Structuur in varianten van voornamen en achternamen
(Clariah NAMES project)

Gerrit Bloothooft, UiL-OTS Utrecht
David Onland, UiL-OTS Utrecht
Martin Reynaert, Tilburg School of Humanities
Katrien Depuydt, INT Leiden
Mathieu Fannee, INT Leiden
Jauco Noordzij, Huygens ING Amsterdam

doelstelling

- beschrijven van de variatie in voornamen en achternamen
- gebruikt voor eenzelfde persoon of binnen een familie (dezelfde entiteit)
- niet noodzakelijk "variant" door gemeenschappelijke grondvorm

Elisabeth

Elly ⟷ ? ⟷ Betje

gebruik

- historische gegevenskoppeling op naam
- na-correctie OCR fouten (digitalisatie)
- zoekfuncties op naam
- naamkunde

domein

- 19e eeuw (wie-was-wie akten burgerlijke stand)
  - 189.000 voornamen
  - 564.000 achternamen
- referentie: basisregistratie personen 2014/2007
  - 608.000 voornamen
  - 320.000 achternamen

referentie materiaal

- voornamen
  - van der Schaar (19.000), met ca 3000 lemma's, en varianten
  - voornamenbank (608.000), bevat Van der Schaar
- achternamen
  - Winkler (10.000), weinig varianten, met globale indeling
  - CBG familienamen (320.000), met netwerkstructuur

lastig te formaliseren

methoden

naamvarianten groeperen onder een 'standaard'

welke standaarden?

- ideaal: naam aan één standaard toekennen
- lukt niet: ambiguïteiten (Jo*, *mina), en onzekerheden

hoe deel je in?

- namen zijn gebruikt voor eenzelfde persoon of binnen een familie (dezelfde entiteit)
- klein spellingsverschil toestaan

varianten van Elisabeth

aanpak

- begin met een flinke deelverzameling
  en probeer die zo goed mogelijk te beschrijven
  - begin met automatische methoden
  - controle door experts
  - ideaal: verzameling zo groot dat alle groepen vertegenwoordigd zijn
    - 44.000 voornamen (98.6% van tokens)
    - 126.000 achternamen (89.0% van tokens)
- voeg de resterende namen automatisch aan de bestaande groepen toe

automatische hulpmiddelen

- (semi) fonetische transcriptie
  - c, k, s
  - i, ie, y, ei
  - s, z
  - f, v
  - f, ph
  - q, k, kw
  - h (ch)
  - …
- edit distance

edit distance

- krachtige methode om overeenkomst tussen twee namen weer te geven met een getal
  - aantal letters *toevoegingen, verwijderingen, veranderingen*

Jan – Joan = 1
Jan – Ian = 1
Johan – Jan = 2
Johannis – Jan = 5

+ extra aandacht eerste letters
+ lengte naam meenemen

inexacte record linkage

- koppelen van akten van de burgerlijke stand
  - akten betreffen dezelfde personen
  - persoon, vader, moeder
    - geboorte, huwelijk, overlijden
  - voornaam persoon
  - voornaam vader
  - voornaam moeder
  - achternaam vader
  - achternaam moeder
  - aktejaar

4 van de 5 namen exact = uniek

(en tijd klopt)

Johannes, zoon van *Pieter Anema* en *Tjitske IJlstra*,
geboren in 1843

Johannes, zoon van *Petrus Anema* en *Tjitske IJlstra*,
huwt in 1868 op 25 jarige leeftijd

vijfde naam genereert variantpaar

Pieter – Petrus (edit distance = 4)

voorbeeld familienaam

- Geboren: *Sophia Joanna Maria* (1842)

  vader: *Karel Eduard Dijk*
  moeder: *Cornelia Sophia Ouburg*

- Huwelijk: *Sophia Joanna Maria Dyk* (22 jaar in 1864)

  vader: *Karel Eduard Dyk*
  moeder: *Cornelia Sophia Ouburg*

  edit distance = 2

voorbeeld verkorte voornaam

- Geboren: *Splinter* (1815)

  vader: *Willem van der Horst*
  moeder: *Johanna van Rossum*

- Overlijden: *Splinter* (1815)

  vader: *Willem van der Horst*
  moeder: *Hanna van Rossum*

  edit distance = 2

voorbeeld Latijnse vorm

- Geboren: Hubertus Hoofwijk (1874)

  vader: *Franciscus Hoofwijk*
  moeder: *Mechtildis Veders*

- Overlijden: *Hubertus Hoofwijk* (14 jaar in 1888)

  vader: *Frans Hoofwijk*
  moeder: *Mechtildis Veders*

  edit distance = 5

naamparen controleren

- fouten in de registraties

  > *Pieter*, geboren in 1808 als zoon van *Jacob Houtlosser* en *Aafje Spruit*, maar zijn overlijdensakte meldt *Grietje Spruit* als zijn moeder wat leidt tot het naampaar *Aafje / Grietje*

  > achternaam *Van Heekeren van Well* vs *Van Heekeren van Kell* als verschrijving of typefout , geen serieuze variant

  > maar ook hoogfrequent *Jacob / Jan, Jacobus / Johannes, Willem / Jan*

  > leesfouten betreffen vaak
  T – F, T – P, T – J, T – S, T – K, F – P, F – J, I – J, M – H, M – W, M – Al
  lastig als dat leidt tot andere bestaande naam

  > typefouten eveneens lastig, naburige toetsen
  Bos - Vos

vraagt handmatige controle

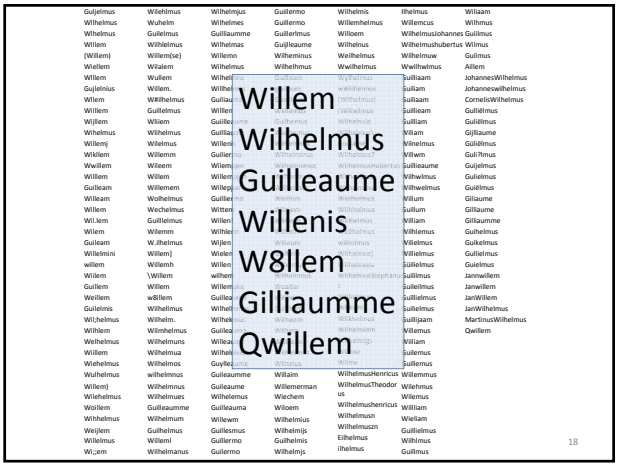
naamparen clusteren

namen die onderling veel als variant optreden
kunnen geclusterd worden tot een groep

problemen

- namen die tot meerdere groepen gerekend zouden kunnen worden
- onvoldoende naamparen beschikbaar om groepen samen te nemen
  (vooral bij achternamen)

vraagt handmatige controle



Names project

(1) startverzameling als 'gouden standaard'
- handmatige controle LINKS standaarden
- handmatige controle relatie naam – standaard

(2) statistisch beschrijven van de varianten (TICCL, Martin Reynaert)

(3) overige namen zo goed mogelijk automatisch aan standaard(en) toewijzen met TICCL (op basis van edit distance)

(4) mogelijkheden onderzoeken voor detaillering van de standaarden

(5) resultaten als linked-open-data beschikbaar stellen

crowd sourcing

- variantparen verzamelen in Voornamenbank en CBG familienamen
  - + tijdindicatie
  - + brongegevens
  - + toelichting

als er tijd is

naamstandaardisatie

vier kwaliteitniveaus

1 – 'gouden standaard': de naam is bekend uit de standaardisatieprocedure op basis van inexacte matching (Bloothooft & Schraagen, 2015)

2 – de naam heeft een *semi-fonetische* vorm die dezelfde is als een naam onder 1)

3 – de naam voldoet aan de volgende eisen (*semi-fonetische* vorm):
a) de lengte moet groter dan 5 zijn
b) kleine Levenshtein afstand met namen onder 1+2 met dezelfde grondvorm

4 – minstens 4 beginletters van de naam (*semi-fonetische* vorm) stemmen overeen met namen uit 1+2 met dezelfde grondvorm

standaardisatie van voornamen

	aantal namen	% namen	aantal tokens	% tokens
type 1	43,673	23.1	62,165,022	**98.60**
type 2	27,134	14.3	123,052	0.20
type 3	26,879	14.2	130,328	0.21
type 4	34,454	19.2	228,585	0.36
som	132,140	70.8	62,646,978	**99.37**
rest	57,036	29.2	390,788	0.63
totaal	189,176	100.0	63,037,766	100.00

standaardisatie van familienamen

	aantal namen	% namen	aantal tokens	% tokens
type 1	126,469	22.4	48,663,863	**89.05**
type 2	80,637	14.3	848,140	1.55
type 3	21,395	3.8	212,436	0.39
type 4	97,647	17.3	390,588	0.71
som	336,148	57,8	50,115,027	**91.71**
rest	237,869	42.2	4,529,551	8.29
totaal	564,017	100.0	54,644,578	100.00