NAMES

a Clariah research project

CLARIAH
Common Lab Research Infrastructure for the Arts and Humanities

Gerrit Bloothooft, David Onland & Richard Oosterlaken
UiL-OTS Utrecht
Martin Reynaert, CS&AI / Tilburg University
Katrien Depuydt & Tanneke Schoonheim, INT Leiden

PROBLEM

spelling variation, errors, and variants of person names in (historical) documents

AIM

standardization of person names for

richer search results
OCR post-processing
nominal record linkage
onomastic research

FULL MATERIAL

564.000 surnames and 189.000 first names (NAMES corpus)
from 19th century sources (63 million tokens from Catch LINKS project - Wiewaswie)

AVAILABLE SEED

proven variants (from true person resolution in LINKS project)
based on 328.411 surname variant pairs and 134.220 first name variant pairs

EXTRACTION OF VARIANTS AND STANDARDS FROM EXTERNAL SOURCES

surnames in Belgium (Debrabandere, 125.000 names, 49.000 in NAMES, 19.000 standards)
corpus of Dutch surnames (CBG, 320.000 names, 59.000 in NAMES, 11.800 standards)
first name dictionary (van der Schaar, 20.000 names, 12.500 in NAMES, 2.400 standards)

EXPERT REVIEW OF SEED

surnames: 127.000 into 11.539 standards
first names: 42.000 into 926 gender-independent standards (or ~500 first syllable based)

APPLICATION OF CLARIN TICCL (Text-Induced Corpus CLean up)

learning statistics from proven variant pairs
automatic standardization of remaining 417.000 surnames and 147.000 first names
- by adding names to expert standards
- to cluster names into additional standards

ISSUES

Errors: dependent on writing, reading, typing (with different statistics)

Ambiguity: possibility to assign a variant to various standards

Edit distance: needs weighting (to be learned from proven variants) in context

Standardization: optimization of ratio of number of proven variant pairs within and between standards

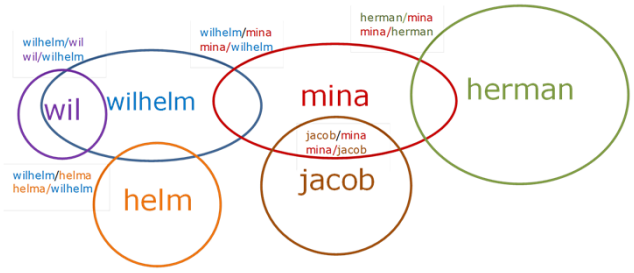
Different levels of standardization: high recall versus high precision

(COMPLEX) EDIT OPERATIONS IN CONTEXT

but how to weigh these in the computation of edit distance

character(s)	context	occurrences	fraction	example
female first name - insertion				
N	E,#	24959	0,48	Aaltje – Aaltjen
E	I,N	5774	0,16	Rina – Riena
H	O,A	5364	0,45	Joanna – Johanna
OH	J,A	3039	0,26	Janna – Johanna
DA	I,#	1493	0,14	Ali – Alida
HA	T,R	1376	0,13	Catrina – Catharina
female first name - substitution				
S>Z	I,A	15085	0,59	Lisa – Liza
A>E	N,#	14557	0,12	Anne – Anna
J>I	T,E	5373	0,05	Antje – Antie
A>TJE	N,#	4348	0,04	Anna – Antje
LE>HEL	L,M	2581	0,21	Willemien – Wilhelmien
ET>IG	N,J	2126	0,37	Annetje – Annigje
female first name - transposition				
I,E	R,N	133	0,04	Rientje - Reintje
R,E	D,K	75	0,32	Henderkien – Hendrekien
A,N	I,#	70	0,47	Christian - Christina
surname - insertion				
N	E,#	15417	0,11	Linde – Linden
S	R,#	8127	0,16	Kuiper – Kuipers
J	I,N	7401	0,31	Bruin – Bruijn
EN	S,#	2158	0,05	Jans – Jansen
CH	S,#	1666	0,54	Bos – Bosch
DAL	N,#	11	0,02	Bloemen – Bloemendal
surname - substitution				
R>N	E,#	3887	0,05	Kortleven – Kortlever
T>D	R,#	3488	0,24	Weert – Weerd
G>K	N,#	2644	0,13	Wenting – Wentink
G>CH	E,T	823	0,46	Knegt – Knecht
Z>JS	I,E	616	0,08	Keizer – Keijser
RITS>SEN	R,#	2	0,00	Gerrits – Gersen
surname - transposition				
R,U	B,G	230	0,11	Verbrug – Verburg
S,T	R,#	83	0,02	Voorts – Voorst
I,E	R,N	75	0,05	Rienders – Reinders

OVERLAP BETWEEN STANDARDS TO BE ACCOMMODATED IN DATA STRUCTURE



MULTIPLE STANDARD OPTIONS

name	*NAMES-standard(s)*
Hermina	**herman**/mina
Helma	**helm**/wilhelm
Mientje	**mina**/herman/wilhelm/jacob
Mina	**mina**/herman/wilhelm/jacob
Mynou	**mijnou**
Willemina	**wilhelm**/wil/mina
Wil	**wil**/wilhelm
Wim	**wilhelm**
Willem	**wilhelm**/wil
Guillaume	**wilhelm**

RESULTING NAMES CORPUS

in RDF format for Linked Open Data
& lexicon service