CLARIN-D

DTA::CAB – a Field Spotter's Guide

Bryan Jurish
jurish@bbaw.de

Universität Graz, Zentrum für Informationsmodellierung
19th March, 2019

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Overview

The Big Picture

- The Situation
- The Problem
- The Approach

Canonicalization Methods

- Type-wise Conflation
  - Lexicon, Transliteration, Phonetization, Rewrite Cascade
- Token-wise Disambiguation
  - Dynamic Hidden Markov Model
- Problems & Workarounds

Experiments

- Generative Canonicalization Methods
- Alignment-based Lexicon

CLARIN-D

— The Big Picture —

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Deutsches Textarchiv (DTA)

CLARIN-D

DFG Project *(2007–2016)*

- digitize ∼ 1,300 print volumes, printed ∼ 1600-1900
  - first editions, detailed metadata, highly accurate transcriptions
- TEI-XML corpus encoding & storage
  - DTABf dialect http://www.deutschestextarchiv.de/doku/basisformat
- linguistic analysis (automated)
  - tokenization, normalization, PoS-tagging, lemmatization
- online search (DDC) http://www.ddc-concordance.org
  - lemma-based, PoS-sensitive, spelling-tolerant
- ongoing maintenance & support through CLARIN-D *(2017–present)*

In Numbers

(2019-03-11)

4,425	transcribed works
+ 603	additional works in DTAQ
214,549,119	tokens
3,305,283	surface types
958,331	digitized pages
1,918,576,860	unicode characters

Deutsches Textarchiv (DTA)

http://www.deutschestextarchiv.de

The Situation

Historical Text $\not\ni$ Orthographic Conventions

- also applies to OCR text, E-Mail, SMS, Tweets, ...
- High variance of graphemic forms

fröhlich *"joyful"*	frölich, fröhlich, vrœlich, frœlich, frölich, fröhlich, vrölich, fröhlig, frölig, ...
Herzenleid *"heart-sorrow"*	hertzenleid, herzenleit, hertzenleyd, hertzenleidt, herzenlaid, hertzenlaid, hertzenlaydt, ...

Conventional NLP Tools $\Longrightarrow$ Strict Orthography

- IR systems, PoS taggers, lemmatizers, morphological analyzers, ...
- Fixed lexicon keyed by (ortho)graphic form
  - Extant lexemes only $\rightsquigarrow$ OOV failures (tool lexicon)
  - Attested lexemes only $\rightsquigarrow$ sparsity & noise (corpus lexicon)

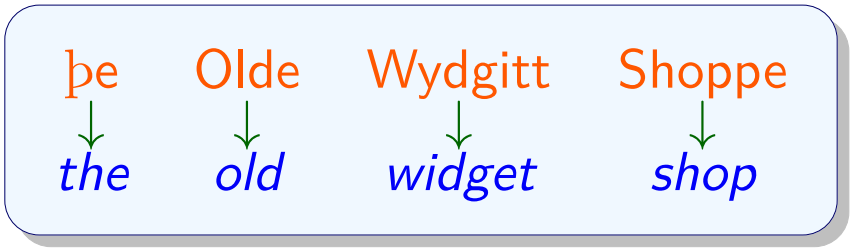
The Problem

	Conventional	Tools
⊕	Historical	Corpus
=		***Soup***

- Corpus variants ***missing*** from application lexicon
  - ***low coverage*** (many unknown types)
  - ***poor recall*** (relevant data not retrieved)
  - ***spurious "noise"*** (poor model fit)
  - *. . . and more!*

The Approach: Canonicalization

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



In a Nutshell

- *Map* each word w to a unique canonical cognate $\widetilde{w}$
- *Defer* application analysis to canonical forms

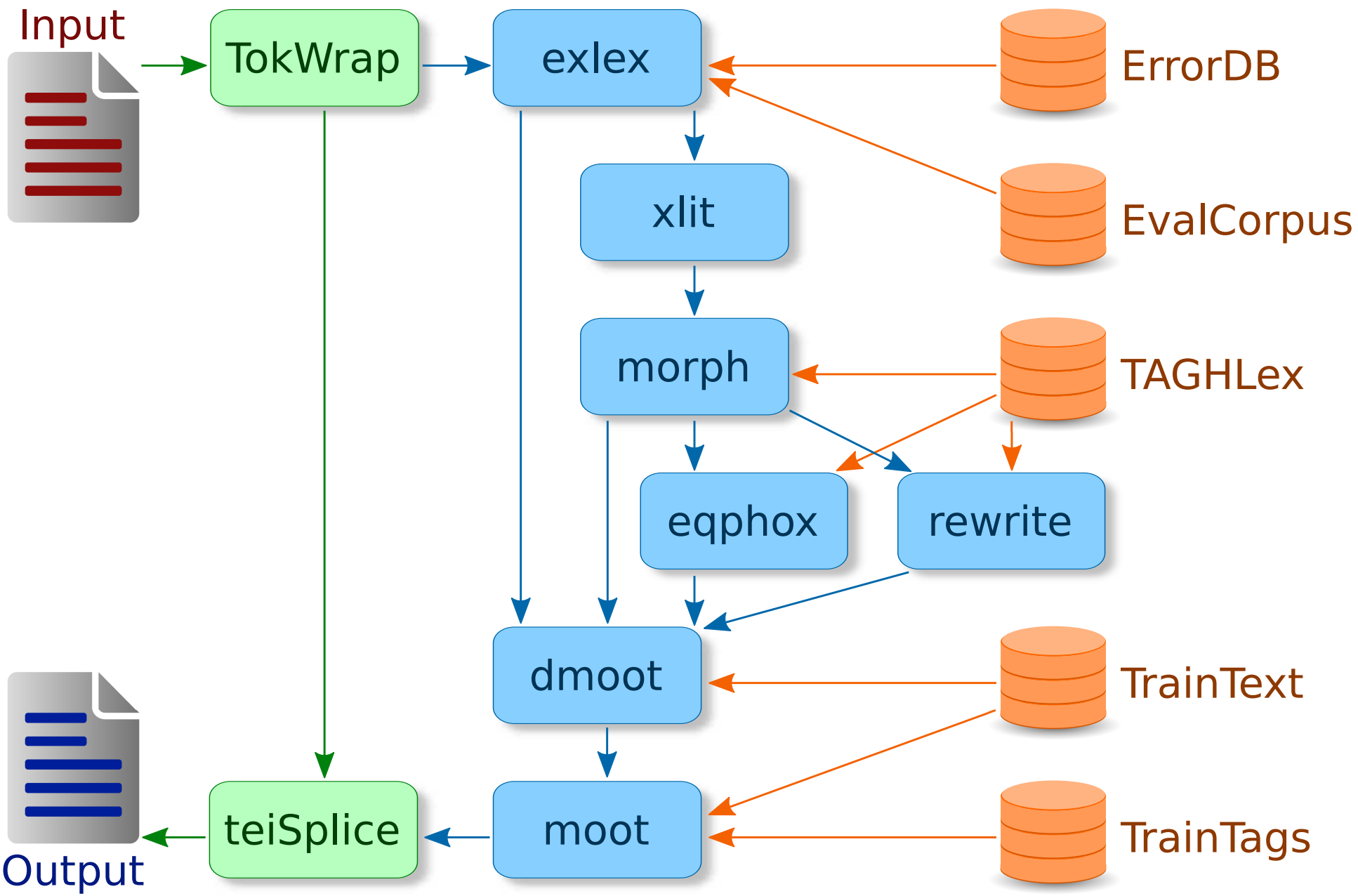
Canonical Cognates

- Synchronically active "extant equivalent(s)" $\widetilde{w} \in \mathrm{Lex}$
- Preserve both root and relevant features of input

Conflation Relation $\sim_r$

- *Binary relation* on strings (words) in $\mathcal{A}^*$ $(\sim_r \subseteq \mathcal{A}^* \times \mathcal{A}^*)$
- Prototypically a true *equivalence relation* *(reflexive, transitive, symmetric)*
  - $w \sim_r v$ holds iff w and v are conflated by relation $\sim_r$
  - equivalence class $[w]_r = \{v \in \mathcal{A}^* \mid w \sim_r v\}$

DTA::CAB System Architecture



CLARIN-D

— Afternoon Workshop —

http://kaskade.dwds.de/~jurish/cab-workshop/

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

CLARIN-D

— Canonicalization Methods —

. . . and where they fail

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Input & Tokenization

garbage in $\rightsquigarrow$ garbage out

Graphical Errors

(cf. Haaf, Wiegand & Geyken, 2013)

Print	wit $\mapsto$ *wett	$\neq$ mit	*"with"*
Transcription	vft $\mapsto$ *fest	$\neq$ und (vn̄)	*"and"*
Whitespace	daßſie $\mapsto$ *Dash|sie	$\neq$ daß sie	*"that she"*

Tokenization Errors

(cf. Jurish & Würzner, 2013)

Abbrev	Durchl[.] $\mapsto$ *torkle	$\neq$ Durchl.	*"highness"*
	Superint[.] $\mapsto$ *Super|rind	$\neq$ Superint.	*"superintendent"*
	vnterthän[.] $\mapsto$ *Unter|teen	$\neq$ untertän.	*"subserviently"*
Space	ists $\mapsto$ *Ists$_{\texttt{NN.gen}}$	$\neq$ ist es	*"is it"*
	ers $\mapsto$ *Airs	$\neq$ er es	*"he it"*
Newline	tau[\n]sent $\mapsto$ *Tau sehnt	$\neq$ tausend	*"thousand"*

‘Exception’ Lexicon (exlex)

Sketch

$$\text{exlex} : \mathcal{A}^* \to \mathcal{A}^* : w \mapsto \widetilde{w}$$

- Finite, deterministic type-wise mapping
- Overrides all other canonicalizers

Sources

(Jurish, Drotschmann & Ast, 2013)

- Alignment-based corpus `http://odo.dwds.de/~jurish/software/dtaec`
  - 126 volumes (1780-1901), 5.6M tokens, 212k types
- Online error database `http://kaskade.dwds.de/demo/caberr`
  - includes rudimentary inflection grammar
  - 36k database records $\rightsquigarrow$ 437k surface mappings

Weaknesses

(cf. Kempken et al., 2006; Gotscharek et al., 2009b)

- can’t handle any *ambiguity*
- can’t handle *productive morphological processes*
- alignment-based bootstrapping *(Jurish & Ast, 2015)*

$\rightsquigarrow$ bogus identity alignments $(w \mapsto w)$

exlex: Bogus Identities

- EvalCorpus assumption: modern edition $\Longrightarrow$ strict orthography
- Implicitly accepted **identity pairs** $(w \mapsto w)$
  - *ca.* 59% types, 87% tokens identical modulo transliteration
- Not always justified by the editions used ***(oops)***

Letter Case	bruder $\mapsto$ *bruder	$\neq$ Bruder	*"brother"*
	trost $\mapsto$ *trost	$\neq$ Trost	*"comfort"*
Extinct Forms	ward $\mapsto$ *ward	$\neq$ wurde	*"was"*
	däuchte $\mapsto$ *däuchte	$\neq$ dünkte	*"seems"*
Prosodic Foot	andre $\mapsto$ *andre	$\neq$ andere	*"other"*
	eignen $\mapsto$ *eignen$_{\mathrm{V}}$	$\neq$ eigenen$_{\mathrm{ADJ}}$	*"own"*
Dialect	kömmt $\mapsto$ *kömmt	$\neq$ kommt	*"comes"*
	nich $\mapsto$ *nich	$\neq$ nicht	*"not"*
Apostrophes	in's $\mapsto$ *in's	$\neq$ ins	*"into the"*
	s'ist $\mapsto$ *s'ist	$\neq$ es ist	*"it is"*

Deterministic Transliteration (xlit)

Sketch

$$w \sim_{\text{xlit}} v :\Leftrightarrow \text{xlit}^*(w) = \text{xlit}^*(v)$$

- **Idea:** account for *extinct characters* *(Jurish 2008, 2010b,c)*
- **Implementation:** $\mathcal{O}(1)$ character lookup table
- mostly useful as *preprocessor* for subsequent methods

Successes

Long-'s'	Abſtand ↦ Abstand	*"distance"*
Superscript-'e'	nȩotig ↦ nötig	*"necessary"*
Diacritics	Hochzît ↦ Hochzeit	*"wedding"*

Failures

Diacritics	wôl ↦ *wol	≠ wohl	*"well"*
Extant Characters	Thür ↦ *Thür	≠ Tür	*"door"*
	vn̄ ↦ *vn (⇝ Vene)	≠ und	*"and"*

TAGH Morphology (morph)

Sketch

(Geyken & Hanneforth, 2006)

- Model (modern) morphological processes as WFST
- Provides (infinite) *weighted target language* for subsequent methods
  - analysis cost $\approx$ derivational complexity
- "Modern-wins" filtering

$$w \mapsto w :\Leftarrow w \in \pi_1(M_{\text{morph}})$$

Overgeneration: 'modern' form *shouldn't* always win

andre $\mapsto$	*André$_{\texttt{NE}}$	$\neq$ andere	*"other"*
from $\mapsto$	*from$_{\texttt{FM.en}}$	$\neq$ fromm	*"pious"*
hiebei $\mapsto$	*Hieb|ei	$\neq$ hierbei	*"hereby"*
Zeugnuß $\mapsto$	*Zeug|nuß	$\neq$ Zeugnis	*"report"*
keyserlich $\mapsto$	*Keys$_{\texttt{NE}}$|erl$_{\texttt{NE}}$|ich	$\neq$ kaiserlich	*"imperial"*
Proceſs $\mapsto$	*Process	$\neq$ Prozeß	*"process"*

- **Workaround:** *safety heuristics*, e.g. *$_{\texttt{FM}}$, *$_{\texttt{NE}}$, *-Ei, *-Nuß, ...

Phonetic Equivalence (eqpho)

Sketch

$$w \sim_{\text{pho}} v :\Leftrightarrow \text{pho}(w) = \text{pho}(v)$$

- **Idea:** conflate words by *phonetic form* *(Jurish, 2008, 2010b)*
- **Implementation:** text-to-speech rule-set *(Möhler et al., 2001)*
  - modified & compiled as FST M_{pho}
  - online $k-$best equivalence cascade search *(Jurish, 2010a)*

$$C_{\text{eqpho}} := M_{\text{pho}} \circ M_{\text{pho}}^{-1} \circ \pi_1(M_{\text{morph}})$$

Problems

Kopfe	↦ *Kopfweh	≠ Kopf	*"head"*
wes	↦ *Wehs	≠ wessen	*"whose"*
gewegen	↦ *Geh|wegen	≠ gewogen	*"weighted"*
maiestat	↦ *Mais|tat	≠ Majestät	*"majesty"*
Moyſe	↦ *Mäuse	≠ Mose(s)	*"Moses"*
Troglodyt	↦ *Troglo|duett	≠ Troglodyt	*"troglodyte"*

- **Workarounds:** target language pruning, cascade lookup cutoffs

Rewrite Cascade (rw)

$$\begin{aligned} \mathrm{best}_{\mathrm{rw}}(w) &:= \arg\min_{v \in \mathcal{A}^*} [\![M_{\mathrm{rw}} \circ \pi_1(M_{\mathrm{morph}})]\!](w, v) \\ w \sim_{\mathrm{rw}} v &:\Leftrightarrow \mathrm{best}_{\mathrm{rw}}(w) = \mathrm{best}_{\mathrm{rw}}(v) \end{aligned}$$

Sketch

- **Idea:** map words to *nearest* extant type *(Jurish, 2010b,c)*
  - generalized string edit distance *(Damerau, 1964; Levenshtein, 1966)*
  - computable even for infinite lexica *(Mohri, 2002; Jurish, 2010a)*
- **Implementation:** online $(k = 1)-$best cascade search
  - editor WFST M_{rw} compiled from ca. 300 SPE-style rules
  - EM-style weight optimization (Viterbi)
  - weight interpolation constants λ_{rw} and λ_{morph}

Problems

ehligen	↦ *Elchen	≠ ehelichen	*"marital"*
gewüntschten	↦ *Gewinde|sekten	≠ gewünschten	*"desired"*
Predigampt	↦ *Birdie|gambit	≠ Predigamt	*"ministry"*
Verdamnuß	↦ *Dominus	≠ Verdammnis	*"damnation"*
Weſſier	↦ *Wessi	≠ Wesir	*"vizier"*

- **Workarounds:** rule adjustments, punitive target weights, lookup cutoffs

Token-wise Disambiguation (hmm)

Idea

(Mays et al., 1991; Brill & Moore, 2000; Jurish, 2010c)

- Allow high-recall overgeneration at type level (xlit, eqpho, rw)
- Recover precision using token-level context

Implementation: Dynamic Hidden Markov Model (HMM)

- **States** are word-conflator pairs

$$\mathcal{Q} = (\mathcal{W} \cup \{\mathbf{u}\}) \times R$$

- **Observations** are input strings

$$\mathcal{O}_S = \bigcup_{i=1}^{n_S} \{w_i\} \subset \mathcal{A}^*$$

- **Transitions** *(static)*

$$A\left(\langle \widetilde{w}_i, r_i \rangle_{i=1}^{m}\right) \approx p\left(\widetilde{w}_m \middle| \widetilde{w}_1^{m-1}\right)$$

- **Lexicon** *(dynamic)*: Maxwell-Boltzmann distribution

$$B\left(\langle \widetilde{w}, r \rangle, w\right) \approx \frac{b^{\beta d_r(w,\widetilde{w})}}{\sum_{r' \in R} \sum_{\widetilde{w'} \in \downarrow[w]_{r'}} b^{\beta d_{r'}(w,\widetilde{w'})}}$$

  - b, β are global model parameters ($b \geq 1, \beta \leq 0$)
  - $d_r(w, \widetilde{w})$ depends on conflator r

- **Lookup** (moot): Viterbi Algorithm *(Viterbi, 1967)*

HMM Problems

Sparse and/or Inappropriate Training Data

- many "normal" words treated as *unknown* (u)
- poor handling of historical *syntax*
  - *In deym dienst bestendig bleyben / die trubsall vnns nicht abtreiben*
    M. Luther et al.: Eyn Enchiridion oder Handbuchlein, 1524
  - *Vnd mich deucht / das er den Hyacinthum auff einer ſeite lieb hatte*
    G. Rollenhagen: Vier Bücher wunderbarlicher . . . himmel, 1603
  - *Daher newlichs einer geſagt: die tortur iſt allmächtig*
    F. von Spee: Gewissens-Buch, Von Processen Gegen die Hexen, 1647
  - *kaum ſieht er hinein, ſo erblickt er ein Medaillon*
    Goethe: Wilhelm Meisters Lehrjahre, Bd. 4, 1796

Edelgebohrnen	↦	*Tele|gebahren	≠ edelgeborenen	*"noble-born"*
potz [blitz]	↦	*Potts	≠ potz$_{\text{ITJ}}$	*"gee [whiz]"*
Secretärsstelle	↦	*Secretärsstelle	≠ Sekretärsstelle	*"secretary's job"*
unwiſend	↦	*anwesend	≠ unwissend	*"unknowing"*
Phaſmate	↦	*Faß|matte	≠ Phasmate	*"phantasms"*

- **Workarounds:** language guessing, PoS heuristics, exlex, . . .
- **Cunning & Devious Plan:** incorporate PoS-tagger model (moot)

“Real” Problems: Target Lexicon

Extinct or Missing Target Lexemes

elektroskopische	↦ *Elektro|ska|piefke	*“electroscopic”*
Sakramentierer	↦ *Sakrament|irrer	*“sacramentists”*
glorwürdigen	↦ *Chlor|würdigen$_{\texttt{VV}}$	*“glory-worthy”*

Extinct or Missing Morphological Processes

aller-	aller|größte	↦ *grüßest	*“greatest of all”*
hier-	hier|nächst	↦ *hin|nagst	*“subsequent”*
ob-	ob|angeregter	↦ *Oboen|gerechter	*“aforementioned”*
-e	Kopf|e	↦ *Kopf|weh	*“head”*
-ig	standhaft|ig	↦ *stoned|hastig	*“steadfast”*
-lich	auflös|lich	↦ *auslöschlich	*“soluble”*

Lexical Shift / Codification

bälder	↦ *Bälde$_{\texttt{NN}}$	*“sooner”*
neulicher	↦ *neulich$_{\texttt{ADV}}$	*“recent”*

More “Real” Problems: Interference

Proper Names

NE ↦ *Lex	Carlowitz	↦ *Gorilla|witz$_{NN}$	≠ Carlowitz
	Philipson	↦ *File|bison$_{NN}$	≠ Philipson
	Moyſe	↦ *Mäuse$_{NN}$	≠ Moses
NE +flect	Jeſ|um	↦ *Jass|ohm	≠ Jesus
	Mathilde|n	↦ *modelten	≠ Mathilde
	George|ns	↦ *Chargen	≠ Georg

Non-Lexical Material

from	↦ *from$_{FM.en}$	≠ fromm	*“pious”*
medicinæ	↦ *Medizin|ah	≠ medicinae	*“medical” (FM.lat)*
mon Dieu	↦ *Mohn Dia	≠ mon Dieu	*“my God” (FM.fr)*
Zn	↦ *Zen	≠ Zn	*“zinc” (chem.)*

Ambiguity / Divergence

wit ↦ mit, wie, weit, wir, wit
widder ↦ wider, wieder, weder, Widder

CLARIN-D

— Experiments —

*. . . or, "how bad is it **really**?"*

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Experiment 1: Generative Canonicalizers

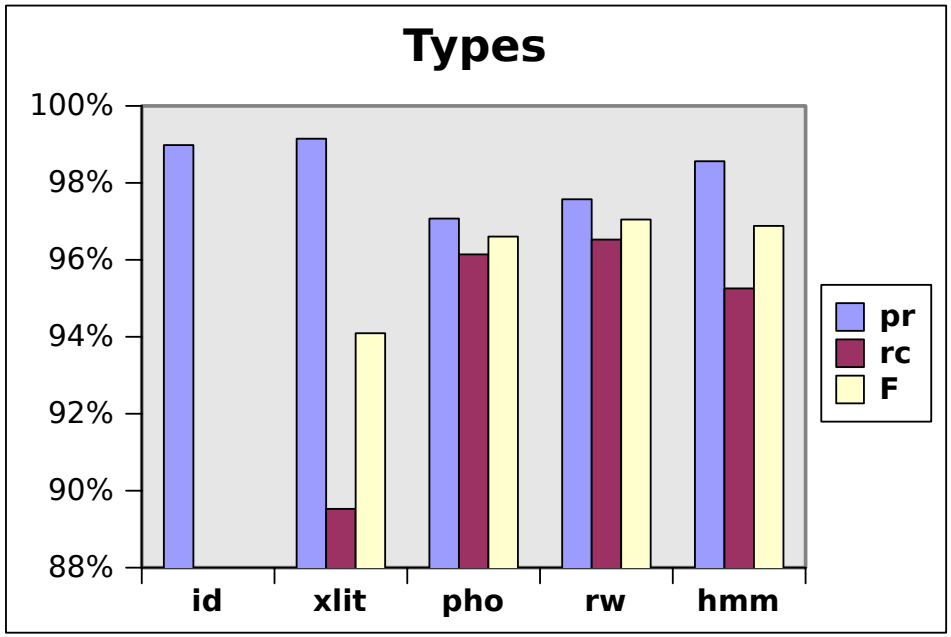
DTA Evaluation Corpus *(Jurish, Drotschmann & Ast, 2013)*

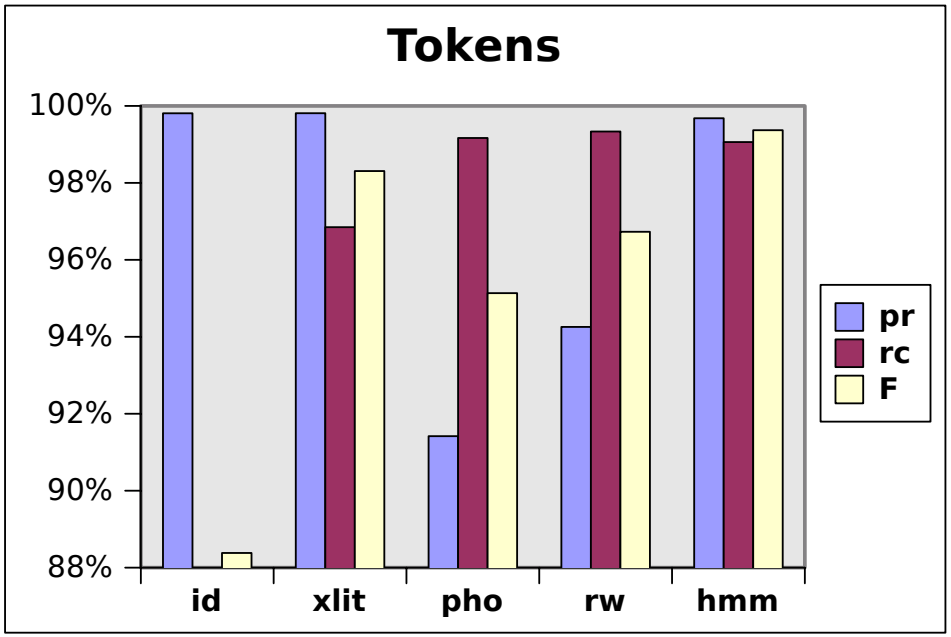
- Gold standard subset of *DTA* Phase 1
  - 13 volumes *(1780 – 1880)*
  - 114,542 tokens; 10,070 types *(alphabetic only)*
- *Canonical cognate* assigned to each token
  - automatic alignment with conventional edition
  - 3-pass manual review
    - Type-wise *(conservative)*
    - Token-wise *(unverified tokens only)*
    - "Suspicious" pairs *(heuristic selection)*

Evaluation Measures

- Simulated information retrieval task
- Type- and token-wise precision (pr), recall (rc), and F

Experiment 1: Results





	% Types			% Tokens		
	pr	rc	F	pr	rc	F
id	99.0	59.2	74.1	99.8	79.3	88.4
xlit	***99.1***	89.5	94.1	***99.8***	96.8	98.3
pho	97.1	96.1	96.6	91.4	99.2	95.1
rw	97.6	***96.5***	***97.0***	94.3	***99.3***	96.7
hmm	98.6	95.3	96.9	99.7	99.1	***99.4***

Experiment 2: Alignment-based Lexicon

'Evaluation' Corpus $\rightsquigarrow$ Ground-Truth Relevance

$$\mathrm{relevant}(w, \widetilde{w}) := \{(v, \widetilde{v}) : \widetilde{v} = \widetilde{w}\}$$

- Most thoroughly annotated corpus subset
- 13 volumes $\sim$ 320k tokens $\sim$ 28k types (words only)

'Training' Corpus $\rightsquigarrow$ Canonicalization Lexicon (lex)

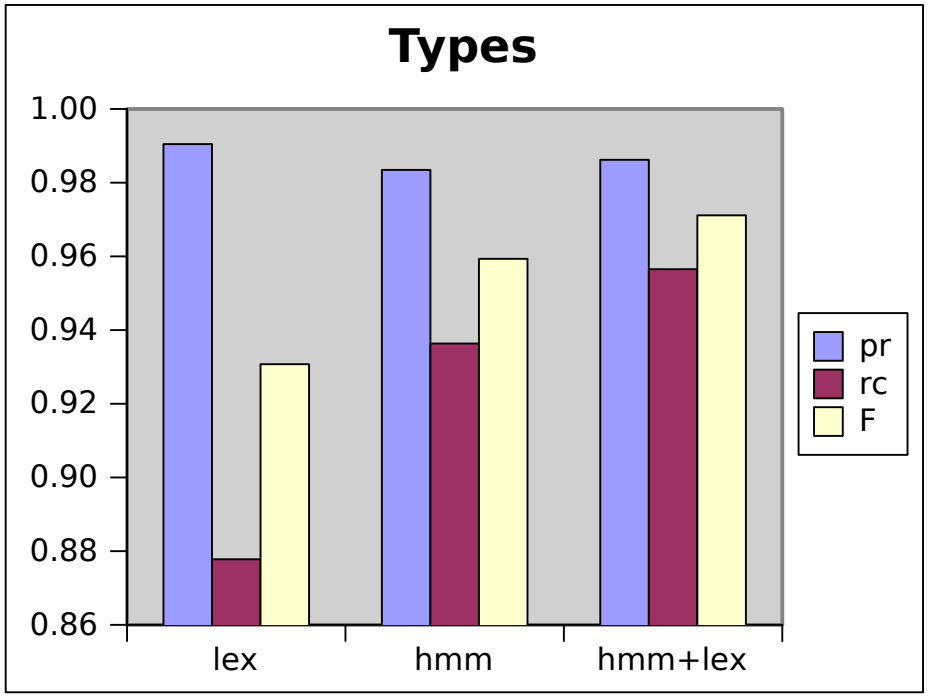
$$\mathrm{lex}(w) = \begin{cases} \arg\max_{\widetilde{w}} f(w, \widetilde{w}) & \text{if } f(w) > 0 \\ w & \text{otherwise} \end{cases}$$

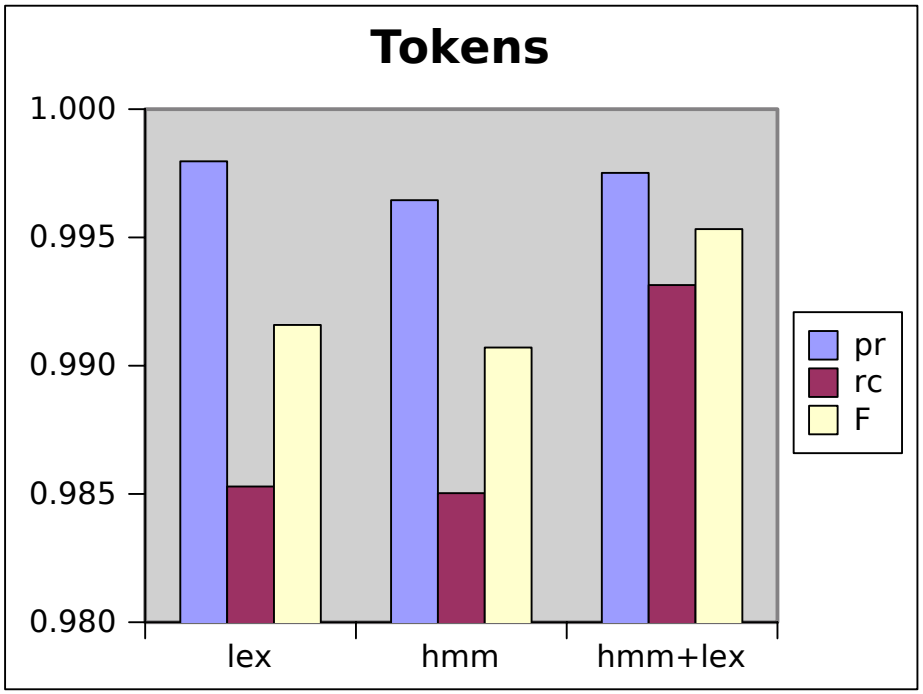
- Strictly disjoint from test corpus (by author)
- 101 volumes $\sim$ 3.5M tokens $\sim$ 158k types (words only)

Evaluation

- Simulated information retrieval $(\mathrm{pr}, \mathrm{rc}, \mathrm{F})$ *(van Rijsbergen, 1979)*
- Tested methods: id, lex, hmm, $\mathrm{hmm{+}lex}$

Experiment 2: Results





	% Types			% Tokens		
	pr	rc	F	pr	rc	F
id	***99.1***	55.7	71.3	***99.8***	78.5	87.9
lex	99.0	87.8	93.1	***99.8***	98.5	99.2
hmm	98.3	93.6	95.9	99.6	98.5	99.1
hmm+lex	98.6	***95.7***	***97.1***	***99.8***	***99.3***	***99.5***

Concluding Remarks

Historical Text and Conventional Tools

don't play together nicely "out of the box"

Canonicalization Methods

- static lexicon — *∼ effective but sparse*
- transliteration — *∼ quick and dirty*
- phonetic equivalence — *∼ elegant but coarse*
- rewrite cascade — *∼ flexible but costly*
- HMM disambiguator — *∼ precision recovery*

Advanced Topics

- parameter optimization — *≥27 free parameters*
- rewrite editor — *induction from training pairs*
- HMM smoothing — *PoS, semantics, morphs, . . .*

CLARIN-D

þe Olde Laſt Slyde

("The End")

http://www.deutschestextarchiv.de/demo/cab/

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN