Analyse van gender bias in word embeddings van de Nederlandse taal op basis van beroepsnamen

Bachelor Kunstmatige Intelligentie
Universiteit Utrecht
Bachelor Scriptie (7.5 ECTS)

Pascal Verkade
6045057

Begeleider: Dr. D.P. Nguyen
Tweede beoordelaar: Dr. F.W. Adriaans

10 februari, 2020

Abstract

Eén van de belangrijkste elementen binnen Natural Language Processing (NLP) zijn word embeddings. Hierbij is elk woord gerepresenteerd door een vector en uit deze vectoren kunnen verschillende relaties worden gehaald die gebruikt kunnen worden binnen de analysemethoden en toepassingen van NLP. Binnen onze taal bevindt zich echter ook menselijke bias zoals vooroordelen en stereotypen waardoor woorden bepaalde associaties krijgen met een groep, gender of ras. Aangezien word embeddings worden geleerd op basis van onze taal middels grote corpora met teksten, zal de bias vermoedelijk ook in deze word embeddings aanwezig zijn. In dit onderzoek zal de gender bias in word embeddings van de Nederlandse taal op basis van beroepsnamen onderzocht worden. Deze beroepsnamen zijn in de Nederlands taal bijzonder moeilijk te analyseren op bias. In het Nederlands wordt er namelijk meestal een onderscheid gemaakt tussen mannelijke en vrouwelijke termen. Na het creëren van een gender subspace en een set met genderneutrale beroepsnamen, zal de directe gender bias geanalyseerd worden middels een cosinusgelijkenis tussen de embeddings van de beroepsnamen en de zelf gecreëerde gender richting uit die subspace. De resultaten hiervan bevestigen ondanks de vrij kleine set van genderneutrale beroepsnamen dat er inderdaad een gender bias in de word embeddings zit. Wel is er vervolgonderzoek nodig om met behulp van andere methoden en een bredere scope de gender bias (of een andere bias) duidelijk in kaart te brengen en te analyseren.

Inhoudsopgave

1 Introductie

Word embeddings

Word embeddings zijn in de laatste jaren een steeds belangrijkere rol gaan spelen in *Natural Language Processing (NLP)* en vormen daarin een essentiële bouwsteen: ze liggen ten grondslag aan vrijwel elk neuraal netwerk dat taal analyseert, wat te danken is aan de betere resultaten en de betekenisvolle representatie ten opzichte van andere (vergelijkbare) representaties. (Pennington, Socher, & Manning, 2014). Elke word embedding is een woord $w \in W$, gecodeerd als een d-dimensionale vector $v_w \in R^d$ (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). De gehele set van word embeddings vormt daarbij een semantisch woordenboek voor programma's en algoritmen die de semantiek en betekenissen van woorden willen gebruiken.

Elke vector bevat waarden die gebaseerd zijn op co-occurrence, wat gezien kan worden als het voorkomen van het gecodeerde woord in de omgeving van een ander woord (Lund, & Burgess, 1996). Hierdoor bevatten de vectoren analyseerbare semantische betekenis: woorden met (bijna) gelijke semantiek en betekenis zullen ook (bijna) gelijke vectoren hebben. Zo zal de vector van 'Playstation' sterk op die van 'Xbox' lijken. Naast gelijkenissen kunnen ook relaties tussen woorden gevonden en uitgedrukt worden binnen word embeddings. Neem bijvoorbeeld het rekenkundig verschil tussen de vectoren van 'Paris' en 'France'. Dit zal parallel zijn aan het verschil tussen de vectoren 'London' en 'England'. Het belang hiervan komt nog meer naar voren als we deze relaties toepassen op analogieën (Mikolov, Yih, & Zweig, 2013)**.** Neem de analogie "*Man* is to *king* as *woman* is to *x*", welke te noteren is als *man:king :: woman:x*. Met een correcte word embedding wordt hier gevonden *x=queen*. Er zal namelijk te zien zijn dat wanneer we de vector van *man* aftrekken van de vector van *king,* daar vervolgens de vector van *woman* bij optellen, we uitkomen bij de vector van *queen*.

Gender bias

Met dit laatste voorbeeld komen we in de buurt van het onderwerp van dit onderzoek. *King* en *queen* zijn namelijk beide genderspecifiek volgens hun definitie en dus is er met het antwoord op de analogie geen probleem. Het probleem treedt op als we genderstereotype hij/zij analogieën bekijken. Nu zal het systeem last hebben van genderstereotypering: er wordt *x=nurse* in de analogie "*father* is to a *doctor* as a *mother* is to a *x*" gevonden. Dit komt door de de training op basis van co-occurrence: er zijn niet alleen in het algemeen, maar ook in onze data simpelweg meer mannelijke doktoren en meer vrouwelijke verpleegsters. Het is echter niet wenselijk om dit te encoderen in de word embeddings, omdat dit dus resulteert in een stereotyperend antwoord. Dit resultaat illustreert hiermee één van de grootste problemen in machine learning dat zeker bij NLP negatieve gevolgen heeft: *human bias*. Sinds het toenemen van het gebruik van word embeddings, zijn verschillende algoritmen hiervan onderzocht en geperfectioneerd. Het onderzoek naar human bias in word embeddings heeft daarentegen pas in de laatste drie jaar meer aandacht gekregen.

Sinds het merendeel van deze aandacht (logischerwijs) vooral uitging naar de Engelse corpora, ontbreekt veel onderzoek naar de human bias in anderstalige corpora. Daarom wordt in dit onderzoek gekeken naar de rol van die human bias in Nederlandstalige word embeddings. Dit is met name interessant in de Nederlandse taal, omdat deze op een specifiek onderdeel verschilt met het Engels. Het gaat hierbij om het volgende (versimpelde) onderscheid: beroepsnamen en rollen met zowel een

mannelijke als vrouwelijke vorm tegenover die met slechts een enkele vorm. In het eerste geval gaat het om woorden zoals ‘bestuurder - bestuurster’, ‘student - studente’ en ‘leraar - lerares’. Bij deze woorden zal met de vrouwelijke vorm altijd een vrouw worden bedoeld en met de mannelijke vorm (bijna) altijd een man. In het tweede geval gaat het om woorden als ‘arts’, ‘griffier’ en ‘verpleegkundige’. Deze laatste groep woorden, die in theorie gender neutraal zouden moeten zijn, zijn de woorden waar de stereotyperende gender bias in te vinden zou moeten zijn. Bij deze woorden is er namelijk een mannelijke of een vrouwelijke associatie aanwezig, ondanks dat zowel mannen als vrouwen deze beroepen en functies uitoefenen. Deze gender bias is bij deze woorden dan ook vermoedelijk terug te vinden in de word embeddings.

Het onderzoek

Om een antwoord te vinden op de onderzoeksvraag *“In hoeverre is er gender bias op basis van beroepsnamen aanwezig in Nederlandse word embeddings”*, wordt er in dit onderzoek gekeken naar de op Nederlandse corpora getrainde word embeddings van Tulkens, Emmercy, & Daelemans, 2016. De Nederlandse corpora zullen onder andere Wikipedia pagina’s bevatten, maar ook nieuwsberichten, websites en artikelen. De word embeddings zijn getraind middels modellen van *word2vec*. Word2Vec is een verzameling van modellen die 2-laagse neurale netwerken gebruiken om word embeddings te produceren. Om de daadwerkelijke gender bias te analyseren zal in dit onderzoek gebruik worden gemaakt van de wiskundige eigenschappen van vectoren zoals de cosinusgelijkenis. Er zal voortgebouwd worden op de methoden die beschreven staan in Bolukbasi, Chang, Zou, Saligrama, & Kalai, 2016a, waarmee een *‘gender direction’* wordt gegenereerd die het stereotyperende verschil tussen de twee genders en daarmee een gender bias representeert. Hiermee worden vervolgens de in theorie genderneutrale beroepsnamen met behulp van de cosinusgelijkenis vergeleken, om zo te bepalen in hoeverre deze woorden overeenkomen met onze representatie van gender bias. Vervolgens kan de gehele set van resultaten vergeleken worden met die van al uitgevoerde onderzoeken en analyses van word embeddings in de Engelse taal om zo de uitkomsten van dit onderzoek in een breder kader te plaatsen.

2 Gerelateerd werk en achtergrond

Word embeddings

Word embeddings zijn woorden die gerepresenteerd zijn in een vectorruimte. Elk woord wordt gerepresenteerd door een vector die geleerd is uit een groot corpus op basis van co-occurrence. Het idee hierbij is dat woorden in vergelijkbare context een vergelijkbare betekenis hebben (Miller and Charles, 1991). Er zijn verschillende trainingsmethoden, hoewel Word2Vec misschien wel de meest bekende is (Mikolov, Sutskever, Chen, Corrado, Dean, 2013). Om word embeddings te creëren wordt Word2Vec gebruikt. Word2Vec is een verzameling van modellen van neurale netwerken met 1-hidden layer. Binnen deze modellen zijn de belangrijkste de Skip-Gram (SG) en Continuous-Bag-of-Words (CBOW) modellen, waarbij de eerste de context probeert te voorspellen gegeven een woord en waarbij de tweede juist probeert het woord te voorspellen gegeven de context. De waarden van de vector worden hierbij als waarschijnlijkheden gebruikt om vervolgens daaruit semantische betekenis te halen. De uitkomst van de word2vec methoden vormt een tabel met word embeddings, waarin elk woord gerepresenteerd is door een vector. Met de wiskundige eigenschappen van vectoren kunnen dan relaties worden aangetoond. Hierbij is de cosinusgelijkenis een methode om de gelijkheid van woorden en vectoren in de vectorruimte te meten (Schütze, 1998). Bolukbasi et al, 2016a ontdekten echter dat die cosinusgelijkenis niet alleen de semantische betekenis en relaties aantoonde, maar ook de gender bias die in de data aanwezig was.

Gender bias in Nederlandse taal

Gender bias in taal wordt al decennia lang bestudeerd en onderzocht in allerlei contexten. Eén van de belangrijkste resultaten voor dit onderzoek is die van Greenwald, McGhee, en Schwartz, 1998. Zij vonden via de *'Implicit Association Tests'* bij mensen een gender bias waarvan ze zich soms niet eens bewust waren. Een voorbeeld hiervan is het volgende welbekende scenario en raadsel:

> *"Een vader en zijn zoon krijgen een ernstig auto ongeluk. De vader overlijdt ter plekke, zijn zoon wordt naar het ziekenhuis gebracht en de dienstdoende arts staat aan het bed van die jongen en zegt: "ik kan deze jongen niet behandelen, hij is mijn zoon!"*

Veel mensen kijken bij dit raadsel even raar op, omdat een arts vaak als mannelijk wordt geassocieerd. Dit wordt tevens bevestigd door Nosek, Banaji, Greenwald, 2002 die ontdekten dat vrouwen met kunst en familie gelinkt worden en mannen met wetenschap en carrière. Bovendien worden woorden zoals 'directeur' standaard geassocieerd met de dominante groep, terwijl de vrouwelijke versie 'directrice' slechts van het mannelijke is afgeleid en niet als hoofdterm wordt gezien (Jakobson, Waugh, Monville-Burston, 1990). Daarbovenop, zo wijzen Kay, Matuszek, and Munson, 2015; Wagner, Garcia, Jadidi, and Strohmaier, 2015; en Ross en Carter, 2011 erop, zit de gender bias ook in online context zoals nieuwsberichten, web search en wikipedia.

Als laatste moet opgemerkt worden dat voorafgaand aan de feministische beweging in Nederland het grootste deel van de beroepen alleen naar mannen verwees. Tijdens en na de feministische beweging zijn daarvoor ook vrouwelijke versies gecreëerd (Kool-Smit, 1967). Het gevolg hiervan is dat in het Nederlands er een neiging is om van iemand die een beroep uitoefent zijn/haar sexe aan te geven (Romein-Verschoor, 1975). Hierdoor hebben heel veel beroepen zowel een mannelijke als een

vrouwelijke vorm. Desondanks liet Romein-Verschoor zien dat bij een neutrale term als arts, mensen toch een man in gedachten hadden.

Door de jaren heen is dit tot het volgende gender-onderscheid uitgegroeid (Taaladvies van de Taalunie, 2018):

1. Binaire beroepsnamen met een exclusieve mannelijke variant (boer-boerin)
2. Binaire beroepsnamen met een inclusieve mannelijke variant (directeur-directrice)
3. Niet binaire mannelijke beroepsnamen (ambtenaar, hoogleraar, minister)
4. Niet binaire vrouwelijke beroepsnamen (naaister, vroedvrouw)
5. Genderoverkoepelende beroepsnamen (leerkracht, hoofd, leidinggevende)

Hierbij moet opgemerkt worden dat bij drie en vier soms wel degelijk een tegenhanger van ofwel het mannelijke ofwel het vrouwelijke bestaat, maar deze is verouderd, wordt niet meer gebruikt of wordt simpelweg niet erkend. Zo bestaat de term ambtenares, maar wordt deze bijna nooit (meer) gebruikt en dit maakt het onderscheid tussen twee en drie niet altijd even scherp.

Analyse van gender bias

Er zijn inmiddels verschillende methoden om stereotypering en bias in word embeddings te analyseren. Een belangrijke basis hiervan is de '*Word Embedding Association Test'* (WEAT), een statistische test gecreëerd door Caliskan, Bryson en Narayaan, 2017. Deze maakt gebruik van twee sets van genderneutrale *'target words'*, waarbij elke set vermoedelijk een specifieke gender bias heeft (programmeur en loodgieter tegenover verpleegkundige en verloskundige). De word embeddings uit deze sets worden dan vergeleken met de word embeddings uit twee '*attribute sets'*: sets die per definitie een bias hebben (man, vader en presentator tegenover vrouw, moeder en presentatrice). Deze vergelijking gebeurt op basis van de cosinusgelijkenis. Vervolgens kan met behulp van hypothese tests de bias in de word embeddings gedetecteerd worden. Een andere veelgebruikte manier om bias te detecteren is de methode die wordt beschreven door Bolukbasi, Chang, Zou, Saligrama, & Kalai, 2016a. Hierin wordt een '*gender direction'* gecreëerd met de woorden die per definitie een bias hebben. Deze gender direction is dan een vector die in grote lijnen de gender bias in de word embeddings vastlegt. Met slechts één enkele genderneutrale dataset kan dan het gemiddelde over de cosinusgelijkenis van deze gender direction met alle genderneutrale woorden uit de set berekend worden. De uitkomst hiervan geeft dan de directe gender bias aan van de gebruikte word embeddings.

3 Data en Methoden

Data

Word embeddings

In dit onderzoek worden twee datasets met word embeddings gebruikt van Tulkens, Emmery en Daelemans, 2016 (https://github.com/clips/dutchembeddings). De eerste set, waarnaar verwezen zal worden als de gecombineerde set, is een set die getraind is op 803 miljoen woorden uit een combinatie van het Roularta corpus (corpus van Belgische wetenschappelijke artikelen), een getokeniseerde Wikipedia dump en het SoNaR corpus (corpus van o.a. nieuwsberichten, boeken, magazines en kranten). De tweede set, waarnaar verwezen zal worden als de COW set, is de dataset die getraind is op het COW corpus, dat automatisch gegenereerd is uit .be en .nl top level domeinen en 4 miljard woorden bevat. Op de data is bovendien een pre-processing gedaan om spellingsverschillen tussen verschillende dialecten eruit te halen, contextarme zinnen te verwijderen en grammaticaal foutieve meervouden te corrigeren. Met deze twee datasets hopen we een een relatie te kunnen zien tussen de structuur en grootte van de set en de gevonden gender bias.

Uit beide datasets zijn respectievelijk 1442950 en 3110718 word embeddings gecreëerd. De word embeddings van beide sets zijn 320-dimensionaal en gecreëerd met het Skip-Gram model van word2vec. Er wordt in dit onderzoek gebruik gemaakt van het Skip-Gram model, omdat dit model goed werkt in combinatie met zeldzame woorden. Dat is in dit onderzoek nodig, omdat er vrouwelijke versies van beroepsnamen zijn die niet veel voorkomen. Het mannelijke is in veel gevallen namelijk inclusief en wordt ook voor vrouwen gebruikt.

Woordensets

Voor dit onderzoek zijn als laatste nog drie woordensets gegenereerd. De eerste set is een set van woordparen die per definitie genderspecifiek zijn zoals broer-zus of jongen-meisje (te vinden in Appendix A). Hiervoor is dezelfde set van 24 woordparen als Tulkens et al, 2016 gebruikt met slechts één kleine wijziging: het woordpaar secretaris-secretaresse is vervangen door boer-boerin. De reden voor deze wijziging is het feit dat secretaresse niet de vrouwelijke versie is van secretaris. Dit is een in het Nederlands veel gemaakte fout, maar de functie van secretaresse is een andere dan die van secretaris. Deze woordparen zullen gebruikt worden om een gender direction te creëren.

De tweede woordenset die voor dit onderzoek nodig is, is een lijst met in theorie genderneutrale beroepen (te vinden in Appendix B). Deze lijst bestaat uit 55 beroepen waar niet zowel een mannelijke als een vrouwelijke term voor is en deze zouden dus genderneutraal moeten zijn. Over deze lijst zal de directe bias berekend worden.

De derde woordenset is een lijst van 48 Nederlandse stopwoorden (te vinden in Appendix E). Deze lijst zal alleen gebruikt worden om te controleren of de creatie van de gender direction wel juist is.

Methoden

Creëren van de gender direction

Om de directe gender bias te testen zal dit onderzoek gebruik maken van de door Bolukbasi et al, 2016a beschreven methoden. Als eerste wordt daarvoor een gender direction gecreëerd. Dit kan met de set van 24 Nederlandse gender paren die al eerder gecreëerd en geselecteerd zijn door Tulkens et al,

2016. Deze zijn per definitie gender biased en zullen bestaan uit soortgelijke woordparen zoals (zoon-dochter). Van elk paar wordt vervolgens het verschil van de word embeddings genomen en uit de combinatie van alle paren worden de *'prinicpal components'* (PC) berekend. Deze PCs zijn vectoren van de projecties van datapunten (=word embeddings) op de '*principal axes'* in de data. Deze principal axes zijn als het ware de assen van de belangrijkste elementen uit de vectoren. Een van die elementen zal het genderverschil zijn tussen de woordparen. Uit die projectie volgt een richting die het merendeel van de variantie van de gender paar vectoren beschrijft. Bovendien, zo wordt in Bolukbasi et al, 2016a beschreven, kan de eerste PC vermoedelijk gebruikt worden om de variantie te verklaren, mits het verschil tussen de eerste PC en de rest groot genoeg is. Deze eerste en hoogste PC zal in dat geval de beoogde gender direction zijn. Om dit te onderzoeken wordt er als controle experiment ook vergeleken met een set van stopwoorden. De verdeling van de varianties over de PC's van deze lijst stopwoorden zou geleidelijk aflopend moeten zijn zonder die hoge eerste PC die wel te verwachten is bij de woordparen set.

Selectie van genderneutrale beroepsnamen

Om de directe bias over Nederlandse beroepsnamen te kunnen bepalen is een lijst genderneutrale beroepen nodig. Het creëren van deze lijst is een bijzonder lastige opgave, omdat er geen eenduidig antwoord bestaat op de vraag of een beroepsnaam genderneutraal is.

Om dit probleem duidelijk te illustreren wordt er teruggegrepen naar het raadsel van de verongelukte vader en zoon. In het oorspronkelijke raadsel wordt de zoon de operatiekamer binnengereden en staat daar een chirurg in plaats van een arts. En hoewel de gender bias bij chirurg nog sterker is, is er een term voor een vrouwelijke chirurg: 'chirurge'. Ondanks dat het discutabel is of dit een correct en gangbaar woord is uit de Nederlandse taal, wordt dit woord wel gebruikt. Dat is de reden dat chirurg in het raadsel vervangen was door arts

Het gebruik van een (al dan niet erkende) vrouwelijke variant van de beroepsnaam is simpelweg niet inzichtelijk of te voorspellen. Zeker niet als over het algemeen de inclusieve mannelijke versie wordt gebruikt. Om te controleren of de vrouwelijke versie officieel bestaat is tijdens dit onderzoek gebruik gemaakt van woordenlijst.org; een website met een door de Nederlandse Taalunie geactualiseerde '*Woordenlijst Nederlandse Taal*'. De vraag blijft echter of die niet-binaire en inclusieve woorden zoals 'ambtenaar' genderneutraal (moeten) zijn als ook het vrouwelijke maar niet erkende 'ambtenares' zo nu en dan ook gewoon gebruikt wordt.

Omdat er op deze vraag geen eenduidig antwoord is en om te voorkomen dat de resultaten vertekend zouden worden door vrouwelijke termen die blijkbaar wel in de data voorkwamen, is er in dit onderzoek gewerkt met de (beperkte) lijst van niet binaire beroepsnamen waarbij een vrouwelijke variant niet te vinden is. Veel woorden waarvan de vrouwelijke versie officieel niet erkend of juist is, zoals architect, bakker en schipper, konden niet gebruikt worden, omdat de vrouwelijke term hiervan wel degelijk voorkomt (architecte, bakkerin en schipperse). Het aantal beroepen dat overbleef was gering en slechts een klein deel daarvan leek een vrouwelijke associatie te hebben.

Directe Bias berekenen

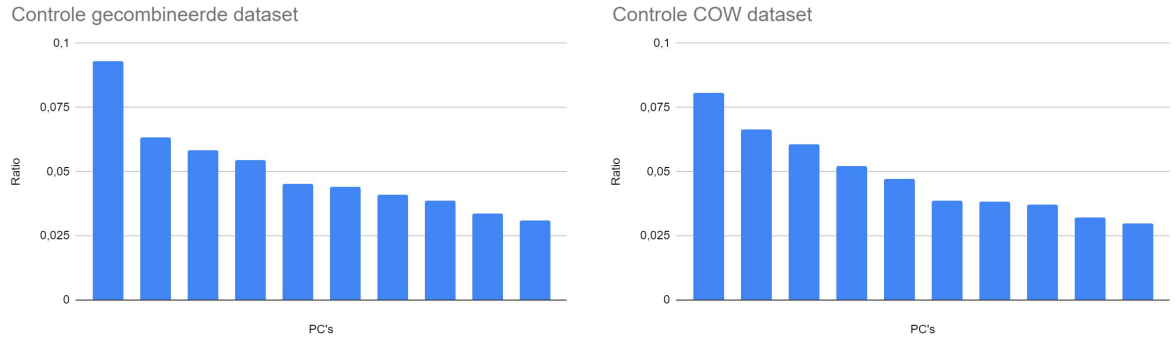
Met deze genderneutrale woorden (N) en de eerder gedefinieerde gender direction (g) kan voor elke woord embedding (w) de directe bias uit Bolukbasi et al, 2016a berekend worden met de formule: $\frac{1}{|N|}\sum_{w\in N}|cos(w,g)|$. Hierin is $cos(x,y)$ de cosinusgelijkenis tussen twee vectoren x en y gedefinieerd als: $\frac{x\cdot y}{||x||\,||y||}$. Met de formule voor directe bias wordt het gemiddelde genomen over de cosinusgelijkenis tussen de gender direction en elk genderneutraal woord. Het geeft aan in hoeverre een woord 'lijkt' op de gender direction en daarmee dus gender biased is.

4 Resultaten

Bovenstaande methoden zijn toegepast op beide datasets. Om inzicht te krijgen in de resultaten zullen beide sets in deze sectie eerst apart besproken worden om vervolgens de overeenkomsten en verschillen op een rijtje te zetten. De detectie van de gender richting is gedaan zoals eerder omschreven: de richting van het verschil tussen de mannelijke en vrouwelijke woordparen wordt berekend om vervolgens de principal components (PC) daarvan te berekenen. Bij beide wordt gekeken naar de hypothese van Bolukbasi et al, 2016a dat de eerste PC (die de meeste variantie verklaart) veel groter is dan de andere die een stuk lager zijn en geleidelijk afnemen. Hierdoor zou de eerste PC als het ware de gender direction zijn en kan deze gebruikt worden voor de daadwerkelijke berekening van de directe bias. Om te kijken of deze aanname juist is wordt het controle experiment uitgevoerd: de variantie verdeling wordt nog gecontroleerd met die van een set stopwoorden. Hierna volgen van beide datasets enkele resultaten van de cosinusgelijkenis om meer inzicht te geven in de data. Hierna volgen de resultaten van de directe bias berekening.

De tussenresultaten van de cosinusgelijkenis tussen elk woord en de gender richting staan in Appendix C en hier zal regelmatig naar worden verwezen om verschillen, overeenkomsten en bijzonderheden aan te duiden. Hieruit valt vooraf al op te merken dat de mannelijke bias negatief lijkt te zijn en de vrouwelijke bias positief. Dit blijkt vooral heel sterk uit de woorden met een vrouwelijke bias zoals verpleegkundige, verloskundige en verzorgende die alle drie sterk positief zijn. Bovendien moet opgemerkt worden dat er weinig woorden zijn met positieve waarden, wat de eerder beschreven verwachting dat er weinig woorden met een vrouwelijke associatie waren, lijkt te bevestigen.

Figuur 1. De diagrammen van de verdeling van de varianties van gender directions van de gecombineerde dataset (links) en de COW dataset (rechts)



Figuur 2. De diagrammen van de verdeling van de varianties van de sanitycheck woordenlijst van de gecombineerde dataset (links) en de COW dataset (rechts)

Gecombineerde dataset

Figuur 1 laat duidelijk zien dat er in de gecombineerde dataset net als bij Bolukbasi et al, 2016a een groot verschil in variantie zit tussen de eerste PC en de anderen. Daarmee verklaart de eerste PC zoals verwacht significant een stuk meer van de variantie. Dit klopt ook als er een controle experiment wordt gedaan op de lijst met stopwoorden. In figuur 2 is duidelijk te zien dat bij een lijst willekeurige stopwoorden de eerste PC niet zo'n groot verschil heeft met de rest van de PC's. Bovendien is de waarde van de eerste PC ook zodanig laag dat hij eigenlijk overeenkomt met de tweede PC van onze gender paren set. De assumptie dat de hoogste PC gebruikt kon worden voor de gender richting was dus juist. Wel moet opgemerkt worden dat het verschil niet zo groot is als bij Bolukbasi et al, 2016a, maar wel significant groot genoeg en tevens gelijkend op soortgelijke onderzoeken zoals die van Basta, Costa-jussà, en Casas (2019), die met dezelfde Engelse word embeddings als Bolukbasi et al, 2016a op een waarde van 0.3 uitkwamen bij de hoogste PC.

Zoals eerder genoemd lijkt de gender bias (de uitkomst van de cosinusgelijkenis) per woord negatief voor een mannelijke bias en positief voor een vrouwelijke bias. De woorden met de sterkste mannelijke bias waren in deze set: ingenieur (-0.25), executeur (-0.20) en magistraat (-0.20). De drie woorden met de sterkste vrouwelijke bias waren: verzorgende (0.11), verloskundige (0.092) en verpleegkundige (0.091). De gemiddelde sterkte van de bias (deze waarde is absoluut) was in deze set 0.093 en deze had een standaarddeviatie van 0.056. Deze vrij hoge standaarddeviatie komt met name door de woorden met een sterke mannelijke bias, maar ook door de woorden waar geen bias lijkt te bestaan: dierenarts (-0.0053), butler (-0.0077) en conciërge (-0.0127).

De berekening van de directe bias over deze waarden gaf een waarde van 0.092, die, net als bij Bolukbasi et al, 2016a waar een waarde van 0.08 gevonden werd op de Engelstalige w2vNEWS dataset, bevestigt dat veel beroepen een substantiële component langs die gender richting hebben en er mag bij deze word embeddings dus worden gesproken van gender-bias.

COW dataset

Ook de COW dataset laat met figuur 1 duidelijk zien dat de hoogste PC significant meer variantie verklaart dan de andere. In figuur 2 is bovendien te zien dat van de lijst met stopwoorden het verschil tussen de hoogste PC en de rest niet veel groter is dan tussen de andere PC's. Ook is te zien dat de hoogste PC zelfs nog lager is dan de tweede PC bij onze gender paren set. Wel is in figuur 1 te zien dat het verschil met de andere PC's iets kleiner is dan bij de gecombineerde dataset. Maar aangezien de hoogste PC bijna even groot is als hierboven kan inderdaad worden aangenomen dat ook deze gezien kan worden als de gender richting.

In deze dataset zijn de woorden met de sterkste mannelijke bias de volgende: mecanicien (-0.21), smid (-0.21) en metselaar (-0.20). De woorden met de sterkste vrouwelijke bias waren dezelfde als bij de gecombineerde set, maar ze hadden in deze set een sterkere bias: verloskundige (0.16), verzorgende (0.12) en verpleegkundige (0.11). De gemiddelde sterkte lag bovendien ook iets hoger en had een waarde van 0.10. De standaarddeviatie verschilde niet veel en kwam uit op 0.056. Verrassend genoeg zijn de waarden van de woorden waar nauwelijks een bias lijkt te bestaan positief: orthodontist (0.0011), ober (0,0053) en huisarts (0.0096).

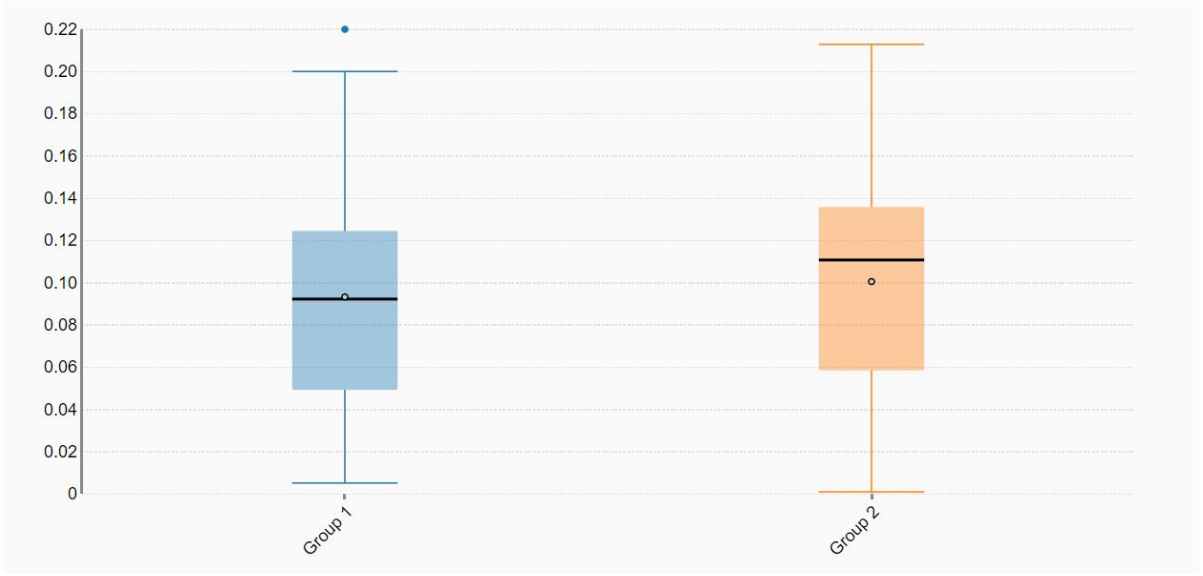
De directe bias die berekend is op de word embeddings uit de COW set heeft een waarde van 0.099. Dit bevestigt, net als bij de gecombineerde set, dat binnen veel beroepen er een gender richting is en dat maakt de word embeddings van deze beroepen inderdaad gender biased.

Verschillen en overeenkomsten

Kijkend naar verschillen en overeenkomsten tussen beide datasets vallen als eerste de woorden op die in de ene dataset een negatieve waarde hebben en in de andere dataset een positieve. Het gaat hierbij om de woorden dierenarts, huisarts, orthodontist, sensor en tolk. Hierin zijn met name huisarts en tolk interessant, de rest van de woorden hebben in beide datasets een vrij lage bias (< 0.03) en dus is het niet heel bijzonder dat ze een andere bias hebben in de datasets. Huisarts en Tolk zijn echter in de gecombineerde dataset redelijk sterk met de respectievelijke waarden -0.060 en -0.065. In de tweede set hebben ze ineens een vrouwelijke bias, maar deze bias is erg laag: 0.0096 en 0.013. Maar ook woorden die in beide sets dezelfde bias hebben vallen op door het verschil in sterkte. Het gaat hier om woorden zoals smid, een woord dat in de gecombineerde set een waarde van -0.10 had en in de tweede set een waarde die 2x zo sterk is: -0.21.

Behalve deze speciale gevallen en outliers komt het merendeel van de gevonden waarden uit beide datasets redelijk overeen. Zeker als gekeken wordt naar de boxplots van beide datasets in figuur 3 (waarbij de absolute waarden genomen zijn, zodat er hier gekeken kan worden naar de sterkte van de bias). Hierin is te zien dat, hoewel de data uit de tweede set iets meer varieert en de eerste set een outlier heeft, de verdeling van de data redelijk gelijk is. Gezien ook de directe bias niet zoveel verschilt, lijkt er is dus geen significant verschil tussen beide datasets te zijn.

Data Summary								
Groups	N	Min	Q_1	Median	Q_3	Max	Mean	SD
Group 1	55	0	0.0494	0.0923	0.1246	0.2501	0.0933	0.0567
Group 2	55	0	0.0586	0.1108	0.1359	0.2129	0.1006	0.0558



Figuur 3. De boxplots van absolute waarden van de gender biases van elk woord uit de gecombineerde set (links / Group 1) en de COW set (rechts / Group 2)

5 Discussie

De data, methode en resultaten

Het eerste discussiepunt dat aangekaart moet worden heeft te maken met de grootte van de sets van woordparen en genderneutrale beroepen. Deze sets zijn wegens het probleem van genderneutrale beroepen in het Nederlands vrij klein gebleven, hoewel de woordparen set bij Bolukbasi et al, 2016 en Tulkens et al, 2016 niet groter was (respectievelijk tien en 24). De genderneutrale beroepen set bleek tijdens het opstellen vrij problematisch, maar hierbij is gekozen voor het eerder genoemde strikte onderscheid dat bij het voorkomen van een vrouwelijke term, de in theorie genderneutrale term uitgesloten wordt. Bij een minder strikt onderscheid had iets van een regel van het aantal keer voorkomen van die vrouwelijke term gedefinieerd kunnen en moeten worden, maar gezien dit neer zou komen op een eigen intuïtief en willekeurig getal, is hier niet aan begonnen. Dit resulteerde dus in de set van 55 genderneutrale beroepsnamen waarvan er slechts 5 een vrouwelijke bias bleken te hebben. Dit maakt de data ietwat eenzijdig en daarom mogelijk minder geschikt om het geheel van gender bias te onderzoeken. Met de in dit onderzoek gebruikte data lag de focus namelijk op mannelijke gender bias.

De methoden die in dit onderzoek gebruikt zijn, vormen het tweede discussiepunt. Er is gekozen voor die van Bolukbasi et al, 2016. Niet alleen zijn deze methoden sterk en in de literatuur erkend, ze waren bovendien ideaal voor het bovengenoemde probleem. De eerder beschreven WEAT methode van Caliskan et al, 2017 is net zo sterk en erkend, maar hierbij zouden er in plaats van een enkele genderneutrale set, twee sets genderneutrale woorden nodig zijn met in elke set een gender bias. Hierbij zou de vrouwelijke gender bias set in de Nederlandse taal een probleem zijn, omdat van de weinige genderneutrale woorden die geaccepteerd waren er slechts enkele een vrouwelijke gender bias hebben. De set zou vervolgens te klein zijn om er een correcte analyse mee te doen.

Behalve dat de beperkte beroepen set vermoedelijk een invloed heeft gehad op de resultaten, moet ook nog opgemerkt worden dat het verschil in variantie in dit onderzoek minder groot was dan in het origineel van Bolukbasi et al, 2016. Desondanks is hij wel groot genoeg en groter dan in enkele andere onderzoeken, zoals die van Basta et al, 2019. Het verschil tussen beide datasets is niet significant genoeg om te bepalen of het gebruik van meer en gevarieerde data een effect heeft op de resultaten. Wel zijn er in beide datasets resultaten gevonden die op elkaar lijken met een vrij hoge directe bias van 0.092 en 0.099 voor de gecombineerde en de COW dataset. Deze waarden bevestigen dat in de Nederlandse word embeddings wel degelijk een gender bias op basis van beroepsnamen zit.

Vervolgonderzoek

Gezien er positieve resultaten zijn gevonden ondanks de hierboven genoemde discussiepunten is er nog meer onderzoek nodig naar gender bias in de Nederlandse taal om dit te controleren en verder in kaart te brengen. Hierbij zou de scope verbreedt kunnen worden om ook andere (in theorie) genderneutrale woorden toe te voegen. Denk hierbij aan bijvoorbeeld objecten, gereedschappen of karakter eigenschappen. Dit is met Engelse word embeddings al wel gedaan in het merendeel van de besproken onderzoeken, maar nog niet (genoeg) met de Nederlandse word embeddings. Het bijkomend voordeel hierbij is dat de beperkte set van genderneutrale beroepen dan ineens uitgebreid kan worden met meer woorden zoals koekenpan, schroevendraaier, hark en naald. Dit zal een

betrouwbaarder resultaat geven en bovendien kan uit de nieuwe veel grotere dataset dan vermoedelijk wel twee losse datasets gecreëerd worden, waarbij de ene dataset een vrouwelijke bias heeft en de andere een mannelijke. Hierop kan dan wel de eerder besproken WEAT methode van Caliskan et al, 2017 toegepast worden om zo ook op een andere manier de gender bias te analyseren.

Ook kan in vervolgonderzoek gekeken worden naar hoe we deze Nederlandse word embeddings van hun bias ontdoen. Bolukbasi et al, 2016 zijn hierin sterk begonnen en er zijn verschillende 'genderneutrale' methoden ontwikkeld, zoals die van Zhao et al, 2018. Wel moet daarbij opgemerkt worden dat het verwijderen van die bias misschien niet eens mogelijk is volgens Gonen en Goldberg, 2019. Zij vonden dat we met veel van die methoden de bias slechts verhullen, in plaats van verwijderen. Bovendien zouden de analysemethoden voor bias onvoldoende zijn en andere meer diepgewortelde aspecten van gender bias niet aan het licht brengen.

Ondanks dat NLP inmiddels overal ter wereld al gebruikt wordt voor verschillende analyses en applicaties, is er dus nog een hoop te doen om bias, wat niet alleen binnen NLP een probleem is, maar ook een maatschappelijke kant heeft, in o.a. Nederlandse word embeddings te analyseren en in kaart te brengen.

6 Appendices

Appendix A: Woordparen

Hieronder staan de 24 woordparen uit Tulkens et al, 2016 die gebruikt zijn voor het definiëren van de gender direction. Hierbij is het paar secretaris-secretaresse vervangen door het paar boer-boerin.

oom-tante	jongen-meisje	broer-zus
broers-zussen	vader-moeder	papa-mama
grootvader-grootmoeder	opa-oma	kleinzoon-kleindochter
bruidegom-bruid	hij-zij	zijn-haar
man-vrouw	koning-koningin	neef-nicht
prins-prinses	zoon-dochter	leraar-lerares
vriend-vriendin	kapper-kapster	directeur-directrice
presentator-presentatrice	boer-boerin	atleet-atlete

Appendix B: Beroepen

Hieronder staan de 55 genderneutrale beroepen die gebruikt zijn om de directe bias uit te rekenen.

accountant	auteur	baggeraar	bemanning
beul	butler	cipier	coach
conciërge	deurwaarder	developer	dierenarts
elektricien	engineer	entrepreneur	etaleur
executeur	goochelaar	griffier	hovenier
huisarts	imker	ingenieur	jongleur
kelner	loodgieter	magistraat	manager
matroos	mecanicien	metselaar	militair
model	monteur	notaris	ober
orthodontist	plaatwerker	postbode	premier
professor	programmeur	rechter	ruiter
schipper	sensor	slager	smid
tandarts	tolk	verloskundige	verpleegkundige
verzorgende	visser	wachter	

Appendix C: Cosinusgelijkenissen van de beroepsnamen

Hieronder staan de waarden van de cosinusgelijkenis tussen de beroepsnamen en de gender direction, de deviatie en de directe bias voor beide datasets (de tabel gaat door op de volgende pagina):

Words	Gecombineerd	COW
accountant	-0,05694637773	-0,1154570127
auteur	-0,1260782382	-0,13081597
baggeraar	-0,06074804548	-0,05532683275
bemanning	-0,01749287514	-0,03166255349
beul	-0,1066295395	-0,1337109147
butler	-0,007691830283	-0,01408312818
cipier	0,03262606745	-0,06938984203
coach	-0,1092780208	-0,1560859189
conciërge	-0,01266166777	-0,03720711387
deurwaarder	-0,05113971171	-0,0519568693
developer	-0,02726934027	-0,1017726071
dierenarts	**-0,00527277175**	**0,01559008598**
elektricien	-0,1084379607	-0,1298357577
engineer	-0,1244159308	-0,136791882
entrepreneur	-0,09049040003	-0,1316067754
etaleur	-0,0641502139	-0,06504802868
executeur	-0,2000546198	-0,125869112
goochelaar	-0,1122209657	-0,0362641278
griffier	-0,139282072	-0,09344436489
hovenier	-0,1247038091	-0,09731165588
huisarts	-0,05966317568	0,009576838095
imker	-0,0432793425	-0,06906986209
ingenieur	-0,2500569623	-0,1675240612
jongleur	-0,112374775	-0,1142296945
kelner	-0,0560219452	-0,06187399965
loodgieter	-0,02123855042	-0,1071095643
magistraat	-0,1977417868	-0,1293980384
manager	-0,1371664282	-0,1349930309
matroos	-0,04459258276	-0,1198054648
mecanicien	-0,1387788855	-0,2128876363

metselaar	-0,1807457194	-0,1966652054
militair	-0,1890698926	-0,109062718
model	0,06550010092	0,06674656042
monteur	-0,1749252262	-0,1179088326
notaris	-0,1316522139	-0,1189770521
ober	0,04761299038	0,00525101198
orthodontist	**-0,02743514678**	**0,001075510868**
plaatwerker	-0,0952858554	-0,1571015965
postbode	-0,1095501836	-0,02306047869
premier	-0,1635302208	-0,100446224
professor	-0,1750149562	-0,1590940399
programmeur	-0,1199668614	-0,1773359022
rechter	-0,1211467956	-0,07982714186
ruiter	-0,09181231731	-0,1652792749
schipper	-0,05218926906	-0,1600276731
sensor	**0,02102136965**	**-0,02733645916**
slager	-0,1228648321	-0,1384541932
smid	-0,1049710129	-0,20525766
tandarts	-0,03263276996	-0,02314181257
tolk	-0,06537602426	0,01319163355
verloskundige	0,09234159902	0,1595661552
verpleegkundige	0,09105358609	0,1108376674
verzorgende	0,1103643506	0,1193776535
visser	-0,08499234225	-0,1456043919
wachter	-0,02386525229	-0,09748223087
gemiddelde over absolute waarden	0,0933350142	0,1006146877
standaarddeviatie	0,05615871879	0,05526674958
directe bias	0,09160967996	0,09930827926

De volgende kleurmarkeringen zijn aangebracht in de tabel:
woord: woorden die in beide datasets een positieve waarde hebben (vrouwelijke bias)
woord: woorden die in beide datasets een sterke negatieve waarde (< -0.1) hebben (mannelijke bias)
woord: woorden die zowel een negatieve als een positieve waarde hebben
waarde: woorden die in beide datasets een lage absolute waarde hebben (< 0.03)

Appendix D: Code

Hieronder staan de python code en methoden die gemaakt en gebruikt zijn voor dit onderzoek. Deze code en methoden zijn gebruikt met de bovenstaande data.

De modellen laden met gensim

```
import gensim
from gensim.models import word2vec
wv = gensim.models.KeyedVectors.load_word2vec_format("data/320/combined-320.txt",
binary=False)
wv2 = gensim.models.KeyedVectors.load_word2vec_format("data/big/cow-big.txt",
binary=False)
```

PCA berekenen (van Bolukbasi et al, 2016a)

```
#from Bolukbasi, Chang, Zou, Saligrama, and Kalai (2016)
def doPCA(pairs, embedding, num_components = 24):
    matrix = []
    for a, b in pairs:
        center = (embedding[a] + embedding[b])/2
        matrix.append(embedding[a] - center)
        matrix.append(embedding[b] - center)
    matrix = np.array(matrix)
    pca = PCA(n_components = num_components)
    pca.fit(matrix)
    # bar(range(num_components), pca.explained_variance_ratio_)
    return pca
```

Direct Gender bias berekenen

```
#calculated using numpy
def directBias(genderNeutralWords, genderdirection, embedding):
    i=0
    sum=0
    for w in genderNeutralWords:
              sum = sum + abs(dot(embedding[w],
              genderdirection)/(norm(embedding[w])*norm(genderdirection)))
              i=i+1
    return(sum/i)
```

Appendix E: Stopwoorden

Hieronder staat de woordenlijst van stopwoorden die gebruikt is om de controle mee uit te voeren op de woordparen set.

de	het	een	die
deze	dit	alles	anders
omdat	maar	terwijl	ook
en	of	daarom	tegen
voor	hoe	waarom	want
wie	wat	wanneer	elk
nu	toen	iemand	iets
niets	niemand	dus	dat
zou	zelfs	zo	welke
weinig	veel	met	zonder
liever	minder	misschien	waarschijnlijk
nog	ooit	nooit	even

Referenties

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016a). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in neural information processing systems*, 4349-4357.

Bolukbasi, T., Chang, K. W., Zou, J., Saligrama, V., & Kalai, A. (2016b). Quantifying and reducing stereotypes in word embeddings. *Proceedings of the Workshop on #Data4Good: Machine Learning in Social Good Applications @ICML 2016*, 41-45.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), E3635–E3644. https://doi.org/10.1073/pnas.1720347115

Gerritsen, M. (2002). Towards a more gender-fair usage in Netherlands Dutch. In *Gender Across Languages*, 81-109.

Gonen, H., & Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, *1*, 609–614. https://doi.org/10.18653/v1/n19-1061

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning word vectors for 157 languages. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (LREC 2018), 3483-3487.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, 74(6), 1464–1480. doi:10.1037/0022-3514.74.6.1464

Jakobson, R. & Waugh, L. R. & Monville-Burston, M. (1990). *On language.* Cambridge, Mass: Harvard University Press. ISBN: 978-0674635364

Kay, M., Matuszek, C., & Munson, S. A. (2015). Unequal representation and gender stereotypes in image search results for occupations. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15,* 3819-3828. doi:10.1145/2702123.2702520

Kool-Smit, J. E. (1967). Het onbehagen bij de vrouw. *De Gids 130* (9/10), 267-282.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28(2), 203–208. doi:10.3758/bf03204766

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems 26 (NIPS 2013),* 3111-3119.

Mikolov, T., Yih, W. T., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. *In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* 746-751.

Miller, G. A., & Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1), 1–28. doi:10.1080/01690969108406936

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice,* 6(1), 101–115. doi:10.1037/1089-2699.6.1.101

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532-1543. doi:10.3115/v1/d14-1162

Romein-Verschoor, A. 1(975). Over taal en seks, seksisme en emancipatie. *De Gids 138* (1/2), 3-36.

Ross, K., & Carter, C. (2011). Women and news: A long and winding road. *Media, Culture & Society*, 33(8), 1148–1165. doi:10.1177/0163443711418272

Schütze, H. (1998). Automatic word sense discrimination. *Computational Linguistics,* 24(1), 97-123.

Taalunie, Vrouwelijke beroepsnamen, http://taaladvies.net, geraadpleegd op 22-01-2020

Taalunie, Woordenlijst Nederlandse Taal, http://woordenlijst.org, geraadpleegd op 30-01-2020

Tulkens, S., Emmery, C., & Daelemans, W. (2016). Evaluating unsupervised Dutch word embeddings as a linguistic resource. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 4130-4136.

Wagner, C., García, D., Jadidi, M., & Strohmaier, M. (2015). It's a man's wikipedia? Assessing gender inequality in an online encyclopedia. *Ninth International AAAI Conference on Web and Social Media (ICWSM.2015)*, 454-463.

Zhao, J., Wang, T., Yatskar, M., Cotterell, R., Ordonez, V., & Chang, K.-W. (2019). Gender bias in contextualized word embeddings. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* 1, 629-634. doi:10.18653/v1/n19-1064

Zhao, J., Zhou, Y., Li, Z., Wang, W., & Chang, K. W. (2018). Learning gender-neutral word embeddings. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP),* 4847-4853. doi:10.18653/v1/D18-1521