Large scale harvesting of variants of proper names

Gerrit Bloothooft, UiL-OTS, Utrecht University
Marijn Schraagen, LIACS, Leiden University
The Netherlands

g.bloothooft@uu.nl
m.p.schraagen@liacs.leidenuniv.nl

name variants

- different versions of a name, that can denote the same object
- requires a proof that the same object is involved (in at least one example)
  - not always easy
  - rarely explicitly provided

proper names in historical sources

Lots of variation

- spelling variation — *Dirk* — *- Dircк*
- suffix variation — *Willem* — *- Willempje*
- abbreviation — *Willem* — *- Wim*
- translation — *Willem* — *- Guillaume*; *Willem* — *- Wilhelmus*
- typos (digitization) — *Willem* — *- Aillem*
- …

variation!

challenge

name variation is difficult to model, therefore:

- learn variation in person names from use of names in real life
  (let data speak for itself)
- automatically from big data

required

- big data
  - with many references to individuals
- true person resolution
  - proof that the same individual is concerned
  - even with data that contain name variants

big data

- Dutch vital registration (who-was-who 2011)
  1811- early 20th century
  - 4.1 million birth certificates (~30%)
  - 3.1 million marriage certificates (~90%)
  - 7.6 million death certificates (~65%)

  55 million name references to persons

source names

1,052,000 different full first names (composite)
Jan, Johanna Maria Cornelia

111,900 different female first names (singular, *Maria*)
82,700 different male first names (singular, *Jan*)

681,000 different surnames (prefixes included)
Bakker, de Vries
600.000 different surnames (prefixes excluded)
Vries

information per person

- first name person (child, bride or groom, deceased)
- first name father
- surname father
- first name mother
- surname mother (always maiden name in The Netherlands)
- age person

person resolution

- assumption: the available information identifies a person uniquely (if there is exact matching)
- relaxed assumption: one of the first names and surnames of the mother or father is not needed for true person resolution

example

Johanna Endt

- marries in 1858 as 29 years old daughter of *Gerrit Endt* and *Dorothea Kerbert*
- dies in 1882 as 54 years old daughter of *Gerrit Endt* and *Doortje Kerbert*

~1829, *Johanna*, *Gerrit*, *Endt*, *Kerbert*, *Dorothea*
~1828, *Johanna*, *Gerrit*, *Endt*, *Kerbert*, *Doortje*

test of assumption
(of true person resolution)

- consider all matches between birth and death certificates with exact matching of all information
- leave out one name per match
- count number of multiple matches

result:
only 85 out of 1,107,162 matches are not unique

harvesting name variant pairs
(procedure)

- identify all record pairs of individuals (over birth, marriage and death certificates) that exactly share
  - first name of the individual
  - approximate year of birth
  - three out of four names of parents (first names and surnames)
- collect pairs of the remaining name, if different

  Christiena – Christina
  Bloothooft - Bloothoofd

harvesting name variant pairs
(results)

female first names	48,600 pairs	246,500 tokens
male first names	31,900 pairs	183,000 tokens
surnames	177,000 pairs	374,900 tokens

average:
first names: 5 to 6 tokens per variant pair
surnames: 2 tokens per variant pair

so far so good, but

- the original certificates are not error-free

  > found variants can be due to errors in the source, during transcription or to typos

- theoretical issue:
  what is a name variant, and what is an error?

example

in the source documents:

Pieter
born as son of *Jacob Houtlosser* and *Aafje Spruit*,
died as son of *Jacob Houtlosser* and *Grietje Spruit*

variant *Aafje – Grietje ?*

variants and errors

distinction is difficult to make

- variants share the same lemma and errors do not

requires onomastic expertise
(which we would like to avoid, let the data speak for itself)

variants and errors

- Variants
  Willem - Wilhelm
  Willem - Guillaume
  Willem - W8llem (no indication of different lemma)
- Errors
  Grietje - Aafje
  Fijtje - Sijtje (understandable reading error but different lemma)

methods for cleaning

- using name dictionaries with lemmas
  - to accept name pairs
- using known non-variants
  - to reject name pairs
- rules
  - to accept name pairs

all with manual intervention (< 2%)

cleaning | name dictionaries

- dictionary of Dutch first names (20,000), but
  - lemmas too detailed
  - names with multiple lemmas
  - only 8% of all first name pairs share lemma in dictionary (43 % of tokens)

results, in variant pairs

- female first name pairs
  34,800 accepted 13,900 errors (29%)
- male first name pairs
  22,500 accepted 9,400 errors (29%)
- surnames pairs
  120,100 accepted 57,100 errors (32%)

very many variant pairs (*Willemina*)

name clusters

- variant pairs (are interconnected)
  Jan - Johannes
  Jan - Joannes
  Jan - Johan
  Johannes – Johan, etc
- create cluster Jan {Jan, Johannes, Johan}

name clusters

- male first names 1.221 (16.487 names, 20%)
- female first names 1.530 (23.816 names, 21%)

compares to number of lemma's in Dutch dictionary of first names, vd Schaar 1964

- surnames 11.686 (93.839 names, 17%)

compares to number in Dutch surnames overview (without many variants), Winkler 1885

conclusions

- person name variants need proof from true person links
- expert knowledge necessary because errors cannot be distinguished fully automatically from true variants (but < 2%)
- final results are promising as a starting point to create a national repository of proven name variants