Querying the *Deutsches Textarchiv*

Bryan Jurish, Christian Thomas, Frank Wiegand
Deutsches Textarchiv / Berlin-Brandenburgische Akademie der Wissenschaften

{jurish,thomas,wiegand}@bbaw.de

Beyond Single-Shot Text Queries:
Bridging the Gap(s) Between Research Communities (MindTheGap’14)

Berlin, 4th March 2014
Overview

The Big Picture
- The Situation
- The Problem
- The Approach

Methods
- Corpus Preprocessing
  - canonicalization, PoS-tagging, lemmatization, . . .
- Index Structure
  - token attributes, equivalence caches
- Runtime Query Interpreter
  - term expansion pipelines

Burning Questions

Summary & Conclusion
— The Big Picture —
Deutsches Textarchiv (DTA)

"German Text Archive" ["DAY–TAY–AH"]

http://www.deutschestextarchiv.de

Deutsches Textarchiv

Grundlage für ein Referenzkorpus der Neuhochdeutschen Sprache


Um den historischen Sprachstand möglichst genau abzubilden, werden als Vorlage für die Digitalisierung in der Regel die Erstausgaben der Werke zugrunde gelegt: Das elektronische Volltextkorpus des DTA ist über das Internet frei zugänglich und dank seiner Aufbereitung durch (computer-)linguistische Methoden schreibweisentolerant über den gesamten jeweils verfügbaren Bestand durchsuchbar. Sämtliche Texte stehen zum Download zur Verfügung.

mehr...
DTA: Project Sketch

- **Goal:** digitize >1300 historical print volumes
  - earliest accessible editions of respective works (1600–1900)
  - detailed metadata, highly accurate transcriptions

- **TEI-XML corpus encoding & storage**
  - DTA base-format (DTABf) dialect

- **Corpus Extensions & Cooperations (DTAE)**
  - CLARIN-D KurPro, Dingler’s *Polytech. Journal, Grenzboten*

- **Linguistic annotation** *(CAB)*

- **Spelling-tolerant online full-text search** *(DDC)*

**Some Numbers** *(DTA+DTAE 2014-02-25)*

- 1,727 transcribed works
- 494,248 digitized pages
- 817,187,143 unicode characters
- 117,574,654 tokens (alpha-numeric)
- 2,855,830 types (alpha-numeric)
The Situation

Historical Text ≠ Orthographic Conventions

- high variance of graphemic forms
  - fröhlich “joyful”
    - frölich, fröhlich, vroelich, froelich, frölich, frölich, vroelich, frölich, frölich, . . .
  - Herzenleid “heart-sorrow”
    - hertzenleid, herzenleit, hertzenleyd, hertzenleidt, herzenlaid, hertzenlaidt, hertzenlaydt, herzenleyd, . . .

Conventional Tools ⇒ Strict Orthography

- Fixed lexicon keyed by (ortho)graphic form
- Extant lexemes only
  - bareword modern-form queries

The Problem

- corpus variants missing from application- or user-lexicon
- low coverage, poor recall, spurious “noise”, etc.
The Approach: Canonicalization


In a Nutshell

- Map each word \( w \) to a unique canonical cognate \( \tilde{w} \)
- Defer application processing to canonical forms

Canonical Cognates

- Synchronically active “extant equivalent(s)” \( \tilde{w} \in \text{Lex} \)
- Preserve both root and relevant features of input

Conflation Relation \( \sim_r \)

- Binary relation on strings (words) in \( \Sigma^* \)
- Prototypically a true equivalence relation
— Methods —
Corpus Preprocessing

Serialization & Tokenization
- Linguistically salient units (words, sentences)
- Serial order ≠ XML document order (footnotes, headings, . . .)

Canonicalization
- “Exception” Lexicon
  \[ w \mapsto \arg \max \tilde{w} f(w, \tilde{w}) \]
- Deterministic Transliteration
  \[ w \mapsto \text{xlit}^*(w) \]
- Phonetic Equivalence
  \[ w \mapsto \{ \tilde{w} : \text{pho}(w) = \text{pho}(\tilde{w}) \} \]
- Rewrite Cascade
  \[ w \mapsto \arg \min \tilde{w} \llbracket M_{rw} \circ M_{Lex} \rrbracket (w, \tilde{w}) \]
- HMM Disambiguator

Conventional NLP Analysis
- Part-of-Speech Tagging
  \( (\text{DeRose}, 1988) \)
- Morphological Analysis, Lemmatization
  \( (\text{Geyken & Hanneforth}, 2006) \)
- Named Entity Recognition
  \( (\text{Didakowski}, 2008) \)
Index Structure

Quasi-Independent Token Attributes

- **Utf8Token** \( (u) \): raw token text (narrow transcription)
- **Token** \( (w) \): transliterated token text (keyboard-friendly)
- ** CanonicalToken** \( (v) \): estimated canonical cognate
- **Pos** \( (p) \): part-of-speech tag (STTS)
- **Lemma** \( (l) \): citation form
- ... diverse assorted administrative and text-structural attributes

Equivalence Caches

\[
\tilde{f}_{a,b} = (f^{-1}_a \circ C) \circ f_b : A \rightarrow \varnothing(f_b(C))
\]

- Map actual attribute values to associated word types
- Also suitable for use with conventional text-only index
- Used to compute ‘extensional’ equivalence with respect to \( a \)

\[
\sim_a = f_a \circ \tilde{f}_{a,\text{ext}} : w \mapsto \{ \tilde{w} \in \text{text}(C) : f_a(w) = f_a(\tilde{w}) \} 
\]
Token-Level Dependencies

- Allow Boolean conjunctions over distinct attributes
  \((@\text{Cooper} \text{ WITH } $p=\text{NN}) \text{ vs. } (@\text{Cooper} \text{ WITH } $p=\text{NE})\)
- Allow set-valued query terms (disjunction)

\[
[T] = \bigcup_{t \in T} [t]
\]

Term Expansion Pipelines

- Map user query terms to corpus equivalents (term conflation)
- External term-set operators (HTTP)
- Implicit attribute-dependent defaults
  - for naïve bareword queries
  - efficient computation using equivalence caches
- Allow explicit overrides for expert users
- Allow shell-style “pipelines” for flexibility

\[
[x | y] = [x] \circ [y]
\]

Default text-based expander

\[
x_{\text{Token}} = f_{\text{Lemma}} \circ \tilde{f}_{\text{Lemma,Token}}
\]
Accessibility Extensions

Query Auto-Completion

- Attribute-sensitive prefix search; *Lemma* attribute by default

\[ \text{Zeitg} \mapsto \{ \text{Zeitgenosse, Zeitgeist, Zeitgeschichte, ...} \} \]

- Inserts appropriate query syntax ($l=@$)
- Likely responsible for high ratio of lemma searches (31%)

Query “Lizard”

- Expansion pipeline visualization, fine-tuning, debugging
- Not very popular: only 321 accesses in 5 months [beta]
Burning Questions

Does it work?

[source: Jurish & Ast, 2012]

<table>
<thead>
<tr>
<th></th>
<th>% Types</th>
<th></th>
<th></th>
<th></th>
<th>% Tokens</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>pr</td>
<td>rc</td>
<td>F</td>
<td>pr</td>
<td>rc</td>
<td>F</td>
<td></td>
</tr>
<tr>
<td>pr</td>
<td>99.1</td>
<td>55.7</td>
<td>71.3</td>
<td>99.8</td>
<td>78.5</td>
<td>87.9</td>
<td></td>
</tr>
<tr>
<td>rc</td>
<td>99.0</td>
<td>87.8</td>
<td>93.1</td>
<td>99.8</td>
<td>98.5</td>
<td>99.2</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>98.3</td>
<td>93.6</td>
<td>95.9</td>
<td>99.6</td>
<td>98.5</td>
<td>99.1</td>
<td></td>
</tr>
<tr>
<td>hmm</td>
<td>98.6</td>
<td>95.7</td>
<td>97.1</td>
<td>99.8</td>
<td>99.3</td>
<td>99.5</td>
<td></td>
</tr>
<tr>
<td>hmm+lex</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- artificial task on \(\sim114\)K-word evaluation corpus (1780–1880)
- **yes**: eliminates >90% type-wise recall errors (>96% token-wise)

Do we really need it?

- “Ambiguity rate” of modern forms: 1.29 (all) vs. 2.62 \((f \geq 10)\)
- Of 29,410 user queries (September 2013 – January 2014):
  - 19,299 (66%) were barewords or bareword phrases
  - 9,219 (31%) were explicit searches of the Lemma attribute
- **yes**: \(\sim97\%\) naïve queries, only 56% string-identities (79% tok)
Concluding Remarks

Historical Text and Conventional Tools

*don’t play together nicely “out of the box”*

Full-Text Search

- canonicalization preprocessing $\leadsto$ modern forms
- equivalence caching $\leadsto$ corpus lexicon
- token-level dependencies $\leadsto$ expert queries
- expansion pipelines $\leadsto$ flexible term conflation

Future Work

- estimate utility for actual user searches
- compare canonicalization preprocessing to expansion-only method(s)
- semantic term expanders
he Olde Laste Slyde
(“The End”)

Thank you for listening!