Learning name variants from true person resolution

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

the problem

- True name variation with large edit distance
  - suffix variation *Antje - Annigje*
  - abbreviation *Jan - Johannes*
  - translation *Willem - Guillaume*
- Name errors with small edit distance *Bos – Vos*
  - misreading as known name in transcription

the solution

- List of true name variants
  - seen for the same individual
  - no errors in this list
- Edit distance / similarity measures for small variation
  - used critically

learning from data

- Name variation that is found for *the same* person
  - requires true person resolution

data

- Dutch civil 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 references to persons

information per person

- First name person
- First name father
- Surname father
- First name mother
- Surname mother (allways maiden name)
- Age person

Assumption: *a name 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

- 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: 85 out of 1,107,162 matches

harvesting name variants

- parsing composite names
  Anna Christiena Elizabeth / Christina Elizabeth

female first names	48,684 pairs	246,519 tokens
male first names	31,885 pairs	183,050 tokens
surnames	177,258 pairs	374,901 tokens

- much less surname variant tokens
- order insensitive

variants and errors

- Name variation in true person resolution can be erroneous
  - in the source document:
    Pieter born as son of *Jacob Houtlosser* and *Aafje Spruit*, but died as son of *Grietje Spruit*
    > *Aafje* / *Grietje*
  - misreading / mistyping during digitisation

variants and errors

- distinction is difficult to make automatically
- onomastics: variants share same lemma (and errors do not)
  - requires expertise
- true variation is difficult to model
- typo is considered a true error

frequency is of little help

- errors can be frequent
  - *Jan / Jacob, Gerrit / Cornelis, Willem / Hendrik*
- true variants can be rare

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

cleaning | name dictionaries

- dictionary of first names (20,000), but
  - lemmas too detailed
  - names with multiple lemmas

  - 3,610 female first name pairs share lemma
  - 2,877 male first name pairs share lemma
  - 8% of all pairs (43 % of tokens) accepted

cleaning | non-variants

- assumption: no variants in composite name
- harvest all name combinations from composite names (*Anna Maria Helena* > *Anna / Maria, Anna / Helena, Maria / Helena*)
- assumption incorrect:
  Jan / Johannes, Neeltje / Cornelia, Arie / Adrianus
  manual correction needed

4,701 first name pairs identified as errors
6 % of all pairs (4% of tokens) rejected

cleaning | rules

- semi–phonetic conversion
  Jozeph > JOSEF
  takes away most true variation in initial
- acceptance rules
  - Levenshtein distance (1- 5)
  - requirements on number of equal initials
  - compares Jaro-Winkler but more relaxed
- manual correction needed

results in variant pairs

- female first names
  34,818 accepted 13,866 errors
- male first names
  22,478 accepted 9,408 errors
- surnames
  120,115 accepted 57,143 errors

29% first name pair errors (11% tokens)
32% surname pair errors (21% tokens)
note: *these are errors in true person resolution (and include typos)*

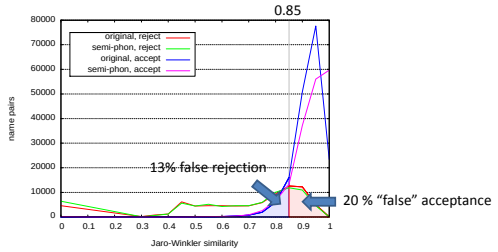
results

compared to Levenshtein distance

- accepted pairs and Lv > 2 (*false rejection*)
  - female first names 15.7 %
    male first names 11.4 %
    surnames 7.0 %
- erroneous pairs and Lv <3 (*"false" acceptance*)
  - female first names 39.0 %
    male first names 49.0 %
    surnames 43.0 %

results

compared to Jaro-Winkler similarity (all pairs)



names involved

- female first names 47% of 61,873
  male first names 38% of 52,964
  surnames 22 % of 569,063
- not detected:
  - names not seen under requirements
  - names always written the same way without variant per individual

conclusions

- knowledge of name variants is needed aside use of edit distance or similarity measures
- true name variants (that can be generalized) can only be derived from true person links
- expert knowledge cannot be missed because errors cannot be distinguished automatically from true variants

conclusions

- many true name errors (including typos) can be found in true person resolution
- but *they can be circumvented* if there is sufficient identifying information (as in this research)

future work

- cluster names from pairs and *standardize*
- apply standards to all names in civil registration
- *harvest new name pairs*
- test name standards in linkage tasks