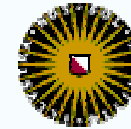




Koppelen van persoonsnamen uit historische bestanden

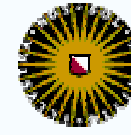
Gerrit Bloothoof
UiL-OTS
Universiteit Utrecht



Het probleem

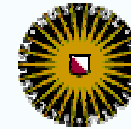
Bijeenbrengen van

- vermeldingen van één persoon,
- in één of meerdere bronnen,
- die elektronisch beschikbaar zijn,
- waarbij de naam niet steeds op dezelfde manier geschreven is



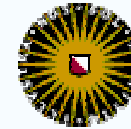
Interdisciplinair karakter

- Naamkunde
- Geschiedenis
- Taalwetenschap
- Computacionele taalkunde
- Kunstmatige intelligentie
- Databasetechnologie



Bronnen van variatie

- Opgave van naam door individu
 - Afhankelijk van doel (kerkelijk, zakelijk, informeel, ...)
 - Sociale wenselijkheid
- Op schrift stellen
 - Luisterfouten, interpretatiefouten
 - Spellingsvariatie
- Lezen van de bron
 - Leesfouten (overeenkomst geschreven letters)
 - Interpretatiefout (verwarring van naamonderdelen)
- Intypen van de naam in database
 - Typefouten (misslagen toetsenbord)



Een voorbeeld

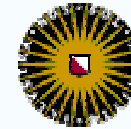
Een familienaam in 17e eeuwse bronnen

Beecque,	del	Delbeecque
Becque,	de la	Delbeque
Beke,	de le	Delbeke
Beque,		
Be r que,		
Beecq,		
Beeck,		



Invalshoeken

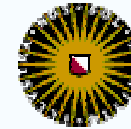
- naamstandaardisatie
 - ambiguiteiten
 - onbekende namen en varianten
- kansberekening
 - wat is de kans twee namen op dezelfde persoon betrekking hebben
 - oplossingen in context



Onbekende namen

% voornamen in bron genoemd in vd Schaar:

- Poorters Amsterdam 1531-1611 **38%** 1.725 / 24k
- DTB Amsterdam 1776-1811 **19%** 11.826 /343k
- Brabant var. 17-19e eeuw **39%** 5.674 /584k
- 1947 volkstelling steekproef **64%** 4.584 / 86k

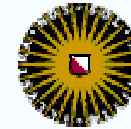


Zinvolle standaardisatie

normaliseert naamvarianten die *in de praktijk*
- in dezelfde tijd - door elkaar gebruikt worden

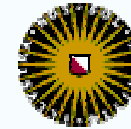
probleem

- grondvorm is dan lang niet altijd zinvol
- er is nauwelijks informatie over contemporaine varianten



Beperkte standaardisatie is nuttig

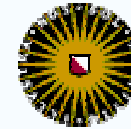
- Spellingsvereenvoudiging
 - semi-fonetisch
- Suffixreductie
 - affixreductie bij voornamen
 - patroniemreductie tot voornaam
 - aangehecht voorvoegsel familienamen



Semi-fonetische vorm

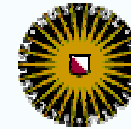
(voorbeelden)

- j Rajmund **raimund**
- tweeklank Andreas **andræs**, Oaries **årys**
- ph Alphons **alfons**
- z Atze **atse**
- q Quirina **kwirina**
- c Oscar **oskar**, Alice **alise**
- x Alexia **aleksia**
- ck Rijcklof **rijklof**
- ch Christiaan **kristiaan**, Aardsche **aardse**
- h Theun **teun**, Deborah **debora**
- #i Ian **jan**



Resultaat semi-fonetische conversie

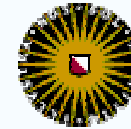
- winst
reduceert het aantal varianten met 1/3
- probleem
sommige significante verschillen gaan verloren
- maar overgeneralisatie is meestal geen probleem en zelfs wenselijk
 - de onderzoeker beslist uiteindelijk zelf
 - overgeneralisatie biedt keuzen



Morfologische decompositie van voornamen

- voornamen
 - Mannen: 541 affixen + 2516 regels
 - Vrouwen: 1337 affixen + 5496 regels
- voorbeeld
 - regel: tse# -> t-se
Aetse -> **Aet-se**

- problemen
 - stam-affix scheiding in trainingsmateriaal
 - vinden van een efficiënte regelmethode
 - informatieverlies (affix kan identificeren)



Patroniemreductie

>> 223 – 771 varianten

Probleem: is affix patronymisch of niet

-man, -ma en -ing vormen

- Voorbeeld

Bruyns

Bruyning

Bruyne

Bruynes

Bruyninckx

Bruynel

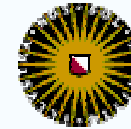
Bruynensz

Bruyninsen

Bruynsma

Bruyneszoonsz

Bruyntsz



Parsing van familienaam voorvoegsels

Familienaam repertorium (1947):
135 typen losse voorvoegsels

Vandenberg

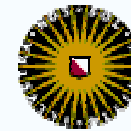
Van den | Berg

Lermite

L' Hermite

probleem

- spellingsvorm hoofdnaam hoeft niet los te bestaan



Oplossing in context

- Wanneer meer dan alleen de naam bekend is in één record
 - Voornaam
 - Patroniem
 - Familiennaam
 - Beroep
 - Herkomst
 - Datum
 - ...

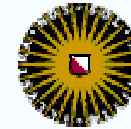


Parsing van velden

Probleem door onzekere interpretatie en invoer in database

Een voorbeeld (een koopman)

voornaam	patroniem	famnaam
Jan	Jacobsen	Benning
Jan, de jonge		Benningh
Jan	(Benning	Coeckebakker)
Jan	Jacobsz. (Benninck	Coeckebakker)
Jan	Jacobsz	(Banninck Couckebakker)



Gegevenskoppeling met complexe relaties

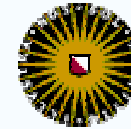
- Meerdere namen en familierelaties in één records
 - man, vrouw, ouders, kinderen (in één record)
 - met soms onzekerheid over die relaties
- Expertsysteem met geïntegreerde naamanalyse
 - laat vaak ruimere naaminterpretatie toe



Disciplines

1

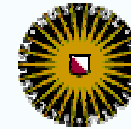
- Naamkunde
 - Kennis van naamvariatie
 - Kennis van contemporaine varianten
- Geschiedenis
 - Kennis van personen, achtergronden
- Taalwetenschap
 - Methoden voor fonetische standaardisatie en morfologische decompositie



Disciplines

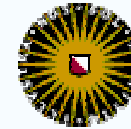
2

- **Computationele taalkunde**
 - Algoritmen voor regelafleiding en regeltoepassing
 - Waarschijnlijkheidsberekeningen
- **Kunstmatige intelligentie**
 - Combineren van kennis
 - Redeneren en interpretatie
- **Database technologie**
 - Efficiënte opslag van informatie
 - Zoekstrategiën



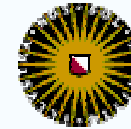
Noodzakelijke basisgegevens

- Verrijkte databases (als kennisbron en voor training van probabilistische systemen)
 - voornamen en patroniemen
 - stam - affix
 - familienamen
 - voorvoegsel – hoofdnaam
 - beroepen en toponiemen
 - *bekend* als familienamen
- Voor alle namen
 - *bewezen* varianten
 - bron, plaats, datum, standaardvorm(en)



Nodig

- Nationale gegevensverzameling
 - Uit verspreide bronnen
 - Verrijking
- Onderzoek naar koppelings / zoekstrategiën
 - Algoritmen
 - Kansberekening
- Gecoördineerde samenwerking vanuit verschillende disciplines



Belang

- Toegankelijkheid van historisch materiaal dat in grote hoeveelheden elektronisch beschikbaar is en komt

En

- Problematiek vrijwel gelijk bij koppeling van moderne databases