

Woordfrequenties

Het woord 'rokjesdag' is sinds de jaren '70 in onze taal vooral bekend geworden nadat Martin Bril het in columns in *Het Parool* diverse keren had gebruikt om de lentedag aan te duiden waarop vrouwen, als bij geheime afspraak, opeens allemaal tegelijk met blote benen in het zonlicht verschenen. In een soort navolging daarvan had ik dit opstel 'grafiekjesdag' willen noemen, maar bij nader inzien leek me dat iets te vergezocht en derhalve gaf ik vrijwillig de kans op om door een eigen woordvondst eeuwige roem te verwerven.

De frequentie waarmee woorden in een taal gebruikt worden is, indien in het perspectief van de tijd bekeken, op vele manieren zeer leerzaam. De introductie van een nieuw woord, of een oud woord dat opeens een andere betekenis krijgt, is het eenvoudigste voorbeeld. Neem het woord *boche*, dat is het Franse scheldwoord voor een Duitser, een *mof*. In een van de vele taalprojecten van Google, Google NGram, is het bijstaande grafiekje te vinden, dat loopt van de jaren 1700 tot 2000:

Évolution historique de l'usage du mot « boche »

Source : Google Ngram, application linguistique permettant d'observer l'évolution de la fréquence d'un ou de plusieurs mots ou groupe de mots à travers le temps dans les sources imprimées.



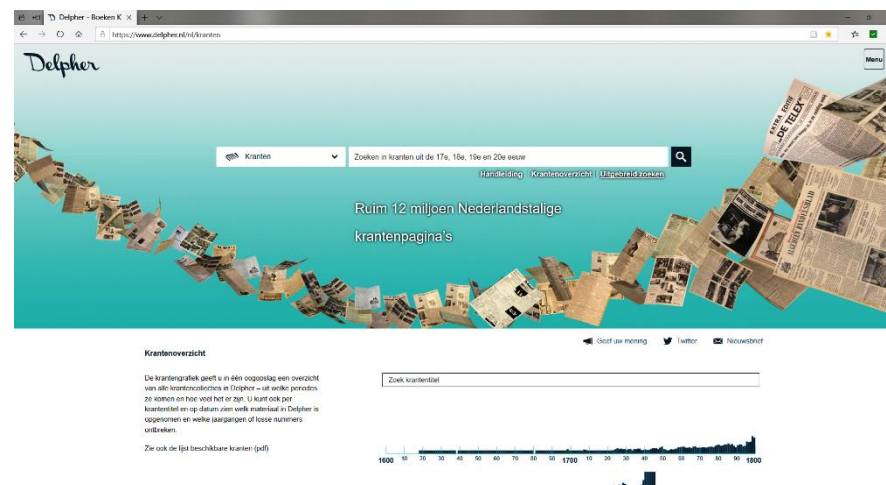
De piek rond de jaren 1914-1918 en, veel lager, 1940-1945 is bijzonder duidelijk te zien. Tijdens de eerdere Frans-Duitse oorlog, 1870-1871, is er echter zelfs geen klein piekje. Google baseert dit soort grafieken op miljoenen bladzijden van gedigitaliseerde teksten in het Frans. Die teksten moeten niet zo maar gescand zijn, maar op woordniveau doorzoekbaar (ook oude teksten in afwijkende spelling en moeilijke lettertypen), dus het is een gigantische klus, die nog lang niet voltooid is. Sinds 2004 heeft Google in samenwerking met enkele grote bibliotheken ongeveer 30 miljoen boeken gedigitaliseerd en dat moeten er 130 miljoen worden.

Maar wie zulke enorme databestanden eenmaal heeft, kan dit soort grafiekjes in luttele seconden op het scherm toveren. Dat is een zegen voor de wetenschap, maar ook voor nieuwsgierige leken van allerlei pluimage.

Voor de Nederlandse taal zijn zulke bestanden ook al geruime tijd in ontwikkeling (deels in samenwerking met Google), waarbij vooral de Koninklijke Bibliotheek in combinatie met diverse universiteiten, ook in België, de drijvende kracht is. Dat heeft al geleid tot de digitale beschikbaarstelling van ruim 80.000 boeken uit de 18de en 19de eeuw; het worden er ruim 160.000. De 50.000 boeken van vóór 1700 worden gedigitaliseerd in samenwerking met ProQuest. Uiteindelijk hoopt de KB haar complete boekenbezit online te zetten.

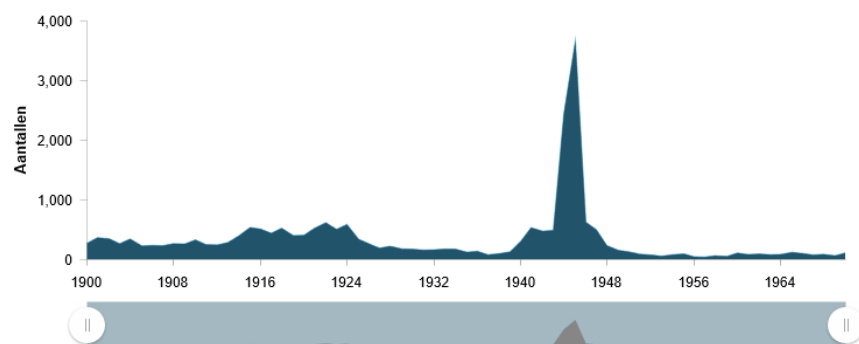
Van al die projecten noem ik nu alleen **Delpher** (*delpher.nl*), dat online geheel en al gratis te raadplegen is (al moet de belangstellende voor sommige faciliteiten eerst even online lid worden van de KB, wat nog geen twee tientjes per jaar kost). In dit project zijn Nederlandstalige bronnen gescand en op woordniveau doorzoekbaar gemaakt: boeken, tijdschriften en kranten. Het gedeelte **kranten** omvat inmiddels 12 miljoen pagina's uit de jaren

1618-1995. Wie naar de website van Delpher gaat, kan daar de categorie 'kranten' kiezen.



The screenshot shows the Delpher search interface for 'kranten'. The search term is 'moffen'. Below the search term, there is a section for 'Historische spelling' with suggestions: '+ mof', '+ moff', and '+ mofjens'. The 'Periode' is set to '01-01-1900' to '31-12-1995'. Other filters include 'Soort bericht', 'Verspreidingsgebied', 'Krantentitel', and 'Plaats van uitgave'. A 'Zoekformulier wissen' button is at the bottom right.

Watertandend kunnen we dan doorklikken naar 'uitvoerig zoeken', waarna het bijgaande zoekscherm verschijnt. We vullen dan de zoekterm in en de periode waarin we willen zoeken. De uitkomst verschijnt op het scherm als een lijst van treffers (klik rechtsonder op het vergrootglas) dan wel een grafiek (klik daar op het grafiekje). Het grafiekje kan meteen worden opgeslagen als .png of .jpg.



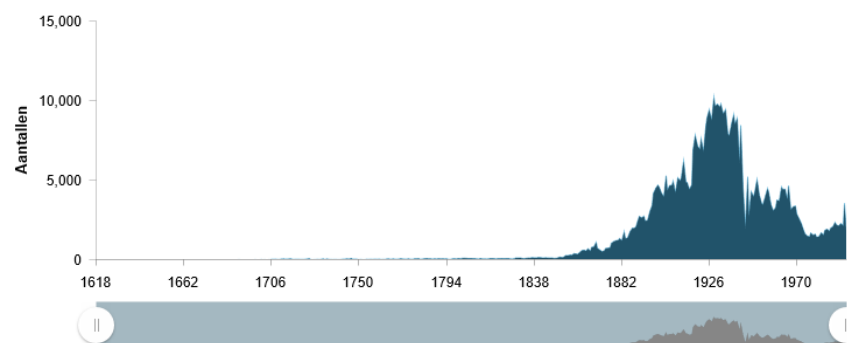
'Moffen', 1900-1970 (Delpher), 23.421 treffers.

Bij het invullen van de zoekterm kunnen we een paar trucs gebruiken. De belangrijkste is dat we twee of meer woorden samen tussen dubbele aanhalingstekens kunnen zetten om ze in die volgorde en letterlijk te laten zoeken, dus "gerard reve" en niet 'gerard' spatie 'reve', want de laatste zoekopdracht levert ook alle keren 'gerard' los en 'reve' los. Buitengewoon handig is voorts het gebruik van een joker (het teken *) voor een of meer ontbrekende

letters. Zo zoeken we met 'voetbal*' tevens op voetbalwedstrijd, voetbalstadion, voetbalshirt enz.

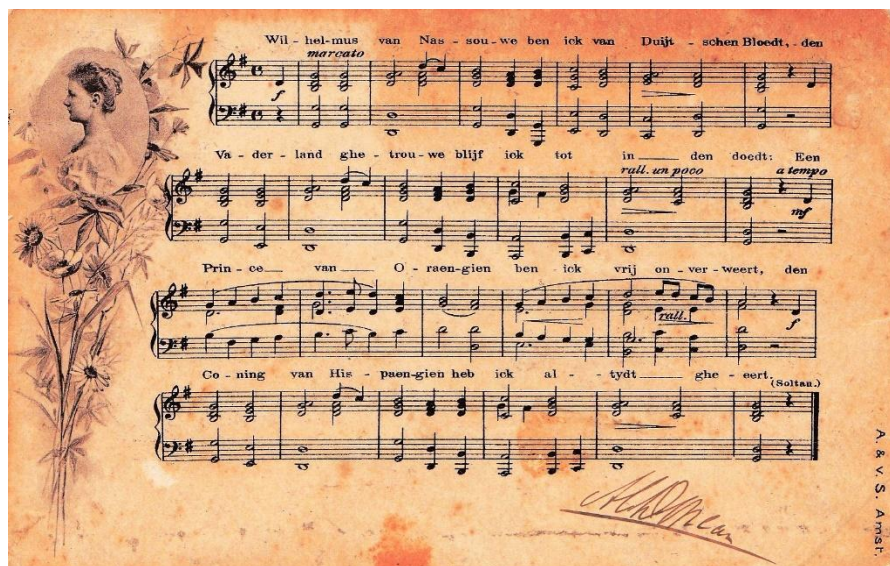
Als we in Delpher op het woord 'moffen' zoeken in de periode 1900-1970 levert dat 23.421 treffers op en het bijgaand afgebeelde grafiekje. We zien de te verwachten piek, maar het leed was kennelijk al weer snel geleden.

Een ander willekeurig voorbeeld is 'chocola*' (asterisk voor -de etc.), nu van de vroegste tot en met de laatste gescande kranten.



'Chocola*', 1618-1995 (Delpher), 523.307 gtreffers.

Nu het principe van het zoeken op woordfrequentie duidelijk is, en de uitvoerbaarheid in de praktijk als het ware een fluitje van een cent is geworden, wordt het tijd voor creatieve toepassingen. De mogelijkheden zijn schier eindeloos, maar ik kan een beginnetje maken door enkele mogelijkheden te laten zien.

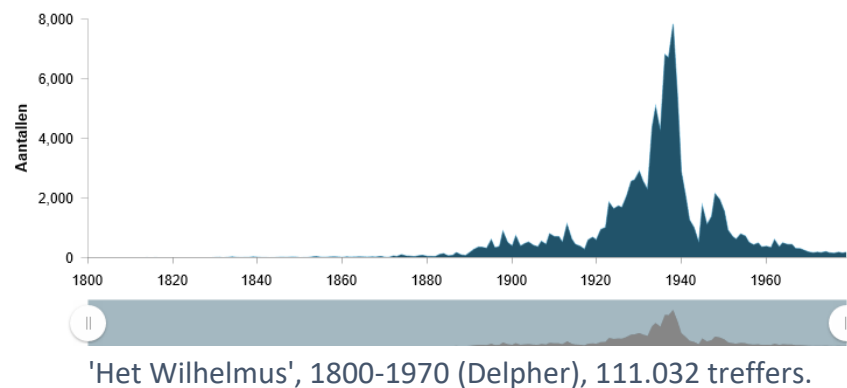


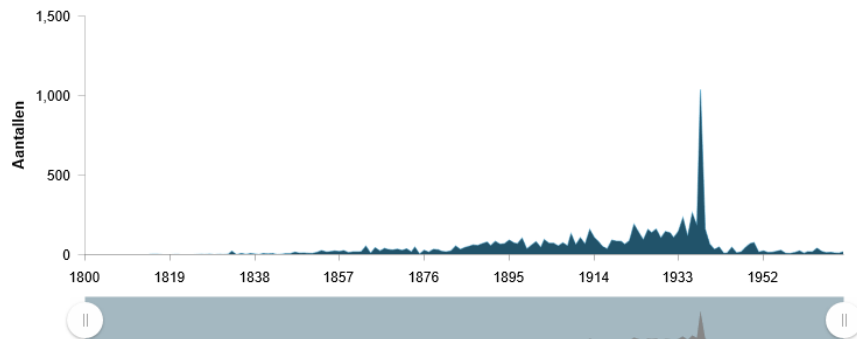
1 Ansichtkaart, 1900.

Laat ik beginnen met drie enigszins controversiële onderwerpen. Allereerst het *Wilhelmus*. In een eerder opstel heb ik de achtergrond en betekenis van het volkslied uitvoerig behandeld. Het is en was onder andere controversieel, omdat het in katholieke (en zuidelijke) kring te gereformeerd 'klonk'. Nadat het in ettelijke Geuzenliedboeken herdrukt was, zocht men aan het begin van de 19^{de} eeuw een alternatief dat ook voor de Belgen in ons kortstondige gezamenlijke koninkrijk aanvaardbaar zou zijn. Die (mislukte) poging leverde van 1815 tot 1932 het volkslied

Wien Neerlands bloed door d'adren vloeit op, dat zelf al meteen controversieel was wegens de - overigens snel gewijzigde - poëtische strofe 'van vreemde smetten vrij'. Dat werd niet als poëtische beeldspraak gezien, maar als raciaal vooroordeel. Dat het *Wilhelmus* ten slotte toch in 1932 de status van officieel volkslied heeft gekregen, naar mijn persoonlijke mening overigens volkomen terecht, is vooral het werk geweest van koningin Wilhelmina, die er bij haar inhuldiging en later nadrukkelijk voor koos.

Laten we daarom in de periode 1800-1970 eens zoeken op 'het wilhelmus' (met lidwoord) en 'neerlands* bloed' (met * wegens -ch). Dat levert twee grafiekjes op.





'Neerlands* bloed', 1800-1970 (Delpher), 8185 treffers.

Opvallend is de toenemende frequentie van de titel 'Het Wilhelmus' in de jaren 1925-1940. De titel 'Wien Neerlands bloed' piekt alleen rond 1940, maar legt het, ook los van die piek, qua aantal vermeldingen (8185) af tegen het huidige volkslied (111.032).

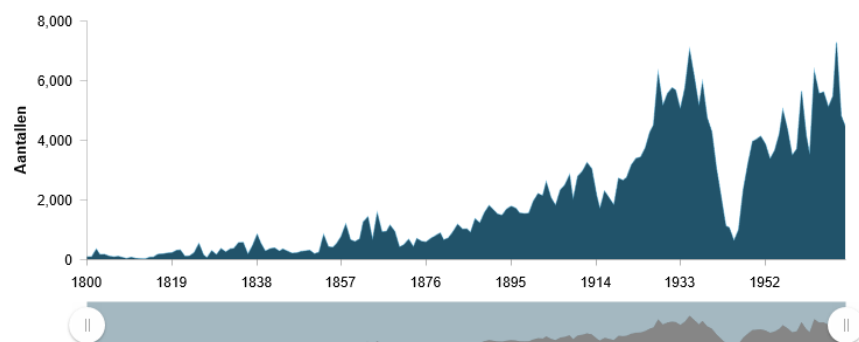
Een tweede controversieel onderwerp dat op deze manier bekeken kan worden is de kwestie Zwarte Piet. De Amerikaanse gevoeligheid met betrekking tot hun binnenlandse slavernijverleden en de ook daarna nog lang voortgezette rassendiscriminatie in niet alleen de zuidelijke staten wordt door sommigen één op één geprojecteerd op de Nederlandse situatie in het gelijktijdige verleden. Hoewel Nederlandse kooplieden tot ongeveer 1860 hebben meegedaan aan de internationale handel in slaven, net als alle handelsnaties en de Afrikaanse naties zelf niet

in de laatste plaats, is verzet tegen de Zwarte Piet-traditie om *die* reden mijns inziens niet op rationele gronden gestoeld. Er woonden tot circa 1970 in ons land vrijwel geen negers, ze behoorden niet tot het straatbeeld.

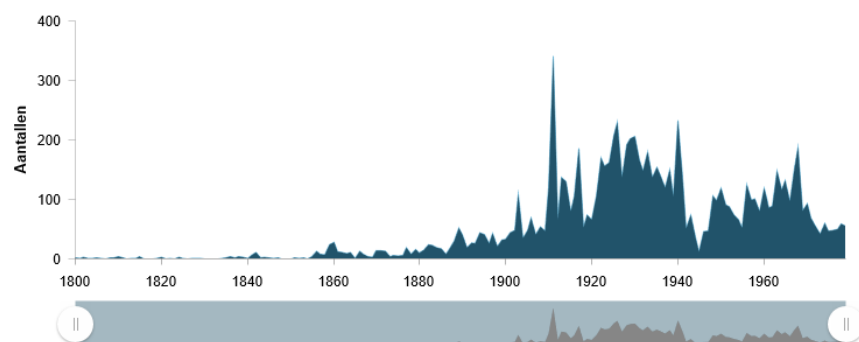
Natuurlijk werd er met een scheef oog naar gekeken, zoals naar alle voyante minderheden in die vervlogen tijden (zoals joden, zigeuners, kermisvolk, losse protestanten in katholieke dorpen, losse katholieken in protestantse dorpen, ongehuwde moeders, reizende schoorsteenvegers en noem maar op), maar raciale discriminatie was dat niet. Institutioneel waren het hier de katholieken en de vrouwen die gediscrimineerd werden, en zeker ook de joden. Er wáren hier geen negers om te discrimineren. Het waren trouwens de westerse landen, waaronder het onze, die uiteindelijk de slavenhandel hebben afgeschaft, omdat er vanaf de aanvang in veel kringen grote weerstand van morele, religieuze en juridische aard tegen bestond. De Nederlandse bevolking achteraf een collectieve schuld aanpraten wegens 'ons' slavernijverleden is mijns inziens een omkering van de werkelijkheid. De westerse landen hebben circa drie eeuwen meegedaan aan een schandalige praktijk van mensenhandel, een commerciële en morele uitwas die

tevooren en daarna eeuwenlang heeft bestaan en voortbestaan. Dat is waar. En sedertdien wordt er al anderhalve eeuw streng tegen alle vormen van slavernij en mensenhandel opgetreden. Dat is ook waar.

In dit verband heb ik gekeken naar de frequentie van 'neger*' en 'nikker*' in de kranten vanaf 1800.



'Neger*', 1800-1970 (Delpher), 325.837 treffers.



'nikker*', 1800-1970 (Delpher), 9278 treffers.

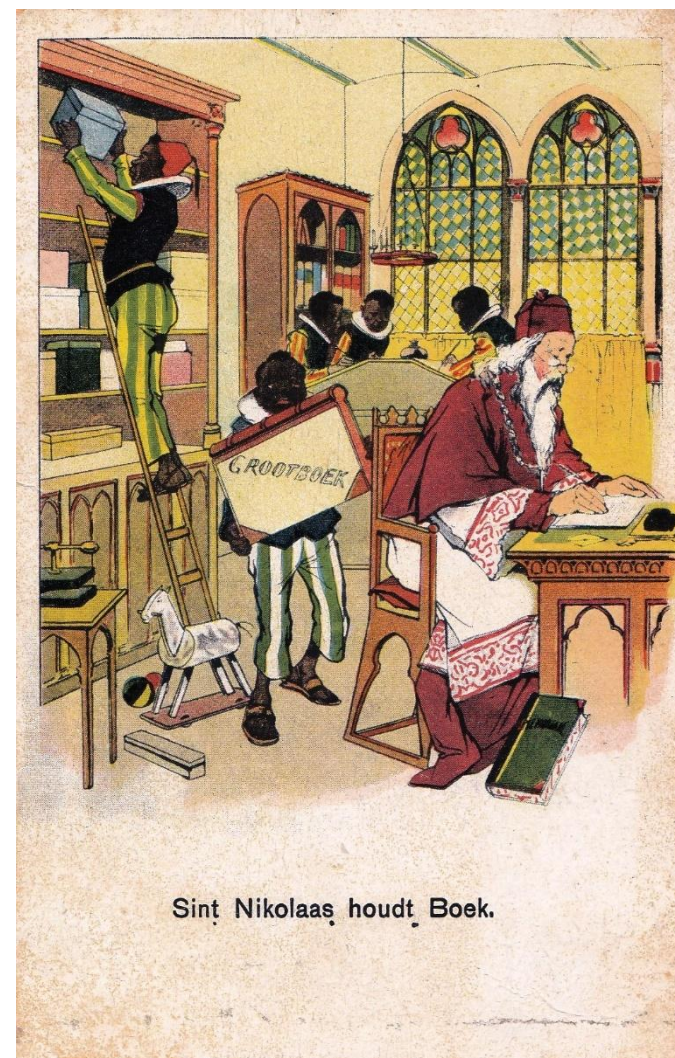
Uit deze twee grafiekjes blijkt dat 'neger' 325.837 keer voorkwam en 'nikker' slechts 9278 keer (waarvan trouwens 1787 keer in het verkleinwoord nikkertje). In beide gevallen zie ik geen opvallende pieken of dalen. Het waren allebei heel gewone woorden. Zo kwam het woord 'jood' (of 'jode') 448.074 keer voor, en 'chinees' 913.366 keer. Daarentegen zien we 'pindachinees' slechts 514 keer, vaak ook als verkleinwoord, maar pas vanaf 1930, toen iemand het kennelijk bedacht heeft. Dit woord was natuurlijk neerbuigend, maar niet beledigend bedoeld.

De betekenis van zulke woorden wordt misschien verhelderd als we er een paar adjectieven bij betrekken. We weten dat het gaat om kranten, dus openbare uitingen. Scheldpartijen trof men daarin niet veel aan. Laat ik vijf positieve en vijf negatieve adjectieven nemen: goede, slimme, mooie, rijke, beroemde resp. slechte, domme, lelijke, arme, onbekende.

Als uit de tabel met deze cijfers over adjectieven al conclusies te trekken zijn over stereotypen in krantenartikelen, dan toch zeker dat de imago's van de vier groepen niet opvallend afwijken van elkaar, behalve dat de Chinees wel erg vaak rijk of slim wordt genoemd en de jood vaak goed. De vraag in hoeverre krantenclichés indicatief

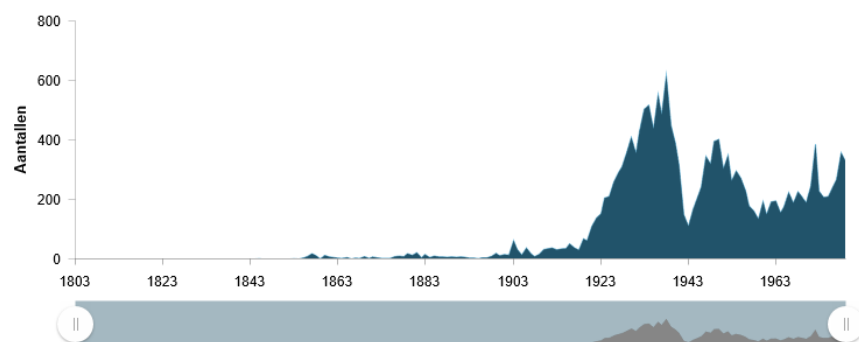
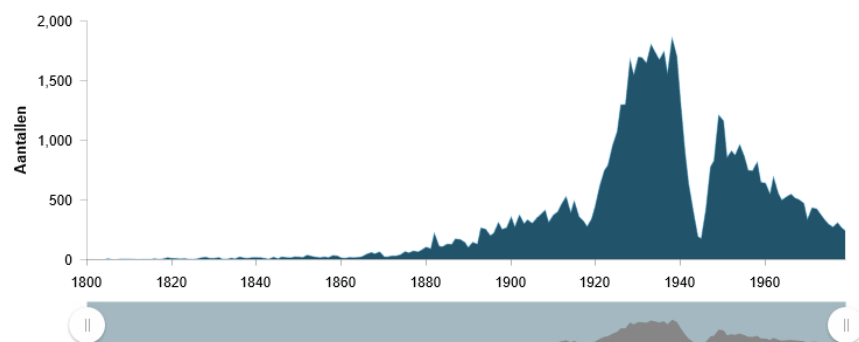
zijn voor discriminatie in de praktijk kan ik niet voor de vuist weg beantwoorden, maar de gegeven cijfers over de relatief schaarse combinaties die in 12 miljoen krantenpagina's van 1800 tot 1970 zijn aangetroffen, geven geen enkele aanleiding om discriminatie van zwarte of bruine mensen te verwachten.

	jood	neger	chinees	nikker
goede	212	29	54	7
slimme	31	25	191	0
mooie	4	39	5	64
rijke	220	34	485	1
beroemde	15	118	7	1
slechte	30	4	5	2
domme	8	36	15	2
lelijke	4	2	0	4
arme	254	236	242	34
onbekende	52	36	210	0
	830	557	1214	65



Ansichtkaart, 1949.

Als derde en laatste controversieel onderwerp dat op basis van de databank van Delpher kan worden belicht, al is deze nog niet helemaal klaar, noem ik Sinterklaas en Zwarte Piet. Om te beginnen kunnen we gewoon even kijken hoe vaak de woordcombinaties 'Sint Nicolaas' (met inbegrip van Sinterklaas) en 'Zwarte Piet' voorkomen.



'Zwarte Piet', 1800-1970 (Delpher), 17.499 treffers.

We zien dat beide woordcombinaties eigenlijk pas na 1900 en vooral na 1920 met behoorlijk hoge frequentie voorkomen, hoewel het beroemde boekje van Jan Schenkman, *Sint Nikolaas en zijn knecht*, al in 1850 was verschenen en vele malen was herdrukt. Na het diepe dal in de grafiek rond de Tweede Wereldoorlog lijkt het alsof het noemen van Sinterklaas geleidelijk afneemt, terwijl Zwarte Piet na 1960 wat frequenter vermeld lijkt te worden. Dat staat dus los van de Pietendiscussie in de afgelopen vijf jaar. Het zou te maken kunnen hebben met het frequenter plaatsvinden van feestelijke intochten van de Sint, op een stoomboot, met vele Pieten in zijn entourage. Dan halen die Pieten de krant in verhouding ook vaker dan vroeger.

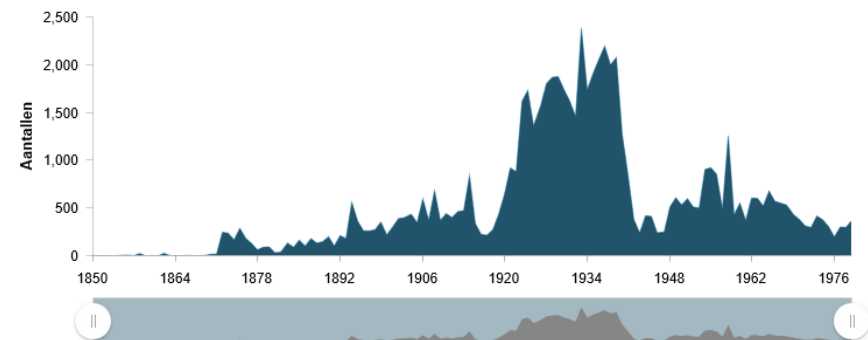
Bij de beoordeling van dit soort cijfers behoort ook eigenlijk een correlatieberekening, al is het maar om een voor de hand liggend verband te bevestigen. Zo heb ik in dit geval de cijfers per decennium voor Amsterdam (14,6 miljoen treffers) vergeleken met die voor Sint Nicolaas (63.672 treffers). Voor de decennia 1850-1860 tot en met 1960-1970 bleek de correlatie .84 te bedragen, dat is zeer hoog. De correlatie tussen de cijfers voor Sint Nicolaas en Zwarte Piet was zelfs .89. Hieruit is indirect af te leiden, dat de frequentie van die namen in al die decennia relatief gezien op het gelijke niveau bleef.

Het laatste is iets wat we ook kunnen constateren als we de decenniumcijfers voor 'Amsterdam' met bijvoorbeeld die voor 'nikker' naast elkaar zetten. Hoewel het bij de tweede groep slechts om 9278 treffers in al die decennia gaat, is de correlatie met 'Amsterdam' maar liefst .92. Dat betekent dat het woord 'nikker' in al die tijd een vrijwel constante aanwezigheid in de kranten heeft gehad, zonder uitschieters naar boven of beneden. Weer geen zichtbare aanwijzing voor een of meer perioden van mogelijke 'discriminatie' dus.



Franse ansichtkaart (Poitiers), 1908.

Er zijn natuurlijk allerlei losse woorden die vragen om een frequentiegrafiek. Neem bijvoorbeeld het Franse dorp Lourdes, waar de veertienjarige Bernadette Soubirous in 1858 diverse verschijningen van de Heilige Maagd kreeg. In 1925 werd ze door de paus zalig, in 1933 heilig verklaard. Hoe snel werd dit een beroemde plek waar honderdduizenden op bedevaart naar toe gingen?



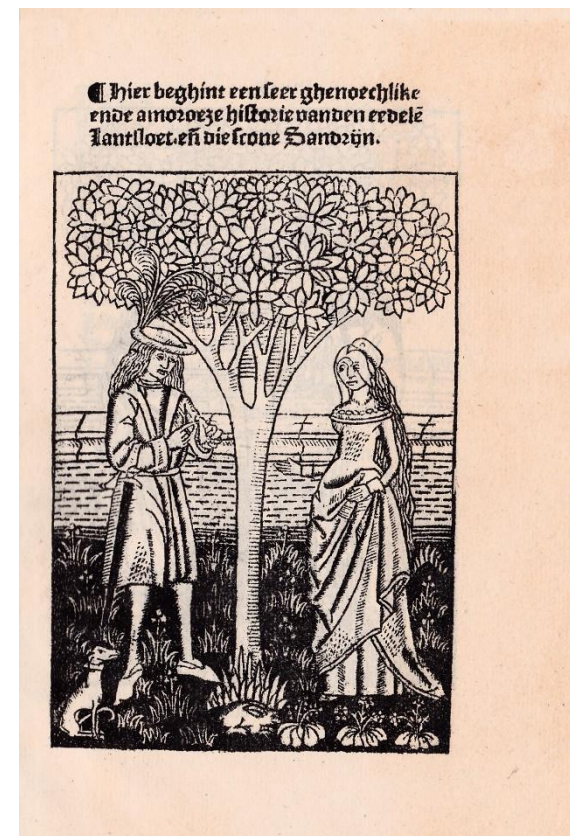
'Lourdes', 1850-1976 (Delpher), 67.868 treffers.

Een heel andere toepassing is voorts het gebruik van de database om kenmerken van een grammaticaal of idiomatisch karakter van een tekst te analyseren of te vergelijken met andere teksten. Heel lang geleden, rond 1990, had ik het in mijn hoofd gehaald om de Goudse druk van *Lantsloet ende Sandrijn* (circa 1490), waarvan Martinus Nijhoff in 1902 een facsimile had uitgegeven, in detail te vergelijken met het abele spel *Lanseloet van Denemerken*

zoals dat in het Hulthemsde handschrift (circa 1410) staat. Ik heb toen beide teksten met de hand overgetypt en daarna een database van de afzonderlijke woorden gemaakt (beide keren bijna 6000). Dat leidde tot constateringën zoals: 'De enkele e voor de lange e-klank komt in LG alleen voor in *crech* (517), *me* (545) en *twe* (702), maar in LH veertien keer, in *her*, *sprect*, *stect*, *nemt*, *ment*, *ontvremt* en dergelijke. De spelling *ei* voor deze klank (in LH vier keer *meinen*) komt in LG niet voor. Wel zien we een keer *meyn*, een keer *alleyn* en zes keer *lytteyken*.' Zo heeft LH tien keer *seldi*, *selen*, maar LG elf keer *suldi*, *sullen*. En zo voorts, zeer gedetailleerd.

Mijn lange artikel bij een voorgenomen editie van de incunabel heeft verder tot niets geleid, mede omdat me kort daarna bleek dat J. Beckers, of een pupil van hem, een soortgelijke studie had ondernomen (ik zag het in *Literatuur*, 1989, nr. 4). Wel heeft die studie - waarvoor ik nog vele andere teksten op hun woordfrequenties en andere kenmerken had bekeken - niet alleen het onderwerp van mijn latere dissertatie mede bepaald, maar bovendien mijn promotor, Herman Pleij, ervan overtuigd dat ik enig verstand van zaken had weten te krijgen. Ik moest dat natuurlijk aantonen, zo had ik voorzien, door een soort doctoraalscriptie in te leveren, want ik had noch

een opleiding als Neerlandicus, noch als geschiedkundige om naar te verwijzen. Wel had ik door mijn studie sociologie kennis van de methoden en technieken van onderzoek, waaronder statistiek. De moeite voor *Lanseloet* is dus niet voor niets geweest en het handmatig *doen* van dit soort dingen is bovendien een uitstekende leerschool.



De digitalisering van teksten is wat oude drukken en handschriften betreft nog lang niet volmaakt. Kies liever de gescande versie van moderne edities, ook al nopen zelfs die nog tot voorzichtigheid. Maar laten we blij zijn met de mogelijkheid om hele lappen tekst, ja, hele oeuvres van soms anonieme auteurs, met enkele toetsindrukken te onderzoeken op woordfrequenties, stijlfiguren, spellingkwesties, eigennamen, geografische details en wat de gek maar kan bedenken.

De redactie van Delpher spreekt zelf terecht van een goudmijn. Ik verwacht dat ik nog in menig opstel uit deze mijn delfstoffen zal putten. Of dat goud zal wezen, moeten we nog maar afwachten. Voorlopig wil ik afsluiten met de vermelding dat ik in dit opstel van 2567 woorden zeven keer 'dus', drie keer 'derhalve' en twee keer 'daarom' heb gebruikt. Dat kan ik in *Word* binnen enkele minuten tellen en gelukkig ook aanpassen (zonder dat iemand het merkt). Maar van de in Delpher opgenomen bronnen kan zoiets nu ook moeiteloos worden geteld in teksten, op *miljoenen pagina's*, die we niet zelf getypt hebben. En dat is een heerlijk vooruitzicht. Dank u, Sinterklaasje!

Peter Cuijpers, 1 juli 2019



Franse ansichtkaart (Ceko), Rotterdams stempel 1931.