RSNSR: Rule based search in text databases with nonstandard orthography

or

In the beginning there was the word…
…but who knows how it was spelled?

Andrea Ernst-Gerlach
Thomas Pilz
University of Duisburg-Essen
Overview

• Project environment and origins of RSNSR
• The problem of spelling variants
• The rule-based approach
• Rule derivation
• Outlook
The environment

- **Projekt Nietzsche-CD**
  - University of Duisburg-Essen
  - Reception of Nietzsche (1865 – 1945)

- **Deutsch Diachron Digital**
  - Humboldt-University of Berlin
  - German textual tradition (– 1900)

- **Deutsches Rechtswörterbuch**
  - Heidelberg Academy of Sciences
  - Juristic Vocabulary (– 1800)

- **Compact Memory**
  - RWTH Aachen
  - Jewish periodicals (1806 – 1966)
Origins of RSNSR

- Projekt
- A web-based system for assisted literature research
- Partial text recognition of German Fraktur documents; document image de-warping
What is RSNSR about?

• Development of a
  – rule-based
  – fuzzy
  search-engine for
  – literature research of historical texts


Berlinische Privilegirte Zeitung, 23.01.1748
Variant spellings

Frequency of variant spellings in historical text documents

Year

%
Variant spellings

gehéiligôt  giheiliget  geheiliget  geheiliget  geheiliget

1000  1200  1300  1400  1500  1600  1700  1800  1900

ghehilghet  geheyliget

Himele  Hemmelen  Hymel  Himmel

Himele  Himile  Hymel  Himmel
Where does the variability come from?

• German texts prior to 1901 or quotations of those are not orthographically standardized.
• OCR of German Fraktur font is error-prone

*Niehsches Ringen um Gott.*

Zur Beleuchtung des erneuten Einflusses Friedrich Niehsches.

Von Dr. Eberhard Arnold, Berlin-Wilmersdorf.
What are we dealing with?

- Inconsistency of spellings even of the same author
- n:m relations between $n$-graphs
  \[ \text{Dompropst} - \text{Thuemmbbröbst} \]
- Circles
  \[ \text{Fluss} \rightarrow \text{Fluß} \rightarrow \text{Fluss} \]
- Phonetic classification of some $n$-graphs is time-dependent \[ \text{allzu} - \text{allzvo} \]
What are we dealing with?

• Orthographical variation to a certain degree depends
  – on phonetic realisation,
  – on the authors origin (time / place).

• OCR-Errors depend
  – on graphical similarities
The rule-based approach

Query

Dictionary

Statistical preprocessing

Rule-based variation

Distance-measure
What do the rules look like?

• Rules
  – Standard regular expressions (java.util.regex)
  – Separation of context and transformation
  – Utility constants (*vowel sound*...)
  – Reiteration-factor
  – Weights
  – Metadata (Timeframe, Location)
What do the rules look like?

• Allograph-transformations
  – Adjustable direction/productivity
  – Historical vs. modern
  – Unicode support
  – Correctness
  – Bi-/Tri-/n-graphs

example
Which problems are involved?

• Different paths lead to same variants
  ursache – ursahe – ursah_
  ursache – ursach_ – ursah_

• Indefinite repetition of application
  Fluß – Fluss – Fluß – Fluss ...

• Rule sideeffects
Where do the rules come from?

Manual derivation

- Retrieve evidences (EviDancer)
- Sort evidences in groups
- Search for similarities in context
- Generate rules
- Re-evaluate rules: visualisation
Where do the rules come from?

Automatic derivation

- Retrieve evidences
- Generate rule core
- Generate rule candidates
- Rule pruning
Generate rule core

• Training set of triplets
  – Contemporary word form
  – Full word form
  – Collection frequency

• Find necessary transformation e. g.
  unnütz → unnuts
  _unn(ü,u)t  t(z,s)_
Generate rule candidates

• Successively adding context to rule cores
  – e. g. unnütz → unnuts
    ü → u  nü → nu  üt → ut

• Abstracting of context
  – Consonant sounds (C)
    e. g. Cü → Cu
  – Vowel sounds (V)
  – Word beginning (^) / ending ($)
    e. g. z$ → s$
Rule Pruning: PRISM Algorithm

- **PRISM:**
  - Classifies set of instances into set of classes
  - Instances are fixed sets of attributes with values from nominal scale
  - Tries to generate high precision values for each class \( C \) identifying instances belonging to \( C \)

- **Extension necessary, because:**
  - Perfect rules are not aimed at
  - PRISM generates rules independent from examples
Rule Pruning: PRISM Extension

- Generate negative examples
- Sort instances by rules itself
- Search for words where rule $i$ matches
- Calculate occurrence frequency $q_i$
- Calculate precision $p_i$
- Remove all instances

where $p_i < p_{\text{min}} \lor q_i < q_{\text{min}}$
Evaluation: Collection

• 64290 word tokens
• 11326 different word types
• 717 types with historic spelling
• 2/3 training set
• 1/3 test set
Evaluation results

• Recall 0.5
• Precision 0.77
• Frequently used rules:

<table>
<thead>
<tr>
<th>Rules</th>
<th>Frequency</th>
<th>Precision</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>t → th</td>
<td>113</td>
<td>0.99</td>
<td>Einteilung – Eintheilung</td>
</tr>
<tr>
<td>ä → ae</td>
<td>42</td>
<td>0.98</td>
<td>Ämter - Aemter</td>
</tr>
<tr>
<td>s → ß</td>
<td>32</td>
<td>0.94</td>
<td>aus – auß</td>
</tr>
<tr>
<td>k → c</td>
<td>24</td>
<td>0.96</td>
<td>Kollegien – Collegien</td>
</tr>
<tr>
<td>ü → ue</td>
<td>19</td>
<td>0.86</td>
<td>Übertragung - Uebertragung</td>
</tr>
<tr>
<td>ä → ai</td>
<td>18</td>
<td>1</td>
<td>souverän - souverain</td>
</tr>
</tbody>
</table>
Outlook: Automatic Derivation

- Application of more than one rule
- Combination with linguistic approaches
e. g. o → od or t → dt
- Devide compound search terms / syllables
- Utilisation of a part of speech tagger
Outlook

• Development of a consistent user-interface
• Enhancement of the automatic derivation
• Evaluation of variant generation in contrast to distance-measures
• Collection of more evidences
• Fuzzy logic for rule weights
Thank you for your interest!

Any questions?