Finding Canonical Forms for Historical German Text

Bryan Jurish

jurish@bbaw.de

Berlin-Brandenburgische Akademie der Wissenschaften
Jägerstrasse 22/23 · 10117 Berlin · Germany

September 30, 2008

Overview

The Big Picture

- **The Situation**: unconventional text corpora
- **The Problem**: conventional tools $\rightsquigarrow$ low coverage
- **The Proposal**: conflation & canonical form(s)

Conflation Methods

- Phonetic Identity
- Lemma Instantiation Heuristics

Concluding Remarks

- Next Steps
- Summary

The Big Picture

The Situation: Corpora

"Unconventional" Text Corpora

- ***Historical text***
- Spoken language transcriptions
- OCR output
- Non-standard dialects

Lexical "Conventions"

- Extinct or dialect-specific lexemes
- Require manual attention

Orthographic Conventions

- Extinct or dialect-specific lexical variants
- Can be handled automatically (to some extent)

The Situation: Text Technologies

Conventional Text Technologies

- Document indexers
- Part-of-speech taggers
- Word stemmers
- Morphological analyzers

Common Characteristics

- Fixed lexicon accessed *via* orthographic form
- Extant lexemes only

Desideratum

- Apply existing tools to “unconventional” corpora

. . . but . . .

The Problem

Conventional Tools + Unconventional Corpus = Soup

- Corpus variants missing from application lexicon
- Low coverage, poor recall, degraded accuracy, . . .

Examples

Source: *Deutsches Wörterbuch (DWB)*: Bartz et al., 2004

- *ir keinr nam war,*
  wa ieder lag am rangen
- *da sah ich sitzen siben frawen*
  radweisz umb einen külen brunnen.
- *vil manige sêle er zuhte*
  dem tiuvel ûʒ sînem rachen.
- *genuoge wurden verbrant,*
  versteinet und mit swerte erslagen

The Proposal

Conflation & Canonical Form(s)

- Collect variant forms into equivalence classes
- Represent classes by (extant) canonical elements

Analysis by Disjunction

- Analyze "extinct" form w by disjunction over extant members of its equivalence class $[w]$:

$$\text{analyses}(w) := \bigcup_{v \in [w]} \text{analyses}(v)$$

- Expect improved recall, some loss of precision

. . . A Case in Point

Base Corpus

- Verse quotations from DWB (Bartz et al., 2004)
- 6,581,501 tokens of 322,271 graphemic types
- Indexed with TAXI corpus indexing system

Preprocessing & Filtering

- UTF-8 $\longrightarrow$ ISO-8859-1 (*e.g.* $œ \mapsto oe$, $\overset{e}{o} \mapsto ö$, $ô \mapsto o$, . . .)
- ***removed*** non-alphabetic & foreign material
- 5,491,982 tokens of 318,383 graphemic types

Conventional Analysis

- TAGH morphology FST (Geyken & Hanneforth, 2006)

Conflation Methods

Phonetic Conflation: Sketch

Idea

- Map each word w to a unique phonetic form $\mathrm{pho}(w)$
- Conflate words with identical phonetic forms

$$[w]_{\mathrm{pho}} := \{v : \mathrm{pho}(v) = \mathrm{pho}(w)\}$$

Phonetization: Letter-to-Sound (LTS) Conversion

- Well-known in text-to-speech (TTS) research
- `ims_german` LTS rule-set (Möhler et al., 2001)
  - for `festival` TTS system (Black & Taylor, 1997)
  - slightly modified for historical input
  - converted to finite-state transducer (FST)

$\rightsquigarrow$ over ***5.5 times faster*** than `festival`

Phonetic Conflation: Coverage

	Types	Tokens
Total	318,383	5,491,982
+TAGH	42.4 %	83.7 %
+TAGH / pho	***54.6 %***	***91.5 %***
Error Reduction	21.1 %	48.2 %

Phonetic Conflation: Problems

Insufficient **(too permissive)**

- Phonetic Identity $\not\Rightarrow$ Lexical Equivalence
- ***Precision Errors*** (conflated but not equivalent)
  - *(hân–Hahn), (niht–Niet), (vil–fiel), (usz–Uhus), . . .*
- Not too dangerous (yet)

Unnecessary **(too strict)**

- Phonetic Identity $\not\Leftarrow$ Lexical Equivalence
- ***Recall Errors*** (equivalent but not conflated)
  - *(guot–gut), (pflag–pflegte), (tiuvel–Teufel), (umb–um), . . .*
- This is the **more severe** of the two problems!

Lemma Instantiation: Sketch

Idea

- Exploit dictionary-corpus structure
- Assume each quote contains an instance of the associated dictionary lemma

String Edit Distance (Levenshtein, 1966; Baroni et al., 2002)

- Relax strict identity criterion

Pointwise Mutual Information (McGill, 1955; Church & Hanks, 1990)

- Filter out “random” phonetic similarities

Restrict Comparisons

- Compare only lemma-instance pairs
- Over ***10 thousand times faster*** (*vs.* all word pairs)

Lemma Instantiation: Coverage

	Types	Tokens
+TAGH	42.4 %	83.7 %
+TAGH / pho	54.6 %	91.5 %
+TAGH / li	***66.7 %***	***94.4 %***

Error Reduction

▸ *vs.* **TAGH / pho**	26.7 %	33.8 %
▸ *vs.* **TAGH**	42.2 %	65.8 %

Examples: Phonetic Conflation

sieben

da sah ich sitzen siben frawen
radweisz umb einen külen brunnen.

kühlen

*viel, *fiel* *Seele*

vil manige sêle er zuhte
dem tiuvel ûz sînem rachen.

aus *seinem*

ihr *nahm* **war* *wahr*

ir keinr nam war
wa ieder lag am rangen.

**Ider*

verbrannt

genuoge wurden verbrant,
versteinet und mit swerte erslagen

Examples: Lemma Instantiation

sieben

da sah ich sitzen siben frawen
radweisz umb einen külen brunnen.

radweise kühlen

viel, *fiel Seele zuckte

vil manige sêle er zuhte
dem tiuvel ûz sînem rachen.

Teufel aus seinem

ihr nahm *war, wahr

ir keinr nam war
wa ieder lag am rangen.

*Ider, jeder

verbrannt

genuoge wurden verbrant,
versteinet und mit swerte erslagen

?Schwert

Concluding Remarks

Next Steps

Test Corpus

- Manually constructed gold standard
- *circa* 11,000 tokens; 4,000 types
- Quantitative analysis: precision & recall
- **Status:** 99% done (pending expert review)

Robust Rewrite Cascades

- Weighted finite-state transducer cascades
  - Generalized edit distance
- "Lazy" best-path lookup
- **Status:** Beta (`gfsmxl`, TAXI/DTA)

Summary

Problem

- Historical text corpora and conventional tools *don't play together nicely*

Proposal

- Conflate lexical variants into equivalence classes
  - . . . by phonetic identity
  - . . . and/or by lemma-instantiation heuristics

Results

- 94.4% tokens covered $\rightsquigarrow$ 65.8% fewer errors

The End

Thank you for listening!