelt die als aanvulling dienen op de klassieke focalisatietheorie zoals aan het begin van dit hoofdstuk beschreven. De conceptualisering die wordt genoemd in het bovenstaande citaat slaat op het gegeven dat dezelfde situaties op verschillende manieren kunnen worden verpakt in een narratief, bijvoorbeeld door subject en object van een zin om te draaien. Volgens Herman zijn concepten als vertelperspectief en

viewpoint niet toereikend om deze verschillen te beschrijven, ze worden namelijk ook bepaald door 'temporal, spatial, affective, and other factors associated with embodied human experience.' (p. 128). De benaderingen die hij toevoegt, bepalen de aard en kwaliteit van de waarneming, bijvoorbeeld 'of ze gedetailleerd of schetsmatig is, of ze object of subjectief is, of ze een wijde of een nauwe focus heeft, of ze synoptisch (statisch) of sequentieel (dynamisch) is enzovoort' (De Coux 2012, p. 64).

Met de introductie van *granularity* drukt Herman de mate van detail uit waarmee het gefocaliseerd object wordt beschreven. Het dient als aanvulling op de afstand in tijd en ruimte tussen focaliserend subject en gefocaliseerd object, hoewel de mate van detail samenhangt met de afstand in tijd en ruimte: 'Scenes are also "sighted" from particular temporal and spatial directions, and viewpoints on scenes can be distal, medial, or proximal, that is, range from being far away to being up close. Each such distance increment, further, may carry a default expectation about the degree of granularity (or level of detail) of the construal.' (Herman 2009, p. 130). Gewoonlijk zal een beschreven situatie met een kleine afstand in tijd en/of ruimte een hoge mate van detail kennen: 'Closer perspectives on scenes generally yield finer-grained (=more granular, more detailed) representations' (130-131), en vice versa: 'more distant perspectives generally yield coarser-grained (=less granular, less detailed) representations' (p. 131). Het aspect van gedetailleerdheid van het gefocaliseerde object geeft dus uitdrukking aan de temporele en ruimtelijke afstand tussen focalisator en gefocaliseerd object. Tegelijkertijd omvat het ook andere vormen van afstand. Wanneer namelijk een personage vluchtig en met weinig detail wordt waargenomen, duidt dit binnen de verhaalwereld mogelijk op een machtsverhouding tussen subject en object van focalisatie.

De interpretatie van Marco Caracciolo in *Slow Narrative and Nonhuman Materialities* (2022), die gedetailleerdheid opvat als 'the level of detail and particularity with which characters' mental states are verbally portrayed' (p. 45) legt mijns inziens al te zeer de nadruk op de (interne) focalisator en de niet-waarneembare aspecten ervan, met als gevolg dat het aspect van representatie onderbelicht blijft. In mijn opvatting van gedetailleerdheid – en ik meen hiermee dichter bij Herman te blijven – wil ik juist de nadruk op representatie leggen. Hierdoor beschrijft de mate van detail namelijk het relationele aspect van de afstand in tijd en ruimte tussen focalisator en gefocaliseerd object.

Gedetailleerdheid is enerzijds interessant als dimensie van focalisatie vanwege de samenhang met de afstand tussen de focalisator en het gefocaliseerde, anderzijds stelt het de lezer in staat uitdrukking te geven aan eventuele dynamiek die optreedt in de focalisatie, die zowel kan gelden voor een veranderende afstand als voor een veranderende blik van

bijvoorbeeld een groep naar een individu. Concreet kan deze dimensie de volgende gradaties hebben: in navolging van Herman gebruik ik de verdeling hoge, gemiddelde en lage mate van detail. Voor de annotatietaak maak ik een vijfpuntsschaal van dit aspect, om het graduele aspect van gedetailleerdheid te benadrukken. Vergelijk ter illustratie de volgende twee voorbeelden, waarvan de eerste een lage mate van detail kent en het tweede een hoge:

> ‘Er staat al een groepje te wachten bij de bushalte. Het ziet er treurig uit.’
>
> ‘Daan staat al te wachten bij de bushalte. Hij heeft grote wallen onder zijn ogen, waarschijnlijk heeft hij weer niet geslapen.’

In het eerste voorbeeld wordt geen onderscheid gemaakt tussen de verschillende personen in het groepje bij de bushalte. Ook wanneer het geheel een eigenschap wordt toegekend, krijgen we niet meer te weten over de losse onderdelen van de groep. Dit voorbeeld heeft dus een lage mate van detail. In het tweede voorbeeld wordt één persoon beschreven, van wie zowel een eigenschap van zijn uiterlijk als zijn gedrag wordt beschreven. Daarom heeft dit voorbeeld een gemiddelde tot hoge mate van detail. Hoewel ook bij dit aspect waarschijnlijk de grootte van een fragment van invloed is bij de beoordeling, maken deze voorbeelden mogelijke variaties in de mate van detail duidelijk.

In deze eerste paragraaf van dit hoofdstuk, waarin ik het theoretische kader van deze scriptie beschrijf, heb ik eerst het concept focalisatie theoretisch ingebed om tot een definitie te kunnen komen. Op basis daarvan heb ik vervolgens in het tweede deel vier aspecten onderscheidden: de positie van de focalisator, de waarneembaarheid van het gefocaliseerde object, het aantal waargenomen objecten (standvastigheid) en de mate van detail van het gefocaliseerde object. Deze aspecten vormen een selectie van het concept van focalisatie; het gaat hier immers niet om een uitputtende beschrijving. Het eerste deel sloot ik af met drie vragen waarmee Mieke Bal focalisatie als relatie tussen het focaliserende subject en het gefocaliseerde object beschrijft. Deze eerste vraag (‘wat focaliseert het personage: op wat is het gericht?’) komt terug in het tweede, derde en vierde aspect, die alle een andere kant van het gefocaliseerde object beschrijven. De tweede vraag (‘Met welke attitude wordt er gekeken naar dingen?’) keert vooral in het vierde aspect terug doordat de mate van detail uitdrukking geeft aan een vorm van hiërarchie. Van een expliciete attitude is echter geen sprake. De tweede vraag van Bal is in mijn beschrijving van focalisatie dus in mindere mate vertegenwoordigd. De derde vraag over wie er focaliseert zit vervat in het eerste aspect en indirect in de andere drie aspecten aangezien waarnemingen wellicht meer over de focalisator vertellen dan over het gefocaliseerde object.

In het methodehoofdstuk ga ik nader in op de annotatietaak die voortbouwt op deze paragraaf. Daarin stel ik op basis van de vier aspecten die ik heb beschreven een annotatieschema op waarin ik laat zien hoe de aspecten concreet worden ingezet bij de vertaalslag van narratologisch concept naar numerieke representatie. Waar in deze paragraaf een literatuurstudie is uitgevoerd, dient het annotatieschema als operationalisering ervan. Hieronder volgt eerst nog een paragraaf waarin ik het modelleren van een letterkundig concept theoretiseer. Nu het concept zelf is gedefinieerd en aspecten zijn geselecteerd kunnen we ons namelijk gaan richten op de stappen die nodig zijn om focalisatie uiteindelijk automatisch te kunnen herkennen. Theorie over het model dat uiteindelijk hiertoe zal dienen is dan ook onmisbaar.

2.2 Modelleren

In de inleiding schreef ik al over Ted Underwood die beweert dat de vooruitgang in onderzoek onder de noemer 'distant reading' niet zozeer met technologische vooruitgang te maken heeft. Hij stelt namelijk dat dit vooral te maken heeft met nieuwe ideeën over modelleren en interpretatie. In zijn boek *Distant Horizons* (2019) is het uitgangspunt dan ook dat letterkundigen niet in staat zijn om zogeheten 'century-spanning trends' (ix) te bestuderen zo lang ze zich richten op een aaneenschakeling van bewegingen en periodes. Underwood laat daarom zien wat er zichtbaar wordt al we ons richten op het grotere plaatje, zoals de manier waarop *science fiction* zich als genre heeft ontwikkeld en hoe dit label niet altijd dezelfde betekenis heeft gehad. Er heeft volgens hem in het computationele onderzoek een verandering plaatsgevonden van het meten van variabelen naar 'framing models of literary concepts' (xi-xii). Onderzoek aan de hand van modellen bestudeert de relatie tussen verschillende variabelen en richt zich dus minder op losstaande feiten. Het gaat Underwood er dus om dat we als letterkundigen nadenken over hoe we de metingen die we verrichten in verband brengen met andere variabelen, zoals lezerspubliek, genre of personages: 'Instead of directly measuring the text, predictive models describe a relationship between social and textual evidence.' (p. 24) Aangezien modelleren een centrale plek inneemt in het denken over literaire concepten binnen een computationele context, is het van belang een kader te schetsen rondom dit concept. Daar richt ik me dan ook op in deze paragraaf. Ik bespreek drie artikelen waarin respectievelijk Richard Jean So (2017), Andrew Piper (2017) en Ted Underwood (2020) reflecteren op de betekenis van modellen in een letterkundige context. Het beeld dat hieruit naar voren komt zal ik vervolgens verhouden tot het model dat ik in deze scriptie wil gebruiken om focalisatie te

onderzoeken. Door het model dat ik in deze scriptie hanteer in te bedden in bestaande literatuur schets ik een theoretische achtergrond bij het methodehoofdstuk en de keuzes die ik daarin maak.

Richard Jean So begint zijn polemisch getitelde artikel '"All Models Are Wrong"' (2017) met het aforisme uit zijn titel, dat afkomstig is van de statisticus George E. P. Box. Volgens Box zijn modellen namelijk slechts nummers die de complexe werkelijkheid niet kunnen representeren. Een model stelt de onderzoeker echter in staat om bepaalde aspecten van een interessant fenomeen te isoleren om op die manier 'certain properties of such aspects' (p. 669) te ontdekken. Dit model kan daarop worden herzien om nieuwe inzichten te genereren. So ziet dit herhalende onderzoeken, waarbij een model steeds aanleiding geeft tot reflectie op hoe het een fenomeen representeert, als een cruciale eigenschap van een model. Het gaat volgens hem dan ook niet over de vraag of een model gelijk heeft. Het is belangrijker om te begrijpen welke fouten een model maakt: 'to understand how it is wrong' (p. 671). Inherent aan deze benadering is de wisselwerking tussen close en distant reading. Het analyseren van de fouten van een model kan immers alleen door de teksten te bestuderen waarmee het model de mist in gaat: 'close reading here is inseparable from recursively improving one's model' (p. 671). Fouten zouden niet vermeden moet worden, veeleer zouden ze onderdeel moeten zijn van het onderzoeksproces. Daarom is *close reading* volgens So een onmisbaar onderdeel van het modelleren.

Andrew Piper gaat in zijn reflectie 'Think Small: On Literary Modeling' (2017), die samen met de bijdrage van Richard Jean So werd gepubliceerd in een themanummer van *PMLA*, ook in op representatie door modellen. Door de mate van representatie centraal te stellen, wordt de 'unproblematic relation between data and the world' (p. 652) op losse schroeven gezet. Vervolgens worden relevante vragen mogelijk over hoe betekenis tot stand komt en over de relatie tussen de representatie en dat wat wordt gerepresenteerd. Piper stelt dat in de letterkunde de minste informatie verloren gaat wanneer een tekst letterlijk wordt geciteerd. Daar staat echter tegenover: 'a tremendous amount of information is lost in all the other aspects of the work that are not cited […]. This loss is close reading's greatest weakness' (p. 653). Een letterkundig model kent volgens Piper vijf 'lagen' waarvan *implementation* (hoe kunnen concepten worden gemeten?, p. 654) voor de huidige paragraaf de interessantste overwegingen bevat. Daarover schrijft hij namelijk dat er geen vrees zou moeten zijn voor reductie bij het meten van een fenomeen. In plaats daarvan zou de onderzoeker zich juist rekenschap van reducties moeten geven, omdat ze een noodzakelijk onderdeel zijn van generalisaties die uiteindelijk worden gedaan op basis van de uitkomsten (p. 654). Daarbij maakt hij een vergelijking met traditionele

generalisaties in de letterkunde: ‘Measurement is no more or less reductive than selecting a passage from a single author and having it stand for all European literature’ (p. 654). Het verschil zit volgens Piper daarin dat modellen inzichtelijk maken hoe de onderzoeker tot zijn beweringen komt.

Ted Underwood licht zijn kijk op modellen toe in ‘Machine Learning and Human Perspective’ (2020). Hij legt uit dat de scheidslijn tussen kwantitatieve en interpretatieve methodes wellicht niet zo vaststaat als vanuit de traditionele letterkunde wordt gedacht. Dit komt door de manier waarop algoritmes informatie ‘leren’: die zijn veeleer afhankelijk van voorbeelden in plaats van vaststaande definities, waardoor ze kunnen worden ingezet om onderliggende aannames te bestuderen (p. 93). Het gaat volgens Underwood dus om de manier waarop verschillende perspectieven kunnen worden verkend met behulp van *machine learning*. Refererend aan de reflecties van So en Piper stelt Underwood dat het doel niet is om vast te stellen wat doorgaat voor significante bevindingen. Het gaat erom een model te definiëren dat een ‘*relation* between measurements’ uitdrukt die betekenis krijgt door sociale context (p. 93). Zodoende kunnen de grenzen tussen categorieën worden bestudeerd. Underwood geeft een voorbeeld van grenzen van gender in taalgebruik, maar de categorieën van bijvoorbeeld de positie van de focalisator kunnen ook op deze manier worden onderzocht. Het modelleren van een relatie maakt het mogelijk om categorieën die geen vaststaande definitie hebben te onderzoeken: ‘The point of machine learning is exactly to model practices of categorization that lack a definition and can be inferred only from examples.’ (p. 97) Hij geeft het ‘klassieke’ voorbeeld van een spamfilter dat wordt getraind om de juiste mails als spam te markeren op basis van tekstuele kenmerken die specifiek zijn voor spam. We kunnen misschien geen precieze definitie van spam geven aan het model, maar door het trainen weet het uiteindelijk toch menselijk gedrag te reproduceren. De inherente bias die hierbij optreedt kan volgens Underwood productief worden ingezet door letterkundigen. Door teksten door menselijke lezers te laten labelen kunnen we namelijk (historische) perspectieven modelleren (p. 97). Modellen worden in dat geval ingezet om een object (genres, gender, focalisatie) te representeren als ‘practices of reception’ (p. 98), wat zoveel betekent als het oordeel van een bepaalde groep lezers inzetten als trainingsmateriaal. Deze kunnen vervolgens worden vergeleken met ‘practices of different eras’ (p. 98). Een model wordt dus op basis van de labels van een lezer getraind om teksten die het niet eerder heeft gezien ook te kunnen categoriseren (p. 98). De gedachte hierachter is dat teksten uit dezelfde categorie vergelijkbare talige patronen vertonen. Hierdoor hoeft het onderzoeksobject niet direct te worden gerepresenteerd, maar volstaat de tekst zelf.

De artikelen van Richard Jean So, Andrew Piper en Ted Underwood belichten verschillende aspecten van modelleren, maar ze komen in grote lijnen overeen qua visie. Voor het model dat ik samenstel om focalisatie te herkennen bieden ze dan ook goede aanknopingspunten. Zo sluit het punt van So over isoleren van aspecten van een fenomeen aan bij mijn benadering van focalisatie; ik kies ervoor vier aspecten te bestuderen waarvan ik veronderstel dat ze nieuwe inzichten kunnen geven over focalisatie. Vervolgens label ik een groot aantal tekstfragmenten door per aspect van focalisatie een categorie toe te kennen aan het fragment. Deze stap dient als trainingsmateriaal voor het model, dat hiermee in navolging van Underwood afhankelijk is van voorbeelden en niet zozeer van vaststaande definities. Uiteraard is de annotatieprocedure ingegeven door bestaande definities van focalisatie, maar door mijn perspectief te modelleren hoop ik te zien in hoeverre deze definities standhouden zodra ze systematisch worden toegepast. Bestaande categorieën van bijvoorbeeld de positie van de focalisator kunnen in het model vanuit een ander perspectief worden benaderd, aangezien het algoritme misschien op basis van andere elementen classificeert dan de menselijke lezer of dan de theorie voorschrijft. Of het algoritme inderdaad vanuit een ander perspectief classificeert, onderzoek ik door middel van deze methode. Hoewel de grenzen tussen categorieën van een aspect vanzelfsprekend lijken op basis van de theorie, omdat we ons neigen te richten op de eenduidige gevallen, is dit wellicht helemaal niet het geval als we de categorieën grondiger onderzoeken met behulp van automatische classificatie. Daarnaast tracht ik mijn particuliere perspectief (mijn eigen 'practice of reception') op focalisatie te overstijgen door mijn annotaties te vergelijken met die van twee andere lezers. Het annoteren behelst in feite een vorm van *close reading* die ruimte laat voor interpretatie, zoals ik in hoofdstuk 3.1 zal betogen. Ik probeer door middel van deze aanpak op een alternatieve manier om te gaan met het informatieverlies dat Piper beschrijft: *close readings* zijn onderdeel van de trainingsdata van het model, terwijl een groter aantal tekstdelen dan in een traditionele *close reading* wordt meegenomen in mijn systematische analyse. Tegelijkertijd vraagt deze nieuwe functie van *close readings* volgens mij om een nieuwe manier van ordenen van het onderzoeksmateriaal. Dit gebeurt bij een *close reading* al min of meer vanzelf: de besproken passages zijn automatisch de relevante passages voor een analyse. Ik vat de opmerking van So over het begrijpen *hoe* een model fouten maakt dan ook op als een mogelijkheid om het onderzoeksmateriaal te ordenen door middel van een analyse (of: *close reading*) van foute classificaties door het model. De positie van *close reading* is sowieso een belangrijk element als het gaat om modelleren in de letterkunde. In het volgende citaat in een blogpost van zijn onderzoeksgroep .txtlab vat Andrew Piper deze tendens goed samen: '[T]o understand how automated language systems work, *close reading* is essential.

Engaging deeply with the nature of texts is at the core of humanistic meachine learning.'[2] In deze scriptie zet ik *close reading* op twee plaatsen in, zij het beide in een andere vorm: bij de annotatietaak door fragmenten in te delen volgens de vier aspecten van focalisatie die ik beschreef en bij de interpretatie van de resultaten van het model door losse fragmenten die door het model zijn gelabeld te interpreteren in het licht van de narratologische theorie.

Tot slot betwijfel ik in hoeverre generaliserende uitspraken over modellen, zoals die van Piper, standhouden als met modellen het punt van generalisatie aan de kaak wordt gesteld. Juist het inzichtelijk maken van gemaakte keuzes tijdens het onderzoek zou namelijk moeten laten zien dat er niet één model kan bestaan dat past in iedere letterkundige context, zoals *close reading* ook niet één manier van lezen is. Daarom beschouw ik dit onderzoek, waarin ik een model probeer te maken dat focalisatie kan classificeren, nadrukkelijk als een verkenning op methodologisch vlak. Deze methode en de keuzes die ik daarbij maak zal ik verder uiteenzetten in het derde hoofdstuk.

[2] https://txtlab.org/2021/11/detecting-narrativity-across-long-time-scales/

3. Methode

Nadat ik in het theoretisch kader heb uitgelegd vanuit welke academische achtergrond ik vertrek, werk ik in dit hoofdstuk de verschillende componenten uit die nodig zijn voor het model waarmee ik focalisatie onderzoek. Ik begin met de bespreking van de specifieke manier waarop ik het model inzet in deze scriptie. Het model dient namelijk niet slechts om het herkennen van focalisatie te automatiseren, op zichzelf genereert het ook al betekenisvolle inzichten. De cruciale stap die dit mogelijk maakt heet 'operationaliseren'.

3.1 Operationaliseren

Franco Moretti (2013) laat in zijn artikel over operationaliseren in de digital humanities zien hoe bestaande (literaire) concepten kunnen worden getransformeerd tot een serie van kwantitatieve acties. Deze stellen de onderzoeker in staat metingen uit te voeren met betrekking tot het bestudeerde concept, waar dit eerder nog niet werd gedaan. In de woorden van Moretti impliceert operationaliseren: 'building a bridge from concepts to measurement, and then to the world' (p. 104). Hierbij worden literair-theoretische concepten dus via kwantificering 'gemeten' in literaire teksten.

Een voorbeeld van een dergelijke operationalisering kan worden gevonden bij Leonardo Impett (2020), in zijn artikel over gebaren in de kunstgeschiedenis. Hij neemt hierbij een bestaand vraagstuk uit de kunstgeschiedenis, namelijk de theorie van de 'Pathosformel' van Aby Warburg over de expressie van extreme emoties in beelden. Door een bestaande theorie als uitgangspunt te nemen, is Impett in staat voort te bouwen op bestaande geesteswetenschappelijke tradities, in plaats van te beginnen van vooraf aan. Daarnaast stelt hij dat de transformatie van een concept naar een algoritme een onderzoeker dwingt om een expliciete computationele lezing te geven van het bestaande concept door de afzonderlijke componenten te benoemen (p. 388).

Vervolgens definieert Impett de 'Pathosformel' opnieuw en met behulp van computervisietechnieken komt hij tot een nieuw antwoord op zijn vraagstuk: de verschillende beelden binnen de 'Pathosformel' blijken meer interne samenhang te vertonen dan tot nu toe werd verondersteld. Bovenal stelt het hem in staat met andere ogen naar het bestudeerde concept te kijken: '[S]uch internal unity […] could only have been visible computationally. This is not because computational techniques could see more than a human observer […] but precisely because they can see less.' (p. 395) Hiermee doelt hij op het gebrek aan kennis van het algoritme over onder meer de kunsthistorische en stilistische context. Met andere woorden,

operationalisering stelt de onderzoeker in staat precies dat onderdeel – en vooral: niet meer dan dat – te bestuderen dat nodig is voor het beantwoorden van de opgestelde onderzoeksvraag.

In de letterkunde zie ik een gelijksoortige benadering, zonder de term 'operationaliseren' expliciet te noemen. Dit is bijvoorbeeld een thema in een artikel van Ted Underwood (2018) over het meten van literaire tijd en in dat van Andrew Piper, Sunyam Bagga, Laura Monteiro et al. (2021) over het meten van narrativiteit doorheen een langere periode. Underwood begint zijn artikel door te wijzen op de aanname in letterkundig onderzoek waarbij anekdotisch bewijs wordt gebruikt om een stelling over een hele tekst te bewijzen. Volgens Underwood ligt zodoende bij de interpretatie van een tekst al te zeer de nadruk op korte tijdspannen, terwijl het niet ongebruikelijk is dat romans meerdere jaren of zelfs decennia bestrijken. De vraag die hij aldus stelt is de volgende: 'Why is experience measured in seconds or minutes more appropriately literary than experience measured in weeks or months?' (p. 342). Hij stelt namelijk dat de gerichtheid op kleine tijdseenheden wordt ervaren als typisch literair en dat onderzoekers als Gérard Genette het modernisme hiervoor als bron aanwijzen.

Underwood onderzoekt deze aanname door, samen met twee anderen, tijdsverloop in literaire teksten te annoteren. Ze schatten ieder het tijdsverloop voor dertig romans (afkomstig uit een periode van bijna driehonderd jaar), waarvan ze zestien fragmenten van ongeveer 250 woorden annoteren. Van iedere roman annoteren ze de eerste en de laatste 500 woorden (i.e. de eerste twee en de laatste twee fragmenten) ('because I was curious about the temporal zooming in or out that might happen there', p. 345), de overige twaalf fragmenten zijn afkomstig uit willekeurige plekken in de romans. Underwood benadrukt dat de fragmenten geenszins generaliseerbare informatie over de gehele roman opleveren en dat een andere *scale of measurement*, oftewel de grootte van het geannoteerde fragment, waarschijnlijk andere resultaten zou opleveren.

Door per roman een gemiddelde van de geannoteerde fragmenten te nemen, is Underwood daarna in staat om op basis van de geannoteerde fragmenten een ontwikkeling van vertelde tijd te schetsen. Hij richt zich dus expliciet op langdurige, geleidelijke ontwikkelingen, die ook gelden als premisse voor zijn boek *Distant Horizons* (2019). Dit stelt hem in staat te laten zien dat de afname van het tijdsverloop in romans al begon aan het begin van de achttiende eeuw en doorloopt tot het einde van de twintigste eeuw, wanneer er zelfs een lichte toename van literaire tijd lijkt op te treden (p. 347-348). Hij merkt op dat dit nog geen aanleiding geeft tot vaststaande conclusies, vanwege de beperkingen van het werken met een gemiddelde van

geannoteerde fragmenten. Nochtans ondersteunen een negatieve correlatie tussen publicatiejaar van een boek en gemiddelde duur van de fragmenten én een grote *effect size* deze waarneming.

Zodoende laat Underwood zien hoe het annoteren van een tekst nieuwe perspectieven kan bieden op bestaande letterkundige vragen. Dit onderzoek sluit dan ook aan bij Franco Moretti's notie van operationaliseren. Een volgende stap zou kunnen bestaan uit het vergroten van de hoeveelheid geannoteerde fragmenten, om vervolgens met behulp van *machine learning* vanuit een ander perspectief tot onderzoeksresultaten te komen, zoals ook gesuggereerd door Underwood.

Het onderzoek van Andrew Piper, Sunyam Bagga, Laura Monteiro et al. (2021) sluit hierbij aan door inderdaad de annotaties te gebruiken als trainingsmateriaal voor de inzet van *machine learning*. In hun paper onderzoeken ze narrativiteit als meetbaar taalkundig fenomeen. Ze nemen daarbij aan dat narrativiteit niet zozeer een binaire categorie (wel of geen narrativiteit) is, maar gradaties kent over meerdere eigenschappen. Vanuit dat perspectief is narrativiteit in verschillende mate, ook binnen eenzelfde tekst, aanwezig in verschillende genres. Piper et al. volgen daarbij David Herman die stelt dat narrativiteit tot stand komt in de interactie tussen de lezer en de tekst (i.e. de linguïstische en semiotische kenmerken van de tekst). Ze onderzoeken vier verschillende genres over een tijdsperiode van twee- tot driehonderd jaar. Hun doel is om te testen in hoeverre narratieven onderhevig zijn aan wat Niklas Luhmann functionele differentiatie noemde, waarbij de domeinen kunst en wetenschap door de tijd heen minder op elkaar gaan lijken (p. 320).

In navolging van David Herman definiëren ze narratieve communicatie aan de hand van vier categorieën.[3] Deze vatten ze op als synthese van een narratologisch raamwerk, 'capturing a good degree of consensus in the field' (p. 321). Hierop annoteren ze 401 passages (inclusief 'codebook') op basis van drie vijf-puntslikertschalen. De schaal loopt van 'strongly disagree' tot 'strongly agree'. In plaats van het annoteren van narrativiteit, annoteren ze dus drie *dimensies* van narrativiteit, waarover ze schrijven: 'We found that this increased reader agreement and allowed for more nuanced understandings of narrative communication. For example, it was not uncommon for some types of discourse to emphasize sequential events but lack an emphasis on agency or building a world' (p. 322). De passages die ze annoteren hebben

[3] Piper et al. (2021) onderscheiden de volgende vier elementen (p. 321): *situatedness* ('narrativity depends on the social context in which it occurs'), *event sequencing* ('narrativity depends on temporally ordered events'), *world making* ('narrativity depends on the fact of disequilibrium such that we can observe a change in the world') en *feltness* ('narrativity captures the *experience* [cursief van het artikel] of events, i.e. "what it is like"').

een lengte van vijf zinnen, omdat ze *local narrativity* ('the extent to which a span of tokens expresses narrative communication', p. 325) onderzoeken. Ze verwijzen hiervoor naar Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He et al. (2016) die een computer een vijfde zin laten voorspellen op basis van vier voorgaande zinnen uit een klein verhaal.

Vervolgens gebruiken Piper et al. de annotaties om met behulp van *machine learning* het volledige corpus bestaande uit 335.245 documenten te voorzien van een 'narrativiteits-score'. Hiertoe vergelijken ze meerdere algoritmes, waarvan *Random Forest* het effectiefst is. Interessant is de verbinding die ze leggen tussen de scores op basis van annotaties en het concept dat ze onderzoeken: 'The question we want to address is how well our models correlate with the *scalar* nature of reader judgements' (325). Vervolgens vinden ze een sterke correlatie tussen de waarschijnlijkheid van een narratief en de annotatie-score. Dit betekent dus dat het model dat ze hebben getraind in hoge mate overeenkomt met de scores van hun eerdere annotaties. Tot slot kunnen ze de waarschijnlijkheid van narrativiteit bekijken voor het gehele corpus.

Dit leidt uiteindelijk tot de vergelijking van narrativiteit tussen vier domeinen, waaruit Piper et al. concluderen dat narrativiteit wellicht onderdeel is van Luhmanns hypothese over functionele differentiatie. Dit houdt in dat narrativiteit een verklarende factor kan zijn voor de hypothese van Luhmann. Een mogelijkheid voor vervolgonderzoek die ze noemen is het verkennen van narrativiteit binnen eenzelfde domein, zoals fictie, aangezien hun model dat nog niet toelaat. Met deze methode zou narrativiteit vanuit een nieuw perspectief kunnen worden onderzocht doordat de ontwikkeling ervan door de tijd heen kan worden bestudeerd.

Interessant aan dit onderzoek is de mogelijkheid die wordt opgeworpen om via het herformuleren of formaliseren van een bestaand narratologisch concept (narrativiteit) de werking ervan te onderzoeken vanuit een nieuw perspectief. Het werken met annotaties van verschillende dimensies van een concept vraagt namelijk om na te denken over de bouwstenen waaruit het bestaat, om vervolgens op basis van deze reflectie het concept opnieuw op te bouwen. Hiermee volgen Piper et al. de benadering zoals die wordt voorgesteld en uitgewerkt door Franco Moretti en Leonardo Impett, waaraan ze de toepassing van longitudinaal onderzoek toevoegen.

Aansluitend op de notie van operationaliseren biedt het artikel 'Blue eyes and porcelain cheeks: Computational extraction of physical descriptions from Dutch chick lit and literary novels' van Corina Koolen en Andreas van Cranenburgh (2018) een aanvullend perspectief door de nadruk te leggen op het proces van annoteren, in plaats van de longitudinale schaal. Ze beogen *chick lit* en literaire romans te vergelijken door zinnen die beschrijvingen van het fysieke voorkomen

van personages bevatten automatisch te extraheren. Dit doen ze door voor beide genres een boek te voorzien van handmatige annotaties die aangeven of een zin een beschrijving dan wel geen beschrijving van een personage bevat. Ze gebruiken hiervoor een corpus bestaande uit 32 boeken, waarvan er twee volledig handmatig worden geannoteerd en 30 worden gebruikt als test-set (de eerste 500 zinnen van iedere tekst worden handmatig geannoteerd). Het corpus bevat zowel ik- als personale vertellingen en (voor zover mogelijk) evenveel mannelijke als vrouwelijke auteurs. Het doel van het annoteren is als volgt: 'to make an inventory of possible variations in descriptions of physical appearance.' (p. 61) Met andere woorden, de geannoteerde zinnen als geheel dienen als representatie van de verschillende manieren waarop een fysieke beschrijving van een personage kan voorkomen in het corpus. Ze kiezen voor annotaties op zinsniveau omdat beschrijvingen van het fysieke voorkomen relatief zeldzaam zijn. De zinnen zijn binair geannoteerd, dus ze bevatten ofwel een beschrijving ofwel geen beschrijving. Hierin verschilt dit onderzoek van dat van Piper et al. (2021) in die zin dat Koolen en Van Cranenburgh (2018) onderzoeken *of* een fragment een beschrijving bevat – waarmee dit een herkenningstaak is – terwijl Piper et al. veronderstellen dat narrativiteit hoe dan ook aanwezig is in een fragment en dat nog slechts de mate van voorkomen moet worden vastgesteld.

Na het annoteren automatiseren Koolen en Van Cranenburgh de extractie van hun zinnen met behulp van *machine learning*-technieken en handmatige lexicaal-syntactische *queries*. Het lastige aan het herkennen van fysieke beschrijvingen van personages bestaat eruit dat ze geen vast patroon van woorden kennen dat duidelijk maakt dat het om een beschrijving gaat (p. 60). Om die reden testen ze verschillende methodes voor de extractie ervan. De meest relevante in het kader van mijn onderzoek is het *machine learning*-algoritme dat ze gebruiken, de Support Vector Machine (SVM). Ze kiezen voor dit algoritme omdat het in staat is te werken met *high dimensionality* (veel features, bestaande uit woorden) en *overfitting* (gevaar van niet-generaliseerbaar model) (p. 65). Als features gebruiken ze *bag-of-words*-representatie. Ze schrijven dat een classificatietaak gewoonlijk op document-niveau gebeurt, maar ze buigen dit om naar classificatie van zinnen, waarbij zinnen worden opgevat als document en de roman als verzameling van documenten (p. 65).

De resultaten zijn vervolgens niet overweldigend, hetgeen ze wijten aan enerzijds de complexiteit van de taak; er is meer context nodig voor het herkennen van beschrijvingen dan voor bijvoorbeeld het herkennen van woordsoorten. Anderzijds verwachten ze een verbetering van hun model als ze beschikken over meer trainingsdata. Beschrijvingen in *chick lit* worden gemakkelijker herkend dan die in literaire romans, hetgeen ze verklaren door te suggereren dat beschrijvingen van personages in *chick lit* meer uniform zijn.

Interessant aan het artikel van Koolen & Van Cranenburgh is het aanknopingspunt dat het biedt voor een synthese tussen *close* en *distant reading*, zoals verondersteld door Sven Vitse (2023): 'een onderzoeksdesign waarbij een zelflerend algoritme getraind wordt op basis van voorbeelddata die handmatig zijn geanalyseerd en geannoteerd' (p. 11). De handmatige annotaties kunnen worden ingezet als vorm van tekstanalyse waarbij in het proces van formaliseren al een gedeeltelijke tekstinterpretatie schuilt, Vitse schrijft hierover: 'Ten minste een deel van de narratologische tekstanalyse kan in beginsel […] worden geherformuleerd als een classificatietaak.' (p. 11) De inzet van een *machine learning*-algoritme heeft vervolgens twee functies: het dient zowel ter evaluatie van de annotatietaak als ter bestudering van het functioneren van het geformaliseerde concept. Met andere woorden, de combinatie van annoteren met daaropvolgend automatiseren biedt een mogelijkheid om enerzijds de kwaliteit van de annotaties te bekijken en anderzijds het bestudeerde concept te operationaliseren, zoals we in bovenstaande artikelen al zagen.

Het zou dan ook interessant zijn de annotatietaak uit te diepen ten einde een synthese tussen 'klassieke' en computationele tekstanalyse nader te onderzoeken. Hierbij valt te denken aan het inkaderen van de teksten die worden geannoteerd om vervolgens exemplarische fragmenten uit te lichten en te bespreken in het licht van de annotatietaak. In dit proces wordt een voortschrijdend tekstbegrip gecombineerd met een groeiend inzicht in het functioneren van het concept dat wordt geannoteerd.

Ik zal dit toelichten aan de hand van mijn eigen onderzoek. De annotaties dienen hierin als beschrijving van focalisatie. Een fragment dat is geannoteerd als interne focalisatie draagt bij aan het pallet aan mogelijke voorkomens van interne focalisatie. Alle passages vormen schakeringen van een categorie, bijvoorbeeld interne focalisatie. Samen zijn de passages een verzameling van de verschillende verschijningsvormen van interne focalisatie. In deze scriptie stel ik me ten doel de mogelijkheden van het operationaliseren van focalisatie te bestuderen. Dit is dus een methodologische verkenning waarmee ik onderzoek welke resultaten het oplevert als een letterkundig concept wordt benaderd vanuit een computationeel perspectief. In het vervolg van dit hoofdstuk werk ik daarom de verschillende stappen uit om focalisatie te operationaliseren.

3.2 Probleembeschrijving

Alvorens over te gaan op het operationaliseren, beschrijf ik eerst de classificatietaak in formele termen van een classificatie*probleem*. Wat houdt het binnen een computationele context

namelijk in om de vier aspecten van focalisatie te ‘leren’ aan een computer? In dit onderzoek beschouw ik het categoriseren van focalisatie als documentclassificatieprobleem, omdat focalisatie vaak enige context vereist die het zinsniveau overstijgt. Dit heeft als voordeel dat het categoriseren van focalisatie waarschijnlijk accurater gebeurt, omdat het model meer context krijgt. Daarnaast is het wat focalisatie betreft niet de vraag *of* het voorkomt in een verhaal, aangezien er in principe voortdurend wordt waargenomen. Het nadeel hiervan is dat de begrenzing van een document enigszins arbitrair is. De lengte van de tekstfragmenten voor de annotatietaak bespreek ik daarom in de corpusbeschrijving (hoofdstuk 3.3).

In deze scriptie heb ik te maken met een vierledig classificatieprobleem, waarbij steeds één aspect wordt geclassificeerd. Dit betekent dat ik per taak een aparte *classifier* train en die bovendien los van elkaar optimaliseer en evalueer. Onder meer Folgert Karsdorp (2016, p. 35-39) laat zien dat één *classifier* alle vier de aspecten tegelijk kan labelen. Ik kies er echter voor om dichter bij de menselijke annotatietaak te blijven, waarbij voor ieder tekstfragment aspect voor aspect een label wordt toegekend. Voor alle aspecten geldt dat steeds maar één label kan worden toegekend, in tegenstelling tot zogeheten *multi-label* classificatie waarbij meerdere labels aan hetzelfde document kunnen worden toegekend. Het gefocaliseerde object kan bijvoorbeeld niet tegelijk waarneembaar en niet-waarneembaar zijn.

Ik onderscheid twee soorten classificatieproblemen: een binair classificatieprobleem en een *multi-class* classificatieprobleem. Voor beide gevallen geldt: *d* betekent een fragment en *C* het corpus van geannoteerde fragmenten. Elk fragment d_i krijgt voor iedere classifier een label y_i toegekend, vier in totaal dus. Het binaire classificatieprobleem geldt voor het aspect van waarneembaarheid, dat bestaat uit twee categorieën. Het doel bestaat eruit ieder document (tekstfragment) d_i een label toe te kennen dat ofwel label 0 ofwel label 1 is. Hierbij vertegenwoordigen beide mogelijke waardes een categorie van het aspect van waarneembaarheid.

Het tweede soort classificatieprobleem, het *multi-class*probleem, geldt voor de aspecten van positie van focalisator, aantal waargenomen objecten en de mate van detail van het waargenomen object. Met ‘multi-class’ wordt bedoeld dat de set labels bestaat uit meer dan twee labels. In deze gevallen wordt voor ieder document d_i één label voorspeld uit een set labels $Y = \{y_1, y_2, \ldots, y_n\}$, waarbij n staat voor het aantal labels (categorieën) waaruit een aspect bestaat. Ik kies hier dus voor een classificatiebenadering en niet voor bijvoorbeeld een ranking-benadering, waarbij een ranglijst wordt opgesteld van de meest geschikte labels. Voor deze classificatietaak verkies ik deze benadering boven een rangschikking, omdat een rangschikking voor deze classificatietaak om twee redenen niet relevant is. Als eerste sluiten categorieën van

aspecten elkaar uit; er kunnen niet tegelijkertijd twee én drie objecten worden gefocaliseerd. Ten tweede is de volgorde op basis van het meest waarschijnlijke label niet betekenisvol, aangezien bijvoorbeeld de hoogste mate voor detail waarschijnlijk wordt gevolgd door de één-na-hoogste categorie.[4]

3.3 Corpus

In deze paragraaf beschrijf ik het corpus dat ik gebruik in deze scriptie. Om het corpus gereed te maken voor gebruik doorloop ik enkele stappen die ik hier zal bespreken, evenals de keuzes die ik hierbij maak en de implicaties die hiermee gepaard gaan.

In dit onderzoek maak ik gebruik van een corpus bestaande uit alle 170 inzendingen voor de Libris Literatuurprijs 2013 zoals beschreven in Van der Deijl, Pieterse, Prinse & Smeets (2016) en Smeets (2021). Dit corpus biedt daarmee een dwarsdoorsnede van de literaire productie in één jaar. Daardoor betreft dit een relatief homogeen corpus in het genre van literaire romans, waarop ik me in dit onderzoek richt. Daarnaast heeft dit corpus als voordeel dat het relatief recente boeken uit hetzelfde jaar betreft, waardoor eventuele veranderingen in de taal zijn uitgesloten. Concreet maak ik een selectie van vier romans uit het corpus, te weten: *Dorst* (2012) van Esther Gerritsen, *Niemand in de stad* (2012) van Philip Huff, *Strak blauw* (2012) van Renée van Marissing en *Euforie* (2012) van Christiaan Weijts. De stappen die ik hieronder beschrijf, voer ik dan ook enkel uit op deze romans. Bij de keuze voor deze romans zijn variaties in vertelvorm (ik- en personale vertellingen) en gender van de auteurs gebalanceerd.

De geselecteerde romans worden opgedeeld in ongeveer even grote fragmenten, waarbij er rekening wordt gehouden met zins- en alinea-eindes. Vervolgens heb ik uit alle vier de romans 250 willekeurige fragmenten geselecteerd (in totaal werk ik dus met 1000 fragmenten) die ik zal gebruiken voor de annotatietaak. Per roman heb ik daarnaast een subset gemaakt van 25 fragmenten (in totaal 100 fragmenten) die worden gebruikt voor het testen van de inter-beoordelaarsbetrouwbaarheid. Het corpus van 1000 fragmenten zal, voorzien van annotaties voor alle vier de aspecten van focalisatie, dienen als trainingsmateriaal voor de algoritmes die ik train om focalisatie te herkennen. De keuze voor vier romans uit 2012 maakt dat het algoritme wordt getraind op een ‘samenballing’ van focalisatie in 2012. Dit heeft als voordeel dat het de kans vergroot dat het model focalisatie zal kunnen herkennen voor romans uit 2012 en daarmee

[4] Voor bijvoorbeeld het label 3 zou het meest waarschijnlijke label zowel 2 als 4 kunnen zijn. Desondanks zou deze ranking redelijk voorspelbaar zijn, aangezien het label 3 nooit gevolgd kan worden door 1 of 5.

dat het model ‘werkt’. Anderzijds zal het model door deze keuze waarschijnlijk minder effectief zijn in het herkennen van focalisatie in andere jaren of periodes, want ik veronderstel dat de tekstuele verschijningsvorm historisch gesitueerd is. Ik beperk me met deze corpuskeuze dus tot focalisatie in romans uit 2012. Dit sluit aan bij de experimentele en verkennende aard van dit onderzoek. Een model dat focalisatie kan herkennen in romans over een langere periode en dat diachroon onderzoek mogelijk maakt, is een logische volgende stap.

Hoe groot dienen de te annoteren fragmenten te zijn? Dit is uiteraard volledig afhankelijk van de onderzoeksvraag. Zo gebruiken Koolen & Van Cranenburgh (2018) één zin als eenheid om beschrijvingen van het fysieke voorkomen van personages te annoteren, terwijl Underwood (2018) zijn teksten opdeelt in fragmenten van 250 woorden om vertelde tijd te bestuderen. Andrew Piper, Sunyam Bagga, Laura Monteiro et al. (2021) merken op dat het slechts van belang is dat kan worden aangenomen dat ‘completed narratives’, die ze onderzoeken, aanwezig zijn. Ze halen hiervoor Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He et al. (2016) aan die stellen dat fragmenten van vijf zinnen hiertoe van voldoende grootte zijn. Ze benadrukken echter dat dit nader onderzocht dient te worden. In de context van deze scriptie is de keuze voor het niveau van annoteren van grote invloed op de beoordeling van focalisatie in een passage, zoals ik al vaststelde in de inleiding. Sterker nog, door deze invloed op de beoordeling van focalisatie is de keuze voor een fragmentgrootte een leidende factor in de structurering van mijn onderzoeksdata; een analyse van focalisatie op hoofdstukniveau levert andere resultaten op dan wanneer fragmenten van enkele zinnen worden geannoteerd. Hiermee leg ik dus de context van het operationaliseren vast, waarbij een deel van een roman (een fragment) wordt voorzien van informatie (een label) over de focalisatie in die passage.[5]

Voor de bestudering van focalisatie kan een fragment niet eenduidig worden afgebakend als wordt afgegaan op wisselingen tussen categorieën van een aspect. Een verandering in

[5] Een letterkundig onderzoek waarbij fragmentgrootte een soortgelijke structurerende functie heeft voor de interpretatie is *Lezen om te schrijven* (1988) van Hugo Bousset. Hierin hanteert hij een zogeheten ‘progressieve lectuur’ (p. 13) waarbij de roman is opgedeeld in leeseenheden om te analyseren. Hij ontleent hiertoe het concept ‘lexie’ aan *S/Z* (1974) van Roland Barthes, die lexieën definieert als ‘units of reading’ (Barthes 1974, p. 13). Voor het opdelen van de tekst om focalisatie te kunnen bestuderen, hoef ik niet zo gedetailleerd te werk te gaan als Bousset door lexieën op te delen in sublexieën en passages. Ik beperk me tot een definitie van lexie die ik geschikt acht voor dit onderzoek, temeer daar Roland Barthes in *S/Z* het arbitraire ervan benadrukt: ‘This cutting up, admittedly, will be arbitrary in the extreme’ (Barthes 1974, p. 13). Dit laat zien dat de lengte van de lexie niet als vaststaand gegeven moet gelden. Het dient vooral als middel om te denken in kleinere eenheden dan de tekst als geheel, zoals gebruikelijk is in de letterkunde. Barthes geeft hier invulling aan door te kijken naar de dichtheid en het wisselen van connotaties (p. 13). Bousset bedient zich van meer mogelijkheden door de lexieën ook weer op te delen.

bijvoorbeeld de mate van detail gaat immers niet altijd gepaard met een wisseling van de positie van de focalisator of van een van de andere aspecten die ik onderscheid bij het beschrijven van focalisatie. Daarom besluit ik in dit onderzoek de tekst te segmenteren door fragmenten te definiëren als tekstdelen met een vaststaand woordenaantal van 150 woorden, waarbij ik rekening houd met paragrafen en hoofdstukken. Dit woordenaantal is een afweging tussen enerzijds de veronderstelde beperkte variatie van focalisatie op zinsniveau en anderzijds de nuances die verloren gaan zodra een geheel hoofdstuk als één enkel fragment wordt geannoteerd. Een nadeel deze fragmentgrootte is dat de positie van het focaliserende subject kan wisselen binnen 150 woorden; de kans dat dit gebeurt binnen één zin is minder groot, hoewel de focalisator ook binnen een enkele zin kan wisselen. Een extra middel om dergelijke wisselingen te verminderen, is het 'bijknippen' van fragmenten door de eerste of laatste zin waar relevant te verwijderen als dit de eenduidigheid van het geannoteerde fragment bevordert. Een voordeel van fragmenten van 150 woorden is dat een beschrijving van een object vaak meer dan één zin beslaat, waardoor de waarneembaarheid en de mate van detail beter te labelen zijn. Enkele experimenten met verschillende fragmentgroottes tijdens het annoteren bevestigen dat met fragmenten van 150 woorden focalisatie redelijkerwijs kan worden beschreven. Zodoende verdeel ik de vier te annoteren romans in fragmenten van gelijke grootte, die worden gebruikt voor de training van het model.

3.4 Annoteren

In de paragraaf over operationaliseren stelde ik in navolging van Franco Moretti (2013) dat operationaliseren uit een serie van kwantitatieve acties bestaat. Het annotatieschema is een volgende component in dit proces van operationaliseren. In het theoretisch kader gaf ik al een definitie van focalisatie en onderscheidde ik vier aspecten die focalisatie beschrijven, te weten: de positie van de focalisator, de waarneembaarheid van het gefocaliseerde object, de standvastigheid van het gefocaliseerde object en de mate van detail van het gefocaliseerde object. Ieder aspect heb ik vervolgens onderverdeeld in een aantal categorieën. In het annotatieschema verbind ik de aspecten en bijbehorende categorieën met de handmatige annotatietaak zoals al kort beschreven in de paragraaf over operationaliseren. Het schema bestaat uit een korte toelichting per aspect, waarna de annotatieprocedure en enkele overwegingen worden besproken. Eerst zal ik echter de positie van de handmatige annotaties binnen deze scriptie beschrijven.

3.4.1 Handmatige annotaties

De annotaties fungeren als menselijke *benchmark*. Dat wil zoveel zeggen als ‘een getrainde lezer zou dit fragment beoordelen als X’. Om generaliserende uitspraken over de geannoteerde fragmenten te kunnen doen, is het van belang dat de verschillende aspecten die worden geannoteerd een bepaalde mate van repliceerbaarheid bezitten. Het zou namelijk onwenselijk zijn wanneer blijkt dat de ene beoordelaar een fragment annoteert als waarneembaar, terwijl een ander het tegenovergestelde beweert. In een dergelijke situatie laat een aspect waarschijnlijk te veel ruimte open voor individuele interpretaties met als gevolg dat er onvoldoende consensus bestaat. Om de mate van overeenstemming te controleren voor ieder aspect, zal ik in de hierop volgende paragraaf de zogeheten *inter-rater reliability* (‘inter-beoordelaarsbetrouwbaarheid’) berekenen. Indien hieruit blijkt dat de beoordelaars voldoende overeenkomen, hebben we een menselijke *benchmark* die ons vertelt dat we redelijkerwijs mogen veronderstellen dat de geannoteerde fragmenten door andere beoordelaars op ongeveer dezelfde wijze zouden zijn geannoteerd. Dit dient tevens als referentiepunt voor het model dat ik uiteindelijk zal testen, waarbij het streven is het model ongeveer even hoog te laten scoren als de inter-beoordelaarsbetrouwbaarheid. Met andere woorden, het is onwenselijk als het model veel beter of veel minder hoog scoort dan de menselijke *benchmark*. Voordat de overeenstemming tussen de verschillende beoordelaars kan worden berekend zal ik eerst per aspect de overwegingen voor het annotatieproces bespreken, daarna stel ik het annotatieschema op waarmee focalisatie per fragment kan worden beoordeeld.

3.4.2 Aspecten van focalisatie

Eén van de aspecten leent zich voor een graduele beschrijving, een aspect wordt dan beoordeeld als zijnde in hogere of lagere mate aanwezig in een fragment. Dit geldt voor de mate van detail. Het is daarbij van belang om tot een intuïtieve betekenis van deze aspecten te komen. Hiermee bedoel ik dat het aspect dusdanig geformuleerd dient te worden, dat een tweede of derde beoordelaar tot (ongeveer) dezelfde beoordeling kan komen. Hiertoe vond tijdens de eerste fase van het annoteren een proces plaats waarbij fragmenten met dezelfde positie op een schaal met elkaar werden vergeleken om een zekere mate van systematiek tussen de fragmenten te ontdekken. Wanneer dit in hogere mate het geval is, wordt het immers waarschijnlijker dat een ander persoon tot dezelfde beoordelingen zal komen. Tijdens het proces van annoteren kende het annotatieschema om die reden nog enige wijzigingen die plaatsvonden in de formulering van het schema en daarmee de formalisering van enkele aspecten. Zo heb ik naar aanleiding van de eerste ronde van annoteren de vraag toegevoegd wat het gefocaliseerde object is, om

vervolgens pas de beoordelaar de gerelateerde aspecten te laten annoteren. Dergelijke toevoegingen, waarbij in het annotatieschema wordt verduidelijkt wat er precies wordt geannoteerd, verhogen de kans op overeenstemming.

Daarnaast heb ik naar aanleiding van de eerste rondes van annoteren besloten om enkele aspecten niet te annoteren op een vijfpuntsschaal, maar als nominale data. Hoewel het aanlokkelijk is alle aspecten te benaderen als graduele schaal – om nuance na te streven – is dit niet voor ieder aspect betekenisvol. Zo kent het onderscheid tussen interne en externe focalisatie ruimte voor ambiguïteit, maar een interne focalisator kan in principe niet meer of minder intern zijn dan een andere interne focalisator. Hieronder breng ik voor ieder aspect kort de definitie zoals opgesteld in het theoretisch kader in herinnering, waarna waar relevant een toelichting volgt met betrekking tot het annoteren van het betreffende aspect. Vervolgens bespreek ik de annotatieprocedure.

Bovendien besteed ik voor ieder aspect aandacht aan de beoordeling van de directe rede, meestal voorkomend in de vorm van een dialoog. De analyse van de directe rede is een bekend probleem in de narratologie, omdat die vaak ruimte laat voor interpretatie (zie onder meer Bal 1981 en Bal 1983). Aangezien de fragmenten in dit onderzoek zonder context beoordeeld zouden moeten kunnen worden, leveren dialogen mogelijk enkele moeilijkheden op. Indien een fragment naast dialoog nog tekst bevat die informatie geeft over de focalisatie, kan de dialoog op basis van die aanwijzingen worden geannoteerd. Denk hierbij bijvoorbeeld aan een focalisator van wie de gedachten over de dialoog worden weergegeven. Voor de fragmenten die slechts bestaan uit dialoog is het annoteren uiteraard lastiger, omdat de vertelling in dat geval een laag naar beneden is geschoven. In dat geval is het onduidelijk of het fragment moet worden geannoteerd als dialoog die als geheel wordt waargenomen dan wel of de inhoud van de dialoog moet worden geannoteerd als de waarneming die beschreven wordt in de intradiëgetische vertelling. In het eerste geval kan een dialoog worden opgevat als een (nagenoeg) zuivere weergave van de waarneming die altijd waarneembaar is. In het tweede geval kunnen dialogen (in verschillende fragmenten) in hoge mate van elkaar afwijken als bijvoorbeeld de gefocaliseerde objecten van elkaar verschillen. Dit houdt onder meer in dat een dialoog een of meer eigen gefocaliseerde object(en) heeft. Het valt überhaupt niet mee generaliserende uitspraken te doen over dialogen en wellicht valt het zelfs buiten het bereik van deze scriptie om een uitputtend annotatieschema op te stellen voor dialogen. Ik zal per aspect toelichten hoe ik in dit schema wil omgaan met dialogen, maar voor nu wil ik duidelijk maken dat ik kies voor de eerste benadering van dialogen waarbij de dialoog op het hogere niveau en dus als geheel wordt geannoteerd.

Focaliserend subject

In de paragraaf over focalisatie heb ik voor het focaliserende subject in navolging van Mieke Bal een onderscheid gemaakt tussen interne en externe focalisatie. Dit berust op het verschil in de positie die de waarnemer inneemt binnen de verhaalwereld: een interne focalisator neemt als personage deel aan het vertelde, een externe focalisator staat daarbuiten en is niet direct aan een personage gebonden (Bal 2009, p. 152). Dit aspect beschrijft dus de positie van de focalisator.

Luc Herman en Bart Vervaeck (2009a) wijzen erop dat externe en interne focalisatie elkaar constant afwisselen in veel verhalende teksten (p. 77). Wanneer de externe focalisator dusdanig nauw aansluit bij de waarnemingen van een personage, spreekt men dan ook van consonantie (Bal 2009, p. 163). Soms kunnen externe en interne focalisatie binnen de beperkte ruimte van één zin wisselen of zelfs ogenschijnlijk samenvallen wanneer niet eenduidig aan te wijzen is welke instantie focaliseert. Herman & Vervaeck benadrukken echter het principiële onderscheid tussen beide posities, ook als ze samenvallen (p. 77).

Vanuit die gedachte is het interessant om fragmenten waarin niet eenduidig één van beide focalisators valt aan te wijzen, te annoteren als tussenpositie. Het gaat dan bijvoorbeeld om fragmenten waarin een ik-nu terugkijkt op gebeurtenissen die deze eerder heeft meegemaakt. Deze gebeurtenissen worden in dat geval waargenomen door een ik-toen. Wanneer echter ook de waarnemingen van de ik-nu doorsijpelen in de herinnering, treedt er consonantie op. Een voorbeeld hiervan is te vinden in het volgende citaat, afkomstig uit *Strak blauw* (2012) van Renée van Marissing:

> ‘Bejaardentehuis? Dat doe ik niet.’ ‘Waar dan ook.’ ‘En jou? Zou ik jou ook nog gek maken als ik krom loop, en er een halfuur over doe om naast je in bed te komen liggen?’ ‘Misschien ben ik dan wel aan het dementeren, dan heb ik niet in de gaten dat het een halfuur duurt en zeg ik elke minuut tegen je: hè gezellig, kom je in bed liggen?’ We lachten. ‘Dat zou toch leuk zijn?’ zei ze. ‘Ja. Ja, dat zou heel leuk zijn.’ Ik deed mijn ogen dicht, en liet de zon mijn huid ouder maken. An kuste me. Nu, zo’n twee jaar later, weet ik het zo net nog niet, met dat samen oud worden. Ik weet het ook zo net nog niet met dat oud worden van mezelf. Het dak is nu ook mijn dak, ik woon hier, in dit huis. (p. 55)

Het citaat opent met een dialoog waarin twee geliefden vooruitblikken op hoe ze samen ouder zullen worden, waarop ze besluiten dat ze dit wel samen zouden willen doormaken. Vervolgens laat de waarnemende ik-toen zich zien, die zont en zich laat kussen door haar geliefde. Als daarop volgt ‘Nu, zo’n twee jaar later’, wordt duidelijk dat de lezer de gedachten volgt van de ik in het heden terwijl ze reageert op de herinnering waarin ze besloot dat ze oud wil worden met haar geliefde. In het heden twijfelt ze namelijk over hun gezamenlijke toekomst. Zo bezien wordt ineens minder duidelijk vanuit wie de dialoog aan het begin van het fragment werd

waargenomen; vanuit de ik-nu of de ik-toen? Zodoende treedt hier ambiguïteit op en dient dit fragment als dusdanig te worden beoordeeld.

Een vereiste om deze categorie op te nemen in het annotatieschema is echter dat er voldoende fragmenten als tussenpositie worden geannoteerd. Daarnaast voeg ik een restcategorie toe voor fragmenten die te weinig duidelijkheid geven over de positie van de focalisator. De fragmenten in deze categorie zullen niet worden meegenomen in de trainingsdata en dienen om te voorkomen dat de tussencategorie zal gelden als *rest*categorie en daarmee veel ruis zou bevatten.

Dialogen – en dan met name binnen fragmenten ingebedde dialogen – kunnen waar mogelijk volgens het schema worden geannoteerd, zoals het citaat al liet zien. Wanneer de waarnemer van een dialoog echter niet duidelijk is omdat de dialoog het gehele fragment beslaat, wordt het betreffende fragment ingedeeld in de aparte categorie van 'dialoog' (label '99'). Het zou namelijk onnodig veel ruis veroorzaken wanneer deze fragmenten onder de restcategorie worden gerekend. Niet alle fragmenten die dialogen bevatten, zullen dus onder deze categorie vallen. Het is immers niet mijn doel om een model te trainen dat gericht is op het automatisch herkennen van dialogen.

Waarneembaarheid

De vraag of een gefocaliseerd object waargenomen kan worden hangt samen met de vraag of het object binnen ('in het hoofd') of buiten ('in de wereld', zichtbaar voor andere personages) een personage ligt. Een interne focalisator kan, naast voor alle personages waarneembare objecten, alleen niet-waarneembare objecten van *zichzelf* focaliseren, zoals eigen gedachten of dromen. Herman & Vervaeck (2009a) gaan in een eindnoot in op de moeilijkheden met classificeren als een interne focalisator speculeert over de binnenwereld van een ander personage en suggereren dat rechtlijnige regels in dit verband niet altijd opgaan (p. 186-187). In de praktijk betekent dit dat ik voor dit annotatieschema vasthoud aan beperkt waarnemende interne focalisators met de aantekening om dit waar nodig flexibel te interpreteren.

Wat dialogen betreft, die zijn altijd waarneembaar; onafhankelijk van de positie van de focalisator. Dialogen zijn in feite een vorm van bewustzijnsweergave van personages en daarmee waarneembaar in de ruimte waarin ze worden uitgesproken. Vaak zullen dialogen echter worden afgewisseld met de gedachten van het waarnemende personage, zoals in het onderstaande citaat uit *Dorst* (2012) van Esther Gerritsen (p. 107-108):

> Coco gaat aan de keukentafel zitten en zegt: 'Ik wil híér zijn.' Blok. Het houten stopblok van de speelgoedtrein. Zo is Coco's zin. Groter blok nu: 'Ik wil dat je gaat.' Het is

> gezegd. Het is gedaan. Vecht er niet tegen. Gefaald. Nu de woorden afwachten, alles gaat voorbij. Laat het verhaal maar ontstaan: Tussen de moeder en de dochter kwam het nooit meer goed. Coco blijft kalm. 'Ga.' Schreeuwen zal haar dochter toch wel. Nog steeds stilte. Kom maar. 'Ik heb nu even geen huis.' Wat gebeurt er? Waar is het schreeuwen, het waaien, het stampen? 'Ik heb nu even geen huis.' 'Ja, dat zei je.' 'Mijn kamer. Ik moest eruit.' 'Je moet je kamer uit?' 'Ik ben eruit. Ik moest er al uit. Heb ik toch gezegd? Dat ik mijn kamer uit moest?'

In dit citaat heeft Elisabeth een gesprek met haar dochter Coco. Als lezer kijken we door Elisabeths ogen terwijl Coco gaat zitten en een gesprek begint. Na Coco's eerste zin volgen we echter eerst nog de gedachten van Elisabeth alvorens ze reageert op Coco ('Ik wil dat je gaat'). Deze gedachten blijven terugkeren en domineren deze scène. Hierdoor wordt de dialoog tussen Elisabeth en Coco naar de achtergrond geschoven terwijl de lezer vooral een beeld krijgt van de gedachten van Elisabeth en het ongemak dat ze ervaart tijdens het gesprek. Dergelijke passages, waarbij het gefocaliseerde objecten nu eens waarneembaar (dialoog) en dan weer niet-waarneembaar (gedachten) zijn, dienen te worden geannoteerd als '99' waarmee uitdrukking wordt gegeven aan de vervlechting van waarneembare en niet-waarneembare objecten. Ik voeg dit label toe als extra mogelijkheid, omdat voor de training van het algoritme fragmenten zo eenduidig mogelijk geannoteerd dienen te worden.

Standvastigheid

De combinatie van focalisator en gefocaliseerd object kan in een tekst variëren wanneer een andere waarnemer optreedt of een ander object wordt waargenomen. Voor dit aspect richt ik me op de vraag in hoeverre één of meerdere *objecten* worden waargenomen. Ik vat standvastigheid voor dit onderzoek dus op als het aantal gefocaliseerde objecten in een fragment, geannoteerd als één, twee of meer dan twee objecten. Indien een fragment meer dan twee objecten bevat, vergroot dit de kans dat de mate van detail van de waarneming niet eenduidig kan worden geannoteerd. De objecten zijn dan mogelijkerwijs in verschillende mate van detail gefocaliseerd. Ik kom hierop terug bij de bespreking van het aspect van detail.

Bij het bepalen van het aantal objecten zal niet zelden sprake zijn van interpretatie. Als bijvoorbeeld een lichaam wordt gefocaliseerd doordat de afzonderlijke lichaamsdelen worden waargenomen, kan het object in dat fragment bestaan uit zowel het lichaam (één object) als de losse delen (meer dan twee objecten). Ik kies er in deze en vergelijkbare gevallen voor om uit te gaan van het hoogste niveau van waarneming, de eerste optie dus. Bovendien kan iets abstracts als een ruimtebeschrijving of een sfeerschets ook gelden als één gefocaliseerd object.

In het geval van een beoordeling van '99' voor het aspect van waarneembaarheid, vat ik het waarneembare object en het niet-waarneembare object op als twee verschillende objecten.

Met andere woorden, indien de waarneembaarheid als ‘99’ wordt geannoteerd, dan is de beoordeling van standvastigheid minimaal een ‘2’ (voor twee objecten).

Gedetailleerdheid

Met gedetailleerdheid wordt de mate van detail waarmee een gefocaliseerd object wordt waargenomen bedoeld. Dit geeft dus uitdrukking aan de aandacht die een waarnemer besteedt aan een object. Dit aspect zal vaak samenhangen met zaken als afstand (in tijd en/of ruimte) tussen het focaliserende subject en het gefocaliseerde object, en het belang dat de focalisator hecht aan dat object. De beoordeling dient daarbij af te hangen van het object zelf, en niet van het fragment als geheel. De beoordelingen reageren op de stelling ‘Het gefocaliseerde object in dit fragment wordt gedetailleerd waargenomen.’ Mogelijke reacties lopen van ‘helemaal oneens’ (zeer lage mate van detail) tot ‘helemaal eens’ (zeer hoge mate van detail). De middelste categorie (‘niet zeker’) slaat op het antwoord op de stelling, dus zoiets als ‘niet laag, niet hoog’ en mag niet worden verward met een restcategorie. Ik kies voor een vijfpuntsschaal omdat ik denk dat er te veel informatie over de fragmenten verloren gaat wanneer wordt beoordeeld op basis van een driepuntsschaal. Met vijf punten kan er namelijk nog een onderscheid worden aangebracht in de mate van hoge en lage detaillering. In het geval van meerdere gefocaliseerde objecten die niet dezelfde mate van detail hebben, dient het fragment voor dit aspect te worden geannoteerd als ‘99’, waarmee ‘niet-annoteerbaar’ wordt uitgedrukt. Dit zal vaak samenhangen met de mate van standvastigheid; lagere standvastigheid (dus meerdere gefocaliseerde objecten) vergroot de kans op het niet eenduidig kunnen annoteren van de gedetailleerdheid van een fragment.

Een fragment kan om verschillende redenen een bepaalde beoordeling krijgen. Een eerste reden betreft het aantal woorden dat aan een object wordt besteed: een hoger aantal betekent gewoonlijk een hogere mate van detail. Ook de mate waarin een waarneming bijdraagt aan een completer beeld van een object kan een indicatie voor de mate van detail zijn. Ter verheldering laat ik hier twee fragmenten zien, afkomstig uit *Strak blauw* (2012) van Renée van Marissing:

1. Ik doe mijn ogen dicht en glijd onder water, gedimde lichten, rumoer op de achtergrond, een sterke stroming die zorgt dat ik moeite moet doen rechtop te blijven zitten. Laat los, laat me achterover zakken en laat het water me grijpen, me meenemen naar een plek waar ik aanspoel, overeind kom, eerst op mijn knieën ga zitten, dan ga staan. Ik voel het water in een klein stroompje uit mijn kleren lopen, daarna druppel voor druppel. Ik kijk om me heen en denk, dit ja, zo moest het zijn.
   (p. 80)

2. Ik kijk door het raam naar een plek waar ik niet kan komen en het raakt me niet. Het doet me niks. Dat ik er niet kan komen, dat ik word tegengehouden door het glas, heb ik nooit erg gevonden, dat is iets waar ik überhaupt nooit over heb nagedacht, sommige dingen zijn zoals ze zijn, maar de plek zelf, het uitzicht, raakt me ook niet, voor het eerst niet. Ik kijk naar het grasveld, de bomen, het bankje en ik voel geen verlangen, geen rust, geen aandacht en geen paniek, dus ik besluit een tijdje geen adem meer te halen. Maar dat doe ik wel, natuurlijk, ik haal wel adem, kijk maar naar de ruit, die af en aan beslaat, waarna de condens keer op keer van buiten naar binnen wegtrekt. Ergens gaat een deur dicht. Er ligt een doffe steen in mijn maag, een steen ingepakt in dik vilt. (p. 123)

Het eerste fragment kent een hoge mate van detail. De waarnemingen vinden plaats vanuit een ik die in dit verhaal bijna altijd de waarnemer is. In dit geval gaat het om een niet-waarneembare situatie waaraan meerdere eigenschappen (‘gedimde lichten’, ‘rumoer op de achtergrond’, ‘sterke stroming’) worden toegedicht die het gevoel van onder water zijn invoelbaar maken. Ook de daaropvolgende waarnemingen kennen een hoge mate van detail, waarbij bijvoorbeeld verschillende stappen van opstaan en water dat uit de kleren loopt (‘druppel voor druppel’) worden waargenomen. Daarom krijgt het gefocaliseerde object van deze dagdroom of ‘aanspoel-scène’ een annotatie van 5.

In het tweede fragment wordt daarentegen met een lage mate van detail waargenomen; het uitzicht van de focalisator wordt gefocaliseerd aan de hand van oppervlakkige waarnemingen (‘het grasveld, de bomen, het bankje’). Deze waarnemingen dienen de groeiende fixatie van de focalisator op zichzelf te benadrukken: in de loop van het verhaal vervaagt de buitenwereld steeds meer. Zelfs haar binnenwereld blijft raadselachtig met waarnemingen als ‘ik voel geen verlangen, geen rust, geen aandacht en geen paniek’, waaruit vooral onwetendheid blijkt. Aan het einde van het fragment keert deze afzondering van de buitenwereld kort terug, wanneer deze doordringt tot de waarneming als een deur die ‘ergens’ dichtslaat. Deze elementen maken dat de combinatie van de binnen- en buitenwereld die worden waargenomen in dit fragment worden geannoteerd als lage mate van detail (annotatie van 1). Overigens is dit een complexe passage met zowel waarneembare als niet-waarneembare objecten. Ik zal daarom de andere drie aspecten langslopen. De positie van de focalisator is intern in de gehele passage, zoals wordt duidelijk gemaakt door het kijken van de waarnemende ik (‘ik kijk door het raam’, ‘ik kijk naar het grasveld’) en het gebrek aan gevoelens dat ze bij zichzelf opmerkt (‘ik voel geen verlangen’, ‘raakt me ook niet’). Door de afwisseling in waarnemingen van de buitenwereld (een ruit, bomen, een deur die dichtslaat) en eigen gedachten, valt de waarneembaarheid niet eenduidig te labelen (label ‘99’). Het aantal objecten is daardoor minimaal twee, maar naast de gedachten wordt onder meer ‘een doffe steen in mijn maag’

waargenomen. Het aspect van waarneembaarheid zou daardoor beoordeeld worden als meer dan twee objecten.

3.4.3 Annotatieschema

Het annotatieschema is te vinden in Bijlage 1.

De focalisatie in een fragment zal op grond van vier aspecten worden gekarakteriseerd. Naast het gegeven dat ieder aspect op zichzelf een betekenisvolle eigenschap van de focalisatie omvat, zit de waarde vermoedelijk vooral in de samenhang van de verschillende aspecten. Zo geeft de mate waarin een externe focalisator gebruik maakt van zijn mogelijkheid om de gedachten van personages te focaliseren mogelijk meer inzicht in het functioneren van de externe focalisator. We weten immers dat het focaliseren van de binnenwereld van een personage kan leiden tot een mogelijke 'informatie-ongelijkheid' tussen personages (Bal 2009, p. 157). Wanneer bijvoorbeeld twee personages een conflict hebben en de lezer van slechts één van beiden de beweegredenen kent, beïnvloedt de externe focalisator het oordeel van de lezer over de conflictsituatie. Zodoende zullen wellicht meer van deze combinaties van aspecten als bron van inzicht kunnen dienen.[6]

3.5 Inter-beoordelaarsbetrouwbaarheid

Een subset van 100 fragmenten (25 per roman) is door drie getrainde lezers (inclusief mezelf) voorzien van annotaties op basis van het annotatieschema. Dit deden ze door per fragment het schema te doorlopen en aan de hand daarvan het fragment te voorzien van vier labels (voor ieder aspect één).

Nu ik drie sets van honderd annotaties heb verzameld, wil ik ze naast elkaar leggen om de mate van overeenstemming te kunnen vergelijken. In hoeverre komen mijn eigen annotaties

[6] Voor alle vier de aspecten kort ik fragmenten eventueel handmatig in door aan het begin en/of einde van het fragment enkele zinnen te verwijderen als dit de eenduidigheid van de annotatie ten goede komt. Ik denk bijvoorbeeld aan het voorkomen van een '99-beoordeling', die een fragment zou uitsluiten voor dat aspect. Daarnaast hebben eenduidige fragmenten een grotere kans om op dezelfde manier geannoteerd te worden door verschillende beoordelaars en krijgt het model 'schonere' voorbeelden als input, zodat een toegekend label wordt gerepresenteerd door het gehele fragment. Deze vrijheid van schrappen acht ik gerechtvaardigd omdat ik geen representatie wil maken van de focalisatie in de *roman* die ik annoteer. Het doel van de annotaties is namelijk de *focalisatie* te representeren en in dat geval bevordert het de kwaliteit om zo nu en dan de fragmenten een klein beetje handmatig op te schonen. Indien een fragment een scènewisseling bevat, wordt 'schoonmaken' van het fragment gecompliceerder. Hoewel ik bij het automatisch opdelen van de romans dit zoveel mogelijk heb proberen te voorkomen, bleek dit onvermijdelijk. Dergelijke fragmenten heb ik waar mogelijk handmatig opgedeeld. Indien dit niet mogelijk was heb ik besloten deze fragmenten uit te sluiten van de annotaties.

voor de vier aspecten overeen met die van de andere twee beoordelaars? De onderliggende vraag is in hoeverre mijn annotaties van de verschillende aspecten van focalisatie generaliseerbaar zijn, dus in hoeverre mijn beoordelingen een particuliere interpretatie overstijgen. Ik bekijk daarom de mate van overeenstemming voor honderd fragmenten tussen drie beoordelaars, om conclusies te kunnen trekken over de generaliseerbaarheid van de duizend fragmenten die ik in totaal heb geannoteerd.

Hierbij neem ik aan dat mijn beoordelingen van de fragmenten consistent zijn. Dit heb ik proberen te bewerkstelligen door allereerst mijn criteria voor annoteren zo helder mogelijk te formuleren en dit als leidraad te laten gelden tijdens het annoteren. Vervolgens heb ik, na de aanvankelijke ronde waarin ik de fragmenten voor het eerst beoordeelde, een tweede controleronde ingelast waarin ik de fragmenten voor een tweede keer bekeek, waarbij ik lette op de consistentie van mijn beoordelingen. In een latere fase van annoteren heb ik daarnaast nog steekproefsgewijs enkele delen van de data gecontroleerd. Daarom zal ik me in deze paragraaf beperken tot de overeenstemming *tussen* (in tegenstelling tot *intra*) de drie beoordelaars, die ik in het vervolg zal aanduiden met de gebruikelijkere term 'inter-beoordelaarsbetrouwbaarheid'.

In het boek *Best Practices in Quantitative Methods* (2011) bespreken Steven E. Stemler & Jessica Tsai (2011) in hun hoofstuk over inter-beoordelaarsbetrouwbaarheid de meest gebruikelijke benaderingen voor de berekening ervan. De keuze voor een benadering hangt vooral af van de vraag wat het doel is van het uitvoeren van inter-beoordelaarsbetrouwbaarheidstoets. In het geval van dit onderzoek ben ik op zoek naar wat Stemler & Tsai de *consensus estimates* noemen. Deze worden gewoonlijk gebruikt om aan te tonen dat een concept dat wordt beschouwd als subjectief alsnog op ongeveer dezelfde wijze wordt beoordeeld door meerdere personen (p. 32).[7] Hierbij wordt een subjectieve interpretatie gemaakt van bijvoorbeeld een kunstwerk om vervolgens een onderliggend concept (denk aan abstracte zaken als creativiteit of, in deze studie: focalisatie) te beoordelen (p. 30). De veronderstelling hierbij is dat als beoordelaars onafhankelijk van elkaar tot overeenstemming komen over verschillende aspecten van het concept dat ze beoordelen, deze overeenstemming bewijs levert voor 'the existence of the construct' (p. 32). Met dit laatste bedoelen ze dat het

[7] De *consensus estimate* omvat strikt genomen slechts een deel van de inter-beoordelaarsbetrouwbaarheid, die bijvoorbeeld ook kan worden berekend aan de hand van *consistency estimates* (Stemler & Tsai 2011, p. 38). Voor mijn onderzoek is slechts de mate van consensus relevant en acht ik het gerechtvaardigd om zaken als consistentie van de beoordelaars buiten beschouwing te laten. Wanneer ik schrijf over inter-beoordelaarsbetrouwbaarheid doel ik dus op de '*consensus estimate* van de inter-beoordelaarsbetrouwbaarheid'.

subjectieve ‘abstracte concept’ (creativiteit of focalisatie) in het geval van overeenstemming kan worden gevangen in beschrijvende aspecten waarvan een gedeelde interpretatie kan bestaan.

Alvorens ik verder ga met het bespreken van de *consensus estimates* wil ik het bovenstaande concreet maken aan de hand van het huidige onderzoek. Zoals gesteld ben ik voor dit onderzoek geïnteresseerd in de vraag of de focalisatie van een fragment in een roman door verschillende lezers op ongeveer dezelfde wijze wordt geïnterpreteerd. De vier aspecten (positie focalisator, waarneembaarheid, etc.) die de lezers beoordelen vormen samen een representatie van het subjectieve concept ‘focalisatie’. Wanneer er voldoende overeenstemming blijkt te zijn over de interpretatie van de vier aspecten, dan geldt dat als aanwijzing voor de mogelijkheid dat focalisatie kan worden geformaliseerd. Concreet betekent dit dat de 100 fragmenten die door drie lezers zijn beoordeeld, mogelijk aantonen dat de 1000 fragmenten die door mij (één lezer) zijn beoordeeld niet slechts mijn persoonlijke interpretatie van focalisatie uitdrukken. Dit zou voldoende aanleiding zijn om aan te nemen dat die laatste groep van 1000 fragmenten geschikt is als trainingsmateriaal voor het vervolg van dit onderzoek. Daarnaast vormen de uitkomsten van de inter-beoordelaarsbetrouwbaarheidstoets een *benchmark* voor het algoritme dat de fragmenten zal gebruiken als trainingsmateriaal. Kortom, bij het toetsen van de *consensus estimates* is het tot stand brengen van inter-beoordelaarsbetrouwbaarheid een doel *an sich*. De toets laat zien of de beoordelingen ‘werken’ of dat sommige aspecten nog (meer) moeten worden aangepast, willen verschillende lezers focalisatie op dezelfde manier beoordelen (Stemler & Tsai 2011, p. 30). Dit kan door bijvoorbeeld een of meer categorieën van een aspect te herzien. Uiteraard kan uit de toets ook blijken dat er te weinig overeenstemming bestaat rondom een bepaald aspect, waardoor dit in het vervolg van het onderzoek beter buiten beschouwing kan worden gelaten.

De keuze voor een toets voor het berekenen van de consensus is afhankelijk van de soort data die is gebruikt. Stemler & Tsai (2011) wijzen erop dat voor nominale schalen het beste Cohen’s kappa gebruikt kan worden (p. 32). Deze toets drukt de mate van overeenstemming tussen twee beoordelaars uit door de verhouding tot de kans op overeenstemming tot berekenen (Cohen 1960). Daarbij geldt dat beoordelaars bij een kappa-score van 0 niet meer overeenkomen dan op basis van kans zou worden verwacht. In dit onderzoek gebruik ik nominale schalen voor de aspecten van de positie van de focalisator en de standvastigheid van het gefocaliseerde object. Ook de binaire schaal van het aspect van waarneembaarheid geldt als nominaal, aangezien de categorie ‘niet eenduidig te benoemen’ voor een derde optie van beoordelen zorgt. Daarom zal ik ook voor het aspect van waarneembaarheid Cohen’s kappa gebruiken bij het berekenen van

de inter-beoordelaarsbetrouwbaarheid. Een bijkomend voordeel van het gebruiken van dezelfde maat voor deze drie aspecten is dat de scores onderling vergeleken kunnen worden. Voor het toetsen van de overeenstemming tussen *drie* beoordelaars door middel van Cohen's kappa moet echter een extra stap worden toegevoegd. Eerst bereken ik per beoordelaarspaar (dus beoordelaar 1 – beoordelaar 2, beoordelaar 1 – beoordelaar 3, etc.) de kappa-score, om vervolgens (handmatig) het gemiddelde te nemen van de drie scores (Hallgren 2012, p. 28; Light 1971).

Het vierde aspect, dat van de mate van detail, is ordinaal (Likertschaal) met een extra categorie voor fragmenten waarin het aspect niet eenduidig valt te benoemen. De overeenstemming voor dit aspect zal ik berekenen aan de hand van de zogeheten *weighted kappa* (Cohen 1968). De weighted kappa is een variatie op Cohen's kappa met als aanpassing dat het gewicht geeft aan de mate van afwijking wanneer beoordelingen van elkaar verschillen. Op die manier kan naast de vraag of twee beoordelaars van elkaar afwijken dus ook uitdrukking worden gegeven aan de eventuele grootte van de afwijking. Hierbij geldt een verschil van 1 (bijvoorbeeld een fragment dat door twee beoordelaars respectievelijk 4 en 5 krijgt toebedeeld) als een hogere mate van overeenstemming dan een verschil van 4 (bijvoorbeeld annotaties van 1 en 5 voor hetzelfde fragment). Ik zal de score berekenen door wederom het gemiddelde te nemen van de scores van ieder beoordelaarspaar.

Aldus zal ik twee verschillende toetsen uitvoeren, te weten Cohen's kappa en weighted kappa. Hiertoe maak ik gebruik van het Python-pakket *scikit-learn* (Pedregosa et al. 2011). Vooraleer over te gaan naar het uitvoeren van de statistische toetsen, zal ik de data van de verschillende beoordelaars per aspect verkennen aan de hand van de frequenties van iedere categorie.

3.5.1 Frequenties

De frequenties zijn vooral relevant voor de trainfase van het onderzoek, omdat ik dan wil weten of er voor iedere categorie van een aspect genoeg trainingsmateriaal verzameld is. De inter-beoordelaarsbetrouwbaarheid kan immers nog steeds worden berekend als een aspect niet gelijk is verdeeld. Daarom ben ik voor nu geïnteresseerd in de overeenstemming tussen de beoordelaars – om vast te kunnen stellen of er sprake is van *existence of construct* – en niet zozeer in de verdeling van de data. Een verkenning van de frequenties dient daarom in deze paragraaf slechts om een beeld te vormen van de data, zodat de statistische toetsen gemakkelijker geïnterpreteerd kunnen worden.

In tabel 1 zien we de frequentieverdeling voor het aspect van de positie van de focalisator per beoordelaar. Het zwaartepunt van dit aspect ligt voor alle drie de beoordelaars overduidelijk bij de interne focalisatie. Het is opvallend hoe sterk de oververtegenwoordiging van deze positie is en en hoe weinig categorieën als extern en dialoog voorkomen. Wellicht heeft dit te maken met de grootte van de geannoteerde fragmenten, omdat een hoofdstuk dat volledig intern is gefocaliseerd is opgedeeld in een aantal fragmenten. Een andere factor die deze verdeling mogelijk (gedeeltelijk) verklaart is de definitie van de positie van de focalisator zoals vastgesteld in het annotatieschema, waarin voor de focalisator vooral een onderscheid wordt gemaakt tussen de 'gelijktijdig waarnemende' focalisator (intern) en de 'herinnerende' of 'terugblikkende' ik-verteller. Ook kan de corpuskeuze een rol spelen, bijvoorbeeld wanneer de interne focalisator in de hedendaagse Nederlandse literatuur simpelweg zeer dominant is. De frequenties in tabel 1 roepen dus een aantal vragen op over het functioneren van de focalisator en het is de moeite waard om hier in het resultatenhoofdstuk nog op terug te komen.

Wat de overeenstemming tussen de beoordelaars betreft zien de frequenties er op het eerste gezicht ongeveer hetzelfde uit, hoewel de tabel geen informatie geeft over het aantal fragmenten dat door alle beoordelaars op dezelfde manier is geannoteerd. Overigens geven Stemler & Tsai (2011) aan dat Cohen's kappa zich goed leent voor het berekenen van overeenstemming wanneer een groot deel van de data binnen één categorie valt. Het is daarbij van belang op te merken dat de kappa berekent in hoeverre er meer overeenstemming is dan op basis van kans. In het geval van dit aspect, waarbij een categorie oververtegenwoordigd is, is het dus van belang dat ook de weinige fragmenten die in de andere categorieën vallen op dezelfde manier zijn geannoteerd door de beoordelaars. Dit wordt door Cohen's kappa meegenomen in de berekening.

	Extern	Consonant	Intern	Restcat.	Dialoog
Beoordelaar 1	6	4	87	0	3
Beoordelaar 2	7	2	84	3	4
Beoordelaar 3	6	8	80	0	6

Tabel 1. Frequenties per beoordelaar voor het aspect van de positie van de focalisator.

De tabellen met frequenties van waarneembaarheid en standvastigheid zijn daarentegen gelijkmatiger verdeeld (respectievelijk tabel 2 en tabel 3). De waarneembaarheid is door alle beoordelaars het vaakst aangeduid als 'niet eenduidig', maar de verschillen zijn veel minder groot in vergelijking met de frequenties van de positie van de focalisator. Daarnaast lijken de

beoordelingen van de drie beoordelaars redelijk goed overeen te komen. Dit geldt ook voor de standvastigheid in tabel 3, hoewel de beoordelingen iets verder uiteenlopen. Ook komt de categorie van meer dan twee objecten aanzienlijk minder voor. Om die reden zal ik voor het berekenen van de inter-beoordelaarsbetrouwbaarheid ook een toets uitvoeren waarbij de categorieën ‘twee objecten’ en ‘meer dan twee objecten’ zijn samengevoegd. Dit acht ik verdedigbaar temeer daar tijdens het annoteren bleek dat dit onderscheid niet altijd even duidelijk is.

	Niet-waarneembaar	Waarneembaar	Niet eenduidig
Beoordelaar 1	25	31	44
Beoordelaar 2	31	31	37
Beoordelaar 3	25	30	45

Tabel 2. Frequenties per beoordelaar voor het aspect van waarneembaarheid van het gefocaliseerde object.

	Eén object	Twee objecten	> Twee objecten
Beoordelaar 1	49	46	5
Beoordelaar 2	58	33	9
Beoordelaar 3	48	47	5

Tabel 3. Frequenties per beoordelaar voor het aspect van standvastigheid van het gefocaliseerde object.

De frequenties van het vierde aspect, dat van de mate van detail, zijn ook niet gelijkmatig verdeeld (zie tabel 4). Het overgrote deel van de fragmenten is namelijk ingedeeld in de categorieën ‘mee eens’ en ‘helemaal eens’ (gedetailleerd en zeer gedetailleerd). Bovendien is de categorie voor zeer lage mate van detail door geen van de beoordelaars gebruikt, terwijl lage mate van detail ook een lage frequentie kent. Is de mate van detail afhankelijk van de fragmentgrootte of kent een gefocaliseerd object al gauw een zekere mate van detail zodra het überhaupt wordt waargenomen? Hier zal ik eveneens in het resultatenhoofdstuk op reflecteren.

Desondanks kennen de frequenties voor de drie beoordelaars op het eerste gezicht een soortgelijk patroon waarbinnen ze alle drie veel vaker beoordelen als ‘mee eens’ of ‘helemaal eens’ dan de andere categorieën. Tegelijkertijd is er enige variatie tussen de beoordelaars, bijvoorbeeld in de verdeling van de categorieën ‘mee eens’ en ‘helemaal eens’. De inter-beoordelaarstoetsen zullen uitwijzen in hoeverre er daadwerkelijk een hogere mate van overeenstemming bestaat dan op basis van toeval mag worden verwacht.

	Helemaal oneens	Oneens	Niet zeker	Mee eens	Helemaal eens	Niet eenduidig
Beoordelaar 1	0	5	16	36	31	12
Beoordelaar 2	0	11	7	53	13	16
Beoordelaar 3	0	7	11	33	38	11

Tabel 4. Frequenties per beoordelaar voor het aspect van gedetailleerdheid van het gefocaliseerde object, waarbij 'helemaal oneens' staat voor lage mate van detail en 'helemaal eens' voor hoge.

3.5.2 Toetsen

Zoals hierboven gesteld toets ik de inter-beoordelaarsbetrouwbaarheid tussen de beoordelaars om de mate van overeenstemming te bepalen bij het annoteren van vier aspecten van focalisatie. Voor het eerste aspect, dat van de positie van de focalisator, is Cohen's kappa (Cohen 1960) berekend voor ieder beoordelaarspaar en vervolgens gemiddeld om tot één score van overeenstemming te komen (Light 1971). Het resultaat duidt op voldoende tot goede overeenstemming, namelijk $\kappa = 0{,}66$ (B1—B2: $\kappa = 0{,}73$; B1—B3: $\kappa = 0{,}70$; B2—B3: $\kappa = 0{,}56$) (Landis & Koch 1977, p. 165). Dit suggereert dat de beoordelaars voldoende overeenstemmen wat betreft de positie van de focalisator, hoewel er tussen de beoordelingen enige variatie bestaat die meegenomen dient te worden in de vervolganalyses. Desalniettemin blijkt er voor dit aspect voldoende bewijs voor de *existence of construct* te zijn en is het geschikt om te gebruiken in het vervolg van dit onderzoek.

Voor het aspect van waarneembaarheid is eveneens Cohen's kappa (Cohen 1960) getoetst per beoordelaarspaar, waarop het gemiddelde van de drie scores is genomen (Light 1971). Hieruit bleek dat er bijna perfect overeenstemming is, $\kappa = 0{,}83$ (B1—B2: $\kappa = 0{,}88$; B1—B3: $\kappa = 0{,}81$; B2—B3: $\kappa = 0{,}79$) (Landis & Koch 1977, p. 165). Deze score duidt op een hoge mate van overeenstemming voor de waarneembaarheid van het gefocaliseerde object, dus dit aspect leent zich uitstekend voor de hieropvolgende analyse.

Als derde heb ik Cohen's kappa (Cohen 1960) uitgevoerd voor het aspect van standvastigheid. Wederom zijn eerst de scores per beoordelaarspaar berekend om vervolgens het gemiddelde te nemen (Light 1971). Uit de score blijkt voldoende tot goede overeenstemming, $\kappa = 0{,}71$ (B1—B2: $\kappa = 0{,}77$; B1—B3: $\kappa = 0{,}73$; B2—B3: $\kappa = 0{,}63$) (Landis & Koch 1977, p. 165). Hoewel deze score niet direct aanleiding geeft tot aanpassingen, neemt de kappa sterk toe wanneer de categorieën 'twee gefocaliseerde objecten' en 'meer dan twee gefocaliseerde objecten' worden samengevoegd. De scores zien er dan namelijk als volgt uit: κ

= 0,80 (B1—B2: κ = 0,82; B1—B3: κ = 0,82; B2—B3: κ = 0,76). Na deze aanpassing kent dit aspect dus een overeenstemming die sterk is verbeterd en zelfs nagenoeg binnen de categorie 'bijna perfect' valt (Landis & Koch, p. 165). Om die reden kies ik ervoor in het vervolg van dit onderzoek voor het aspect van standvastigheid te werken met de samengevoegde categorieën (dus met een categorie minder), zoals in de laatste kappa-toets. Bovendien liet de frequentietabel van dit aspect al zien dat deze aanpassing de trainfase zal vergemakkelijken.

Tot slot is voor de mate van detail, het vierde aspect, de weighted kappa met lineair gewicht berekend (Cohen 1968). Na de scores per beoordelaarspaar te hebben gemiddeld (Light 1971), bleek de score flink lager uit te vallen dan die van de eerste drie aspecten. Er blijkt weliswaar voldoende tot goede overeenstemming te zijn tussen de beoordelaars, maar de score valt maar nipt binnen deze categorie van overeenstemming: κ = 0,62 (B1—B2: κ = 0,65; B1—B3: κ = 0,73; B2—B3: κ = 0,47) (Landis & Koch 1977, p. 165). Een aanpassing van het aspect waarbij de twee laagste en de twee hoogste categorieën worden samengevoegd (respectievelijk zeer laag en laag, en zeer hoog en hoog), levert zelfs een lagere score op voor Cohen's kappa (Cohen 1960), κ = 0,57. De uitkomsten van deze inter-beoordelaarsbetrouwbaarheidstoets suggereren dat er ook voor dit laatste aspect voldoende overeenstemming is ten opzichte van de kans op overeenstemming. Echter, de scores laten ook zien dat er een aanzienlijke foutmarge zal zitten in het trainingsmateriaal, aangezien de subjectieve beoordelingen in de annotaties deze ruimte laten. Het is van belang dit mee te nemen in het vervolg van dit onderzoek omdat het onderscheidend vermogen van deze laatste toets minder groot is dan dat van de andere drie aspecten.

Aldus wijzen de toetsen voor de inter-beoordelaarsbetrouwbaarheid op (minstens) voldoende overeenstemming voor alle aspecten, waarmee ik voor alle vier tevens de *existence of construct* bewezen acht. Daarmee zijn alle aspecten geschikt om te gebruiken in het vervolg van dit onderzoek, waarbij het aspect van standvastigheid een lichte aanpassing kent.[8]

[8] In het resultatenhoofdstuk wil ik de scores van de toetsen die ik in deze paragraaf heb uitgevoerd kunnen vergelijken met die van de verschillende algoritmes die ik gebruik bij de automatische classificatie. De eerste gelden daarbij als *ground-truth* scores: de annotaties van de lezers zijn het uitgangspunt met betrekking tot het herkennen van de verschillende aspecten van focalisatie. De voorspellingen van de algoritmes en de scores die ze opleveren geven inzicht in hoe dicht het algoritme de menselijke lezer benadert, dus hoe 'goed' het systeem werkt. Om deze vergelijking mogelijk te maken, zet ik de kappa-scores uit deze paragraaf om naar zogeheten *F*1-scores, die ik in het resultatenhoofdstuk ook gebruik om de algoritmes te evalueren. In de evaluatieparagraaf geef ik meer uitleg over deze score en over de verhouding tussen annotaties en automatische classificaties.

3.6 Modellen

In deze paragraaf beschrijf ik de stappen die ik doorloop om de classificatiemodellen aan de computer te 'leren' en de keuzes die ik daarbij maak. Ik maak gebruik van de gevestigde classificatiemethode logistische regressie en het taalmodel BERT. Hoewel beide technieken hun oorsprong vinden in het onderzoeksveld van *natural language processing* (NLP)[9], dat zich bevindt op het grensvlak van taalkunde, computerwetenschappen en kunstmatige intelligentie, kennen beide technieken inmiddels brede toepassingen binnen de geesteswetenschappen. Allereerst zal ik zowel logistische regressie als BERT introduceren, om inzicht te geven in wat het idee is achter de technieken die ik gebruik. Omdat ik in dit onderzoek veeleer geïnteresseerd ben in de toepassing van bestaande classificatietechnieken en de kwalitatieve implicaties ervan – in plaats van in de precieze technische werking van de technieken – richt ik me vervolgens in de theoretische inbedding van BERT vooral op vergelijkbare geesteswetenschappelijke en letterkundige toepassingen waar ik op voortbouw. Tot slot zal ik voor beide algoritmes toelichten welke keuzes ik heb gemaakt bij de toepassing ervan op de dataset en de annotaties met vier aspecten van focalisatie.

Wat de input voor de vier modellen betreft heb ik alle fragmenten die zijn geannoteerd als '99' verwijderd uit het corpus van het betreffende aspect. Dit betekent dat ik per aspect het volgende aantal fragmenten overhoud: 977 voor focaliserend subject, 847 voor waarneembaarheid, 1000 voor standvastigheid en 969 voor gedetailleerdheid.

3.6.1 Logistische regressie

Logistische regressie is een van de meest gebruikte vormen van *machine learning* en geldt volgens Daniel Jurafsky & James H. Martin (2023) als de 'baseline supervised machine learning algorithm for classification' (p. 79). Bovendien stelt Ted Underwood in *Distant Horizons* (2019) dat modellen gebaseerd op relatief simpele woordfrequenties – waaronder logistische regressie – even goed menselijke oordelen over literaire teksten kunnen voorspellen als complexere modellen (p. 21). Het lijkt dan ook logisch om logistische regressie als uitgangspunt te nemen voor mijn classificatiemodel, aangezien het een goed richtpunt zal vormen voor het modelleren van focalisatie.

Ted Underwood (2019) legt logistische regressie op een conceptueel niveau uit (193-194). Stel dat je een aantal datapunten hebt waar je een rechte trendlijn doorheen trekt, dan voer

[9] Een voorbeeld van onderzoek naar automatische tekstclassificatie binnen het veld van NLP is Van der Burgh & Verberne (2019), die aan de hand van een dataset van boekrecensies van *Hebban* het aantal sterren waarmee een recensent een recensie heeft beoordeeld proberen te voorspellen.

je lineaire regressie uit. Deze lijn drukt de relatie uit tussen twee variabelen. Nieuwe metingen van een van beide variabelen kunnen vervolgens op de trendlijn worden geplaatst, waardoor je een inschatting kunt maken van de tweede variabele van het betreffende datapunt. Logistische regressie wordt gebruikt wanneer de variabele die door de trendlijn wordt voorspeld een binaire waarde heeft. In dat geval worden de datapunten die de lineaire functie oplevert getransformeerd naar een waarde tussen de 0 en 1 door middel van een zogeheten sigmoïde (logistische functie) (Jurafsky & Martin 2023, p. 81). Underwood beschrijft vervolgens dat de modellen die gewoonlijk worden ingezet voor tekstclassificatie duizenden predictor-variabelen hebben die de afhankelijke, binaire variabele voorspellen. Daarnaast is het gebruikelijk om regularisatie toe te voegen aan logistische regressie (Underwood 2019, p. 195; Jurafsky & Martin 2023, p. 97-98). Dit voorkomt dat een model slechts het trainingsmateriaal kan herhalen, in plaats van dat het kan generaliseren en onbekende data kan classificeren. Een veel gebruikte vorm van regulariseren is de L2-regularisatie.

Voor mijn eigen toepassing van logistische regressie met L2-regularisatie maak ik gebruik van het Python-pakket *scikit-learn* (Pedregosa et al. 2011). De regularisatiesterkte *C* zet ik op 1. Daarnaast maak ik gebruik van de mogelijkheid binnen *scikit-learn* om het gewicht van de klassen van een aspect te balanceren op basis van de frequenties. Voor de binaire aspecten van waarneembaarheid en standvastigheid (na de aanpassing op basis van de inter-beoordelaarsbetrouwbaarheid) gelden deze instellingen. De positie van de focalisator en de gedetailleerdheid bestaan uit meer dan twee klassen en daarom verander ik de parameter ‘multi class’ voor deze twee aspecten in ‘multinomial’. Verder volg ik Ted Underwood in zijn benadering van logistische regressie als hij stelt dat het maximaliseren van de accuraatheid van het model niet zijn doel is (p. 196). Mijn doel is om een geïnformeerd beeld te krijgen van de mate waarin ik met logistische regressie focalisatie kan herkennen en hoewel ik verschillende mogelijkheden verken, ben ik niet op zoek naar het model dat nét een beetje hoger scoort door de instellingen nauwkeuriger af te stellen. Logistische regressie heeft binnen deze scriptie namelijk de functie van richtpunt of *baseline*. Enerzijds dient deze om te kunnen vergelijken met de handmatige annotaties. Anderzijds veronderstel ik dat het complexere taalmodel BERT beter in staat zal zijn om focalisatie te classificeren. Logistische regressie dient in dat geval als richtpunt dat een indicatie geeft van hoeveel beter een complexer model presteert. Op de precieze verhouding tussen deze drie (annotaties, logistische regressie, BERT) ga ik nog nader in in de volgende paragraaf.

De datasets zijn vanwege de beperkte omvang steeds als volgt verdeeld: 80% van de fragmenten wordt gebruikt voor een zogeheten *10-fold cross-validation*, met de overige 20%

worden de modellen getest. Bij *10-fold cross-validation* wordt de dataset opgedeeld in tien verschillende subsets, de *folds* (Jurafsky & Martin 2023, p. 71). Van de tien subsets dient steeds één subset als test-set, de rest wordt gebruikt als train-set, waarna de score van het model op de test-set wordt berekend (p. 71). Dit wordt in totaal tien keer uitgevoerd met steeds een andere test-set, terwijl de resterende 90% dient als train-set. Op die manier worden dus tien modellen getraind op 90% van de data (p. 71). Vervolgens wordt de 10% van de data die gedurende de training apart is gehouden gebruikt om de *10-fold cross-validation* te testen (p. 71). De scores die hieruit voortkomen vergelijk ik met die van het BERT-model.

Een belangrijke factor die ik nog niet heb besproken is die van numerieke tekstrepresentatie, ook wel vectoriseren genoemd. Met andere woorden, op welke manier kunnen de woorden die als input dienen ook daadwerkelijk betekenis krijgen voor het model? De meest simpele manier van vectorrepresentatie heet een *bag-of-words*-model (BoW-model). Dit model bestaat uit een vocabulaire dat alle unieke woorden uit een corpus bevat. Elk document (in dit onderzoek: fragment) wordt gerepresenteerd door een vector. Deze vector bestaat uit een lijst van frequenties voor ieder woord uit het vocabulaire. Op basis van vectoren kan bijvoorbeeld de afstand tussen twee verschillende vectoren worden gemeten. Ik gebruik deze manier van representeren, omdat BoW-modellen effectief werken voor veel tekstclassificatieproblemen (Underwood 2019, p. 196). Door de simpliciteit van BoW-modellen kennen ze echter ook beperkingen. Om die reden werk ik daarnaast met ‘term frequency-inverse document frequency’ (tf-idf). Een tf-idf-model geeft gewicht aan woordfrequenties in een document (tf) afhankelijk van hoeveel documenten in het corpus die woorden bevatten (idf). Als een woord vaak voorkomt in een klein aantal documenten, krijgt het veel gewicht toebedeeld binnen een vector. Dergelijke woorden gelden als topicale woorden. Daarentegen worden hoogfrequente woorden over veel documenten (zoals functiewoorden) en laagfrequente woorden (zoals spelfouten) gerepresenteerd met minder gewicht.

3.6.2 BERT

Het taalmodel BERT [10], dat staat voor ‘Bidirectional Encoder Representations from Transformers’, werd geïntroduceerd in 2019 en betekende een doorbraak in het onderzoek met grootschalige taalmodellen. Voortaan was het relatief eenvoudig voor onderzoekers om te werken met *state-of-the-art* taalmodellen, ook zonder enorme datasets en zeer krachtige

[10] Het originele BERT-model is geïntroduceerd in een paper van Devlin, Chang, Lee & Toutanova (2019). Wanneer ik het model bespreek en hiervoor informatie van de makers gebruik, haal ik dus dit artikel aan.

computers. In mijn onderzoek fungeert de logistische regressie als referentiepunt om te verkennen in hoeverre een algoritme een notie kan hebben van focalisatie. De resultaten kan ik vervolgens vergelijken met die van BERT, waarvan ik veronderstel dat het in hogere mate in staat zal zijn focalisatie te herkennen in literaire teksten. Om de werking van BERT uit te leggen zal ik hieronder de belangrijkste concepten bespreken. Op die manier hoop ik een brug te slaan tussen enerzijds technisch computationeel onderzoek en anderzijds letterkundig onderzoek dat zich goed leent voor computationele toepassingen. Ik veronderstel namelijk dat BERT geschikt is voor het proces van operationaliseren in een letterkundige context, zoals ik aan het begin van dit methodehoofdstuk heb beschreven.

Representation learning

De *distributional hypothesis* veronderstelt dat woorden die in vergelijkbare contexten voorkomen vergelijkbare betekenissen hebben (Jurafsky & Martin 2023, p. 103). Zo zullen bijvoorbeeld de woorden ‘dokter’ en ‘arts’ beide vaak omgeven zijn door woorden die met een medische omgeving te maken hebben. Dit zou betekenen dat de betekenis van een woord kan worden geleerd aan de hand van de woorden die vaak samen met het woord voorkomen.[11] De *distributional hypothesis* wordt in NLP ingezet in de vorm van zogeheten *vector semantics*. Dit zijn representaties van de betekenis van woorden (deze representaties worden *embeddings* genoemd) bestaande uit omringende woorden (p. 107). Op die manier zullen twee woorden met een vergelijkbare context op een vergelijkbare manier worden gepresenteerd in een vector. De woorden ‘arts’ en ‘dokter’ komen op die manier beide relatief vaak voor in combinatie met bijvoorbeeld ‘patiënt’ en ‘ziek’. Met andere woorden, wanneer een taalmodel de betekenis van woorden leert, wordt in NLP gebruik gemaakt van *vector semantics* om de *distributional hypothesis* in de praktijk in te zetten door representaties van woordbetekenis te leren. De betekenis van het woord ‘dokter’ zou dus gerepresenteerd worden door middel van een vector die woorden als ‘patiënt’ en ‘ziek’ bevat. Deze representaties noemen we zoals gezegd *embeddings*. Deze vorm van leren heet ook wel *representation learning* waarbij op basis van een inputtekst bruikbare representaties worden gedestilleerd door het model (p. 103). Dit kan vervolgens worden gebruikt voor allerhande taken, waaronder classificeren zoals in dit onderzoek.

[11] Hoewel, ‘betekenis’ is in dit verband een abstract begrip zonder heel duidelijke eigenschappen. De woordenboekbetekenis van een woord, bijvoorbeeld, zal in veel gevallen niet volstaan om de volledige betekenis van een woord te vatten (denk aan zaken als connotatie of metaforische betekenislagen).

Contextual embeddings

De *embeddings* kunnen woordbetekenis op een statische manier representeren door alle keren dat een woord voorkomt hiervoor dezelfde vector te gebruiken. Woordbetekenis wordt in dit geval gerepresenteerd als woordtypes, waarbij hetzelfde woord voor iedere context dezelfde *embedding* heeft. Een alternatieve vorm van *embeddings*, en hiervan maakt BERT gebruik, bestaat eruit dat iedere keer dat een woord voorkomt, er een andere representatie van dat woord wordt gemaakt als het zich voordoet in een andere context (Jurafsky & Martin 2023, p. 236-237; Devlin, Chang, Lee & Toutanova 2019, p. 1-2). Deze manier van het representeren van context heet *contextual embeddings*. De vectors waaruit contextuele *embeddings* bestaan vormen een representatie van de betekenis van woordtokens: een specifiek voorkomen van een woordtype in een specifieke context (Jurafsky & Martin 2023, p. 237). De meest gebruikelijke toepassing van *contextual embeddings* is als input voor classificatie door middel van *fine-tuning*, wat de laatste stap vormt in een model dat gebruikmaakt van *transfer learning*, zoals ook het geval is bij BERT. Hierop kom ik nog terug.

Bidirectional encoders

De contextuele *embeddings* maken bij het tot stand komen van representaties gebruik van zogeheten *bidirectional encoders* (Jurafsky & Martin 2023, p. 228-229). Een *encoder* krijgt een hoeveelheid tekst als input en zet deze om naar een contextuele representatie (p. 202). Dit omzetten kan gebeuren met behulp van verschillende soorten encoders, bij BERT is gebruikgemaakt van een *transformer-encoder* (p. 228-229). Met bidirectionele encoder wordt bedoeld dat een encoder voor de contextuele representaties gebruikmaakt van alle informatie op basis van de tekstuele input, in plaats van slechts informatie van links naar rechts gelezen (p. 229). De inzet van een bidirectionele encoder heeft als voordeel dat het model in staat is tot meer complexe taken die meer contextinformatie vereisen, zoals classificatietaken.

Ik zal een voorbeeld geven dat het belang van een bidirectionele encoder verduidelijkt. Wanneer we de zinnen ‘Ik zit op de bank.’ en ‘De bank verstrekt geen leningen.’ vergelijken, weten we dat het verschil tussen beide betekenissen van ‘bank’ kan worden herkend op basis van de context waarin het woord voorkomt. We merken dan bijvoorbeeld op dat de woorden ‘zit’ en ‘op’ vaker voorkomen samen met het meubelstuk, terwijl ‘verstrekt’ en ‘leningen’ waarschijnlijk in dezelfde zin staan als de financiële instelling. Wanneer we echter slechts van links naar rechts kunnen lezen, weten we in de tweede zin alleen nog maar dat het woord ‘bank’ is voorafgegaan door ‘de’ en kunnen we nog maar lastig bepalen met welke betekenis van het woord we te maken hebben. Een encoder representeert woordbetekenis weliswaar op basis van

de gehele context – dus dit voorbeeld is niet volledig toereikend, bovendien is het ook een versimpelde voorstelling van de werking van een encoder – maar het principe werkt ongeveer hetzelfde. Wanneer de encoder van links naar rechts werkt, bestaat de representatie van de input 'bank' in de tweede zin namelijk uit een context die niet duidelijk maakt om welke betekenis van het woord het gaat, aangezien het die informatie niet heeft kunnen zien. Een bidirectionele encoder lost dit probleem op door ieder token uit de input te representeren met context uit de gehele input (p. 229).

Transfer learning

Transfer learning houdt in dat de kennis die een taalmodel heeft vergaard kan worden overgeheveld naar andere domeinen, om het model vervolgens in te kunnen zetten voor andere taken (Jurafsky & Martin 2023, p. 228). Dit gebeurt door middel van twee stappen: *pretraining* en *fine-tuning*. Het eerste houdt in dat de representaties van woord- of zinsbetekenissen door een model worden geleerd op basis van zeer grote hoeveelheden tekst. Daarop volgt de fase van *fine-tuning*, de representaties van het voorgetrainde model fungeren als basis om een model te trainen voor een specifieke taak. Een voorgetraind model wordt in het geval van bidirectionele encoders getraind door steeds een weggelaten woord in een zin te voorspellen op basis van de rest van de zin, dit heet *masked language modeling* (p. 232). De onderliggende aanname van deze vorm van *transfer learning* is dat de stap van *fine-tuning* relatief wordt vereenvoudigd dankzij het voorgetrainde model (p. 228). De toepassing op een specifieke taak kan namelijk worden getraind door bovenop de output van het voorgetrainde model een klein aantal classificatielagen toe te voegen die de toepassing van het model op een specifieke taak mogelijk maakt (p. 237). Door middel van *transfer learning* wordt zodoende de kennis van het heel grote, voorgetrainde taalmodel bewaard. Hier kan in de fase van *fine-tuning* op worden voortgebouwd zodat het model in deze fase dus niet volledig opnieuw alle kennis hoeft te leren. In de praktijk kan één voorgetraind model dus vele toepassingen hebben doordat andere onderzoekers hun eigen *fine-tune*-laag toevoegen.

Dit is dan ook hoe ik mijn model opbouw in dit onderzoek: ik maak gebruik van een voorgetraind model en voeg daar een laatste classificatielaag met mijn eigen data aan toe. Ik gebruik een BERT-model dat is getraind op een Nederlandstalig corpus, genaamd BERTje[12], als mijn voorgetrainde model. Dit model maakt gebruik van dezelfde architectuur en parameters als het 'orginele' BERT-model, maar het is getraind op een grootschalig Nederlandstalig corpus

[12] Dit Nederlandstalige BERT-model staat beschreven in De Vries, Van Cranenburgh, Bisazza et al. (2019). Ik verwijs dus naar hun paper voor details over BERTje.

zodat het door middel van een *fine-tuning*-taak toepassingen kan hebben voor het Nederlands. Het trainingscorpus van BERTje bestaat uit onder meer de volledige Nederlandstalige Wikipedia, een zeer grote hoeveelheid hedendaagse en historische fictie en een corpus van nieuwsberichten. In totaal bestaat de dataset voor de *pretraining* van BERTje uit ongeveer 2,4 miljard tokens (De Vries, Van Cranenburgh, Bisazza et al. 2019, p. 2). Op enkele aanpassingen in het trainingsproces na volgt het model de trainingsstappen van BERT (p. 2-3).

Letterkundige toepassingen van BERT

Eerder letterkundig onderzoek dat gebruik maakt van BERT vinden we bijvoorbeeld bij Matthew Sims, Jong Ho Park en David Bamman (2019), die onderzoek doen naar het herkennen van gebeurtenissen in literaire teksten (*event detection*). Ze stellen dat gebeurtenissen in literaire teksten vragen opwerpen over onder meer de vorm van een tekst en over hoe een tekst is gestructureerd, waarmee ze aansluiten bij een letterkundige traditie die teruggaat tot de Russische formalisten. Tegelijkertijd constateren ze dat er in Natural Language Processing al veel bestaand onderzoek is naar de detectie van gebeurtenissen, maar dan in een context van nieuwsberichten (p. 3623). Ze stellen een letterkundige definitie van een gebeurtenis op en annoteren vervolgens handmatig woorden die een gebeurtenis tot stand brengen ('*event triggers*', p. 3626). Een token wordt dus binair geclassificeerd met ofwel het label *event trigger* ofwel *non-event trigger*. Vervolgens blijken labels voor onbekende tokens het beste voorspeld te kunnen worden met behulp van de contextuele representaties van het voorgetrainde BERT-model (p. 3628-3629).[13] In het laatste deel van het onderzoek passen ze hun model toe om een vergelijking te maken tussen relatieve aantallen evenementen in een corpus met prestigieuze auteurs en een groep niet-prestigieuze auteurs. Dit stelt ze in staat een betekenisvol onderscheid te maken tussen beide groepen (p. 3631). De studie van Sims et al. (2019) maakt dus duidelijk dat het BERT-model succesvol kan worden ingezet voor een binaire classificatietaak met literaire teksten als onderzoeksobject.

In zijn masterscriptie past Joris Veerbeek (2020) het BERTje-model toe met een stap van *fine-tuning* om evaluatieve uitspraken in Nederlandse en Vlaamse dagbladkritiek te herkennen en te categoriseren. Hij zet daarbij BERTje in voor zowel een binaire classificatietaak (evaluatieve uitspraak of niet) als voor een *multi-label* rankingprobleem (voorpellen van aspect- en eigenschapscodes van de uitspraak) (p. 43-44). Dit laatste, *multi-label*, houdt in dat één fragment meerdere labels krijgt toegewezen. Doordat Veerbeek in zijn

[13] Sims et al. (2019) maken gebruik van het voorgetrainde BERT-model en in plaats van *fine-tuning* passen ze het voorgetrainde model aan op basis van eerdere studies naar *event detection* (p. 3629).

onderzoek slechts de bovenste laag van het model door middel van *fine-tunen* traint op zijn eigen dataset, is hij in staat om ondanks relatief weinig taakspecifieke input alsnog te werken met een accuraat model. Vervolgens past hij zijn getrainde model toe op ongeziene recensies, waarna hij Nederlandse en Vlaamse dagbladkritiek tussen 1946 en 2018 kan vergelijken op een groot aantal eigenschappen en aspecten.

Het onderzoek van Veerbeek laat zien dat het Nederlandstalige BERT-model succesvol kan worden ingezet bij het beantwoorden van letterkundige vragen. Ook het *fine-tunen* waarvan hij gebruik maakt dient als voorbeeld voor mijn studie, aangezien het trainen van een beperkt aantal taakspecifieke voorbeelden veel nieuwe mogelijkheden biedt voor letterkundige toepassingen. Bovendien sluit dit aan op het proces van operationaliseren, waarbij een annotatietaak in de buurt komt van een *close reading*. De wens om handmatig teksten te labelen sluit goed aan bij de inzet van het BERT-model, dat een taak kan automatiseren ondanks de relatief geringe hoeveelheid trainingsdata die handmatige annotatie oplevert.

Fine-tunen van BERT

Evenals voor logistische regressie train ik vier BERT-modellen, één voor ieder aspect van focalisatie. Voor zowel de binaire als *multi-class* aspecten maak ik bij het *fine-tunen* van BERT gebruik van *HuggingFace's Transformers* (Wolf, Debut, Sanh et al. 2020) met *PyTorch*-integratie (Paszke, Gross, Massa et al. 2019). Tijdens het *fine-tunen* maak ik gebruik van een *Google Colab*-omgeving, zodat ik mijn scripts kan uitvoeren met een GPU. Ik draai alle modellen voor vijf *epochs*.

3.7 Evaluatie

Beide algoritmes die ik beschreef in de vorige paragraaf dienen geëvalueerd te worden, zodat we weten welk algoritme het geschiktst is voor het classificeren van focalisatie. Ook maakt de evaluatie duidelijk hoe goed een algoritme in staat is om fragmenten te classificeren en, belangrijker nog, waar de valkuilen van het algoritme zitten. Alvorens uit te leggen hoe de prestatie van een algoritme kan worden gemeten en een aanvullende analysemethode te introduceren, licht ik eerst toe hoe de verschillende classificatiemethodes in deze scriptie zich tot elkaar verhouden.

3.7.1 Verhoudingen *classifiers*

De eerste manier van classificeren in deze scriptie werd gedaan door middel van manuele annotaties, die zijn getoetst op voldoende overeenstemming. De labels van deze classificatie gelden als 'gold labels' (Jurafksy & Martin 2023, p. 68) die we de modellen proberen aan te leren. Dit leverde per aspect een score op die dient als *baseline*; dit wil zeggen dat een algoritme dat in staat is om focalisatie te herkennen op ongeveer dezelfde scores zou moeten uitkomen. Het is op zichzelf geen doel om deze *baseline*-scores te evenaren, maar het dient als richtlijn voor de modellen. Als een score ver onder die van de annotaties zit kunnen we immers veronderstellen dat het betreffende algoritme waarschijnlijk behoorlijk afwijkt van een menselijk oordeel. In de zoektocht naar een automatische classificatiemethode die in de buurt komt van de *ground truth*-scores vergelijk ik per aspect logistische regressie (met twee verschillende woordrepresentaties: simpele BoW en BoW met tf-idf) met de Nederlandstalige versie van BERT die gebruik maakt van semantische, contextuele *embeddings*.

Om de scores van de modellen beter te kunnen vergelijken, bereken ik voor ieder aspect een zogeheten *random baseline*. Dit is een model dat labels toekent op basis van frequenties van de categorieën van een aspect, onafhankelijk van de input waarop de modellen zijn getraind. Dit model is met name interessant omdat een aantal aspecten ongelijk is verdeeld en het maakt inzichtelijk in hoeverre een model goed scoort door simpelweg te labelen als de meest frequente categorie. Ik gebruik voor het *random baseline*-model de DummyClassifier (ingesteld als 'stratified'[14]) van *scikit-learn* (Pedregosa et al. 2011). Kortom, de *random baseline* dient ter vergelijking met de modellen. Het beste model kan vervolgens worden afgezet tegen de scores van de handmatige annotaties en nader worden geanalyseerd. Ik veronderstel dat de classificatie met BERT vanwege de complexiteit van het model tot betere resultaten leidt dan die met logistische regressie. Het is echter interessant om logistische regressie als richtpunt te gebruiken voor het taalmodel, om de vergelijking te maken tussen een relatief simpel model en een complex model dat in twee fases getraind is.

[14] De keuze voor het soort *random baseline*-model heeft mogelijk verregaande implicaties voor de interpretatie van de scores van de modellen. Als een *baseline* onterecht zeer hoog scoort, dan lijkt de classificatietaak minder complex dan in werkelijkheid het geval is en lijken modellen die mogelijk lagere scores opleveren onterecht niet goed te scoren. Dit kan bijvoorbeeld voorkomen als 80% van de data in één categorie valt en de *baseline* wordt berekend met de 'most_frequent'-strategie in *scikit-learn*. Een volledig willekeurige *baseline* ('uniform' in *scikit-learn*) zou in hetzelfde voorbeeld echter een te pessimistisch beeld schetsen van de complexiteit van de classificatietaak. De optie 'stratified' geeft een *baseline* die willekeurig classificeert, terwijl het rekening houdt met de verdeling van categorieën binnen een aspect. Hiermee denk ik een *baseline* te gebruiken die daadwerkelijk meer informatie geeft over de scores van de modellen.

3.7.2 Maatstaven

Voor de evaluatie van de prestaties van de modellen maak ik gebruik van drie maatstaven: *precision*, *recall* en *F*1-score. Ik kies voor *F*1-score aangezien verschillende categorieën ongelijk zijn verdeeld, zoals bleek in hoofdstuk 3.4, en omdat Folgert Karsdorp (2016) stelt dat deze score geschikt is als evaluatiemiddel in dergelijke gevallen (p. 62). Deze drie concepten worden op een andere manier berekend voor binaire en *multi-class* aspecten, dus ik bespreek ze beide. Om de scores van de binaire classificatiemodellen te kunnen berekenen, moeten we eerst meer weten over de verhouding tussen het voorspelde label door het model en het label dat is toegekend door de menselijke beoordelaar. Dit kan namelijk op vier manieren. Als beide overeenkomen en positief zijn (categorie van waarneembaar object), dan spreken we van een *true positive*. Voorspelt het model onterecht een positief label, dan geldt dit als een *false positive*. Het model kan ook een negatief label voorspellen (categorie van niet-waarneembaar object). Komt dit overeen met het oordeel van de menselijke beoordelaar, dan is de voorspelling een *true negative*. Als het model ten onrechte een negatief label toekent, is er tot slot sprake van een *false negative*. In tabel 6 staan de vier mogelijkheden in een overzicht.

		Menselijke beoordeling	
		Positief label	Negatief label
Beoordeling model	Positief label	*True positive*	*False positive*
	Negatief label	*False negative*	*True negative*

Tabel 6. De vier mogelijke verhoudingen tussen het label zoals door het model voorspeld en zoals beoordeeld door de menselijke lezer.

Precision richt zich op de fragmenten die zijn geclassificeerd als positief en drukt uit hoeveel van de positieve classificaties ook daadwerkelijk positief zijn. Dit wordt berekend door het aantal *true positives* te delen door de som van *true positives* en *false negatives* (Jurafksy & Martin 2023, p. 69):

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives+false positives}} \quad (1)$$

Recall geeft aan hoeveel fragmenten met het label ‘waarneembaar’ daadwerkelijk op die manier zijn geclassificeerd. Het geeft dus ook een indicatie van de fragmenten die het model heeft gemist. De *recall* wordt berekend door het aantal *true positives* te delen door de soms van *true positives* en *false negatives*. Alleen de laatste component verschilt dus van de berekening van *precision* (Jurafksy & Martin 2023, p. 69):

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives+false positives}} \quad (2)$$

Het laatste concept, de *F*1-score, is het gewogen gemiddelde van *precision* en *recall*. Deze score wordt dus als volgt berekend (Jurafksy & Martin 2023, p. 69):

$$F1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} \cdot \text{recall}} \quad (3)$$

Bij de beoordeling en vergelijking van de prestaties van de modellen zal ik vooral gebruik maken van de *F*1-score, aangezien deze *precision* en *recall* meeneemt in de berekening.

Voor de twee *multi-class* aspecten geldt dat eerst voor ieder label *precision*, *recall* en *F*1-score worden berekend. De *F*1-score voor het gehele aspect bereken ik daarna door het gemiddelde te nemen van de *F*1-scores van de labels (Jurafsky & Martin 2023, p. 70-71), gebalanceerd voor de frequentie van de labels (de optie 'weighted' onder 'average' in *scikit-learn*'s f1_score) (Pedregosa et al. 2011).

3.7.3 Methode van *algorithmic failure*

In de paragraaf over modelleren besprak ik een reflectie van Richard Jean So (2017) waarin hij stelt dat het van belang is te begrijpen welke fouten een model maakt ('to understand how it is wrong', p. 671). De fouten die een model maakt bij een tekstclassificatie dienen als materiaal voor nadere analyse die onderdeel is van het onderzoeksproces. Dit hangt samen met de vraag hoe teksten of onderzoeksmateriaal worden gestructureerd tijdens het interpreteren. Als de fouten van het classificatiemodel hierbij leidend zijn, bepalen deze welke van de fragmenten aanleiding geven tot een analyse. Een voorbeeld van een methode die gebruik maakt van fouten bij het onderzoeken van een algoritme heet de methode van *algorithmic failure*, die ik in deze scriptie opvat als een uitwerking van de door So gesuggereerde methode van evalueren.

De methode van *algorithmic failure* werd geïntroduceerd door Jill Walker Rettberg (2022), die expliciet voortbouwt op een artikel van Anders Kristian Munk, Asger Gehrt Olesen en Mathieu Jacomy (2022). Laatstgenoemd artikel stelt dat foutieve voorspellingen in *machine learning* de interessantste zijn voor een kwalitatieve onderzoeker. Rettberg noemt deze methode *algorithmic failure* (p. 1). Hiermee onderstreept ze dat het de moeite waard is om een richtlijn voor *machine learning* te optimaliseren zodat de verkeerde voorspellingen betekenisvol zijn en niet te wijten zijn aan onnauwkeurige instellingen van het model. Tegelijkertijd laat Rettberg zien dat deze methode ook werkt voor relatief simpele algoritmes, zoals logistische regressie.

In navolging van Munk et al. (2022) stelt Rettberg dat een kwalitatieve onderzoeker niet per se hoeft te weten *waarom* een model tot een foutieve voorspelling is gekomen, interessanter

is de vraag *welke fragmenten* het niet heeft kunnen voorspellen (p. 2). Ze geeft een voorbeeld van een classificatietaak waarbij werkwoorden moeten worden voorspeld die een interactie beschrijven tussen fictionele personages en 'machine vision technologies'. Het model voorspelt of een beschrijvend werkwoord actief of passief is op basis van informatie over het personage dat is verbonden aan het werkwoord. Uit de bevindingen van Rettberg blijkt dat het algoritme vooral verkeerd voorspelt in ambigue gevallen (p. 3). Dit is dan ook een aanleiding voor een nadere kwalitatieve analyse van deze specifieke interacties. Uit deze 'fouten' blijkt dat er een complexe dynamiek bestaat tussen de vorm van werkwoorden en de mate van macht van verschillende betrokken actoren (p. 5). Op deze manier worden de fouten van het algoritme onderdeel van een 'productieve methodologie': 'Embracing algorithmic failure as a productive methodology upends the typical assumption that data science should lead to increased objectivity and accuracy.' (p. 4)

Deze benadering van evalueren bevalt me omdat het rekenschap geeft van de inherente ideologische lading van technologie (in dit geval van *machine learning*). Een algoritme is immers niet neutraal of objectief en is in staat om cruciale onderliggende keuzes schijnbaar onzichtbaar te maken. Door de methode van *algorithmic failure* te omarmen probeer ik in navolging van Jill Walker Rettberg (2022) hiervan rekenschap te geven in het ontwerp van mijn methode. Foutieve voorspellingen verworden op deze manier tot 'machine learning as a collaborator' (p. 4) om de relatie tussen computationele en menselijke beoordelaars beter te begrijpen.

Concreet zal ik *algorithmic failure* bestuderen door me voor alle vier de aspecten te richten op patronen in de fragmenten die een verkeerd label zijn toegewezen door het algoritme. Daarbij probeer ik vast te stellen of er gelijkenissen zijn tussen de fouten, zoals Rettberg een patroon vindt van werkwoorden die onterecht zijn gelabeld als passief. Dergelijke bevindingen zal ik daarna toelichten aan de hand van voorbeelden om ze vervolgens te interpreteren. Op die manier denk ik een model productief in te zetten om meer inzicht te verkrijgen in de bestudeerde fragmenten en het functioneren van het model zelf.

In het volgende hoofdstuk van deze scriptie, het resultatenhoofdstuk, zal ik de verschillende modellen evalueren aan de hand van de methodes die ik in deze paragraaf heb beschreven. Allereerst zal ik dus voor alle vier de aspecten van focalisatie de resultaten van de annotatietaak bespreken. Ten tweede zal ik de resultaten van achtereenvolgens het *random baseline* algoritme, de logistische regressie met L2-regularisatie en BERTje laten zien. Deze resultaten geef ik weer door middel van *precision*, *recall* en de *F*1-score die ik in de huidige paragraaf

introduceerde. Vervolgens kan ik de uitkomsten interpreteren en vergelijken met elkaar en de resultaten van de handmatige annotaties. Als derde zal ik de classificaties van het hoogst scorende algoritme nader bestuderen. Hiertoe besprak ik in de huidige paragraaf de methode van *algorithmic failure*, die meer inzicht kan geven in de classificaties van het algoritme en in hoe die er per aspect en per categorie uitzien.

4. Resultaten

Dit hoofdstuk bestaat uit het presenteren en interpreteren van de resultaten van achtereenvolgens de handmatige annotaties, de scores van de verschillende algoritmes die ik hanteer en de methode van *algorithmic failure*. In het methodehoofdstuk besprak ik de componenten die nodig zijn om tot deze resultaten te komen, zoals de stap van operationaliseren, het corpus en de achtergrond van de algoritmes. Hierop volgt de fase die Andrew Piper (2020) 'validation' (p. 21) noemt. Deze bestaat ten eerste uit het toekennen van scores die uitdrukken hoe goed de modellen in staat zijn voorspellingen te doen van labels van de vier aspecten. Hiermee wordt in feite de vraag gesteld hoe goed een model in staat is om aspecten van focalisatie te labelen. Ten tweede bestaat validatie volgens Piper uit het nader bekijken van de voorspellingen: 'what the model is predicting when it predicts a particular category.' (p. 21) Dit geeft volgens hem inzicht in hoe het model functioneert. Zoals ik beschreef in het methodehoofdstuk, pas ik dit laatste punt van validatie toe door juist de fouten van het model te bestuderen, omdat dit mijns inziens meer inzicht geeft in wat Piper noemt 'the nature of the problem' (p. 21).

4.1 Bespreking manuele annotaties

Allereerst bespreek ik de resultaten van de handmatige annotatietaak. Zoals gesteld, zijn de annotaties zowel methodisch een noodzakelijke stap voor automatische classificaties als op zichzelf resultaten genererend die het bestuderen waard zijn. In deze paragraaf richt ik me op dat laatste. Eerder besprak ik al hoe in de letterkunde teksten veelal worden bestudeerd door relevante passages uit een roman te verzamelen en die te voorzien van een interpretatie. We weten daardoor echter niet wat er aan mogelijk bruikbare informatie niet wordt meegenomen bij een analyse. De annotatietaak dient als systematische lezing van een tekst; op basis van het annotatieschema heb ik focalisatie geïnterpreteerd voor duizend fragmenten.[15] Het is interessant de annotaties nader te bestuderen; hoe zijn de categorieën binnen een aspect verdeeld en wat vertellen die frequenties over een aspect?

In het vervolg van deze paragraaf bespreek ik steeds per aspect eerst de frequenties van de annotatietaak, om vervolgens nader in te gaan op de categorieën van dat aspect. Ik probeer daarbij patronen te ontdekken binnen de annotaties van het betreffende aspect, die ik illustreer aan de hand van exemplarische fragmenten. Hoewel ik voor het trainen van de algoritmes

[15] De resultaten hiervan zijn geenszins representatief voor alle voorkomens van focalisatie, aangezien ik ervoor heb gekozen om fragmenten afkomstig uit vier romans uit 2012 te analyseren.

fragmenten heb verwijderd die zijn gelabeld als ambigu of anderszins niet binnen het annotatieschema pasten (label '99'), voeg ik die fragmenten voor de volledigheid van mijn interpretatie hier weer toe. Alle aspecten hebben dus in totaal duizend labels toebedeeld gekregen.

4.1.1 Focaliserend subject

Label	Frequentie
1	90
2	81
3	806
4	0
99	23

Tabel 7. Frequenties van het aspect van focaliserend subject voor 1000 fragmenten.

Uit bovenstaande tabel blijkt dat het overgrote deel van de fragmenten voor het aspect van focaliserend subject is beoordeeld als categorie 3 (interne focalisator). Bovendien zijn nog eens 81 fragmenten beoordeeld als consonant of ambigu, waarbij de interne focalisator ook betrokken is. De verdeling van de positie van het focaliserend subject valt lastig te generaliseren naar uitspraken over focalisatie in het algemeen, omdat de fragmenten afkomstig zijn uit maar vier romans. Deze bevinding sluit echter wel aan bij eerdere onderzoeken naar perspectief en focalisatie, zoals Ian Watt (1957) die schrijft over een 're-orientation of the narrative perspective' (p. 176) waarna fictie vanaf de achttiende eeuw meer en meer 'subjective and inward' (p. 177) gericht wordt. Ook Manfred Jahn (2007) stelt dat vanaf grofweg de achttiende eeuw een 'psychological turn' (p. 94) heeft plaatsgevonden in de literatuur. Watt (1957) plaatst dit binnen een bredere ontwikkeling van een 'urban way of life' waarbinnen de nadruk ligt op het subjectieve en private van het individu (p. 178). Onderdeel van deze ontwikkeling is de verschuiving van externe naar interne focalisatie als dominante wijze van focaliseren.[16]

De verdeling van focaliserende subjecten die zijn geannoteerd geeft te denken over de verhouding tussen interne en externe waarnemers. Vanuit de narratologische theorie worden ze in principe tegenover elkaar geplaatst, bijvoorbeeld wanneer een externe waarnemer dichter bij

[16] Toegegeven, Jahn (2007) en Watt (1957) situeren deze wending lang voor de 21ste eeuw, maar beiden plaatsen hem in een trend die in ieder geval tot diep in de twintigste eeuw heeft geduurd (Jahn 2007, p. 94) en we mogen aannemen dat de hedendaagse roman in ieder geval wordt beïnvloed door deze langdurige traditie.

de vertellende instantie staat en een personage de interne waarneming gedelegeerd kan krijgen van de verteller. Wanneer in (hedendaagse) romans interne waarnemingen veel dominanter zijn, lijkt de externe focalisator echter ook meer een rol te krijgen van 'één van de' waarnemende entiteiten, in plaats van hét centrum van waarneming. In feite wordt dit 'ontheiligen' van de externe focalisator benadrukt doordat ik in mijn annotatieschema geen onderscheid maak tussen verschillende soorten interne focalisator; voor sommige romans die nagenoeg volledig intern worden gefocaliseerd, is het waarschijnlijk zinvoller om verschillende focaliserende personages binnen de verhaalwereld te kunnen onderscheiden.

Desondanks gaat een interpretatie op basis van bovenstaande frequenties mogelijk voorbij aan het feit dat een externe focalisator soms over meer informatie beschikt dan interne focalisators, aangezien deze niet is gebonden aan een personage binnen de verhaalwereld. Daarom heeft de externe focalisator in potentie toegang tot een hogere laag van de vertelling dan een interne focalisator. Dit roept de vraag op wat de precieze functie is van de externe focalisatie in de hedendaagse roman, waarin hij kwantitatief gezien minimaal aanwezig is. Heeft een externe waarnemer bijvoorbeeld vaker een 'sleutel' tot de plot of een interpretatie van de vertelling in handen, zoals we ook zagen bij de analyse van *The Echo Maker* door Luc Herman & Bart Vervaeck (2009b)? En zijn deze waarnemingen vaker 'belangrijker' voor de relatie tussen de lezer en de verhaalwereld? Deze vragen voeren te ver voor deze scriptie, maar laten zien dat het systematisch bestuderen van het focaliserende subject een nieuw licht werpt op de betekenis van hoe centra waarnemingen binnen een verhaal zijn verdeeld.

4.1.2 Waarneembaarheid

Label	Frequentie
0	172
1	675
99	153

Tabel 8. Frequenties van het aspect van waarneembaarheid voor 1000 fragmenten.

Voor het aspect van waarneembaarheid geldt wederom dat het grootste deel van de fragmenten binnen één categorie valt, namelijk de categorie 'waarneembaar' (zie tabel 8). Ik denk dat dit aspect moet worden beschouwd in relatie tot de alomtegenwoordigheid van de interne focalisator. Naarmate een verhaalwereld meer wordt waargenomen door een personage, ontstaat er wellicht meer aandacht voor de binnenwereld van het individu, waardoor de

categorie ‘niet-waarneembaar’ vaker verwacht zou worden. Volgens Ian Watt (1957, p. 176) zijn interne waarneming van dagelijkse en huiselijke zaken zelfs onlosmakelijk verbonden met aandacht voor de gedachten van personages. We weten dan echter nog niet wat de verhouding is tussen waarneembare en niet-waarneembare objecten. In mijn systematische lezing heb ik beide geteld, waarna blijkt dat waarneembare objecten ruim in de meerderheid zijn. Ik denk dat het interessant is dat een relatief groot deel van de waargenomen objecten niet over de binnenwereld van personages gaat, ondanks de wending naar binnen die we bij het vorige aspect zagen. De restcategorie (‘99’) geldt voor fragmenten die niet eenduidig binnen een van de andere categorieën passen. Voor deze fragmenten geldt dus dat ze ook aandacht aan de binnenwereld besteden. Overigens zijn de resultaten wellicht enigszins vertekend doordat alle dialogen gemarkeerd zijn als waarneembaar, waardoor het aantal fragmenten met binnen deze categorie hoger uitvalt.

In de narratologische theorie wordt beschreven dat het aspect van waarneembaarheid ook gaat over de verdeling van informatie. Heeft een lezer weet van de gedachten van een personage, dan is hij wellicht sneller geneigd mee te voelen met dit personage. Op basis van de frequenties komen we weinig te weten over de informatieverdeling binnen een tekst. Omdat dit wel degelijk interessant is, zal hierop terug te komen bij de *algorithmic failure*-analyse naar aanleiding van de classificaties van het algoritme.

4.1.3 Standvastigheid

Label	Frequentie
1	508
2	492

Tabel 9. Frequenties van het aspect van standvastigheid voor 1000 fragmenten.

De evenwichtige verdeling tussen beide categorieën van standvastigheid, zoals getoond in tabel 9, is vooral interessant in vergelijking met de resultaten van het aspect van de mate van detail. Ik veronderstelde immers dat, op basis van de narratologische theorie over standvastigheid, de kans groter is dat er minder gedetailleerd wordt waargenomen naar mate er meer objecten worden waargenomen. De samenhang tussen standvastigheid en gedetailleerdheid wordt onderstreept door het feit dat tijdens het annoteren bleek hoezeer de keuze bij standvastigheid de mate van detail beïnvloedt; de mate van detail is vaak lager zodra een scène als geheel wordt beoordeeld en er dus meerdere onderdelen van die scène in één fragment worden beschreven,

en wordt waarschijnlijk hoger als één object in een afzonderlijk fragment een beoordeling krijgt. Om die reden zal ik de resultaten van beide aspecten combineren bij de bespreking van gedetailleerdheid.

4.1.4 Gedetailleerdheid

Categorie	Frequentie
1	31
2	87
3	119
4	271
5	461
99	31

Tabel 10. Frequenties van het aspect van gedetailleerdheid voor 1000 fragmenten.

Tabel 10 maakt duidelijk dat de mate van detail ongelijk is verdeeld over de zes mogelijke categorieën. Het zwaartepunt van de annotaties ligt bij de twee hoogste categorieën, die bij elkaar opgeteld bijna driekwart van alle fragmenten omvatten. Deze verdeling suggereert dat een object al gauw gedetailleerd wordt waargenomen, of dat detaillering gauw als hoog opgevat wordt door lezers. Het blijkt, met andere woorden, lastig om een object met weinig detail waar te nemen. David Herman (2009) ziet detaillering van een object als effect van de context waarin ze plaatsvindt (p. 130-131). Onder meer de positie van de waarnemer en de afstand tussen subject en object dragen bij aan de kwaliteit van de waarneming. Uit de resultaten van de systematische lezing blijkt dat de mate van detail van een object niet simpelweg een opstelsom is van de context. In de praktijk lijkt alleen al het feit *dat* een object wordt waargenomen het object van enige detaillering te voorzien. Bovendien is de mate van detail wellicht alleen vergelijkbaar *binnen* een tekst (bijvoorbeeld hoe een personage gedurende het verhaal wordt waargenomen) en is het minder geschikt om te vergelijken *tussen* teksten. Om te onderzoeken hoe een lage mate van detail voorkomt, bespreek ik het volgende fragment dat is gelabeld met een zeer lage mate van detail (label 1). In het fragment, afkomstig uit *Euforie* (2012) van Christiaan Weijts, zijn de losse onderdelen waaruit de scène (dus geannoteerd als één object) bestaat niet gespecificeerd, waardoor ze inwisselbaar worden:

> Aan alle straatkanten ziet hij hetzelfde tafereel. Geparkeerde auto's, bedolven onder een dikke vacht sneeuw, worden schoongeveegd terwijl de motor zachtjes pruttelt, als een pan op laag vuur. Met een schop scheppen de bestuurders sneeuw weg rond de wielen – vastgekoekte ijsbrokken bikken ze los met de steel. Soms duwen passagiers of passanten een auto bij de laadklep, terwijl het gas giert en de banden slippen. Schiet een voertuig uiteindelijk in het spoor van ingereden smurrie, dan heeft dat altijd iets verrassends wat reden is voor een bescheiden feestje van lichtseinen of claxonneren. (p. 33-34)

In de eerste zin wordt gesproken over 'hetzelfde tafereel', waarna een scène wordt geschetst van een winterochtend die overal in de stad zou kunnen plaatsvinden. Ook de waargenomen figuren zijn inwisselbaar; we zien slechts hun handelingen ('worden schoongeveegd', 'lichtseinen of claxonneren') en algemene aanduidingen als 'de bestuurders' en 'passagiers of passanten'. Dit fragment vormt een onderbreking binnen een overpeinzing van de hoofdpersoon over het mondiale financiële systeem. De waarneming in het fragment heeft als effect dat het een contrast schetst tussen de materiële wereld en de abstracte financiële wereld waarvan hij zich afvraagt of 'iemand het werkelijk nog [begrijpt]' (p. 34). De werkende middenklasse wordt in deze context afgebeeld als anonieme speelbal die zich laat leiden door machten van bovenaf. Hiermee is het besproken fragment exemplarisch voor een patroon dat vaker terugkeert in fragmenten die zijn gelabeld als zeer lage mate van detail. Vaak worden namelijk onbekende figuren of plaatsen waargenomen, die de grenzen van de wereld van de waarnemer markeren. Dit sluit in principe aan bij David Hermans (2009) notie van gedetailleerdheid als product van onder meer de afstand tussen waarnemer en waargenomen object.

Een variant op deze generieke vorm van waarnemen is de opsomming, zoals in dit deel van een fragment: 'Er bestond een bloeiend circuit van bedrijfsdagen, symposia, publieksdagen, vaak gesponsord door woningcorporaties, bouwbedrijven of cultuurfondsen.' (Weijts 2012, p. 375) Dergelijke opsommingen beslaan soms een geheel fragment en lopen uiteen van het opsommen van de wereldproblemen tot simpelweg benoemen van de omgeving ('Grasmaaier. Jerrycan, ik til hem op, halfvolle jerrycan. Tuinstoelen. Bal, lek. Kruiwagen. Stenen.', Van Marissing 2012, p. 126). De gemene deler is hier steeds dat objecten op afstand worden gezet door de waarneming te richten op meerdere objecten. Een *Wilcoxon rank-sums test*[17] wijst uit dat fragmenten met één object een significant hogere gemiddelde score voor detaillering hebben dan fragmenten met meerdere objecten ($W = 151182$, $p < 0.01$, eenzijdig). De statistische toets

[17] De *Wilcoxon rank-sums test* is een non-parametrische toets die gewoonlijk wordt gebruikt bij een vergelijking tussen de gemiddelden van twee groepen waarvan de waarden niet normaal zijn verdeeld (Gries 2013, p. 215). In dit geval gaat het dus om mate van detail van twee groepen: één object en meerdere objecten. De mate van detail is voor beide groepen niet normaal verdeeld, waardoor de *Wilcoxon rank-sums test* de meest geschikte toets is.

bevestigt dus het beeld dat ook al naar voren kwam na bestudering van fragmenten met lage detaillering: als er meerdere objecten worden gefocaliseerd, is het waarschijnlijker dat ze worden waargenomen met minder detail.

De gevonden patronen binnen de categorie 'zeer lage mate van detail' bieden nochtans maar een deel van de verklaring voor het feit dat deze categorie weinig gerepresenteerd is in de dataset. Bovenstaande patronen zijn namelijk te specifiek om een volledige categorie te kunnen omvatten. Daarom veronderstel ik dat het aspect van gedetailleerdheid – meer nog dan de andere aspecten – onderhevig is aan interpretatie. Uit de inter-beoordelaarsbetrouwbaarheidstoets bleek weliswaar voldoende tot goede overeenstemming, maar er is waarschijnlijk een uitgebreidere annotatieprocedure nodig om alle categorieën van dit aspect betekenisvol te laten zijn. Nader onderzoek met meer beoordelaars die een groter aantal fragmenten beoordelen op de mate van detail zou meer duidelijkheid kunnen scheppen over de werking van dit aspect. In dat geval is het mogelijk om labels te vergelijken en door middel van overleg ook voor laagfrequente categorieën tot een gedeelde interpretatie te komen. Ik kom hierop terug in het slothoofdstuk van deze scriptie. Desondanks blijkt de systematische lezing die ik in deze scriptie hanteer een nuttig middel om verschillende fragmenten met dezelfde beoordeling te kunnen vergelijken. Deze methode stelt me namelijk in staat om een groep fragmenten apart te bestuderen om meer te weten te komen over de onderlinge samenhang binnen een categorie, zoals bleek uit mijn analyse van fragmenten met een zeer lage mate van detail.

4.2 Scores

Voor het bespreken van de resultaten voor het classificeren van de vier aspecten van focalisatie toon ik steeds per aspect een tabel met daarin de scores van de handmatige annotatietaak en van de verschillende modellen. Naar aanleiding van iedere tabel licht ik de relevante bevindingen voor het betreffende aspect uit. In deze paragraaf ligt de nadruk op het beschrijven van de resultaten, in de paragraaf over *algorithmic failure* richt ik me meer op een analyse van de resultaten.

4.2.1 Focaliserend subject

Tabel 11 toont de scores voor het classificeren van externe focalisatie, consonante of ambigue focalisatie en interne focalisatie. Bij de bespreking van de frequenties bleek dat ongeveer 80%

van dit aspect bestaat uit intern gefocaliseerde fragmenten. Daarom geef ik per categorie de scores weer voor *precision*, *recall* en *F*1.

		Externe focalisator			Consonant/ambigu		
Model	*Features*	*Precision*	*Recall*	*F*1	*Precision*	*Recall*	*F*1
Annotaties		0.74	0.74	0.74	0.58	0.83	0.58
Random baseline		0.14	0.17	0.15	0.09	0.06	0.07
LR	BoW	0.42	0.28	0.33	0.50	0.31	0.38
	Tf-idf	0.39	0.39	0.39	0.46	0.38	0.41
BERTje		0.67	0.22	0.33	0.38	0.19	0.25

		Interne focalisator		
Model	*Features*	*Precision*	*Recall*	*F*1
Annotaties		0.98	0.94	0.96
Random baseline		0.83	0.83	0.83
LR	BoW	0.87	0.94	0.90
	Tf-idf	0.88	0.90	0.89
BERTje		0.86	0.97	0.91

Tabel 11. Resultaten voor het classificeren van de positie van de focalisator per model en features. De bovenste rij toont de resultaten van de handmatige annotatietaak. De gekleurde rij markeert het model met de hoogste gewogen F1-score (annotaties buiten beschouwing gelaten).

De tabel laat grote verschillen zien tussen de drie categorieën van dit aspect.[18] Zo geldt voor zowel de handmatige annotaties, de *random baseline* als de modellen met trainingsinput dat de interne focalisator veruit het accuraatst wordt geclassificeerd. Zo loopt de *F*1-score van de externe focalisator van 0.33 tot 0.39 en die van consonant/ambigu van 0.25 tot 0.41, terwijl de interne focalisator scores heeft van 0.89 tot 0.91. Deze verschillen tussen de categorieën zijn voor de modellen waarschijnlijk deels te verklaren door de hoeveelheid trainingsdata die zeer ongelijk is verdeeld. Hierdoor vallen de scores van dit aspect over het geheel genomen wat tegen. Kijken we daarentegen naar de interne focalisator, dan zitten alle modellen enkele procentpunten boven de *baseline* en zijn ze zelfs niet eens zo ver verwijderd van de scores van de menselijke beoordelaars. Er zijn kleine onderlinge verschillen tussen de modellen, maar het meest opvallende resultaat is dat het BERT-model niet veel beter – en over het geheel genomen zelfs minder goed – presteert dan het veel simpelere *bag-of-words*-model. Het is daarbij opvallend dat het BERT-model vooral moeite lijkt te hebben met de meest ambigue categorie van dit aspect.

[18] Hoewel ik geen inter-beoordelaarsbetrouwbaarheidstoets heb uitgevoerd voor dit aspect wanneer de categorie 'consonant/ambigu' wordt weggelaten, levert trainen en evalueren met weglating van deze categorie slechts een minimale verbetering in *F*1-scores op: 0.36 voor extern en 0.91 voor intern. De score voor intern is zelfs even hoog als de *baseline*-score, die een stuk hoger is wanneer dit aspect uit twee categorieën bestaat.

4.2.2 Waarneembaarheid

Over het aspect van waarneembaarheid bestond de hoogste mate van overeenstemming tussen de menselijke beoordelaars, wat suggereert dat dit aspect wellicht eenduidiger valt te benoemen dan de andere aspecten. De scores van de modellen, zoals te zien in tabel 12, suggereren dat eenduidigheid een positief effect heeft op de classificatietaak. In de vorige paragraaf zagen we dat ook dit aspect ongelijk verdeeld is. De categorie van niet-waarneembare fragmenten kent vrij grote verschillen met *F*1-scores die uiteenlopen van 0.57 tot 0.70. Alle scores zijn echter veel hoger dan de *random baseline* die 0.20 scoort. De scores van het BERT-model, dat het hoogst scoort, zitten ruim onder die van de handmatige annotaties. Dit zou eventueel te maken kunnen hebben met het geringere aantal fragmenten waarop voor de categorie van niet-waarneembaar is getraind. De scores van waarneembaarheid zijn namelijk (iets) hoger dan de *baseline*, voor alle modellen hoger dan de categorie 'niet-waarneembaar' en de modellen lopen minder uiteen (0.90 tot 0.92). Het Nederlandstalige BERT-model scoort ook hier hoger dan de logistische regressie en het zit zelfs maar enkele procentpunten onder die van de handmatige annotaties (0.95 tegenover 0.92).

		Niet-waarneembaar			Waarneembaar		
Model	*Features*	*Precision*	*Recall*	*F*1	*Precision*	*Recall*	*F*1
Annotaties		0.96	0.93	0.95	0.94	0.96	0.95
Random baseline		0.21	0.20	0.20	0.79	0.80	0.80
LR	BoW	0.64	0.60	0.62	0.90	0.91	0.90
	Tf-idf	0.64	0.51	0.57	0.88	0.93	0.90
BERTje		0.71	0.69	0.70	0.92	0.93	0.92

Tabel 12. Resultaten voor het classificeren van de waarneembaarheid van het gefocaliseerde object per model en features. De bovenste rij toont de resultaten van de handmatige annotatietaak. De gekleurde rij markeert het model met de hoogste gewogen F1-score (annotaties buiten beschouwing gelaten).

Een nadere bestudering van de woorden die van grote invloed zijn op de classificatie door de logistische regressie, uitgedrukt in coëfficiëntwaarden[19], levert vanwege het beperkte aantal voorbeelden weinig nieuwe inzichten op. Weliswaar blijken meer abstracte woorden als 'wind', 'boel' en 'schemergebied' kenmerkend voor de categorie 'niet-waarneembaar', maar het gaat steeds om laagfrequente woorden en voor de hand liggende woorden als 'herinneren' of 'denken' komen niet. De meeste woorden lijken eerder toevallig een enkele keer voor te komen in niet-waarneembare fragmenten en niet in waarneembare fragmenten, wat ze een hoge

[19] Bij logistische regressie wordt aan ieder woord een gewicht toegekend dat uitdrukt hoezeer een woord van invloed is op een classificatie. Woorden met een hoge coëfficiëntwaarde zijn dus karakteristiek voor een label. Door de woorden met hoge coëfficiëntwaarden te bestuderen kunnen we mogelijk dus een beeld krijgen van de woorden die het logistische regressie-model sturen bij een classificatie.

coëfficiënt geeft. De categorie ‘waarneembaar’ toont hetzelfde patroon: woorden als ‘hoofdhuid’, ‘spijkers’ en ‘appelwangen’ zijn waarschijnlijk vaak waarneembaar, maar komen niet meer dan drie keer voor in het gehele corpus. De coëfficiëntwaarden maken dus vooral duidelijk dat de modellen waarschijnlijk beter zouden classificeren als ze zouden beschikken over meer voorbeelden.

4.2.3 Standvastigheid

In tabel 13 zien we de resultaten van de modellen voor het aspect van standvastigheid. Voor de fragmenten met één object lopen de scores van de modellen nauwelijks uiteen, terwijl ze ongeveer tien procentpunten boven de *baseline*-score (0.46) zitten, namelijk van 0.55 tot 0.57. Hoewel de trainingsdata voor dit aspect gelijkmatig is verdeeld over beide categorieën, scoren alle modellen hoger voor de tweede categorie ten opzichte van de eerste categorie, met scores van 0.58 tot 0.66. Het BERT-model scoort zelfs elf procentpunten hoger in vergelijking met de andere categorie. Dit suggereert dat de categorie ‘Twee of meer objecten’ meer onderlinge samenhang heeft met meer *features* die een fragment met twee of meer objecten onderscheiden van fragmenten met één object.

		Eén object			Twee of meer objecten		
Model	*Features*	*Precision*	*Recall*	*F*1	*Precision*	*Recall*	*F*1
Annotaties		0.91	0.90	0.91	0.90	0.91	0.89
Random baseline		0.47	0.44	0.46	0.46	0.49	0.47
LR	BoW	0.59	0.54	0.56	0.56	0.60	0.58
	Tf-idf	0.61	0.54	0.57	0.57	0.64	0.61
BERTje		0.67	0.46	0.55	0.58	0.77	0.66

Tabel 13. Resultaten voor het classificeren van het aantal gefocaliseerde objecten per model en features. De bovenste rij toont de resultaten van de handmatige annotatietaak. De gekleurde rij markeert het model met de hoogste gewogen F1-score (annotaties buiten beschouwing gelaten).

4.2.4 Gedetailleerdheid

De scores van het vierde aspect staan tot slot in tabel 15. In de vorige paragraaf bleek al dat de lagere categorieën sterk zijn ondervertegenwoordigd in het trainingscorpus. De scores van de modellen laten dan ook zien dat de hoogste twee categorieën veel beter worden geclassificeerd dan de rest. Opvallend genoeg wordt de laagste categorie door geen enkel model gelabeld, zelfs niet door de *random baseline*. Ik veronderstel dat dit te maken heeft met enerzijds een (groot) gebrek aan trainingsdata en anderzijds de hoge mate van interpretatie waarmee dit aspect te maken heeft. Uit de annotatietaak bleek immers al dat voor dit aspect beduidend minder

overeenstemming bestond tussen de beoordelaars. Het lijkt erop dat het gebrek aan eenduidigheid ook moeilijkheden oplevert voor de modellen.

Wanneer we kijken naar de onderlinge verschillen tussen de twee LR-modellen en het BERT-model, dan zien we dat BERTje voor categorie 2 en 3 zelfs iets minder hoog scoort. Daarentegen weet het categorie 4 en met name categorie 5 (meer dan tien procentpunten verschil) beter te labelen. Een voor de hand liggende verklaring voor dit verschil is wederom het gebrek aan voorbeelden, ook voor de oververtegenwoordigde labels. Een dataset van duizend fragmenten die worden verdeeld over vijf categorieën blijkt namelijk te gering. Als gevolg hiervan neemt de kans toe dat woorden die mogelijk een belangrijke bijdrage leveren aan de classificatie in de testset niet letterlijk in de trainingsset voorkomen. Door het gebrek aan kennis van woorden buiten woordfrequenties om, kan een *bag-of-words*-model dit niet corrigeren en zal een label niet herkend worden. Het BERT-model is hiertoe dankzij de contextuele *embeddings* tot op zekere hoogte wel in staat, waardoor het met relatief weinig voorbeelden alsnog sommige relevante synoniemen herkent.

Wellicht opvallender is het gegeven dat het BERT-model voor de categorie met zeer hoge mate van detail zelfs de menselijke annotaties overtreft. Het is echter lastig om dit resultaat te wegen, aangezien het model voor de rest van de categorieën ruim onder de menselijke richtlijn zit. Gezien de ongelijke verdeling van data is het voor dit aspect immers zelfs mogelijk om over het geheel genomen een goede score te behalen door slechts de vijfde categorie te kunnen herkennen.[20]

		1	2	3	4	5
Model	*Features*	*F*1	*F*1	*F*1	*F*1	*F*1
Annotaties		0.00	0.33	0.22	0.53	0.51
Random baseline		0.00	0.00	0.00	0.26	0.47
LR	BoW	0.00	0.13	0.24	0.36	0.56
	Tf-idf	0.00	0.13	0.10	0.38	0.52
BERTje		0.00	0.11	0.07	0.39	0.68

Tabel 14. Resultaten voor het classificeren van de gedetailleerdheid van het gefocaliseerde object per model en feature. De tabel representeert de volgende vijf categorieën van dit aspect: 1) zeer lage mate van detail; 2) lage mate van detail; 3) niet laag, niet hoog; 4) hoge mate van detail; 5) zeer hoge mate van detail. De bovenste rij toont de resultaten van de handmatige annotatietaak. De gekleurde rij markeert het model met de hoogste gewogen F1-score (annotaties buiten beschouwing gelaten). In Bijlage 2 staan de volledige resultaten van dit aspect.

[20] Uit de inter-beoordelaarsbetrouwbaarheidstoetsen (hoofdstuk 3.5.2) voor dit aspect bleek dat er te weinig overeenstemming bestond voor dit aspect wanneer het is verdeeld over drie categorieën (laag-midden-hoog). Om die reden acht ik het niet geoorloofd om alsnog een model te trainen met de labels gehercodeerd zodat het aspect uit twee of drie categorieën bestaat.

4.3 Algorithmic failure

In de vorige paragraaf bekeek ik per aspect welk classificatiemodel de hoogste score behaalt. Voor de positie van de focalisator blijkt logistische regressie met *bag-of-words features* hier het meest voor geschikt. Voor de andere drie aspecten is het Nederlandstalige BERT-model het effectiefst. In deze paragraaf analyseer ik deze modellen door middel van de methode van *algorithmic failure*, zoals ik beschreef in het methodehoofdstuk. Met deze methode diep ik mijn bevindingen uit, de scores uit de vorige paragraaf geven immers weinig informatie over de aspecten zelf. *Algorithmic failure* dient als een productieve methodologie die aan de hand van de fouten van de modellen tot nieuwe inzichten leidt over de werking van focalisatie.[21] Daartoe zal ik steeds per aspect de fouten van het beste model bestuderen om daarin mogelijk patronen te ontdekken die ik vervolgens uitlicht en interpreteer.

Het is dus niet alleen interessant om de verkeerde voorspellingen te bespreken, zoals gebruikelijk is wanneer een model wordt geëvalueerd. De methode van *algorithmic failure* verbindt deze bevindingen met een kwalitatieve interpretatie over wat deze fouten ons leren over het onderzochte fenomeen om bijvoorbeeld een aanleiding te zijn voor nader kwalitatief onderzoek. Volgens Jill Walker Rettberg (2022) leent ook een model dat geen hoge scores haalt zich voor deze methode: '"bad" machine learning can be quite adequate if the goal is to identify rich cases for analysis' (p. 2). Ook stelt ze in navolging van Munk, Olesen en Jacomy (2022) dat we niet hoeven te begrijpen *waarom* een fragment verkeerd is voorspeld. Het is namelijk interessanter om de fragmenten zelf te onderzoeken: '*which cases* it failed to predict' (p. 2). Overigens vat ik dit niet op als poging tot een uitputtende interpretatie van een model, veeleer nodigt *algorithmic failure* uit tot de bestudering van een model met oog voor typische geesteswetenschappelijke benaderingen als interpretatie, ruimte voor onzekerheid en aandacht voor details. Het gaat er volgens Jill Walker Rettberg (2022) dan ook om door middel van deze methode aanleidingen te vinden voor nader kwalitatief onderzoek.

4.3.1 Focaliserend subject

Het aspect van focaliserend subject bestaat voor het merendeel uit fragmenten in de categorie van interne focalisator. De voorspelde labels komen daarom veelal overeen met deze categorie. Er is echter ook een aantal fragmenten gelabeld als interne focalisator, terwijl het manueel

[21] 'Fouten' heeft in deze context de betekenis van 'niet overeenkomend met de handmatige annotatie'. In het kader van de hoge mate van interpretatie die soms gepaard gaat met het beschrijven van focalisatie, valt niet direct uit te sluiten dat een foutieve classificatie door het model dichter bij de 'werkelijkheid' ligt dan de menselijke beoordeling, die daarop kan worden herzien.

toegekende label dat van een van de andere categorieën is. Overigens wijkt het aspect af van het voorbeeld dat Jill Walker Rettberg (2022) geeft, aangezien ze een binair classificatiemodel bestudeert. In mijn analyse van het aspect van focaliserend subject kijk ik naar fragmenten die handmatig zijn gelabeld als extern of consonant/ambigu en die door het model zijn voorspeld als intern. Aangezien interne focalisatie veruit het vaakst voorkomt, stuiten we bij fragmenten met 'foutieve interne focalisatie' op de grens van deze categorie – het zijn immers de fragmenten die het model onder de dominante groep schaart, terwijl ze eigenlijk tot een andere categorie behoren.

In totaal zijn er twaalf extern gefocaliseerde fragmenten gelabeld als intern, deze vallen dus op de grens met interne focalisatie. De 'foutieve interne focalisatie' omvat veelal fragmenten die op het eerste gezicht niet binnen één categorie vallen; het gaat om fragmenten waarbinnen de verdeling van macht om te waarnemen tussen de verteller die de focalisatie delegeert en het ontvangende personage niet direct duidelijk is. Het volgende fragment heb ik beoordeeld als extern en is door het model geclassificeerd als intern:

> "Een gezonde geest in een gezond lichaam!" Dat was ook wat Max Kanselaar, de gymleraar, had geroepen, meteen aan het begin van het eerste schooljaar. Schouder aan schouder stonden de brugklasjongens te bibberen op het sportveld, allemaal in de voorgeschreven 'korenblauwe korte broek' onder een wit t-shirt. Kanselaar, robuust en kalend, droeg een glimmend trainingspak en inspecteerde de rij. Grofweg waren er twee soorten jongens: kinderlijke, beweeglijke ventjes en lange, lijzige types met de baard in de keel. Johannes Vermeer behoorde, met een handjevol klasgenoten, tot die laatste. Hij tuurde over het hek naar het aangrenzende hockeyveld waar de meisjes bijeen waren, in vergelijkbare broekjes en shirts, al permitteerden zij zich meer afwijkingen en stonden ze ook niet zo strak in het gelid. Een brugklas is altijd een raar gezicht. De fysieke veranderingen voltrekken zich snel maar allerminst simultaan, waardoor je amper gelooft dat zo'n ratjetoe van ukjes en slungels allemaal dezelfde leeftijd heeft.
> (Weijts 2012, p. 18)

Interessanter dan de vraag welke van de twee categorieën daadwerkelijk als het 'werkelijke' label geldt (het fragment is erg ambigu), is de vraag welke elementen aanleiding geven tot twijfel. Het fragment opent met een uitroep, die wordt verbonden met een herinnering aan de gymleraar van Johannes Vermeers burgklas. De herinnerende instantie kunnen we ongeveer gelijkstellen aan de verteller en externe focalisator. Deze herinnering begint vol beschrijvingen van een gymles waardoor we als lezer als het ware in de rij met bibberende brugklasjongens staan, wat we kunnen opvatten als een verschuiving naar de belevende, interne focalisator. Met het citeren van de kleur van het gymbroekje uit vermoedelijk een informatiebrief, maakt echter duidelijk dat het geen waarneming is waarvan we helemaal zeker zijn dat Vermeer die *toen*

heeft gedaan. We hebben waarschijnlijk nog steeds te maken hebben met waarnemingen van de externe focalisator. Dit geldt ook voor de meer beschouwende zin over ‘twee soorten jongens’, waarin een typische blik van buitenaf doorklinkt. Daarop volgt expliciet het ‘tuurde’ naar het hockeyveld waardoor we ineens meekijken met Johannes Vermeer als brugklasser. De interne focalisator maakt echter plaats voor analyserende uitspraken waarin de herinnerende focalisator het overneemt, namelijk over de brugklas die ‘altijd een raar gezicht [is]’ en ‘zo’n ratjetoe van ukjes en slungels’. De macht om waar te nemen lijkt hier samen te hangen met de macht om te herinneren; is een situatie in het verleden leidend voor een herinnering of zijn dat de projecties vanuit het heden? Het feit dat de waarnemingen wisselen in mate van detail (vergelijk het citeren van de kleur sportbroekjes met het anoniem blijven van de klasgenoten) geeft in ieder geval aan dat de herinnering vanuit het heden niet compleet is en wellicht niet volledig betrouwbaar.

Het fragment dat hier werd uitgelicht heb ik beoordeeld als extern gefocaliseerd, omdat de waarneming grotendeels bij het herinnerende personage ligt dat optreedt als externe focalisator. De vele beschrijvingen en het expliciete waarnemen vanuit de brugklasser zouden echter ook redenen kunnen zijn om het tegendeel te beweren. Deze ‘foutieve interne focalisatie’ legt de strijd om waarneming in *Euforie* bloot voor de verhaallijn die zich afspeelt tijdens de middelbareschooltijd van Johannes Vermeer. Deze periode vormt de grondtoon van de verhaallijn in het heden, waarin de waarnemingen van Vermeer ook gebrekkig blijken te zijn. Dit besproken fragment zou daardoor vanuit de methode van *algorithmic failure* als aanleiding kunnen dienen voor een analyse van de betrouwbaarheid van waarnemingen in *Euforie* en in hoeverre de externe focalisator objectiviteit veinst zonder betrouwbaar te zijn.

4.3.2 Waarneembaarheid

Het tweede aspect is voor het grootste deel van de fragmenten handmatig gecategoriseerd als waarneembaar en het model toont dit ook bij zijn classificaties. Er zijn echter veertien fragmenten die het onterecht heeft gelabeld als waarneembare fragmenten. Een analyse van deze fragmenten laat zien dat ze op een paar punten overeenkomen. Opvallend genoeg komt in een aantal fragmenten een dialoog voor. Deze vindt plaats in de gedachten van het focaliserende personage en kan bijvoorbeeld de vorm hebben van een gesprek in het hoofd van een personage: ‘Maar nu ontbreekt je die stormachtige opwinding waar eerdere projecten op dreven en die vanzelf die oprechte arrogantie liet oplaaien.’ (Weijts 2012, p. 110) Fragmenten die volledig uit dialoog bestaan, zijn geannoteerd als waarneembaar, waardoor een intern gesprek wordt aangezien voor een waarneembaar fragment. Dit geldt ook voor fragmenten waarin een

denkbeeldige dialoog wordt gevoerd, soms inclusief aanhalingstekens om spreekbeurten te markeren, zoals in het volgende voorbeeld: '"Laura, je bent onbenaderbaar." Hé, die stem, die ken ik, dat licht geaffecteerde, dat is de stem van mijn moeder. "Volgens mij gaat het niet goed met je." Oe, nog één, we doen een quiz, dat was een vriendin.' (Van Marissing 2012, p. 76-77) Deze passage is onderdeel van de gedachten van het focaliserende personage waarin ze eerdere gesprekken met bekenden oproept en waarop ze tussendoor commentaar levert. Hierdoor ontstaat het effect van een dialoog met tussendoor de gedachten van het personage dat reageert op het 'gesprek'. Deze bevinding laat zien dat het model dialogen herkent als waarneembare fragmenten, wat op zichzelf al interessant is. Vervolgens levert dit dus ook een aantal fouten op doordat niet alle fragmenten die op een dialoog lijken ook als zodanig moeten worden beschouwd.

Daarnaast zijn er meerdere fragmenten die openen met een waarneembaar gefocaliseerd object met daarin een sterke *marker* voor waarneembaarheid, zoals 'ziet' of 'kijkt'. De volgende zin is afkomstig uit een fragment dat volledig niet-waarneembaar is op het begin van deze zin na: 'Dan leunt ook Coco naar achter, kijkt naar het nog schone roze tafelkleed en voelt alweer dat fris verheugen, dat steeds maar terugkeert sinds maandagochtend.' (Gerritsen 2012, p. 33) Als lezer volgen we de blik van Coco, waarna we aan de hand van haar reactie op het roze tafelkleed in de rest van het fragment getuige zijn van haar gedachten. Er is hier in feite sprake van ruis – het fragment is wellicht niet eenduidig genoeg – die wordt 'gecorrigeerd' door het model, dat echter ongevoelig blijkt voor het niet-waarneembare restant van het fragment. Deze foutieve classificatie is interessant omdat de gevolgen van interpretatie duidelijk worden. In de praktijk kunnen waarneembare en niet-waarneembare elementen elkaar voortdurend afwisselen binnen een fragment en – ondanks een extra categorie in mijn annotatieschema voor ambigue fragmenten – blijken zowel de menselijke beoordelaar als het classificatiemodel soms elementen te missen. Vanuit een narratologisch oogpunt is waarneembaarheid interessant omdat deze vaak wordt ingezet om informatieongelijkheid te creëren. De lezer wordt dan beïnvloed in zijn oordeel doordat hij meer te weten krijgt over het ene personage dan over het andere. Als deze mogelijkheid echter zo subtiel wordt ingezet dat een lezer niet altijd doorheeft dat er sprake is van informatieongelijkheid, dan mist hij wellicht cruciale nuances in zijn analyse. Door de menselijke lezer te confronteren met de beoordelingen van de computer toont de methode van *algorithmic failure* de tekortkomingen van de lezer: die ziet het waarneembare object niet en labelt het geheel als 'niet waarneembaar'. Maar ook: de fout die de computer hier maakt zou een lezer ook kunnen maken. Uiteraard geeft deze bevinding aanleiding tot reflectie op de annotatieprocedure; zou het model deze fouten ook hebben gemaakt als meer menselijke

lezers de fragmenten met ‘foutieve waarneembaarheid’ hadden beoordeeld? Wanneer meerdere lezers dezelfde fragmenten beoordelen, zou daarop een analyse kunnen volgen van interpretatieverschillen. Deze kan uitwijzen in hoeverre verschillen ontstaan door bijvoorbeeld ambiguïteit of door gebreken in de interpretatie van een lezer of onduidelijkheden in de definitie van een concept. Voor nu geeft de inzet van *algorithmic failure* voor het aspect van waarneembaarheid aanleiding tot nader onderzoek naar de manier waarop de lezer kennis toebedeeld krijgt door de focalisator.

4.3.3 Standvastigheid

Tijdens het bestuderen van de foutieve classificaties voor het aspect van standvastigheid werden de beperkingen van de methode van *algorithmic failure* duidelijk – en in het verlengde daarvan wellicht die van een computationele benadering van focalisatie. Tijdens de annotatieprocedure en het trainen van de modellen bleek al dat standvastigheid minder hoog scoorde dan het andere binaire aspect. Daarom zijn er bijvoorbeeld meer foutieve classificaties: fragmenten met één object zijn door het model 55 keer gelabeld als twee of meer objecten, fragmenten met meerdere objecten zijn 23 keer gelabeld als één object. Bij de daaropvolgende analyse van de betreffende fragmenten werd ik echter geconfronteerd met het feit dat sommige fragmenten simpelweg niet eenduidig binnen één categorie vallen. Waar dit voor de eerdere aspecten vaak (bijna) direct duidelijk is, bleef het lastig om standvastigheid te categoriseren. Laat ik benadrukken dat dit niet geldt voor alle fragmenten en dat het aspect in deze vorm geschikt is voor analyse. We zagen immers al dat de menselijke beoordelaars voldoende overeenkomen en dat de categorieën samenhang vertonen met die van de mate van detail. Er blijkt echter ook een groep fragmenten te zijn die maar lastig eenduidig valt in te delen binnen het aspect van standvastigheid. Het viel me bijvoorbeeld op dat een aantal ‘foutieve meerdere objecten’-fragmenten een combinatie kent van dialoog voorafgegaan of gevolgd door een waarneming, gedachte of handeling. Dit zijn vaak één of twee zinnen die de dialogen inleiden of afronden, zoals: ‘Ik stond op en pakte twee nieuwe flesjes bier uit de koelkast.’ (Van Marissing 2012, p. 37). Op zichzelf valt wellicht te beargumenteren dat er twee objecten zijn: de dialoog en de handeling daarna. Dergelijke fragmenten komen echter ook regelmatig voor bij de ‘foutieve één object’, waardoor deze analyse geen standhoudt. Bovendien suggereert dit dat de grenzen tussen beide categorieën in sommige gevallen tamelijk vaag kunnen zijn.

Deze bevindingen in ogenschouw nemende ben ik geneigd te concluderen dat dit aspect zich onvoldoende leent voor de methode van *algorithmic failure*, ondanks de veronderstelling dat ‘“bad” machine learning can be quite adequate if the goal is to identify rich cases for

analysis' (Rettberg 2022, p. 2). De methode vereist namelijk een zekere mate van samenhang binnen een categorie en die is er in het geval van standvastigheid te beperkt. Wellicht is deze classificatietaak te complex; ik bestudeer voor de methode van *algorithmic failure* bijvoorbeeld niet één woord, zoals het geval is bij de analyse van Jill Walker Rettberg (2022), maar een fragment van ongeveer 150 woorden. Al met al concludeer ik dat voor het aspect van standvastigheid geen betekenisvolle foutenanalyse kan worden uitgevoerd, maar dit hangt wellicht ook samen met de geringe omvang van de dataset.

4.3.4 Gedetailleerdheid

Eerder in dit hoofdstuk besprak ik al de ongelijke verdeling van categorieën binnen het aspect van gedetailleerdheid. Hieruit concludeerde ik dat een object zodra het wordt waargenomen wellicht algauw een zekere mate van detail toegeschreven krijgt, tenzij een beschrijving onderdeel wordt van een groter geheel, zoals een scène die geldt als één object. Ook bleek dat de modellen (inclusief het *random baseline*-model) amper fragmenten als lage detaillering heeft geclassificeerd – zeer lage mate van detail zelfs geen enkele keer. Enerzijds is dit weinig verwonderlijk gezien het gebrek aan voorbeeldfragmenten, anderzijds is het de moeite waard om wat meer grip te krijgen op deze categorieën. Ik zal daarom de methode van *algorithmic failure* inzetten voor de analyse van fragmenten die ik heb beoordeeld als (zeer) lage mate van detail en die niet als zodanig zijn geclassificeerd door het BERTje-model. In totaal gaat het om 24 fragmenten.

Een exemplarisch fragment dat ik heb beoordeeld als een lage detaillering en het model als zeer hoge, is afkomstig uit *Niemand in de stad* (Huff 2012, p. 77-78):

> Tweede kerstdag vlieg ik, zoals elk jaar, naar Elisabeth en haar ouders in Frankrijk. Mijn moeder gaat met een vriendin naar Stockholm. De skivakanties in Méribel zijn mijn favoriete vakanties. Niet alleen door het skiën, maar ook door de herhaling die in de dagen kruipt: het opstaan, de tafel dekken, het aanmaken van het vuur, met de vader van Elisabeth naar de bakker gaan, het ontbijten in een opgewarmd, houten huis. De auto in en naar de liften. 's Middags lunchen we op de piste, in een klein, warm restaurant met een Franse patron die "goedemiddak" zegt. Chez Kiki heet zijn uitspanning. Na de lunch stap ik op mijn skischoenen naar de wc, voorzichtig op de natte trap.

Mijn redenering is dat de herhaling van zowel de jaarlijkse skivakantie als de dagelijkse handelingen weerklinkt in de detaillering van de beschreven objecten. Als we het fragment opvatten als een verzameling van losstaande objecten, dan zien we dat die objecten nauwelijks meer worden beschreven dan simpelweg benoemen. Zo wordt er gesproken over 'een vriendin', 'de tafel' en 'de liften', waardoor ze iets inwisselbaars hebben. Een kleine uitzondering is het

restaurant, dat een naam krijgt en de patron wordt kort genoemd. Deze observatie raakt wellicht aan de kern van het segmentatieprobleem binnen deze scriptie: hoe baken ik af wat ik als object beschouw? In de huidige opzet heb ik fragmenten voor alle vier de aspecten op dezelfde wijze gesegmenteerd, maar in de praktijk laat dit ruimte open voor interpretatie van hoe bijvoorbeeld een object wordt afgebakend, zoals in bovenstaand fragment.

Beschouwen we anderzijds het fragment uit *Niemand in de stad* als één scène die een skivakantie beschrijft, dan zijn de onderdelen (ochtendritueel, houten huis, lunch) nog steeds inwisselbaar: het hadden de feiten en voorvallen kunnen zijn uit een groot aantal andere skivakanties. Wat betekent het dan als een model veel detail herkent in dit fragment? Wellicht is de grote hoeveelheid objecten in het fragment leidend, er komen bijvoorbeeld specifieke woorden voor als 'patron' en 'skischoenen' die kunnen doorgaan voor het toevoegen van detail. Er zijn meer fragmenten als dit waarbij relatief veel woorden worden besteed aan een situatie of een gedachte, terwijl het object weinig eigenschappen toegekend krijgt. Als lezer voel je bij zulke fragmenten dan ook afstand tot het beschreven object, hetgeen aanleiding kan geven tot een beoordeling van weinig detail. Een voorbeeld hiervan is een fragment uit *Euforie* (2012) waarin de hoofdpersoon de wereldgebeurtenissen uit zijn jeugd als volgt aankondigt: 'De decennia waarin ze opgroeiden waren precies die waarin de meeste hinder de deur uit was gewerkt.' (p. 220) Daaropvolgend somt hij in een paar zinnen oorlogen, natuurrampen en terrorisme op om vervolgens een punt te kunnen maken over dreigingen. De 'decennia' die worden aangekondigd krijgen dus een hele beschouwing met een hoop specifieke woorden, maar echt gedetailleerd wordt het fragment niet (het model kende het label 'zeer hoge mate van detail' toe).

Het feit dat ik dit patroon bij meerdere foutieve classificaties waarneem, zou een aanwijzing kunnen zijn voor het feit dat er een spanning bestaat tussen enerzijds formele tekstkenmerken en anderzijds de interpretatie ervan. Waar een aspect als waarneembaarheid van een object relatief weinig interpretatie vraagt van de lezer, is het aspect van gedetailleerdheid in hogere mate afhankelijk van de interpretatie van de lezer. Dit zagen we eerder al bij de overeenstemming tussen de menselijke lezers en keert – weinig verrassend – terug bij de classificaties van het model. Ik denk daarom dat we classificatietaken op een continuüm zouden kunnen plaatsen tussen formele teksteigenschappen en interpretatie. De modellen zijn daarbij beduidend succesvoller in het classificeren van de eerste. Het is interessant om dit gegeven te confronteren met letterkundige interpretaties. Hoe kan een particuliere interpretatie door middel van argumentatie tot consensus tussen meerdere lezers

worden? De relatie tussen computationele modellen en interpretatie kan ons wellicht met andere ogen naar het menselijke interpretatieproces doen kijken.

5. Conclusie en discussie

In deze scriptie stelde ik me ten doel focalisatie te operationaliseren om het concept computationeel te kunnen onderzoeken. Mijn hoofdvraag luidde als volgt: 'Hoe kan focalisatie computationeel onderzocht worden?' Het was dus methodologisch mijn doel een manier te vinden om het bestaande narratologische concept automatisch te kunnen herkennen, waardoor dit onderzoek verkennend van aard was. Ik heb daartoe een definitie opgesteld van focalisatie en vier aspecten[22] geselecteerd die samen een beschrijving vormen van focalisatie. Vervolgens heb ik een corpus bestaande uit romanfragmenten van 150 woorden handmatig geannoteerd voor deze vier aspecten. Om er zeker van te zijn dat mijn eigen interpretatie van focalisatie die van een particuliere interpretatie voldoende overstijgt, vergeleek ik een deel van mijn beoordelingen met die van twee andere lezers. De inter-beoordelaarsbetrouwbaarheid bleek voldoende aanwezig, hoewel niet voor alle aspecten in even hoge mate. Daarna vergeleek ik verschillende manieren om de fragmenten numeriek te representeren en twee modellen om de fragmenten te categoriseren, namelijk logistische regressie en het Nederlandstalige BERT-model. Voor drie van de vier aspecten (waarneembaarheid, standvastigheid en gedetailleerdheid) was het BERT-model het beste in staat om de classificatietaak uit te voeren, voor één aspect was dit logistische regressie (focaliserend subject). Dit sluit gedeeltelijk aan bij de stelling van Ted Underwood (2019) dat simpele *bag-of-words*-modellen voor classificatietaken niet veel onderdoen voor complexere algoritmes. Ik denk dat dit geldt voor classificatietaken die vooral afhankelijk zijn van formele tekstkenmerken, terwijl voor taken met een hogere mate van interpretatie meer context (en daarmee complexere modellen) vereist is.

De methodologische verkenning die ik in mijn scriptie heb uitgevoerd heeft op twee vlakken een conclusie opgeleverd, die samenhangt met de twee deelvragen die ik heb opgesteld. Bovendien maakt ieder van deze conclusies enkele beperkingen van mijn methode duidelijk. Als eerste trek ik een conclusie uit mijn benadering van de annotatietaak als vorm van *close reading*. Dit deel van mijn scriptie diende – naast het vergaren van mijn corpus met trainingsmateriaal voor de modellen – om de eerste deelvraag te beantwoorden: 'Welke toevoeging aan de narratologie biedt het systematisch onderzoeken van focalisatie?' Het operationaliseren van narratologisch concept naar annotatie vormde namelijk de aanleiding

[22] De vier aspecten waren: positie van het focaliserend subject, waarneembaarheid van het gefocaliseerde object, aantal waargenomen objecten (standvastigheid) en gedetailleerdheid van het gefocaliseerde object.

voor een systematische lezing van vier romans uit 2012; in plaats van de relevante delen van een tekst te betrekken bij een analyse, heb ik een groot deel van de romans systematisch bestudeerd. Hieruit bleek dat interne focalisators ruim in de meerderheid zijn in mijn corpus en dat een 'klassieke' waarnemer die boven het verhaal hangt nagenoeg afwezig is. De gedachten van personages zijn daarentegen niet zo sterk aanwezig als zou worden verwacht van interne waarnemers. Ondanks een veronderstelde wending naar binnen (Watt 1957; Jahn 2007) blijkt de buitenwereld, onder meer in de vorm van dialogen, aanwezig in een groot deel van het corpus.

Een andere interessante bevinding van de systematische lezing bestaat uit het feit dat gefocaliseerde objecten in hoge mate met veel detaillering worden waargenomen. Dit hangt samen met het aantal objecten in fragmenten en draagt bij aan de bevinding dat focalisatie in hoge mate vatbaar is voor interpretatie. In een 'traditionele' *close reading* worden over het algemeen de meest eenduidige passage besproken die een duidelijke interpretatie opleveren. Wanneer we voor een willekeurig fragment de focalisatie beschrijven blijkt ambiguïteit echter eerder regel dan uitzondering. Bovendien is de schaal van analyseren een belangrijke factor; wordt focalisatie bestudeerd op hoofdstukniveau, dan levert dat andere resultaten op dan voor fragmenten van 150 woorden. Dit is tegelijkertijd een beperking van mijn onderzoek. Ik heb immers gekozen voor een vaste fragmentgrootte en die keuze heeft zonder twijfel beïnvloed dat ik bijvoorbeeld veel fragmenten met een hoge mate van detail heb gevonden. Ik denk daarom dat mijn onderzoek het belang van transparantie laat zien – ook voor *close readings* – over de onderliggende keuzes die binnen een onderzoek worden gemaakt, bijvoorbeeld met betrekking tot de schaal van bestuderen en over mogelijke blinde vlekken. Binnen deze scriptie had ik niet de ruimte om mijn annotatieprocedure uit te breiden, maar goede voorbeelden van hoe verschillende beoordelaars samen tot een groter begrip van een concept kunnen komen zijn Evelyn Gius & Janina Jacke (2017) en Krishnapriya Vishnubhotla, Adam Hammond & Graeme Hirst (2022).

Daarnaast zou vervolgonderzoek met een uitgebreider annotatieschema kunnen werken. Een logische toevoeging zou een beoordeling van de attitude van het focaliserende subject tegenover het gefocaliseerde object zijn, zoals gesuggereerd door Mieke Bal (2009, p. 153). Ook zou het aspect van de positie van de focalisator nader gespecificeerd kunnen worden. Ik denk bijvoorbeeld aan strengere definities voor categorieën, zodat onder meer de grote groep van interne focalisators verder wordt uitgediept.[23] Het herkennen van verschillende interne

[23] Om hier nader op in te gaan: een veelvoorkomend probleem met eenduidig labelen van de positie focalisator treedt op wanneer de handelingen of gedachten (waarneembaar of niet-waarneembaar) van een personage worden

focalisators en vertellagen[24] lijken me hoe dan ook een belangrijke vervolgstappen. Verder is het herkennen en categoriseren van dialogen voor een Nederlandstalig corpus een belangrijke aanvulling. Andreas van Cranenburgh (2019) heeft hiertoe met het herkennen van coreferentie al een belangrijke basis gelegd, terwijl Krishnapriya Vishnubhotla, Adam Hammond & Graeme Hirst (2022) dialogen herkennen en toewijzen voor Engelstalige teksten.

Het tweede en derde deel van het resultatenhoofdstuk, waarin ik respectievelijk de scores van de modellen bespreek en de classificaties van de beste modellen per aspect analyseer met behulp van de methode van *algorithmic failure*, dienen beide om de tweede deelvraag te beantwoorden. Deze is namelijk: 'In hoeverre kan focalisatie worden geformaliseerd ten behoeve van digitaal onderzoek?' Waar ik in het resultatenhoofdstuk mijn bevindingen nog per aspect bestudeerde, richt ik me nu op het geheel. Het gaat er nu dus om de resultaten voor de vier aspecten samen te bestuderen om uitspraken te kunnen doen over het concept focalisatie. Met andere woorden, wat kan ik op basis van mijn resultaten overkoepelend concluderen over het automatisch herkennen van focalisatie? Ik keer hiertoe weer terug naar de narratologische theorie, waarin focalisatie wordt beschreven als 'een verhouding tussen een "object" dat waargenomen wordt en een "subject" dat waarneemt' (Herman & Vervaeck 2009a, p. 75). Hoewel in het ideale geval alle onderdelen van die verhouding uitputtend in kaart worden gebracht, bleek dat binnen deze scriptie niet mogelijk. Op basis van het functioneren van de modellen per aspect kunnen we echter wel vaststellen of er onderdelen van focalisatie met succes automatisch kunnen worden herkend.

Wat de positie van het focaliserende subject betreft werden de resultaten beïnvloed door de ongelijke verdeling van voorbeelden per categorie. Zo komt het herkennen van interne focalisators dicht in de buurt van de menselijke oordelen. De categorieën van externe focalisator en ambigue/consonante focalisator scoorden echter niet hoog genoeg op toegepast te kunnen worden op nieuwe teksten, ondanks het feit dat ze de *random baseline*-scores ruimschoots overtreffen. Dit sluit aan bij de conclusie van Corina Koolen & Andreas van Cranenburgh (2018), die schrijven dat hun onderzoeksobject meer 'wereldkennis' vereist en dat ze

waargenomen. Op basis van slechts die waarneming (en vaak ook met meer context) valt veelal niet met zekerheid te zeggen wie de waarnemende instantie is; een externe focalisator die dicht op het personage zit, het (interne) personage zelf of beiden? In mijn scriptie heb ik in dergelijke gevallen vaak gekozen voor de tweede optie. Gezien de verkennende aard van dit onderzoek valt deze keuze te rechtvaardigen, maar vervolgonderzoek zou dit probleem nader kunnen uitwerken.

[24] Pogingen hiertoe zijn al ondernomen in de artikelenreeks 'Annotation Guidelines' van *Journal of Cultural Analytics*. Vreemd genoeg werken onder meer Mats Wirén & Adam Ek (2021) met de indeling van focalisatie die Genette (1980) maakte, terwijl deze in de narratologische theorie al een behoorlijke tijd is herzien (verg. Bal 1983).

onvoldoende trainingsdata tot hun beschikking hebben (p. 66). De wereldkennis waaraan ze refereren vat ik in het geval van focalisatie op als de delen van een tekst die interpretatie vereisen. Het *bag-of-words*-model is in staat om ieder lexicaal element te benutten als *feature*, zowel positieve als negatieve *cues* (p. 65). De impliciete tekstdelen die aanleiding geven tot een interpretatie kunnen daardoor mogelijk alsnog worden herkend door een model, maar dit vraagt om een voldoende aantal voorbeelden. De analyse met *algorithmic failure* wees uit dat de automatische classificaties inderdaad in staat zijn om de menselijke beoordelingen te confronteren met ambigue fragmenten. Enerzijds kan dit dienen als aanleiding tot een nadere *close reading*, anderzijds laat deze bevinding zien dat een algoritme sommige nuances expliciet maakt die een menselijke lezer niet opmerkt. Hoewel de positie van focaliserend subject dus nog niet automatisch kan worden toegepast op nieuwe teksten, denk ik in deze scriptie hiervoor dus de basis te hebben gelegd en ik ben hoopvol met betrekking tot de kans van slagen van vervolgonderzoek.

Het gefocaliseerde object heb ik getracht te beschrijven met de overige drie aspecten, terwijl die uiteraard indirect ook het focaliserende subject karakteriseren. Het aspect van waarneembaarheid kende de hoogste scores en leent zich in principe goed voor toepassing op onbekende teksten. Een mogelijke toevoeging voor dit aspect bestaat uit een aparte categorie voor dialogen, zodat de categorie met waarneembare objecten duidelijker gericht is op waarnemingen van de verhaalwereld. Daarnaast bleek uit de *algorithmic failure*-analyse dat dit aspect wellicht baat zou hebben bij annotaties op zinsniveau, zodat de subtiele wisselingen tussen waarneembare en niet-waarneembare categorieën eenduidiger kunnen worden geclassificeerd.

Het aantal waargenomen objecten bleek met onvoldoende zekerheid automatisch te worden herkend, waardoor het zich vooralsnog niet leent voor verdere toepassing. Hoewel meer voorbeelden dit probleem waarschijnlijk deels verhelpen, bleek uit de inter-beoordelaarsbetrouwbaarheidstoetsen en de *algorithmic failure*-analyse dat tekstfragmenten soms te weinig samenhang hebben om dit aspect eenduidig te kunnen categoriseren. Dit hangt samen met de fragmentgrootte: naarmate een fragment meer woorden bevat, is de kans kleiner dat één object wordt waargenomen. Zo bezien zou dit aspect wellicht vragen om een herkenningstaak die de vraagt stelt waar de beschrijving van een object begint en eindigt. Corina Koolen & Andreas van Cranenburgh (2018) laten echter zien dat ook deze taak behoorlijk complex is. Het is daarentegen opvallend dat dit aspect samenhang toont met het aspect van gedetailleerdheid. Daarmee biedt dit aspect aanleiding tot vervolgonderzoek dat de aard van waarnemingen nader bestudeert door verschillende aspecten die bijdragen aan de mate

van detaillering te onderzoeken in het licht van een cognitieve benadering van narratologie, zoals gesuggereerd door David Herman (2009). Een grotere groep lezers zou een corpus kunnen annoteren om te onderzoeken of er patronen in de waarneming zitten. De uitkomsten van de analyse van *algorithmic failure* suggereren namelijk dat deze relatie complexer is dan een simpele optelsom van factoren zoals het aantal waargenomen objecten. Mijn onderzoek bevestigt daarmee de complexiteit van lees- en begripsprocessen, bijvoorbeeld over waar we grenzen van objecten afbakenen.

Het model voor het aspect van gedetailleerdheid bleek per categorie de minst accurate voorspellingen te doen en is dan ook niet geschikt voor het labelen van onbekende teksten. Wederom heeft dit te maken met een gebrek aan trainingsmateriaal en eenduidigheid binnen de categorieën. Ik denk bovendien dat dit aspect zich goed leent voor een annotatietaak waarbij meerdere beoordelaars betrokken zijn en die tot een gezamenlijke interpretatie van teksten kunnen komen.

Concluderend kan de tweede deelvraag dus niet eenduidig positief worden beantwoord. Slechts het aspect van waarneembaarheid is succesvol geformaliseerd en geoperationaliseerd zodat het digitaal onderzocht kan worden. Ik denk dat de overige aspecten echter voldoende aanleiding bieden om in vervolgonderzoek alsnog digitaal onderzocht te worden. Al met al kan ik gezien de verkennende aard van deze scriptie spreken van een positief resultaat.

Met betrekking tot de veronderstelde objectiviteit van ‘de computer’ hoop ik in deze scriptie inzichtelijk te hebben gemaakt hoeveel keuzes ik heb gemaakt om tot mijn resultaten te komen, zowel in de aanloop naar het model, rondom het model zelf, als tijdens de evaluatie en interpretatie van het model. Wanneer we dus spreken over objectiviteit van computers, gaan we voorbij aan de menselijke hand die voortdurend aanwezig is. Ik zou zelfs durven stellen dat computers minder objectief zijn dan *close readings*, die toch bij uitstek subjectief van aard zijn. Dit heeft ermee te maken dat je als lezer weet dat er tijdens een interpretatie belangrijke keuzes zijn gemaakt die de resultaten ervan beïnvloeden. De geveinsde objectiviteit van computers beweert het tegendeel en verdoezelt die keuzes, hetgeen ik in deze scriptie heb proberen te voorkomen. Hiermee volg ik Andrew Pipers (2017) pleidooi voor modellen die inzichtelijk maken hoe een onderzoeker tot zijn beweringen komt. Juist in de botsing tussen de menselijke lezer en computationele classificaties ontstaat de complementaire waarde die beide benaderingen hebben, zo bleek uit onder meer mijn *algorithmic failure*-analyse. Te midden van een discours waarin technologisch positivisme – dat gericht is op het behalen van de beste onderzoeksresultaten – overheerst, is het van belang als geesteswetenschapper een kritische blik

te behouden ten opzichte van de onderzoeksobjecten die we bestuderen én de onderzoeksmethoden die we daarbij hanteren.

Vervolgonderzoek zou de invloed van fragmentgrootte op de training van algoritmes nader kunnen bestuderen. Mijn resultaten laten zien namelijk zien dat dit wel degelijk een factor is, terwijl onder meer Ted Underwood (2018) en Andrew Piper, Sunyam Bagga, Laura Monteiro et al. (2021) hier weinig aandacht aan besteden. Bovendien beschreef ik op basis van een artikel van Herman & Vervaeck (2009b) over *The Echo Maker* van Richard Powers hoe dit vanuit een narratologisch perspectief ook een aanvulling is. Daarnaast kan de verhouding tussen de subject- en objectpositie van focalisatie explicieter beschreven worden. Dit kan door toevoeging van bijvoorbeeld een aspect van attitude, die de houding (positief of negatief) van het subject ten opzichte van het object categoriseert. Dit betreft echter een aspect dat een hogere mate van interpretatie vereist en wellicht minder sterk aanwezig is aan de oppervlakte van een tekst. Het zou daarom interessant zijn om ook dit aspect als gezamenlijk annotatieproces te benaderen.

Tot slot heeft deze scriptie een weg gebaand voor een computationele benadering van diachrone narratologie. Enerzijds besprak ik in deze scriptie voorbeelden van longitudinaal onderzoek naar literatuur die zich richt op zogenaamde *century-spanning trends* (Ted Underwood 2018; Ted Underwood 2019; Andrew Piper, Sunyam Bagga, Laura Monteiro et al. 2021). Anderzijds is binnen de narratologie al eerder betoogd dat onderzoek naar de geschiedenis van narratieve vormen ons meer kan leren over welke functies narratologische concepten hebben gehad binnen verhalen (Fludernik 2003; De Jong 2014). Met dit verkennende onderzoek denk ik een stap te hebben gezet in de richting van een synthese van beide onderzoekstradities. Om op mijn hoofdvraag terug te komen: al met al heb ik gevonden dat focalisatie computationeel onderzocht kan worden, maar er is nog een weg te gaan om zulk onderzoek te perfectioneren. Moge mijn onderzoek een heldere, eerste stap zijn.

6. Bibliografie

Bal, M. (1981). Notes on Narrative Embedding. *Poetics Today 2*(2), p. 41-59.

Bal, M. (1983). The Narrating and the Focalizing: A Theory of the Agents in Narrative. *Style 17*(2), p. 234-269.

Bal, M. (2009). *Narratology: Introduction to the theory of narrative*. Toronto: Toronto University Press.

Barthes, R. (1974). *S/Z: An Essay*. New York: Hill & Wang.

Bousset, H. (1988). *Lezen om te schrijven. Een progressieve en cumulatieve lectuur van Het boek alfa van Ivo Michiels*. Amsterdam: De Bezige Bij.

Caracciolo, M. (2022). *Slow narrative and nonhuman materialities*. Lincoln: University of Nebraska Press.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Education and Psychological Measurement, 20*(1), p. 37-46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*(4), p. 213-220.

Coux, J. De (2012). 'Horen, zien en schrijven in Brakmans *Pop op de bank*'. In: L. Bernaerts & B. Vervaeck (red.), *Het binnenste buiten. Werk en leven van Willem Brakman*, p. 59-70. Gent: Academia Press.

Cranenburgh, A. van. (2019). 'A Dutch coreference resolution system with an evaluation of literary fiction'. *Computational Linguistics in the Netherlands Journal, 9*.

Deijl, L. van der, Pieterse, S., Prinse, M. & Smeets, R. (2016). Mapping the demographic landscape of characters in recent Dutch prose: A quantitative approach to literary representation. *Journal of Dutch Literature, 7*(1), p. 20-42.

Devlin, J., Chang, M., Lee, K. & Toutanova, K. (2019). 'BERT: Pre-training of deep bidirectional transformers for language understanding'. *arXiv:1810.04805v2 [cs.CL]*.

Fludernik, M. (2003). 'The Diachronization of Narratology'. *Narrative 11*(3), 331-348.

Fludernik, M. (2009). *An Introduction to Narratology.* New York: Routledge.

Genette, G. (1980): *Narrative Discourse. An Essay in Method*. Ithaca, New York: Cornell University Press.

Gerritsen, E. (2012). *Dorst*. Breda: De Geus.

Gius, E. & Jacke, J. (2017). 'The Hermeneutic Prof of Annotation: On Preventing and Fostering Disagreement in Literary Analysis. *International Journal of Humanities and Arts Computing, 11*(2), p. 233-254.

Gries, S. Th. (2013). *Statistics for Linguistics with R: A Practical Introduction.* Berlin/Boston: De Gruyter Mouton.

Hallgren, K. (2012). Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutorials in Quantative Methods for Psychology*, *8*(1), p. 23-34.

Herman, D. (2009). 'Beyond Voice and Vision: Cognitive Grammar and Focalization Theory'. In: P. Hühn, W. Schmid, J. Schönert (Eds.). *Point of View, Perspective and Focalization. Modeling Mediation in Narrative*, p. 119-142. Berlijn/New York: Walter de Gruyter.

Herman, L. & Vervaeck, B. (2009a). *Vertelduivels*. Nijmegen: Uitgeverij Vantilt.

Herman, L. & Vervaeck, B. (2009b). 'Capturing Capgras: *The Echo Maker* by Richard Powers'. *Style 43*(3), 407-428.

Huff, P. (2012). *Niemand in de stad*. Amsterdam: De Bezige Bij.

Impett, L. (2020). Analyzing Gesture in Digital Art History. In: *The Routledge Companion to Digital Humanities and Art History*, p. 386-407. New York: Routledge.

Jahn, M. (2007). 'Focalization'. In: D. Herman (ed.), *Cambridge Companion to narrative* (p. 94-108). Cambridge: Cambridge University Press.

Jong, Irene J. F. de (2014): 'Diachronic Narratology (The Example of Ancient Greek Narrative)'. In: Peter Hühn et al. (Eds.), *The living handbook of narratology*. Via: https://www-archiv.fdm.uni-hamburg.de/lhn/node/95.html.

Jurafsky, D. & Martin, J. H. (2023). *Speech and Language Processing* [online, onafgerond boek, beschikbaar via https://web.stanford.edu/~jurafsky/slp3/]. Stanford: Stanford University.

Karsdorp, F. (2016). *Retelling Stories. A Computational-Evolutionary Perspective* [Proefschrift]. Nijmegen: Radboud Universiteit.

Koolen, C. & Cranenburgh, A. van (2018). Blue eyes and porcelain cheeks: Computational extraction of physical descriptions from Dutch chick lit and literary novels. *Digital Scholarship in the Humanities*, *33*(1), p. 59-71.

Landis, J. R. & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), p. 159-174.

Light, R. J. (1971). Measures of response agreement for qualitative data: Some generalizations and alternatives. *Psychological Bulletin*, *76*(5), p. 365-377.

Marissing, R. van, (2012). *Strak blauw*. Amsterdam/Antwerpen: Atlas Contact.

Meijer, M. (1996). *In tekst gevat. Inleiding tot de kritiek van de representatie*. Amsterdam: Amsterdam University Press.

Meinkema, H. (1980). *De maaneter*. Amsterdam/Brussel: Elsevier Manteau.

Meister, J.C. & Schönert, J. (2009). ‘The DNS of Mediacy’. In: P. Hühn, W. Schmid, J. Schönert (Eds.). *Point of View, Perspective and Focalization. Modeling Mediation in Narrative*, p. 11-40. Berlijn/New York: Walter de Gruyter.

Moretti, F. (2013). ‘Operationalizing’: Or, the Function of Measurement in Literary Theory. *New Left Review (84)*, 103-119.

Mostafazadeh, N., Chambers, N., He, X., Parikh, D., Batra, D., Vanderwende, L., Kohli, P. & Allen, J. (2016). A Corpus and Cloze Evaluation for Deeper Understanding of Commonsense Stories. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 839-849.

Munk, A. K., Olesen, A. G. & Jacomy, M. (2022). The Thick Machine: Anthropological AI between explanation and explication. *Big Data & Society*, *9*(1), p. 1-14.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradburry, J., Chanan, G. et al. (2019). ‘PyTorch: An Imperative Style, High-Performance Deep Learning Library’. *arXiv:1912.01703 [cs.LG]*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M. et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*, p. 2825-2830.

Piper, A. (2016). ‘There will be numbers’. *Journal of Cultural Analytics*.

Piper, A. (2017). ‘Think Small: On Literary Modeling’. *PMLA*, *132*(3), 651-658.

Piper, A. (2020). *Can We Be Wrong? The problem of Textual Evidence in a Time of Data.* Cambridge: Cambridge University Press.

Piper, A. (2021, 16 november). *Detecting narrativity across long time scales*. .txtlab. Geraadpleegd op 10 februari 2023, van https://txtlab.org/2021/11/detecting-narrativity-across-long-time-scales/

Piper, A., Bagga, S., Monteiro, L., Yang, A., Labrosse, M. & Liu, Y. L. (2021). Detecting Narrativity Across Long Time Scales. In: *CHR 2021: Computational Humanities Research Conference*, 319–332.

Powers, R. (2006). *The Echo Maker*. New York: Farrar, Strauss and Giroux (Picador).

Rettberg, J. W. (2022). Algorithmic failure as a humanities methodology: Machine learning’s mispredictions identify rich cases for qualitative analysis. *Big Data & Society*, *9*(2), p. 1-6.

Rimmon-Kenan, S. (2005). *Narrative Fiction*. Londen/New York: Routledge.

Sims, M., Park, J. H. & Bamman, D. (2019). Literary Event Detection. In A. Korhonen, D. Traum & L. Màrquez (Red.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3623-3634). Florence: Association for Computational Linguistics.

Smeets, R. (2021). *Character Constellations. Representations of Social Groups in Present-Day Dutch Literary Fiction*. Leuven: Leuven University Press.

So, R. J. (2017). 'All Models Are Wrong'. *PMLA, 132*(3), 668-673.

Stemler, E. & Tsai, J. (2011). Best Practices in Interrater Reliability Three Common Approaches. In J. Osborne (Ed.), *Best Practices in Quantitative Methods* (p. 29-49). Thousand Oakes: SAGE Publications, Inc.

Underwood, T. (2017). 'A genealogy of distant reading'. *Digital Humanities Quarterly, 11*(2).

Underwood, T. (2018). Why Literary Time is Measured in Minutes. *ELH 85*, p. 341-365.

Underwood, T. (2019). *Distant Horizons: Digital Evidence and Literary Change*. Chicago/Londen: The University of Chicago Press.

Underwood, T. (2020). 'Machine Learning and Human Perspective'. *PMLA, 135*(1), 92-109.

Veerbeek, J. (2020). *De culturele verankering van evaluatie* [Masterscriptie]. Utrecht: UU.

Vishnubhotla, K., Hammond, A. & Hirst, G. (2022). 'The Project Dialogism Novel Corpus: A Dataset for Quotation Attribution in Literary Texts'. *arXiv:2204.05836v1 [cs.CL]*.

Vitse, S. (2023, nog te verschijnen). 'Traditioneel' lezen binnen de computationele letterkunde. *Cahier voor Literatuurwetenschap 14*.

Vries, W. de, Cranenburgh, A. van, Bisazza, A., Caselli, T., Noord, G. van & Nissim, M. (2019). 'BERTje: A Dutch BERT Model'. *arXiv:1912.09582 [cs.CL]*.

Watt, I. (1957). *The Rise of the Novel: Studies in Defoe, Richardson and Fielding*. Berkeley/Los Angeles: University of California Press.

Weijts, C. (2012). *Euforie*. Utrecht/Amsterdam/Antwerpen: De Arbeiderspers.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A. et al. (2020). 'HuggingFace's Transformers: State-of-the-art Natural Language Processing'. *arXiv:1910.03771v5 [cs.CL]*.

Bijlage 1 Annotatieschema

Hieronder volgen de concrete stappen die de beoordelaars voor ieder fragment doorlopen tijdens het proces van annoteren. Per aspect staan de mogelijke beoordelingen genoemd, inclusief bijbehorende codering.

'Wat is in dit fragment de positie van de focalisator ten opzichte van de verhaalwereld?'
Het antwoord op deze vraag kan worden gekozen uit de volgende categorieën:

1. Extern;
2. Ambigu/consonant;
3. Intern;
4. Restcategorie;
99. Dialoog.

Een hulpmiddel bij het onderscheiden van externe en interne focalisatie, zoals gesuggereerd door Shlomith Rimmon-Kenan (2005, p. 76-77), bestaat eruit het betreffende fragment te herschrijven in de eerste persoon. Als dit 'kan' is het fragment intern gefocaliseerd, zo niet dan is het extern gefocaliseerd. Hoewel dit niet geldt als algemene regel, dient het als leidraad voor het maken van een basaal onderscheid tussen beide vormen van focalisatie.

Alvorens over te gaan naar drie stellingen voor de resterende aspecten, moet er eerst een tussenvraag worden beantwoord: 'Wat is het gefocaliseerde object in dit fragment?' Het is bij het bepalen van het object van belang uit te gaan van het hoogste niveau van waarneming.
Nadat het gefocaliseerde object is bepaald, kunnen de stellingen hieronder worden beoordeeld.

'Het gefocaliseerde object in dit fragment is waarneembaar.'
Mogelijke beoordelingen:

0. Nee;
1. Ja;
99. Meerdere objecten die als geheel niet in dezelfde categorie passen.

'Het gefocaliseerde object in dit fragment is niet standvastig.'
Mogelijke beoordelingen:

1. Nee, er wordt één object gefocaliseerd;

2. Ja, er worden twee objecten gefocaliseerd;
3. Ja, er worden meer dan twee objecten gefocaliseerd.

‘Het gefocaliseerde object in dit fragment wordt gedetailleerd waargenomen.’
Mogelijke beoordelingen kunnen worden gegeven op een vijfpuntsschaal:

1. Helemaal oneens;
2. Oneens;
3. Niet zeker;
4. Mee eens;
5. Helemaal eens;
99. Meerdere objecten in verschillende mate van detail.

Bijlage 2 Classificatiescores gedetailleerdheid

		1			2		
Model	*Features*	*Precision*	*Recall*	*F*1	*Precision*	*Recall*	*F*1
Annotaties		0.00	0.00	0.00	0.43	0.27	0.33
Random baseline		0.00	0.00	0.00	0.00	0.00	0.00
LR	BoW	0.00	0.00	0.00	0.17	0.11	0.13
	Tf-idf	0.00	0.00	0.00	0.17	0.11	0.13
BERTje		0.00	0.00	0.00	1.00	0.06	0.11

		3			4		
Model	*Features*	*Precision*	*Recall*	*F*1	*Precision*	*Recall*	*F*1
Annotaties		0.18	0.29	0.22	0.71	0.42	0.53
Random baseline		0.00	0.00	0.00	0.28	0.24	0.26
LR	BoW	0.22	0.25	0.24	0.34	0.39	0.36
	Tf-idf	0.12	0.08	0.10	0.34	0.44	0.38
BERTje		0.20	0.04	0.07	0.34	0.44	0.39

		5		
Model	*Features*	*Precision*	*Recall*	*F*1
Annotaties		0.35	0.92	0.51
Random baseline		0.46	0.48	0.47
LR	BoW	0.57	0.55	0.56
	Tf-idf	0.52	0.53	0.52
BERTje		0.60	0.77	0.68

Tabel 15. Resultaten voor het classificeren van de gedetailleerdheid van het gefocaliseerde object per model en feature. De tabel representeert de volgende vijf categorieën van dit aspect: 1) zeer lage mate van detail; 2) lage mate van detail; 3) niet laag, niet hoog; 4) hoge mate van detail; 5) zeer hoge mate van detail. De bovenste rij toont de resultaten van de handmatige annotatietaak. De gekleurde rij markeert het model met de hoogste gewogen F1-score (annotaties buiten beschouwing gelaten).