Constructing a Canonicalized Corpus of Historical German by Text Alignment

Bryan Jurish,
Marko Drotschmann,
Henriette Ast

Deutsches Textarchiv
Berlin-Brandenburgische Akademie der Wissenschaften
http://deutschestextarchiv.de

New Methods in Historical Corpora
Manchester, 29th-30th April, 2011

Overview

The Big Picture

- Canonicalization
- Desiderata
- Proposal

Construction

- Sources
- Text Alignment
- Manual Annotation

Applications

- Test Corpus
- Canonicalization Lexicon

Conclusion

— The Big Picture —

Canonicalization

a.k.a. (orthographic) 'standardization', 'normalization', 'modernization', . . .

The Problem

- Historical text $\not\ni$ orthographic conventions
- Conventional NLP tools $\Rightarrow$ strict orthography
  - ***Fixed lexicon*** keyed by ***orthographic form***
  - ***Extant*** lexemes only

þe	Olde	Wydgette	Shoppe
↓	↓	↓	↓
the	*old*	*widget*	*shop*

The Approach

- *Map* each word w to a unique canonical cognate $\widetilde{w}$
  - Synchronically active "extant equivalents"
  - Preserve both ***root*** and ***relevant features*** of input
- *Defer* application analysis to canonical forms

Desiderata

Evaluation

- **Compare** various canonicalization functions $c(\cdot)$
- **Task**: information retrieval $\Longrightarrow$ ***(precision, recall)***
- **Retrieval** via canonical equivalence:

$$\mathrm{retrieved}_c(w) := \left[c \circ c^{-1}\right](w) = \left\{v : c(v) = c(w)\right\}$$

- **Relevance** requires manual verification!

$$\mathrm{relevant}(w) := \textbf{?}$$

Ground-Truth Corpus

- Manually verified canonicalization pairs $(w, \widetilde{w})$
- "Gold standard" $\widehat{c}(\cdot)$ for training & evaluation

$$\mathrm{relevant}(w) := \mathrm{retrieved}_{\widehat{c}}(w) = \{v : \widetilde{v} = \widetilde{w}\}$$

- Minimize manual annotation effort

. . . but how?

Proposal

Intuitions

- Contemporary editions of historical works
  $\Longrightarrow$ ***already standardized***
- Expect mostly identity canonicalizations $w = \widetilde{w}$
  (at least for 18th-19th century German)

Construction (Sketch)

- ***Align*** historical text with a contemporary edition
  - maximize identity alignments
- ***Confirm*** or ***Reject*** type-wise alignments
  - exploit Heaps' Law
- ***Manually annotate*** only unconfirmed tokens
  - don't lose "interesting" anomalous material

— Construction —

Sources

Text Resources

- Source texts: *Deutsches Textarchiv (DTA)*
  - *Belles lettres*, drama, verse, philosophy
- Target texts: `gutenberg.org`, `zeno.org`

Prototype Corpus

- 13 volumes, published 1780–1880
- ca. 350k tokens $\sim$ 28k types *(words only)*

Ongoing Construction ('full' corpus)

- 129 volumes, published 1780–1901
- ca. 5.2M tokens $\sim$ 219k types *(words only)*

Text Alignment

Preprocessing

- Tokenization (1 word / line)
- Transliteration e.g. *(ſ ↦ s)*, *(oͤ ↦ ö)*

Basic Alignment

- Token-wise LCS (GNU `diff`)
- $> 77\%$ identity, $> 94\%$ transliterated identity

Heuristic Alignment

- For each change change hunk
  - multi-token alignments e.g. *(zwei und vierzig ↦ zweiundvierzig)*
  - character-wise 'best' match (Levenshtein)

Type-wise Confirmation

Idea

- Manually ***confirm*** or ***reject*** non-identity alignments
- Exploit Heaps' Law
  - vocabulary grows *logarithmically* with corpus size
- Conservative acceptance only

Results (prototype corpus)

- **Available:** 18k tokens $\sim$ 5.8k types
- **Confirmed:** 16k tokens (90%) $\sim$ 4.5k types (77%)

Throughput

- ca. 3.95 seconds / pair $\approx$ 15 words / second

Token-wise Annotation

Idea

- Resolve remaining uncanonicalized tokens (ca. 2%)
- Retain anomalous canonicalization patterns

Preprocessing Filters

- Block pruning ($\approx$ 2.2%)
- Closed-class lexicon

Annotations

- Canonical form + administrative flags
- Expert review for problematic cases
- Throughput (total) $\approx$ 1.3 words / second

— Experiments —

Materials

Prototype Corpus $\rightsquigarrow$ Ground-Truth Relevance

- Most thoroughly annotated corpus subset
- 340k tokens; 29k types (words only)

Full Corpus $\rightsquigarrow$ Canonicalization Lexicon (LEX)

$$\mathsf{LEX}(w) = \begin{cases} \arg\max_{\widetilde{w}} f(w, \widetilde{w}) & \text{if } f(w) > 0 \\ w & \text{otherwise} \end{cases}$$

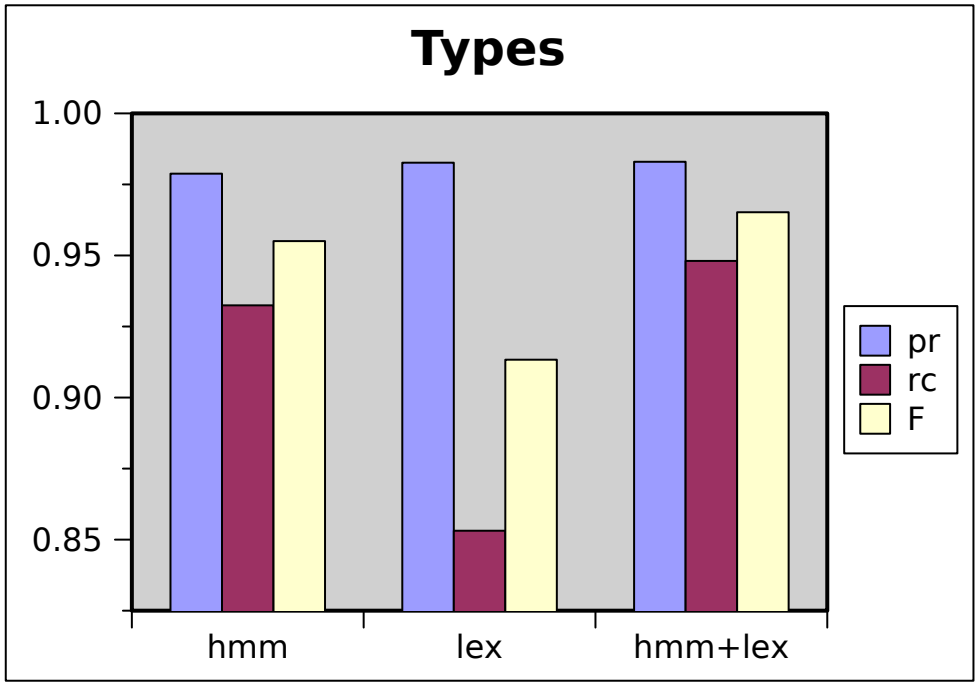
- Strictly disjoint from test corpus (by author)
- Partially annotated (no expert review)
- 2.4M tokens; 140k types (words only)

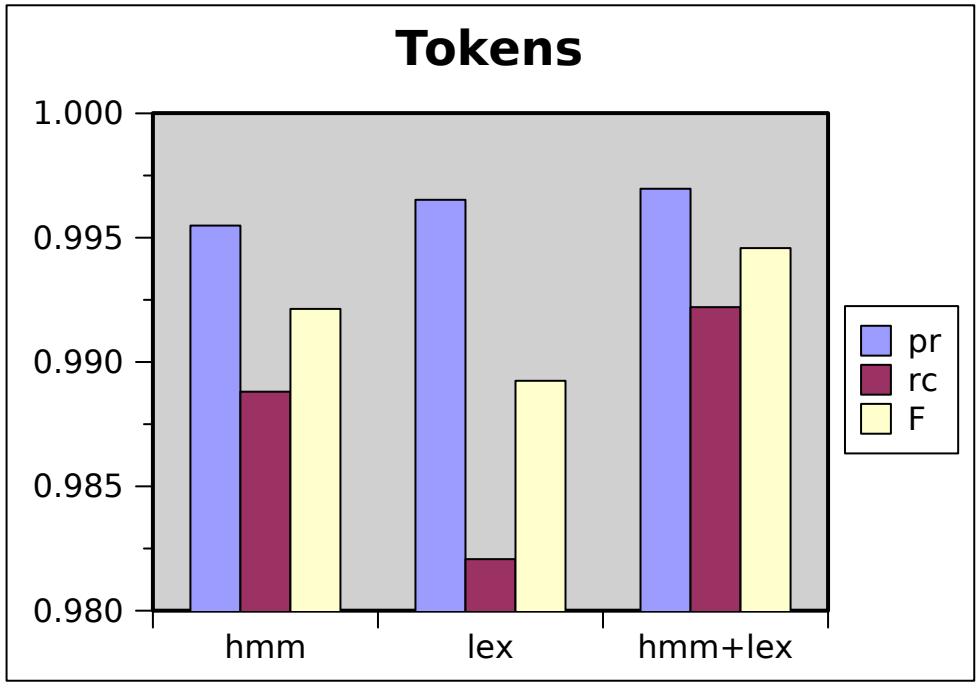
HMM Canonicalization Cascade

(Jurish, 2010c)

- Robust finite-state canonicalizer
- Tested methods: ID, LEX, HMM, HMM+LEX

Results





	% Types			% Tokens		
	pr	rc	F	pr	rc	F
ID	98.3	57.1	72.2	***99.7***	79.1	88.2
HMM	97.9	93.2	95.5	99.5	98.9	99.2
LEX	98.3	85.3	91.3	99.7	98.2	98.9
HMM+LEX	***98.3***	***94.8***	***96.5***	99.7	***99.2***	***99.5***

Conclusion

Construction

- Alignment with contemporary edition
- Type-wise confirmation
- Token-wise annotation

$\rightsquigarrow$ ***minimal-effort corpus bootstrapping***

Applications

- Simple corpus-based lexicon $\Rightarrow$ surprisingly effective
  - very high precision
  - mediocre recall for unknown types (sparse data)
- ‘Exception’ lexicon for HMM canonicalizer
  - best overall performance
  - corpus-based and generative techniques

complement one another

þe Olde Laste Slyde
(“The End”)

Thank you for listening!

— Addenda —

Token Annotation GUI

DTA::EvalCorpus::Editor::Gtk2: phase3.corpus_01.xml *

File Edit Go History View Help

Open Save Back Forward Previous Next Apply Quit

Context

Sapere aude !

Habe **Muth** dich deines eigenen Verſtandes zu bedienen !

iſt al¬ſo der Wahlſpruch der Aufklärung .

General

src: kant_aufklaerung_1784.phase3.q001.xml
wid: w838
sid: s138
page: 17 (56%)
u: Muth
old: Muth
cur: Mut
new: Mut
ext:

Flags

LEX
Valid Sentence
Valid Token
Review Required
Notes

Actions

CAB Query
Browse Source
View XML
Copy Original
Apply

Changed to document data_01/kant_aufklaerung_1784.phase3.q001.xml 100%

des Muthes liegt, ſich ſeiner ohne Leitung eines andern zu bedienen. Sapere aude! Habe Muth dich deines eigenen Verſtandes zu bedienen! iſt al-

GUI: Batch Editor Window

DTA::EvalCorpus::Editor::Gtk2: Batch Editor

Search XPath: .//w[@old="Muth"] XPathifier

Global Search Clear Results Find

n	class	left	old	right	base
1	LEX	... durch Vernunft , frohen	**Muth**	und guten Willen zu über...	goethe_lehrjahre04_1796
2	LEX	...Alter einen ſo geruhigen	**Muth**	und eine ſo ſchöne Nacht...	brentano_kasperl_1838
3	LEX	Habe	**Muth**	dich deines eigenen Verſ...	kant_aufklaerung_1784
4	LEX	... Sprache , und hatte den	**Muth**	nicht ihren Augen zu beg...	goethe_lehrjahre04_1796
5	LEX	Er ſprach ihnen	**Muth**	ein , und ſeine Gründe w...	goethe_lehrjahre02_1795
6	LEX	...ſtecktes Ziel Mit frohem	**Muth**	und ſtrengem Fleiß errei...	goethe_torquato_1790
7	LEX	... , ihm ganz ängſtlich zu	**Muth**	zu werden , daß es ihr n...	spyri_heidi_1880

Eval Perl Code: Perlificator

Select All Remove Apply Dismiss

Administrivia

Class	N	%Edited
LEX	2684	59.22 %
NE	874	19.29 %
JOIN	792	17.48 %
GRAPH	101	2.23 %
SPLIT	72	1.59 %
BUG	40	0.88 %
GONE	8	0.18 %
FM	1	0.02 %