Namen: varianten en fouten

WieWasWie bevat heel veel verschillende voornamen en familienamen. Vaak betreft het varianten van dezelfde naam. Door te onderzoeken welke namen voor dezelfde persoon zijn gebruikt kunnen we deze varianten automatisch op het spoor komen. Dat kan leiden tot een goed onderbouwde standaardisatie van namen.

tekst Gerrit Bloothooft

Bij het opstellen van een genealogie proberen we personen te identificeren aan de hand van alle beschikbare gegevens. Persoonsnamen spelen daarbij een centrale rol. Maar wanneer namen heel algemeen zijn of niet exact overeenkomen door varianten, spellingsfouten, aliassen, of door fouten zelfs heel andere namen zijn, dan wordt de beslissing of gegevens op dezelfde persoon betrekking hebben onzeker.

Dat er veel variatie in namen bestaat kunnen we afleiden uit WieWasWie. Daarin is inmiddels een flink deel van de akten van de negentiende-eeuwse burgerlijke stand opgenomen, waaronder de akten die voorheen via Genlias raadpleegbaar waren. We tellen minimaal 681.000 verschillende familienamen en 209.000 verschillende voornamen (aantallen gebaseerd op Genlias, november 2011).

Dat zijn er aanzienlijk meer dan bekend zijn uit *De Nederlandsche Geslachtsnamen* van Winkler (1885), waarin ongeveer tienduizend familienamen worden beschreven, en het *Voornamenboek* van Van der Schaar (1964, eerste druk) met zo'n twintigduizend voornamen, waaronder veel varianten. En zelfs al zijn deze standaardwerken niet volledig, het daarin opgenomen aantal namen is een stuk kleiner dan het aantal verschillende namen dat we in de praktijk tegenkomen.

In het kader van het LINKS project wordt in deze bijdrage een mogelijkheid onderzocht om naamvarianten automatisch af te leiden uit de akten in WieWasWie.

Naamvarianten

We noemen twee verschillende namen variant van elkaar wanneer ze voor dezelfde persoon zijn gebruikt. Als we zeker weten dat de vermeldingen Jannetje Aardema en Jantien Aerdema dezelfde persoon betreffen, dan zijn Jannetje en Jantien, en Aardema en Aerdema varianten van elkaar. Maar als we op deze manier de naamvarianten in WieWasWie willen vinden, dan moeten we eerst gelijke identiteiten vaststellen. De rijkdom aan gegevens in de akten van de burgerlijke stand biedt daarvoor een mogelijkheid. Bij geboorte, huwelijk en overlijden wordt een persoon in de bijbehorende akte in beginsel altijd genoemd met de ouders erbij. Dat betekent een combinatie van drie voornamen en twee familienamen. We gaan er vanuit dat deze naamcombinatie samen met het geboortejaar een unieke persoon beschrijft. Het geboortejaar wordt daarbij met een kleine marge uit de leeftijd bij huwelijk en overlijden afgeleid.

De volgende aanname is dat dit zelfs geldt wanneer maar vier van de vijf betrokken namen gelijk zijn. Eventuele verschillen in de vijfde naam kunnen dan worden beschouwd als echte naamvariatie (want betrekking hebbend op dezelfde persoon). Dit is getest door alle aktecombinaties te selecteren waarbij vijf namen en geboortejaar overeenstemden, en daaruit vervolgens een naam weg te laten. Dat leidde in slechts vijfentachtig van 1,1 miljoen gevallen tot een niet-uniek resultaat. Dat geeft vertrouwen in de methode.

Automatisch oogsten leverde vervolgens 48.584 verschillende variantparen van vrouwelijke voornamen op met Elisabeth – Elizabeth, Willemina – Wilhelmina en Geertrui – Geertruij aan de top. Voor mannelijke voornamen werden er 31.885 verschillende variantparen gevonden waarbij Johannes – Johannis, Jacob – Jakob en Arie – Arij het meest

Hoe namen tot verwarring kunnen leiden: Pieter Brueghel de Oude, zijn zoons Pieter Breughel de Jonge en Jan Brueghel de Oude, en diens zoon Jan Breughel de Jonge, publiek domein en coll. Rijksmuseum

Pieter Brueghel (1526/1530-1569)

Pieter Breughel (1564-1638)

Jan Brueghel (1568-1625)

Jan Breughel (1601-1678)

voorkwamen. Van de 177.258 verschillende familienaamparen waren Jansen – Janssen, Bruin – Bruijn, Ruijter – Ruyter het meest frequent.

Naamfouten

Alles verliep voorspoedig, tot de resultaten wat preciezer werden bekeken. De op deze manier ontdekte naamparen bleken niet altijd aannemelijke varianten van elkaar te zijn, maar konden ook zijn gebaseerd op registratiefouten. Zo vonden we dat Pieter Houtlosser, die volgens zijn huwelijksakte in 1808 werd geboren als zoon van Jacob Houtlosser en Aafje Spruit, volgens zijn overlijdensakte als moeder Grietje Spruit zou hebben. Dit gaf een automatisch afgeleid naampaar Aafje – Grietje, waarvan onze naamkundige kennis zei dat die fout was. Het betrof evenwel dezelfde persoon omdat zowel de huwelijksakte als de overlijdensakte ook nog melding maakten van Aaltje Kort als vrouw van Pieter Houtlosser. Hoewel we hoopten dat er met de beschreven methode geen extra kennis nodig zou zijn om naamvarianten te oogsten, blijkt dit voor het herkennen van fouten toch onmisbaar. We stuiten hier op het verschil tussen een naamvariant en een naamfout. Naamvarianten kunnen naamkundig worden gedefinieerd door een gemeenschappelijk lemma of grondvorm en ze kunnen in alle gevallen worden geaccepteerd als namen voor eenzelfde persoon. Die naamvarianten kunnen bijvoorbeeld zijn gebaseerd op suffixvariatie, spellingsvariatie en verkortingen, maar ook op spelfouten en typefouten. Ook Corenlis – Cornelis is dan een 'variant' naampaar. Foute naamparen hebben daarentegen geen gemeenschappelijke basis en zijn niet generaliseerbaar. Hooguit kunnen we een verschil, zoals tussen Aafje en Grietje, als een registratiefout interpreteren omdat er heel veel aanvullende evidentie is dat het dezelfde persoon betreft.

De realiteit is zelfs nog wat ingewikkelder, want er zijn ook kleinere fouten die resulteren in een bestaande andere naam. Denk bijvoorbeeld aan een leesfout bij digitalisatie waarbij de transcribent kiest voor een verkeerde interpretatie, zoals Fijgje – Sijgje, of een typefout, zoals in Aafje – Aagje, waarbij f en g naastliggende toetsen op het toetsenbord zijn. Dit zijn lastige fouten omdat beide namen bestaan (in tegenstelling tot Aagje – Aabje, al lijken weinig spellingen onbestaanbaar).

De leesfouten betreffen vaak de combinaties: T – F, T – P, T – J, T – S, T – K, F – P, F – J, I – J, M – H, M – W, M – Al. Dit type paren is allemaal handmatig als naamfout aangemerkt omdat ze niet generaliseerbaar zijn. Wanneer ze in een individueel geval toch worden gelijkgesteld, moet daar sterk aanvullend bewijs voor zijn.

Bij het onderscheiden van echte varianten van fouten biedt de frequentie geen hulp. Je zou kunnen denken dat fouten zoals Aafje – Grietje zelden worden gemaakt, terwijl echte varianten juist vaak voorkomen. Dat blijkt niet het geval te zijn. Zowel fouten als echte varianten kunnen zeldzaam of frequent zijn. De meest voorkomende fouten zijn bijvoorbeeld Jacob – Jan, Willem – Jan, Gerrit – Hendrik, Willem – Hendrik, dus verwarringen van veelvoorkomende namen. →

> Alles verliep voorspoedig, tot de resultaten wat preciezer werden bekeken

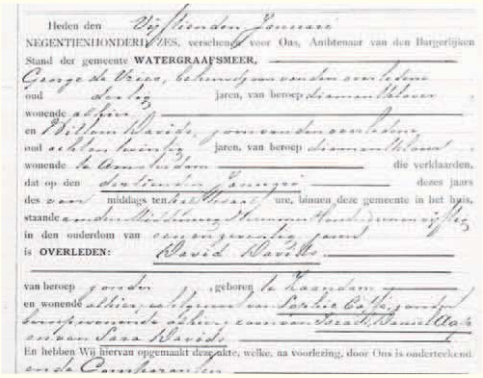

KENNISGEVING.

Bij Koninklijk Besluit van 22 Januari 1886, No. 16, is toegestemd dat de ondergeteekende, DAVID AAP, met al zijn wettige nakomelingen den geslachtsnaam AAP verandere in dien van DAVIDS.

AMSTERDAM, 3 Februari 1886.

(400) D. AAP.


BS Huwelijk met David Aap



Niet iedere afwijking in de spelling is een naamvariant of naamfout. De op 10 februari 1867 in het huwelijk getreden David Aap liet in 1886 bij Koninklijk Besluit zijn familienaam wijzigen. Zijn overlijden te Watergraafsmeer op 13 januari 1906 is daardoor (via de site FamilySearch) terug te vinden onder de naam 'David Davids', coll. CBG en WieWasWie.

Verifiëren van naamparen

Er is een aantal methoden onderzocht om de gevonden naamparen te verifieren. De eerste is om het al genoemde *Voornamenboek* te gebruiken. Dat geeft veel naamvarianten (vaak zonder nadere onderbouwing). Het bevat ongeveer twintigduizend namen die samen 3.737 geslachtsafhankelijke lemma's hebben.

LINKS beoogt een reconstructie van alle negentiende en vroeg twintigsteeeuwse families in Nederland. De basis voor deze reconstructie wordt gevormd door WieWasWie, de toegang op de akten van de burgerlijke stand zoals die in de openbare archieven van Nederland worden bewaard.
De beschikbaarheid van deze dataset biedt een enorm potentieel voor wetenschappelijk onderzoek, mits de individuen aan elkaar worden gelinkt tot families.
Het project is een samenwerking tussen het Internationaal Instituut voor Sociale Geschiedenis, het Leiden Institute of Advanced Computer Science, het Meertens Instituut en de Virtual Knowledge Studio.

Dat verkorte namen vaak meer dan een lemma hebben is niet erg. Lastiger is het dat bepaalde lemma's een onderscheid maken welke in de praktijk niet blijkt, zoals Adagonda, Adelgonde en Aldegonde. Ook associaties van Nelie met Cornelis en Nella met Petronius (via Petronella) lijken te subtiel. Al met al kon het *Voornamenboek* maar vijf procent van onze varianten bevestigen, maar dat zijn wel de meest voorkomende. Een vergelijkbaar familienamenboek (met grondvormen of lemma's) ontbreekt vooralsnog.
Een tweede methode is om het verschil tussen de namen in een paar in een getal vast te leggen. Dat getal geeft het aantal letters dat moet worden veranderd om de ene naam in de andere om te zetten. Bij Dirk – Derck moet een i in een e worden veranderd en een c worden toegevoegd; dat zijn twee veranderingen. De gedachte is dat wanneer het verschil klein is de namen waarschijnlijk variant zijn, en juist niet wanneer het verschil groot is. Dat is overigens ook weer niet altijd waar, zoals we zagen bij de initiaalfouten. Handmatige controle is dus nodig. Deze methode wordt heel veel gebruikt om een inschatting te maken of namen variant van elkaar zijn, maar kan ook soepeler worden gebruikt omdat de geselecteerde naamparen een grotere kans hebben om werkelijk variant van elkaar te zijn. Ongeveer dertig procent van de naamparen kwam niet door deze test heen.

Resultaat

Uiteindelijk resulteerden 34.818 variantparen van vrouwelijke voornamen, 22.478 variantparen van mannelijke voornamen en 120.115 variantparen van familienamen. Omdat verwante namen in allerlei paarcombinaties voorkomen, kunnen uit de naamparen groepen worden gevormd die onderling variant zijn. Dat resulteert in 45.000 voornamen die in circa 2.600 groepen kunnen worden samengenomen, en 115.000 familienamen die in 13.600 groepen zijn opgenomen. De orde van grootte van het aantal groepen komt aardig in de buurt van het aantal voornaamlemma's in Van der Schaar en het aantal familienamen in Winkler.
Dit resultaat kan de basis zijn van standaardisatie van namen. Zo'n naamstandaardisatie is dan gefundeerd in zeer zekere persoonsidentificatie. Dat kan weer veel nut hebben voor het vinden van personen in grote gegevensbestanden en het koppelen van hun gegevens. •

Gerrit Bloothooft is verbonden aan het Utrecht Institute of Linguistics – OTS van de Universiteit Utrecht en participeert in het LINKS project.