Troonredes door de jaren heen

N.T de Visscher
10667474

Bachelor thesis
Credits: 12 EC

Bachelor Opleiding Informatiekunde

University of Amsterdam
Faculty of Science
Science Park 904
1098 XH Amsterdam

Supervisor
Dr. M. J. Marx

ILPS, IvI
Faculty of Science
University of Amsterdam
Science Park 904
1098 XH Amsterdam

2016-06-20

Inhoudsopgave

Abstract

1 Inleiding

Dit onderzoek is een replicatie-onderzoek aan de hand van een eerder uitgevoerd onderzoek dat als paper is gepubliceerd. Het paper in kwestie beschrijft een onderzoek en de resultaten hiervan naar de "State of the Union" speeches uit Amerika van Rule et al. [2015]. In dat onderzoek werd gekeken naar wat er veranderde binnen de speeches sinds ze voor het eerst werden gehouden en hoe de speeches er hedendaags uitzien. Er werd gekeken naar de verschuiving in taalgebruik en inhoud in de speeches. Dit werd gedaan om een beeld te krijgen van wat de belangrijke onderwerpen in een speech waren en welke thema's behandeld werden over de jaren. Hieruit werd gekeken naar het moment dat de verschuiving naar het moderne Amerikaanse politieke bewustzijn plaatsvond. In het onderzoek werd gekeken naar de speeches van 1790 t/m 2014 welke jaarlijks uitgesproken worden door de president.
De State of the Union is niet de enige speech die in deze vorm al lange tijd jaarlijks wordt gehouden. Zo ook de Nederlandse troonredes. Deze troonredes worden sinds 1814 jaarlijks door de koning(in) voorgedragen. De troonredes kunnen gezien worden als een Nederlandse versie van de State of the Union. De troonredes bevatten echter ook nog vaak reflecties op het afgelopen jaar naast van een samenvatting over de staat van het land. De staat van het land wordt ook besproken maar dit is meer gericht op een blik op de positie van Nederland in de wereld. Dit is echter niet altijd het geval geweest. Zo was de inhoud vroeger anders dan nu, net als het taalgebruik. Tijdens het onderzoek zal er gekeken worden naar de veranderingen in de troonredes over de jaren. Het gaat hier vooral om de onderwerpen en thema's die behandeld worden. Dit zal onderzocht worden aan de hand van automatische tekstanalyse technieken. Dit om ervoor te zorgen dat er geen handwerk aan te pas komt, zodat alles controleerbaar is en zoveel mogelijk menselijke fouten worden voorkomen. Als deze technieken goed werken zouden ze toegepast kunnen worden op andere corpora om een snel overzicht te geven van de inhoud of verschuiving van inhoud over tijd van deze corpora. Er zal allereerst niet specifiek worden gekeken naar de verandering in taalgebruik en grammatica, maar dit kan wel gebruikt worden om een beeld te vormen over de veranderingen binnen de troonredes.
De terugkoppeling naar het informatiekunde vakgebied ligt hier bij het automatiseren van verschillende technieken en het weergeven en gebruik van informatie. Doordat deze technieken worden geautomatiseerd kunnen ze mogelijk op vele verschillende plekken worden toegepast. Ook kunnen er betere controles worden uitgevoerd omdat alles uiteindelijk door computers

wordt uitgevoerd. Verder wordt er gekeken naar de mogelijkheid om de technieken toe te passen op andere datasets en andere toepassingen van de technieken. Hierbij kan bijvoorbeeld gekeken worden naar wetenschappelijke literatuur om de verschuiving van wetenschappelijke focus weer te geven. De informatie kan op verschillende manieren worden gebruikt en weergegeven. Hierbij zullen we kijken naar de manier waarop informatie wordt gerepresenteerd en wat men vanuit deze representatie met de informatie kan doen.
Om het onderzoek een duidelijke richting te geven zal gepoogd worden de volgende onderzoeksvraag te beantwoorden:

- Hoe kan "woord co-occurence" als tekstanalyse techniek worden gebruikt om de verschuiving in onderwerpen/thema's over 200 jaar troonredes weer te geven?

We beantwoorden deze vraag met behulp van de volgende deelvragen:

1. Geven de methodes gebruikt in Rule et al. [2015] relevante resultaten over het Nederlandse troonrede corpus?

2. Zijn de resultaten representatief voor de werkelijkheid?

2 Gerelateerd werk

2.1 Latent Dirichlet Allocation (LDA)

LDA is een statistisch model dat uitgaat van observaties. Hierbij kunnen groepen van observaties worden uitgelegd door on-geobserveerde groepen. Dit gebeurt doordat de on-geobserveerde groepen uitleggen waarom sommige delen van de data gelijk zijn aan de geobserveerde data. Binnen de tekstanalyse kan dit als volgt worden gezien: Observaties zijn woorden verzameld in documenten. Hierbij wordt gesteld dat elk document een mix is van een klein aantal onderwerpen en het gebruik van een woord binnen de tekst is toe te wijden aan één van de onderwerpen van het document. Elk woord vloeit dus voort uit een gelimiteerd aantal onderliggende onderwerpen, terwijl er een oneindige hoeveelheid van onderwerpen kunnen zijn.

Voor tekstanalyse kan dit worden gebruikt om een beeld te vormen van de verschillende onderwerpen die een tekst mogelijk representeert. Er moet hier wel rekening gehouden worden dat dit geen exacte indicatie is van de onderwerpen. Dit doordat er geen rekening wordt gehouden met de context van de woorden, maar enkel met de onderwerpen waar de woorden aan toe te wijzen zijn [Blei et al., 2003].

2.2 Latent Semantic Analysis (LSA)

LSA is een techniek vanuit de natuurlijke taalverwerking. De techniek analyseert de relatie tussen groepen van documenten en de termen die deze bevatten. Dit wordt gedaan door te kijken naar concepten die de relatie tussen de documenten en termen weer kunnen geven. LSA is gebaseerd op een stelling. Deze stelling is dat woorden die gelijksoortige betekenissen hebben, zoals "huis" & "woning", in soortgelijke delen van tekst voorkomen. LSA maakt gebruik van een matrix waarin bijgehouden wordt hoe vaak een woord voorkomt per paragraaf. Deze matrix wordt gereduceerd waarbij de gelijkheid tussen paragrafen behouden wordt door middel van singuliere waardenontbinding(SWO). SWO wordt hier gebruikt om de matrix te ontbinden naar de relevante kolommen, welke representatief zijn voor soortgelijke paragrafen.

Na het reduceren worden de woorden uit het matrix vergeleken door de cosinus van de vector van hun rij te nemen. Dit geeft een waarde van 0 tot 1, waarbij een waarde van 1 wil zeggen dat de woorden erg

gelijk aan elkaar zijn en een waarde van 0 aangeeft dat de woorden zeer ongelijk aan elkaar zijn [Dumais, 2004].

2.3 Topic Modelling

Topic modelling is een type statistisch model voor het vinden van abstracte "topics", ook wel onderwerpen, in collecties van documenten. Binnen de tekstanalyse wordt topic modelling vooral gebruikt om verborgen achterliggende semantische structuren tussen documenten weer te geven. Dit wordt gedaan door te kijken naar de samenhang tussen onderwerpen binnen de verschillende teksten. Deze onderwerpen die uit de topic modelling technieken voortkomen zijn uiteindelijk gevormd vanuit clusters van gelijksoortige woorden.

Topic modelling geeft idealiter ook een indruk van het taalgebruik dat men kan verwachten binnen een tekst. Zo verwacht men dat binnen een tekst over voetbal de woorden "keeper" & "buitenspel" vaker voorkomen dan bijvoorbeeld het woord "basketbal". Omdat een tekst echter meerdere onderwerpen kan behandelen kan er worden gekeken naar de proporties van de onderwerpen binnen de tekst. Zo kan een tekst over sport 80% voetbal zijn en 20% basketbal. Men zou hier dan verwachten dat er ongeveer 4 keer zoveel voetbal gerelateerde woorden in de tekst voorkomen dan dat er basketbal gerelateerde termen voorkomen [Sojka, 2010].

2.4 Digital Humanities

Digital Humanities, digitale geesteswetenschappen, is het onderzoeksgebied dat zich bezighoud met de kruising tussen computerwetenschappen en geesteswetenschappen. Hier wordt onderzoek gedaan aan de hand van gedigitaliseerd materiaal en materiaal dat digitaal is gemaakt. Het bestrijkt verschillende onderwerpen, waaronder data mining, data organisatie en informatie verzameling. Hierbij gaat het veelal om grote datasets welke vaak online te vinden zijn.

De definitie van digital humanities wordt constant aangepast, omdat er steeds meer onderzoek naar en binnen gedaan wordt. Ook wordt het steeds duidelijker dat digital humanities een steeds groter gebied bestrijkt. Dit doordat er steeds meer onderzoek wordt gedaan naar de mogelijkheden van het gebruik van computers om taken te automatiseren.

Ook kunnen computers steeds meer taken uitvoeren waardoor er meer digitaal uitgevoerd kan worden [Berry, 2012].

3 Methodology

3.1 Beschrijving van de data

De dataset die gebruikt gaat worden om deze vragen te beantwoorden is een verzameling van troonredes sinds 1814. De gehele dataset bestaat uit 172 verschillende troonredes welke in totaal uit 224.165 woorden bestaan, deze set zullen we vanaf nu het corpus noemen. Deze troonredes zijn terug te vinden op www.troonredes.nl [Herko Coomans, 2015] . Om een duidelijker beeld te geven van de troonredes een korte uitleg van wat de troonredes nu precies zijn en waar ze vandaan komen.

De troonrede wordt nu jaarlijks door de koning(in) uitgesproken op Prinsjesdag, de 3de dinsdag van september. Deze dag is ook de opening van het parlementaire jaar. In de 19de eeuw viel de opening van de Staten-generaal op de eerste maandag van november en later op de 3de maandag van oktober. Voor 1904 werden de troonredes uitgesproken in de vergaderzaal van de 2de kamer. Sinds 1904 worden de troonredes voor de ridderzaal op het binnenhof in Den Haag uitgesproken. De troonrede wordt live uitgezonden op televisie en is na te lezen op de website van de Nederlandse overheid. De eerste troonrede werd in 1814 als een algehele toespraak voor de Staten-generaal gehouden. De troonredes worden vooral gebruikt om wets- en beleidsveranderingen door te geven, als beschouwing op het afgelopen jaar en de staat van het land.

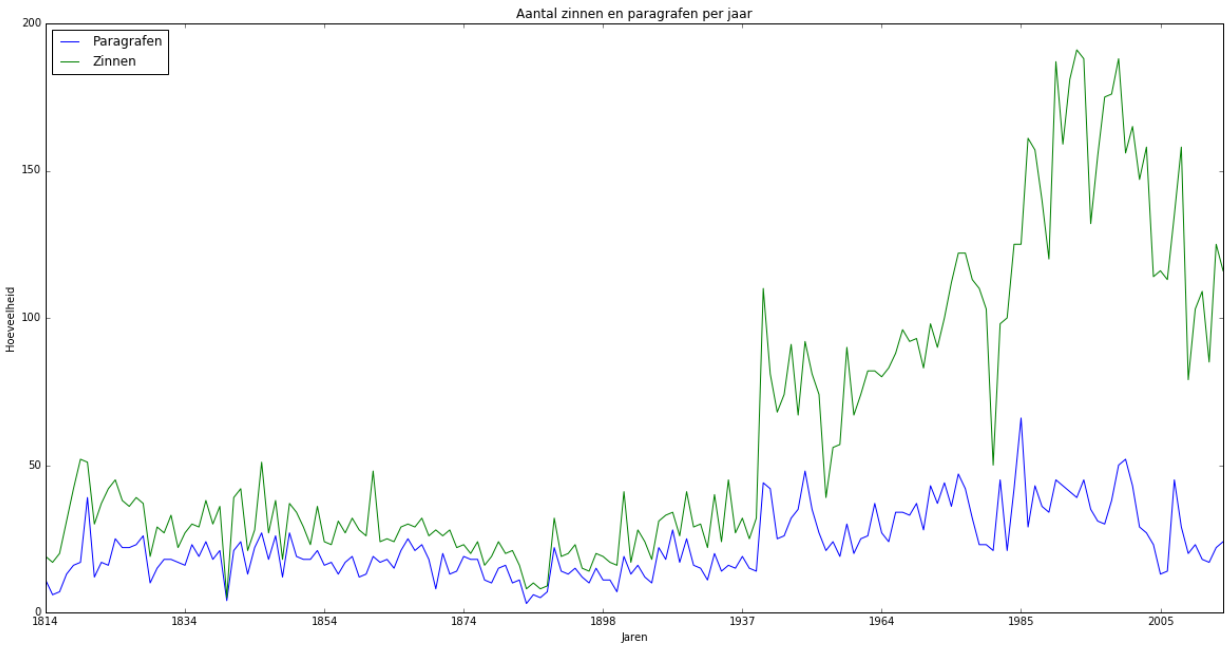
In recentere jaren heeft deze beschouwing zich ook uitgebreid naar gebeurtenissen door de hele wereld die invloed uitoefenen op de Nederlandse staat. Ook wordt sinds 1918 het regeerprogramma voor het komende jaar in de troonredes behandeld. Sinds 1848 worden de troonredes geschreven door ministers vanwege een grondwetsherziening waardoor de ministers verantwoordelijk werden voor al het doen en laten van de koning(in). Hiermee werd het kabinet verantwoordelijk voor de uitspraken die gedaan worden in de troonredes [Rijksoverheid, 2016]. Dit zorgt ervoor dat de troonredes een beeld geven van wat de Nederlandse regering op dat moment belangrijk vindt en zijn ze in die zin een weerspiegeling van de dan heersende politieke urgentie.

Om een duidelijker beeld te geven van de inhoud en omvang van de troonredes geeft tabel 1 een kort overzicht met enkele statistieken. Hierbij is de langste troonrede die van 1993 en de kortste troonrede die van 1880.

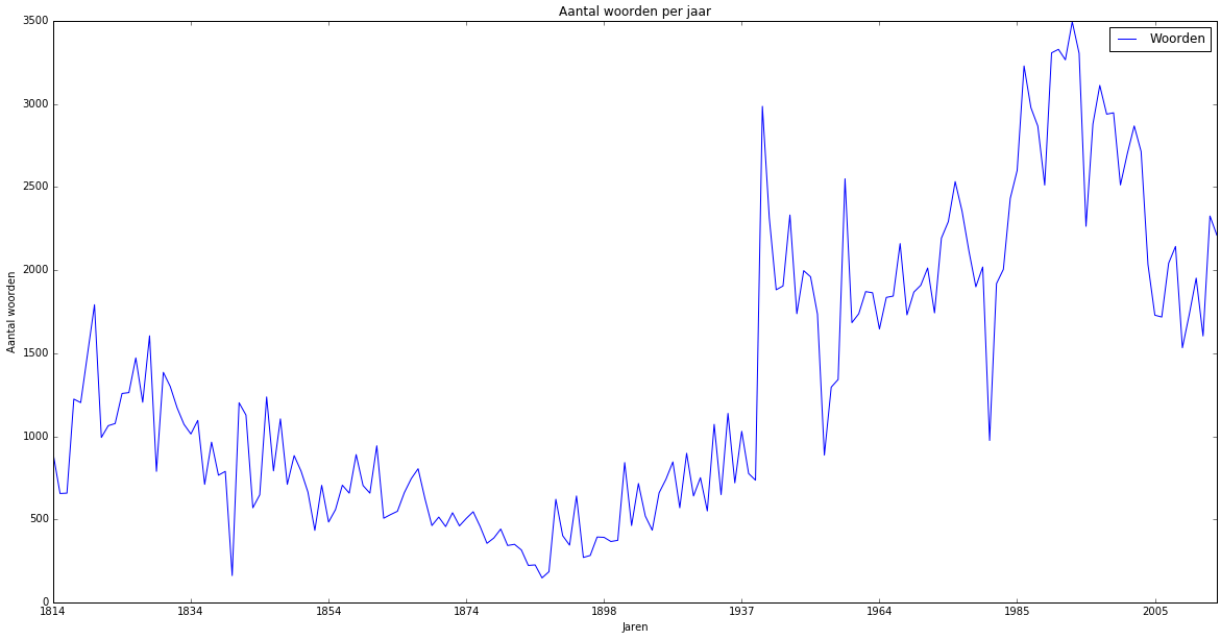
	Paragrafen	Zinnen	Woorden
Gehele corpus	3887	10415	224165
Gemiddeld	22	60	1303
Langste troonrede	39	191	3495
Kortste troonrede	16	20	343

tabel 1: Troonrede statistieken

De grafiek in figuur 1 geeft een beeld van de grootte van de troonredes over de jaren aan de hand van het aantal zinnen en paragrafen per jaar. Figuur 2 toont een grafiek die weergeeft hoeveel woorden de troonrede elk jaar bevatte.



figuur 1: Weergave van het aantal zinnen en paragrafen per jaar.



figuur 2: Weergave van het aantal woorden per jaar.

Opmerkelijk is hier de stijging in het aantal zinnen en woorden per troonrede na 1937. Deze periode correleert met de stijgende onrust in de oploop naar de Tweede Wereldoorlog en de oorlog zelf. Hierdoor kan deze verandering in troonrede grootte rond 1945 mogelijk verklaard worden door het feit dat er in die periode terug is gekeken op de oorlog. Het feit is dat er in die oorlogstijd geen troonredes uitgesproken zijn, hierover verdere informatie in 3.1.1. Omdat er in die periode geen troonredes uitgesproken zijn had men een grotere periode van gebeurtenissen om op te reflecteren. Ook zal er mogelijk veel aandacht besteed zijn aan het herstellen van het vertrouwen van het volk in de regering en hoe Nederland als land verder zou gaan na de oorlog.

3.1.1 Missende data

Er zijn verscheidene jaren waarvan er geen data beschikbaar is, dit doordat er niet elk jaar een troonrede gegeven is. Dit heeft verschillende redenen, zoals oorlogen, angst voor rellen, onvrede over het kabinet of gezondheidsredenen omtrent de koning(in). Dit zorgt ervoor dat niet elk jaar sinds 1814 wordt gerepresenteerd door een troonrede. Omdat de inhoud van de troonredes als verzamelde dataset wordt

gebruikt heeft dit minimaal invloed op de uiteindelijke uitkomsten van het onderzoek dat er jaren missen. Door de missende jaren is het weergeven van de verschuiving in onderwerpen mogelijk niet volledig representatief, omdat er geen uitspraken gedaan kunnen worden over de missende jaren. In tabel 2 een overzicht van de jaren waar geen troonredes van zijn en de reden dat er in die jaren geen troonrede is gegeven:

Jaren	**reden**
1888-1890	Verhindering wegens ziekte van de koning en interne beleidsproblemen
1905-1924	Eerste wereldoorlog en politieke instabiliteit
1940-1947	Tweede wereldoorlog en politieke instabiliteit

tabel 2: Missende troonredes

3.1.2 Kwaliteit van de data

De data waarmee gewerkt zal worden zijn 172 troonredes die in de laatste 200 jaar zijn uitgesproken. Deze troonredes zijn binnengehaald vanaf www.troonredes.nl [Herko Coomans, 2015] en opgeslagen als html pagina's. Alle troonredes die te vinden zijn op www.troonredes.nl die niet digitaal door de overheid werden aangelever zijn handmatig ingevoerd door de beheerders van de website. De troonredes die niet digitaal opgenomen waren zijn overgetypt vanuit het nationaal archief. De troonredes die nu worden uitgesproken worden ook door de Nederlandse overheid online gepubliceerd, waarna deze worden gekopieerd naar de website van www.troonredes.nl.

3.2 Trefwoord extractie

Voor het onderzoek maken we gebruik van een selectie van woorden uit het corpus. Deze woorden worden als trefwoorden uit het corpus gehaald. Er bestaan verschillende manieren om trefwoorden uit teksten te halen. De meeste eenvoudige manier is door simpelweg te kijken naar de waarschijnlijkheid dat een woord voorkomt. Dit wordt gedaan door te kijken naar de frequentie dat een woord voorkomt in een tekst en dit

te delen door het totaal aantal woorden in de tekst. Verder kan men kijken naar een geheel corpus, de meest bekende methode hiervoor is TF-IDF (term frequency–inverse document frequency)[Ramos, 2003]. Hierbij wordt gekeken naar het voorkomen van woorden in een tekst tegenover het voorkomen in het gehele corpus. De TF-IDF score van een woord is hoog voor een specifieke tekst uit een corpus als deze vaak voorkomt in die tekst, maar verder weinig in de rest van het corpus. Door de frequentie van het voorkomen van een woord in een tekst te compenseren met het voorkomen van het woord in de gehele tekst wordt rekening gehouden met woorden die in het algemeen veel worden gebruikt, zoals stopwoorden [Aggarwal and Zhai, 2012]. Daarnaast zijn er methodes gericht op individuele teksten zoals in Matsuo and Ishizuka [2004], waar een algoritme gebruikt wordt om trefwoorden uit een enkele tekst te halen zonder gebruik te maken van een corpus.

3.3 relevantie van co-occurrences

Co-occurrences zijn uiteindelijk simpele combinaties van woorden. Deze combinaties geven aan of twee woorden samen voorkomen in een zin, paragraaf of tekst. Hierbij wordt meestal gekeken naar het samen voorkomen in een zin. Met behulp van co-occurrences kan men verschillende clustering methodes uitvoeren. De clusters gevormd door deze methodes kunnen weergeven welke woorden veel invloed hebben op de tekst doordat deze het middelpunt van clusters zullen zijn. Met behulp van deze clusters kunnen de meest belangrijke termen dus worden afgeleid. Ook kunnen met behulp van co-occurrences veel voorkomende multi termen worden gevonden in teksten. Enkele voorbeelden hiervan zijn: "Verenigde Staten" en "Hoger onderwijs". Deze multi termen zijn individuele woorden die ook als één woord gezien kunnen worden. Met behulp van het bepalen van co-occurrences kunnen deze multi termen makkelijk gevonden worden.

3.4 co-occurrence extractie

Men kan op verschillende manieren co-occurrences uit teksten halen. Om deze manieren te verduidelijken gebruiken we de volgende voorbeeldzin: "Het is erg heet". Vanuit deze zin kunnen op de volgende manieren co-occurences worden gehaald. Door enkel co-occurrences te gebruiken van woorden die exact naast elkaar in een zin staan, "Het,is" & "is,erg" &

"erg,heet". Ook kan het door woorden in een zin volgens de zinsvolgorde met elkaar te koppelen zelfs als er andere woorden tussen zitten, hierdoor zouden "Het,heet" & "Het,erg" & "is,heet" ook co-occurrences zijn. Of het kan door alle mogelijke combinaties van woorden in een zin te vormen, dit gebeurt veelal op alfabetische woordvolgorde. Volgens de laatste manier wordt geen rekening meer gehouden met de volgorde van de woorden. De manier die het beste is om de co-occurrences uit de tekst te halen is afhankelijk van wat men uiteindelijk wil kunnen zeggen met de co-occurrences [Shimohata et al., 1997].

3.5 Invloed van context

Om een onderwerp of meerdere onderwerpen toe te kunnen wijzen aan een tekst moet men rekening houden met meerdere aspecten. De meest belangrijke hiervan is de context van de tekst. De context bepaald namelijk op welke manier woorden geïnterpreteerd worden en wat ze betekenen voor de tekst. Zo zijn er woorden die enkel voor specifieke domeinen betekenis hebben. Er zijn al meerdere methodes ontwikkeld om rekening te houden met de context van een tekst, maar de meeste hiervan hebben hiervoor een corpus nodig om het context netwerk te kunnen bouwen. Om van een individuele tekst de context te kunnen bepalen kan zeer lastig zijn. Er kan rekening gehouden worden met de context van een corpus door gebruik te maken van woordenboeken voor het specifieke domein dat het corpus bestrijkt. Hiervoor moet men echter wel zelf bepalen wat het domein van het corpus is [Aggarwal and Zhai, 2012].

3.6 Methodes

Om uit het corpus te kunnen achterhalen of er overkoepelende onderwerpen zijn en wat deze mogelijk zijn wordt er gebruik gemaakt van tekstanalyse technieken. Specifiek wordt er gebruik gemaakt van POS-methodes en co-occurence aanpakken Callon et al. [1991].

Allereerst wordt er met behulp van POS-methodes(Pattern of Speech) gekeken naar het volledige corpus. Deze methodes kijken naar het corpus als geheel en proberen hier patronen uit te halen. Er is specifiek gebruik gemaakt van "Pattern", een Python module, die getraind is voor Nederlandse tekst. Deze module is getraind op verschillende soorten Nederlandse teksten en kan woorden uit de tekst onderverdelen in verschillende categorieën, zoals lidwoorden en zelfstandige naamwoorden. Met behulp van deze module is het corpus gefilterd zodat enkel de zelfstandige naamwoorden zijn overgebleven. De reden dat we enkel zelfstandige naamwoorden gebruiken is het feit dat deze representatief zijn voor de tekst. Dit omdat vanuit de zelfstandige naamwoorden afgeleid kan worden wat het doel en inhoud van de tekst is, want zelfstandige naamwoorden kunnen grammaticaal gezien als het onderwerp van een zin gezien worden.

Hierna wordt de data van de troonredes gelemmatiseerd. Hierdoor worden de woorden herleid tot hun lemma(stam), waardoor ze kunnen worden geanalyseerd als één term. Enkele voorbeelden hiervan: "steden" wordt herleid tot "stad" en "mensen" kan herleid worden tot "mens". Door alle woorden te herleiden tot hun stam wordt er een duidelijk beeld gevormd van alle unieke woorden die gebruikt worden. Hierdoor zal de relatie tussen woorden sneller duidelijk worden, omdat er nu geen onderscheid wordt gemaakt tussen de verschillende vervoegingen van unieke woorden. Met behulp van deze gelemmatiseerde termen wordt een co-occurence matrix opgesteld. Deze matrix wordt aangemaakt door bij te houden hoe vaak een combinatie van 2 termen samen in een paragraaf voorkomen. De matrix geeft uiteindelijk voor alle combinaties van 2 termen de frequentie dat ze samen in een paragraaf voorkomen weer.

Voor alle combinaties uit deze matrix wordt een nabijheidsscore berekend via de volgende definitie:

$$S(W_1, W_2) \equiv^{\text{def}} \frac{\sum_{c \epsilon W \setminus W_1, W_2\}, PMI(W_1,c)>0} min(PMI(W_1,c), PMI(W_2,c))}{\sum_{c \epsilon W\{W_1, W_2\}, PMI(W_1,c)>0} PMI(W_1,c)}$$

Hierbij is het doel om uit te vinden welke termen relevant zijn en hoe deze zich over tijd met andere termen associëren. De score geeft het verwantschap tussen twee termen "W1" & "W2" aan. Dit verwantschap wordt bepaald aan de hand van PMI(Pointwise Mutual Information) tussen twee termen [Bouma, 2009]. Hierbij geeft de PMI tussen twee termen "X" & "Y" aan hoeveel de termen ons over elkaar kunnen vertellen. Hierbij gaat het om het verschil in de kans dat term "X" of "Y" individueel voorkomt in een paragraaf in de tekst en de kans dat termen "X" & "Y" samen voorkomen in een paragraaf in de tekst. Wiskundig gezien wordt PMI als volgt berekend:

$$PMI(X,Y) \equiv^{\log \frac{p(X,Y)}{p(X)}}$$

Hierbij is P(X,Y) de kans dat termen "X" & "Y" samen in een paragraaf voorkomen en P(X) de kans dat term "X" in een paragraaf voorkomt. Dit wordt berekend door het aantal keer dat de term voorkomt in de tekst te delen door het aantal paragrafen. De PMI is een score die positief of negatief kan zijn. Goede co-occurence paren van termen hebben een hoge PMI score omdat ze net iets minder samen voorkomen dan dat ze individueel voorkomen. Een score van 0 betekent dat de termen onafhankelijk van elkaar zijn en er geen uitspraak gedaan kan worden over het voorkomen van term "X" als "Y" voorkomt en vice versa.

Het PMI wordt berekend via de TF(Term Frequency) van een woord. Dit wil zeggen dat er gekeken wordt naar het aantal keren dat het woord voorkomt in de tekst. Als het woord "minister" bijvoorbeeld 5 keer voorkomt in een troonrede heeft het een TF van 5 voor die troonrede. Op deze manier wordt ook de TF voor alle woorden berekend over het gehele corpus. Hierbij worden alle troonredes als één enkele tekst gezien om te bepalen hoe vaak een woord in het gehele corpus voorkomt. Dit wordt gebruikt om de P(X) als volgt te berekenen:

$$P(X) \equiv \frac{TF(X)}{AantalParagrafen}$$

Hierbij is de TF(X) de Term Frequency van het woord in het gehele corpus en "Aantal Paragrafen" het aantal paragrafen in het gehele corpus.

De co-occurence matrix wordt gebruikt bij het berekenen van P(X,Y). In de co-occurence matrix wordt bijgehouden hoe vaak twee termen samen in een zin voorkomen. In essentie is dit de TF van de combinatie

van termen. Hiermee wordt P(X,Y) op dezelfde manier berekend als P(X), waarbij de TF(X) wordt vervangen door de TF van de co-occurence combinatie vanuit de co-occurence matrix.

De vergelijking voor de nabijheidsscore maakt gebruikt van de termen uit het gehele corpus om context te bepalen. Hiervoor wordt voor elke combinatie termen (W1,W2) gekeken naar de woorden waarmee ze samen in een paragraaf voorkomen. De som van PMI(W,C) voor alle termen (C) uit het corpus waarvoor geldt dat PMI(W,C) >0 wordt voor beide termen berekend. Dit betekent dus dat er enkel wordt gekeken naar combinaties van termen die onafhankelijk van elkaar zijn, met een score van 0, of positief. Als hieruit volgt dat PMI(W1,C) gelijk is aan PMI(W2,C) kan men stellen dat als W1 in een paragraaf voorkomt W2 ook voorkomt en vice versa. Een nabijheidsscore van 0 geeft aan dat de woorden nooit samen in een paragraaf voorkomen. Aan de hand van deze score wordt een gewogen semantisch netwerk gevormd met de termen als nodes en de gewichten op de edges.

Om de termen binnen dit netwerk te kunnen analyseren wordt gebruik gemaakt van een community detectie algoritme. Het specifieke algoritme dat gebruikt wordt is dat van Blondel et al. [2008]. Het doel van dit algoritme is om vanuit het gewogen netwerk clusters te vormen van samenhangende subsets van termen. Om ervoor te zorgen dat enkel de meest relevante termen worden meegegeven aan het community detectie algoritme is een drempelwaarde bepaald. Deze drempelwaarde geeft aan vanaf welke waarde voor de nabijheidsscore de co-occurence combinaties door het algoritme verwerkt worden. De drempelwaarde ligt tussen de 0 en 1 en wordt bepaald door vanaf 0 de score op te hogen. Dit is een handmatig proces waarbij de drempelwaarde is bereikt als er binnen het netwerk een component van meer dan twee nodes verdwijnt. Hiermee worden de zwak verbonden nodes uit het netwerk verwijderd. De drempelwaarde die gevonden is voor ons netwerk is =0.65. Het resulterende netwerk is door het community detectie algoritme verwerkt om de verschillende subsets te kunnen identificeren. Vanuit deze subsets zijn onderwerpen bepaald aan de hand van de termen die er onder vallen. Afhankelijk van de termen binnen deze clusters kunnen deze clusters gezien worden als indicatief voor onderwerpen/thema's binnen het corpus. Dit induceren van onderwerpen moet een overzicht geven van de verschillende termen die tot een onderwerp behoren en hoe sterk de samenhang is tussen termen binnen een onderwerp. Ook zou het een beeld moeten geven van de samenhang tussen verschillende onderwerpen.

4 Evaluatie

4.1 Geven de methodes gebruikt in Rule et al. [2015] relevante resultaten over het Nederlandse troonrede corpus?

Omdat relevant een brede term is zal deze worden toegespitst. Hierbij zal relevant als het volgende worden gezien: de resultaten zijn relevant als ze gebruikt kunnen worden voor het specifiek bestuderen van individuele troonredes en mogelijk leiden naar de meest relevante en belangrijke momenten in tijd/troonredes.

De Nederlandse troonrede is net zoals de Amerikaanse State of the Union een toespraak die jaarlijks vanuit de overheid worden gegeven. In beide gevallen wordt de voordracht door één positie voorgedragen, voor Nederland is dit de koning(in) en voor de VS is dit de president. Voor beide teksten geldt dat ze zijn opgebouwd uit verschillende paragrafen en er veelal voor elke paragraaf één onderwerp centraal staat. Beide hebben een centraal motief gericht op het meedelen van de huidige staat van het land en mogelijk voornemens voor het komende jaar.

Omdat de opbouw en het doel van beide teksten zoveel op elkaar lijken zijn de methodes die gebruikt zijn binnen Rule et al. [2015] goed toepasbaar op het corpus van de Nederlandse troonredes. Hierbij hoeven er maar kleine aanpassingen gemaakt te worden, waarbij het grootste verschil ligt in het feit dat de teksten in verschillende talen zijn geschreven. Voor het Nederlands betekent gaf dit enkele problemen omdat woorden over de jaren op verschillende manieren zijn geschreven en de taal over de jaren is veranderd. Zo werd het woord "de" vroeger als "den" geschreven. Hierdoor zijn enkele woorden die vroeger anders geschreven werden niet opgevangen door de POS-tagging. Op het moment dat dit duidelijk werd is het handmatig aangepast. Op deze manier is er net als in Rule et al. [2015] het corpus gereduceerd tot de relevante zelfstandige naamwoorden. Hierbij is dus gebruik gemaakt van modules die op de Nederlandse taal getraind waren. Hierna zijn de 1000 meest voorkomende termen gebruikt om de co-occurence matrix op te stellen dat gebruikt is om de nabijheidsscore te berekenen waarmee het semantische netwerk is gevormd.

De resultaten die hieruit zijn gekomen geven een overzicht van de verschuivingen van inhoud van de troonredes. Aan de hand van deze resultaten komen mogelijk interessante periodes naar voren doordat een

onderwerp mogelijk meer wordt behandeld. Ook kan aan de hand van de veel voorkomende termen worden gekeken naar individuele troonredes door te sorteren op voorkomen van termen of co-occurences. Hiermee kan dus gesteld worden dat de resultaten volgens onze definitie relevant zijn.

4.2 Zijn de resultaten representatief voor de werkelijkheid?

De verkregen resultaten zijn de geïnduceerde onderwerpen en het daarmee verkregen overzicht van onderwerp verschuivingen. Hiervoor was het belangrijkste dat er vanuit het semantische netwerk door middel van het community clustering algoritme duidelijke clusters gevormd werden. Vanuit de termen binnen een cluster kan een achterliggend onderwerp worden bepaald dat door deze termen wordt geïmpliceerd. Het is dus noodzakelijk dat de clusters duidelijk verschillend van elkaar zijn en de relatie tussen de termen in een cluster zo min mogelijk verschillende onderwerpen impliceren. Hoe duidelijker de relatie is tussen de termen in een cluster hoe makkelijker het is om er een onderwerp aan te koppelen. Dit is het makkelijkst voor clusters die gecentreerd zijn rondom één term, zoals "onderwijs" of "ontwikkeling".

De onderwerpen die gevonden zijn staan in tabel 3:

Onderwerpen
regeringsbeleid
onderwijs
arbeidsbeleid
koloniaalbeleid
wetgeving
ontwikkeling
staatplanning
defensie

tabel 3: Geïmpliceerde onderwerpen

Voor elk onderwerp is er een lijst met termen waarvoor geldt dat als deze voorkomen in een tekst ze dat onderwerp impliceren. Tabel 4 geeft een overzicht van de 5 meest representatieve woorden per cluster:

Staatplanning	Onderwijs	Ontwikkeling	Defensie
regering	regering	regering	regering
ontwikkeling	ontwikkeling	ontwikkeling	veiligheid
maatregel	onderwijs	maatregel	uitbreiding
beleid	mogelijkheid	mens	unie
mens	probleem	gebied	politie

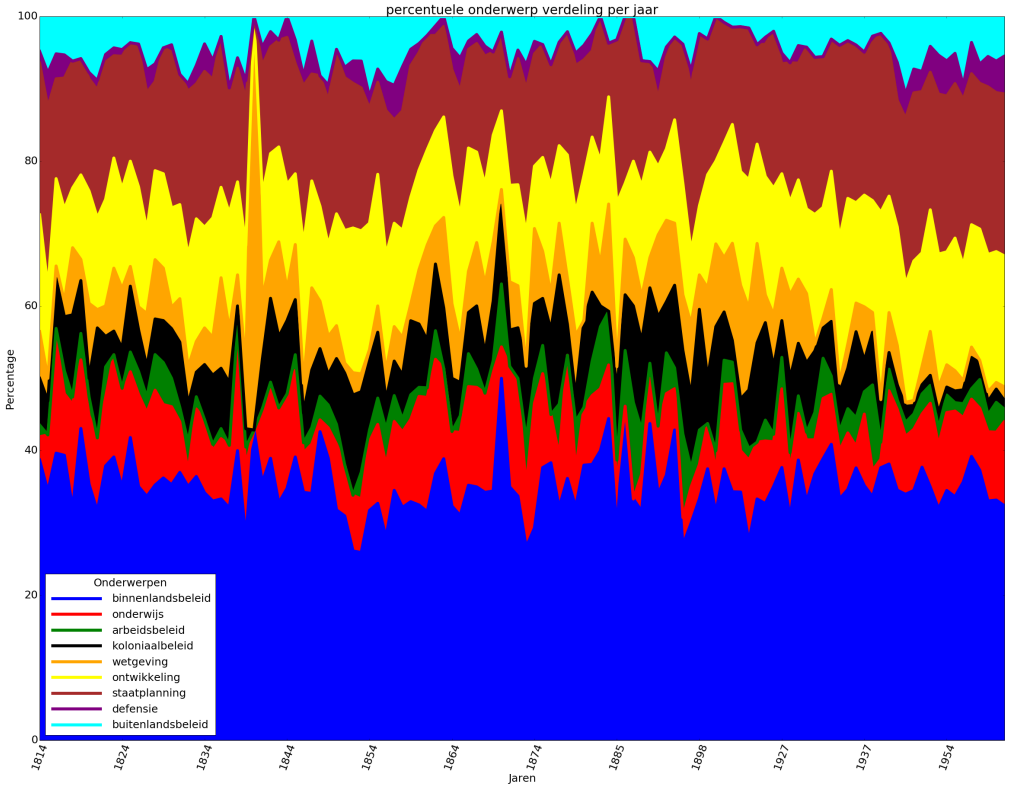
Arbeidsbeleid	Wetgeving	Buitenlandsbeleid
volk	wet	samenwerking
werkgelegenheid	regeling	toekomst
welvaart	wetsontwerp	handel
werk	wijziging	europa
werkloosheid	voorbereiding	wereld

Binnenlandsbeleid	Koloniaalbeleid
regering	verbetering
land	toestand
ontwikkeling	nederlandsch-indië
nederland	rijk
maatregel	curaçao

tabel 4: 5 meest representatieve woorden per cluster

Zoals te zien is in tabel 4 zijn er enkele onderwerpen waar dezelfde termen veel in voorkomen. De meest voorkomende termen zoals "regering" en "ontwikkeling" zijn representatief voor verschillende onderwerpen. Dit doordat deze termen in vele co-occurences voorkomen, maar van de verschillende clusters deel uit maken doordat ze aan een ander woord zijn gekoppeld.

Aan de hand van de termen voor elk onderwerp is een grafiek opgesteld met daarin voor elk jaar de hoeveelheid dat een onderwerp besproken werd. Dit is gedaan door te kijken naar de volledige tekst van dat jaar en voor elk woord na te gaan bij welk onderwerp het hoort. Hierna is door gebruik te maken van een normaal verdeling een grafiek gevormd die per onderwerp procentueel aangeeft hoeveel van de troonrede besteed was aan de behandeling van dat onderwerp. Hieruit volgt de grafiek in figuur 3:



figuur 3: Verschuiving van inhoud van troonredes over de jaren aan de hand van de gevonden onderwerpen.

Vanuit deze grafiek is duidelijk af te lezen welke onderwerpen het belangrijkst waren in een jaar. Zo is bijvoorbeeld duidelijk te zien dat er rond 1839-40 het onderwerp wetgeving sterk naar voren komt, wat terug te leiden is naar de werkelijkheid, omdat in 1840 de officiële scheiding plaatsvond tussen Nederland en België ondertekend in het verdrag van Londen. Door dit verdrag is een groot deel van de Nederlandse grondwet aangepast [Schroeder, 1996]. De resultaten zijn dus zichtbaar terug te leiden naar de werkelijkheid.

5 Conclusies

Het onderzoek had als doel het repliceren van een eerder onderzoek, namelijk dat van Rule et al. [2015], op een andere dataset. De dataset die hiervoor gebruikt is bestond uit Nederlandse troonredes van 1814 tot 2014. Hierbij werd vooral gekeken naar de tekstanalyse technieken voor het weergegeven van verschuivingen van onderwerpen binnen corpora door gebruik te maken van co-occurences.

Er werd allereerst gekeken of de technieken gebruikt binnen Rule et al. [2015] daadwerkelijk toepasbaar waren op onze dataset en relevante resultaten leverden. Hiervoor werd een specifieke definitie gesteld wat voor ons als relevant wordt gezien. Aan de hand van de resultaten is gebleken dat de toegepaste technieken wel degelijk resultaten leverden die relevant voor ons waren. Hierdoor hebben we kunnen stellen dat de technieken gebruikt binnen Rule et al. [2015] ook op onze dataset gebruikt kon worden.

Na dit vast te stellen is er gekeken naar het verband tussen de resultaten van het uitvoeren van de tekstanalyse technieken op het corpus met de werkelijkheid. Hier werd duidelijk dat de resultaten te herleiden waren en een duidelijke link hebben met de werkelijkheid. Hoewel één moment hier zeer uitspringt is het echter moeilijk om andere momenten net zo exact terug te koppelen. Er zijn echter wel duidelijk trends te zien. Deze trends geven weer welke onderwerpen jaarlijks meer of minder ter sprake kwamen.

Uiteindelijk kunnen we hiermee antwoord geven op de hoofdvraag van dit onderzoek, *"Hoe kan "woord co-occurence" als tekstanalyse techniek worden gebruikt om de verschuiving in onderwerpen/thema's over 200 jaar troonredes weer te geven?"*

Woord co-occurences kunnen gebruikt worden om de meest belangrijke woorden en meest voorkomende verbanden tussen woorden in een tekst weer te geven. Hierbij moet er wel een duidelijk onderscheid gemaakt worden tussen het soort woorden dat men gebruikt uit de tekst. De meest indicatieve soort woorden hebben wij hier gevonden als de zelfstandige naamwoorden, maar dit kan voor andere corpora mogelijk anders zijn. Uiteindelijk hebben we aan de hand van de veel voorkomende co-occurences termen aan onderwerpen kunnen toewijzen die behandeld werden in het corpus. Doordat elk van deze onderwerpen een eigen lijst met termen had die het onderwerp impliceerde hebben we kunnen weergeven welke onderwerpen elk jaar het meest aan bod kwamen en hoe de behandeling

van de onderwerpen verdeeld was.

Uiteindelijk is er een duidelijk overzicht gevormd van de verschuiving van onderwerpen over de jaren in figuur 3. De resultaten weergegeven in deze afbeelding zijn teruggekoppeld naar de werkelijkheid waaruit gesteld kan worden dat de resultaten valide zijn.

Voor vervolg onderzoek zou gekeken kunnen worden naar het gebruik van deze technieken op andere corpora. Ook zou er gekeken kunnen worden naar het opstellen van een model waar randvoorwaarden in staan waar teksten binnen een dataset aan moeten voldoen om met deze methoden onderzocht te worden. Dit zouden andere modellen kunnen zijn voor verschillende vakgebieden, omdat daar mogelijk niet alleen gekeken moet worden naar de zelfstandige naamwoorden in de teksten. Vanuit zo'n model zou ook duidelijk opgesteld kunnen worden wat voor uitspraken gedaan kunnen worden aan de hand van de resultaten.

5.1 Acknowledgements

References

Charu C Aggarwal and ChengXiang Zhai. A survey of text clustering algorithms. In *Mining Text Data*, pages 77–128. Springer, 2012.

David M Berry. *Understanding digital humanities*. Palgrave Macmillan, 2012.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10): P10008, 2008.

Gerlof Bouma. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, pages 31–40, 2009.

Michel Callon, Jean Pierre Courtial, and Françoise Laville. Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemsitry. *Scientometrics*, 22(1):155–205, 1991.

Susan T Dumais. Latent semantic analysis. *Annual review of information science and technology*, 38(1):188–230, 2004.

Pieter C. Lagas. Herko Coomans. troonredes.nl, 2015. URL `www.troonredes.nl`.

Yutaka Matsuo and Mitsuru Ishizuka. Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools*, 13(01): 157–169, 2004.

Juan Ramos. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, 2003.

Rijksoverheid. rijksoverheid, 2016. URL `https://www.rijksoverheid.nl/onderwerpen/koninklijk-huis/inhoud/positie-en-rol-staatshoofd/troonrede`.

Alix Rule, Jean-Philippe Cointet, and Peter S. Bearman. Lexical shifts, substantive changes, and continuity in state of the union discourse, 1790–2014. *Proceedings of the National Academy of Sciences*, 112 (35):10837–10844, 2015. doi: 10.1073/pnas.1512221112. URL `http://www.pnas.org/content/112/35/10837.abstract`.

Paul W Schroeder. *The transformation of European politics, 1763-1848*. Oxford University Press, 1996.

Sayori Shimohata, Toshiyuki Sugio, and Junji Nagata. Retrieving collocations by co-occurrences and word order constraints. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 476–481. Association for Computational Linguistics, 1997.

Petr Sojka. Software framework for topic modelling with large corpora. In *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Citeseer, 2010.