Information Acces to Historical Documents from the Early High German Period

Christian Meyer
21.02.2011

Projektseminar „Unlocking the secrets of the past“
Caroline Sporleder

Content

- Introduction
- What's the big Deal?
- The Way to Salvation
- Conclusion

Introduction

Goals:

- Preservation of historical documents
- Facilitation of public access to historical content
- Search the documents

  → Digitization

Problem:

- No normalized spelling

Introduction

Computational Approach:

- Determine orthography variants
- How to adapt methods accordingly
- Merge them into an algorithm

What's the big Deal?

Different Classes of linguistic variations:

- Phonological/Graphical

Grapheme	Variants
<a>	< á, å, ah, aa, ai, ae, a̋ >
<e>	< eh, ee, ei, ey, å, e̋, ä >

- Morphological

  Licht-e vs. Licht-er

- Lexical

  Urlaub: permission → leisure time

- Syntactical

What's the big Deal?

8 types of problems:

- New word form
  handeln → marcken
- Latin words
- Word splitting
  Winterzeit → Winter Zeit
- Partial Variation
  Großteil → Mehrteil
- Variaton of prefixes/suffixes
  Kindchen → Kindlein
- Typesetting variations
- Graphemic-phonetic variations
  Abertheur → Abenteuer
- New character

The Way to Salvation (by example of IR)

Generally:

- Modern to historic dictionary

- Rule-based generative matching

  (i → y)

  Phonetic Normalization

- Matching by Word similarity

→ Merge all this together

The Way to Salvation (by example of IR)

Levenshtein distance:

- Three operations
  - Insertion
  - Deletion
  - Substitution
- Each operation has a weight
- Similarity of two sequences of characters determined by the sum of weights

For our purpose: modify Levenshtein weights

The Way to Salvation (by example of IR)

How the 3 approaches influence each other:

- Derive rules from dictionary
- Derive rules from Levenshtein weights
- Derive Levenshtein weights from rules

Conclusion

- Quite sophisticated Solution
- Pretty good results

What's left?

- Improvement of OCR
- Figuring out a working XML representation and query method