

“Elchen, Elektroskapiefken, vñ Andrés Kopfweh” or: When Canonicalization Algorithms Attack

Bryan Jurish

Deutsches Textarchiv / Berlin-Brandenburgische Akademie der Wissenschaften

jurish@bbaw.de

Ruhr-Universität Bochum, 28th May, 2013

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

— The Big Picture —

Deutsches Textarchiv (DTA)

Primary Goals

- digitize ~ 1,300 print volumes, printed ~ 1600-1900
 - ▶ first editions of respective works
 - ▶ detailed metadata, highly accurate transcriptions
- TEI-XML corpus encoding & storage
 - ▶ DTA base format (DTABf) dialect
- linguistic analysis (automated)
 - ▶ tokenization, normalization, PoS-tagging, lemmatization
- online search (DDC) <http://www.ddc-concordance.org>
 - ▶ lemma-based, PoS-sensitive, spelling-tolerant

In Numbers

(DTA+DTAE 2013-05-02)

1,276 transcribed works
353,245 digitized pages
567,200,587 unicode characters
81,049,777 tokens (alpha-numeric)
1,947,414 types (alpha-numeric)

Deutsches Textarchiv (DTA)

<http://deutschestextarchiv.de>

(beta)

Anmelden (DTAQ)

Texte ▾ Projekt ▾ Dokumentation ▾ Impressum

DTA

suchen

in den Titeldaten im Korpus auf der Webseite [Hilfe](#)

Das DTA in Zahlen (Kernkorpus)

786 transkribierte Werke
288855 digitalisierte Seiten
66068434 fortlaufende Wortformen
464959660 Zeichen (Unicode)
546 weitere Werke in DTAQ

Neue Werke im DTA

[alle Titel ...](#)

Deutsches Textarchiv

GRUNDLAGE FÜR EIN REFERENZKORPUS DER NEUHOCHDEUTSCHEN SPRACHE

Das Deutsche Textarchiv stellt einen disziplinen- und gattungsübergreifenden Grundbestand deutschsprachiger Texte aus dem Zeitraum von ca. 1600 bis 1900 bereit. Die Textauswahl erfolgte auf der Grundlage einer von Akademiemitgliedern erstellten und ausführlich kommentierten, umfangreichen Bibliographie. In Ergänzung wurden einschlägige Literaturgeschichten und (Fach-)Bibliographien ausgewertet. Aus der Gesamtliste der auf diesem Wege ermittelten Titel wurde von der DTA-Projektgruppe ein hinsichtlich der repräsentierten Textsorten und Disziplinen ausgewogenes Korpus zusammengestellt (weitere Informationen zur Textauswahl ...).

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 ...](#)

The Situation

Historical Text Orthographic Conventions

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

fröhlich fröhlich, fröhlich, vroelich, frøelich,

“joyful” frølich, frøhlich, vrölich, fröhlig, ...

Herzenleid hertenleid, herzenleit, hertenleyd,

“heart-sorrow” hertenlaidt, hertenlaydt, ...

Conventional NLP Tools Strict Orthography

- IR systems, PoS taggers, stemmers, lemmatizers, morphological analyzers, parsers, ...
- Fixed lexicon keyed by (ortho)graphic form
- Extant lexemes only

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 \tilde{w}
- *Defer* application analysis to canonical forms

Canonical Cognates

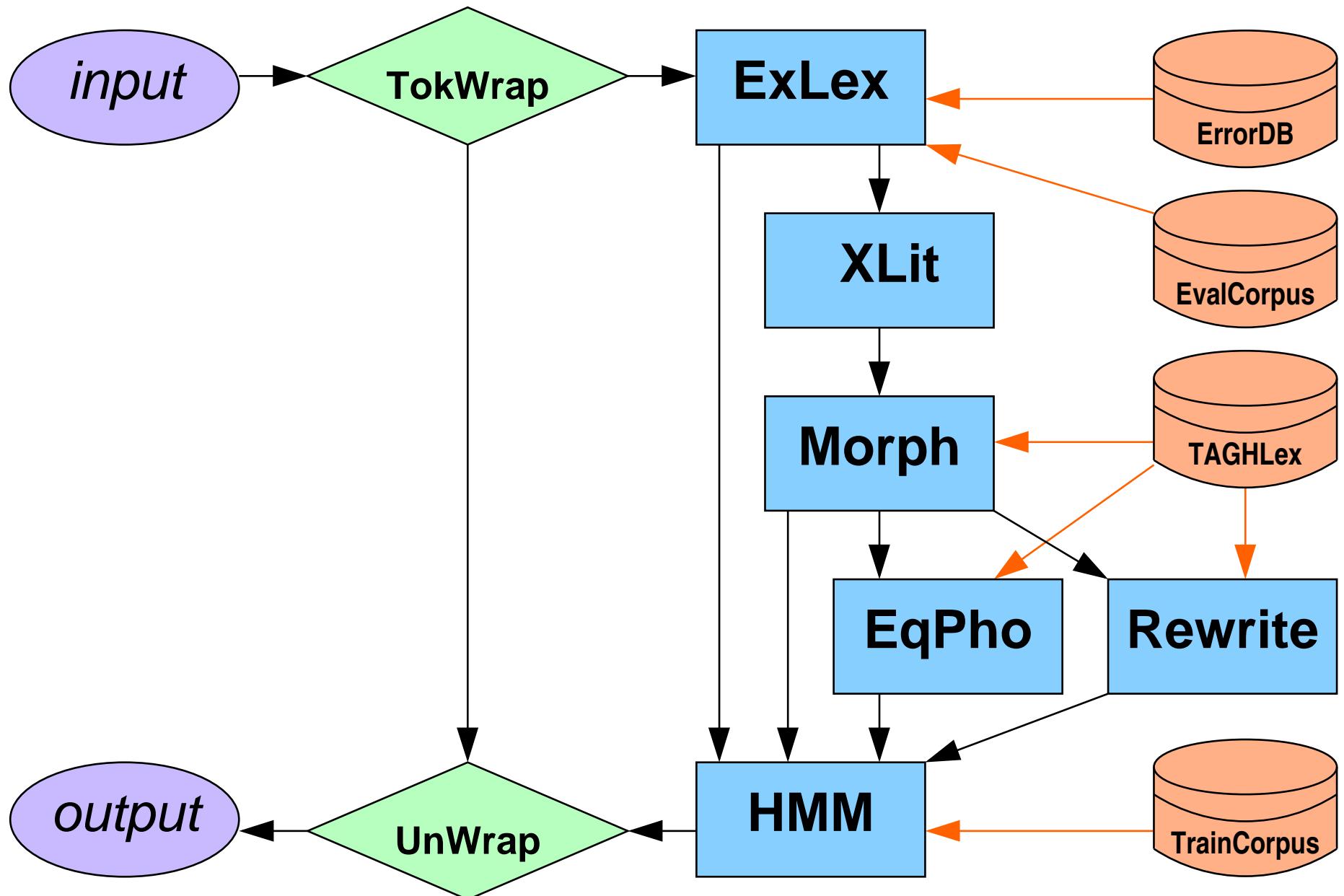
- Synchronously active “extant equivalent(s)” $\tilde{w} \in \text{Lex}$
- Preserve both root and relevant features of input

Conflation Relation \sim_r

- *Binary relation* on strings (words) in \mathcal{A}^*
- Prototypically a true *equivalence relation*

— Canonicalization Methods —

DTA::CAB System Architecture



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 (vñ)	“and”
Whitespace	daß sie \mapsto *Dash sie	\neq daß sie	“that she”

Tokenization Errors

(cf. Würzner, Jurish & Lemnitzer, [forthcoming])

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 _{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

$$\text{exlex} : \mathcal{A}^* \rightarrow \mathcal{A}^* : w \mapsto \tilde{w}$$

- Deterministic type-wise mapping
- Overrides all other canonicalizers

Sources

(Jurish, Drotschmann & Ast, [in print])

- Alignment-based corpus
 - ▶ 126 volumes (1780-1901), 5.6M tokens, 212k types
- Online error database
 - ▶ includes basic parallel inflection paradigms

kaskade.dwds.de/demo/dtaec

kaskade.dwds.de/demo/caberr

Weaknesses

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

- can't handle any *ambiguity*
- can't handle *productive morphological processes*
- alignment-based bootstrapping
 - ▶ bogus identity alignments ($w \mapsto w$)

(Jurish & Ast, [forthcoming])

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

exlex: Bogus Identities

- Assumed: modern edition \implies 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	\neq	eigenen	"own"
Dialect	kömmmt \mapsto *kömmmt	\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 \mapsto Abstand	“distance”
Superscript-‘e’	n ^e tig \mapsto nötig	“necessary”
Diacritics	Hochzît \mapsto Hochzeit	“wedding”

Failures

Diacritics	wôl \mapsto *wol	\neq	wohl	“well”
Extant Characters	Thür \mapsto *Thür	\neq	Tür	“door”
	vñ \mapsto *vn (\rightsquigarrow Vene)	\neq	und	“and”

TAGH Morphology (morph)

Sketch

(Geyken & Hanneforth, 2006)

- Model (modern) morphological processes as WFST
- Provides *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	$\mapsto *André_{\text{NE}}$	\neq	Andere	“other”
from	$\mapsto *from_{\text{FM.en}}$	\neq	fromm	“pious”
hiebei	$\mapsto *Hieb ei$	\neq	hierbei	“hereby”
Zeugnuß	$\mapsto *Zeug nuß$	\neq	Zeugnis	“report”
keyserlich	$\mapsto *Keys_{\text{NE}} erl_{\text{NE}} ich$	\neq	kaiserlich	“imperial”
Proceſs	$\mapsto *Process$	\neq	Prozeß	“process”

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

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	$\mapsto *Kopfweh$	\neq	Kopf	“head”
wes	$\mapsto *Wehs$	\neq	wessen	“whose”
gewegen	$\mapsto *Geh wegen$	\neq	gewogen	“weighted”
maiestat	$\mapsto *Mais tat$	\neq	Majestät	“majesty”
Moyſe	$\mapsto *Mäuse$	\neq	Mose	“Moses”
Troglodyt	$\mapsto *Troglo duett$	\neq	Troglodyt	“troglodyte”

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

Rewrite Cascade (rw)

Sketch

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

- **Idea:** map words to *nearest* extant type (Jurish, 2010b,c)
 - ▶ generalized string edit distance (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
 - ▶ weight interpolation constants λ_{rw} and λ_{morph}

Problems

ehligen	\mapsto *Elchen	\neq	ehelichen	"marital"
gewünschten	\mapsto *Gewinde sekten	\neq	gewünschten	"desired"
Predigampt	\mapsto *Birdie gambit	\neq	Predigamt	"ministry"
Verdamnuß	\mapsto *Dominus	\neq	Verdammnis	"damnation"
Weßier	\mapsto *Wessi	\neq	Wesir	"vizier"

- **Workarounds:** 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
- Recover precision using token-level context

Implementation: Dynamic Hidden Markov Model (HMM)

- States** are word-conflator pairs

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

- Observations** are input strings

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

- Transitions (static)**

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

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

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

- b, β are global model parameters ($b \geq 1, \beta \leq 0$)
- $d_r(w, \tilde{w})$ depends on conflator r
- Lookup** (moot): Viterbi Algorithm

(Viterbi, 1967)

HMM Example

Input

Dete sammlete Steyne im Rockermel

HMM Example

Input Dete sammlete Steyne im Rockermel

xlit *Dete* *sammlete* *Steyne* *im* *Rockermel*

HMM Example

Input Dete sammlete Steyne im Rockermel

xlit *Dete* *sammlete* *Steyne* *im* *Rockermel*

pho	∅	∅	{ <i>Steine</i> }	$\left\{ \begin{array}{l} im, \\ ihm \end{array} \right\}$	{ <i>Rockärmel</i> }
-----	---	---	-------------------	--	----------------------

HMM Example

Input	Dete	sammlete	Steyne	im	Rockermel
xlit	<i>Dete</i>	<i>sammlete</i>	<i>Steyne</i>	<i>im</i>	<i>Rockermel</i>
pho	\emptyset	\emptyset	$\{Steine\}$	$\left\{ \begin{array}{l} im, \\ ihm \end{array} \right\}$	$\{Rockärmel\}$
rw	<i>Tete</i> $\langle 1 \rangle$	<i>sammelte</i> $\langle 5 \rangle$	<i>Steine</i> $\langle 1 \rangle$	<i>im</i> $\langle 0 \rangle$	<i>Rockermehl</i> $\langle 10 \rangle$

HMM Example

Input Dete sammlete Steyne im Rockermel

xlit Dete sammlete Steyne im Rockermel

pho \emptyset \emptyset {Steine} $\left\{ \begin{array}{l} \text{im}, \\ \text{ihm} \end{array} \right\}$ {Rockärmel}

rw Tete⟨1⟩ sammelte⟨5⟩ Steine⟨1⟩ im⟨0⟩ Rockermehl⟨10⟩

hmm Dete sammelte Steine im Rockärmel

HMM Example

Input Dete sammlete Steyne im Rockermel

xlit Dete sammlete Steyne im Rockermel

pho	∅	∅	{ <u>Steine</u> }	$\left\{ \begin{array}{l} \underline{im}, \\ ihm \end{array} \right.$	{ <u>Rockärmel</u> }
-----	---	---	-------------------	---	----------------------

rw Tete⟨1⟩ sammelte⟨5⟩ Steine⟨1⟩ im⟨0⟩ Rockermehl⟨10⟩

hmm Dete sammelte Steine im Rockärmel

Output Dete **sammelte** **Steine** im **Rockärmel**

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

Edelgebohrnen	→ *Tele gebahren	≠ edelgeborenen	“noble-born”
potz [blitz]	→ *Potts	≠ potz _{ITJ}	“gee [whiz]”
Secretärsstelle	→ *Secretärsstelle	≠ Sekretärsstelle	“secretary’s job”
unwifend	→ *anwesend	≠ unwissend	“unknowing”
Phasmate	→ *Faß matte	≠ Phasmate	“phantoms”

- **Workarounds:** language guessing, PoS heuristics, exlex, . . .

“Real” Problems: Target Lexicon

Extinct or Missing Target Lexemes

elektroskopische	↔ *Elektro ska piefke	“electroscopic”
Sakramentierer	↔ *Sakrament irrer	“sacramentists”
glorwürdigen	↔ *Chlor würdigen _{VV}	“glory-worthy”

Extinct or Missing Morphological Processes

aller-	aller größte	↔ *grüßest	“greatest of all”
hier-	hier nächst	↔ *hin nagst	“here next”
ob-	ob angeregter	↔ *Oboen gerechter	“aforementioned”
-e	Kopf e	↔ *Kopf weh	“head”
-ig	standhaft ig	↔ *stoned hastig	“steadfast”
-lich	auflös lich	↔ *auslöslich	“soluble”

Lexical Shift / Codification

bälder	↔ *Bälde _{NN}	“sooner”
neulicher	↔ *neulich _{ADV}	“recent”

More “Real” Problems: Interference

Proper Names

NE \mapsto *Lex

Carlowitz \mapsto *Gorilla|witz_{NN} \neq Carlowitz

Philipson \mapsto *File|bison_{NN} \neq Philipson

Moyſe \mapsto *Mäuse_{NN} \neq Moses

NE +flect

Jes|um \mapsto *Jass|ohm \neq Jesus

Mathilde|n \mapsto *modelten \neq Mathilde

George|ns \mapsto *Chargen \neq Georg

Non-Lexical Material

from \mapsto *from_{FM.en} \neq fromm

“pious”

medicinæ \mapsto *Medizin|ah \neq medicinae

“medical” (FM.lat)

mon Dieu \mapsto *Mohn Dia \neq mon Dieu

“my God” (FM.fr)

Zn \mapsto *Zen \neq Zn

“zinc” (chem.)

Ambiguity / Divergence

wit \mapsto mit, wie, weit, wir, wit

widder \mapsto wider, wieder, weder, Widder

— Experiments —

Experiment 1: Generative Canonicalizers

DTA Evaluation Corpus

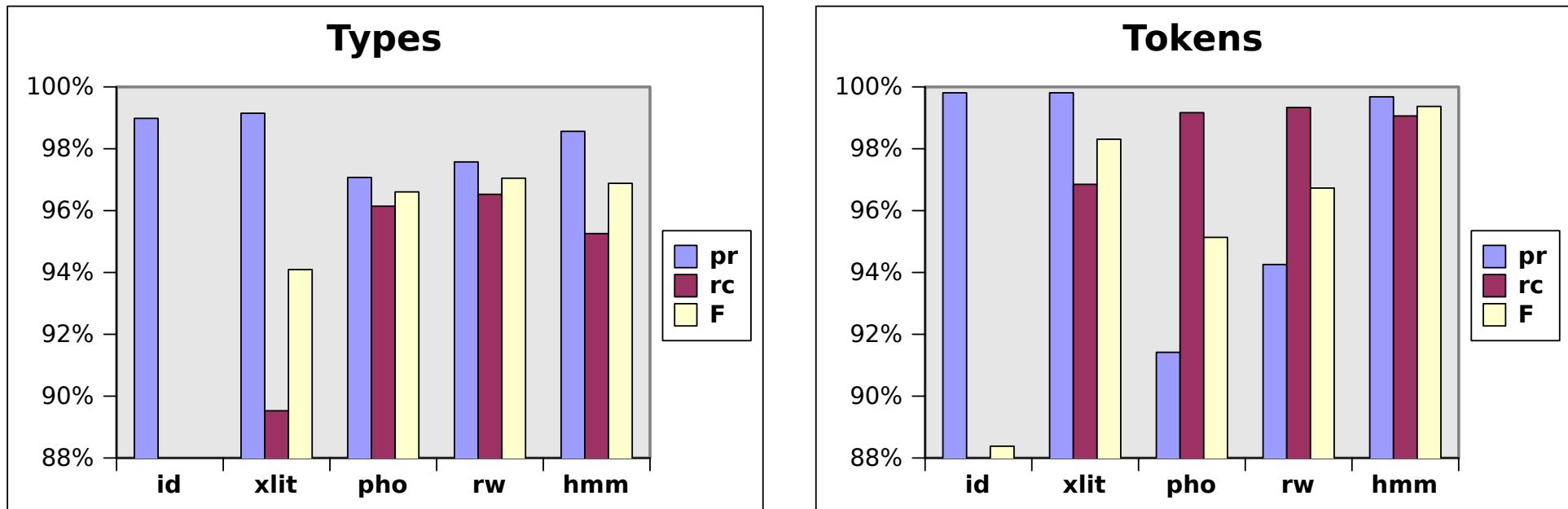
(*Jurish, Drotschmann & Ast, [in print]*)

- 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

$$\text{relevant}(\mathbf{\tilde{w}}, \tilde{w}) := \{(\mathbf{v}, \tilde{v}) : \tilde{v} = \tilde{w}\}$$

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

'Training' Corpus \rightsquigarrow Canonicalization Lexicon (lex)

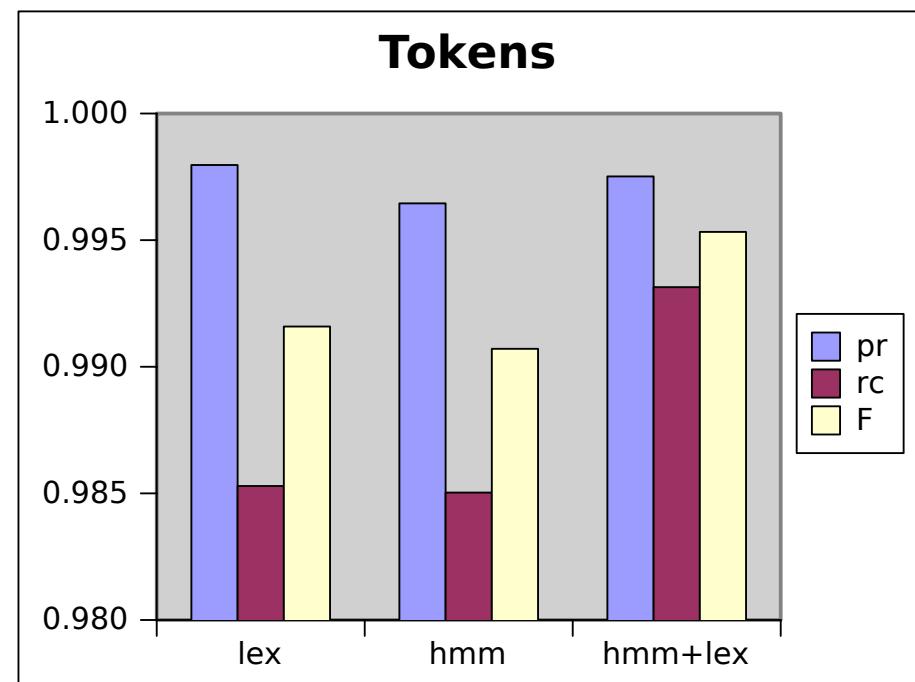
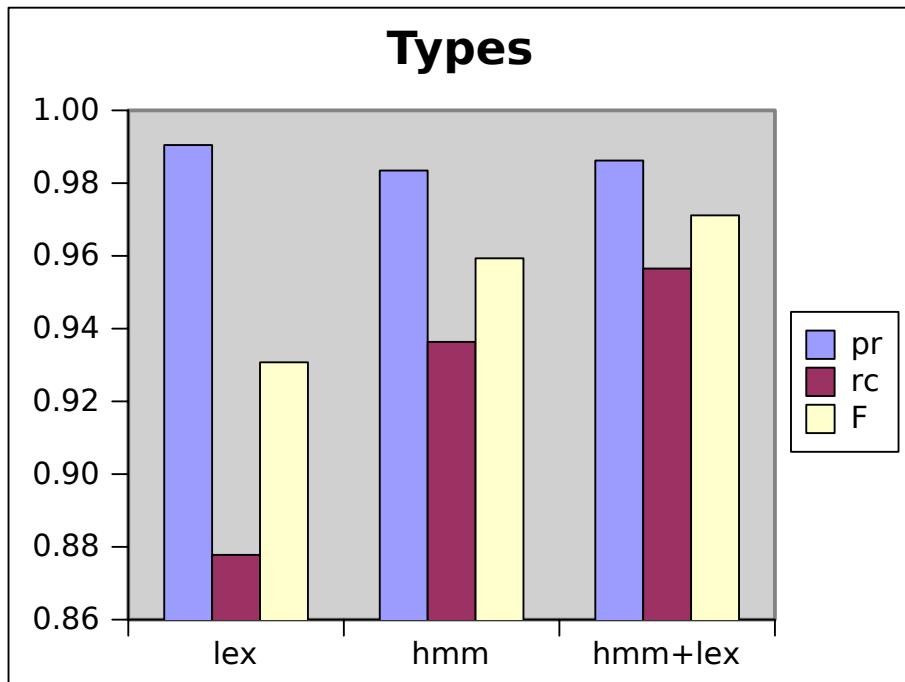
$$\text{lex}(\mathbf{w}) = \begin{cases} \arg \max_{\tilde{w}} f(\mathbf{w}, \tilde{w}) & \text{if } f(\mathbf{w}) > 0 \\ \mathbf{w} & \text{otherwise} \end{cases}$$

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

Evaluation

- Simulated information retrieval (pr, rc, F) (*van Rijsbergen, 1979*)
- Tested methods: id, lex, hmm, 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

Future Work

- parameter optimization ≥ 27 free parameters [WIP]
induce from training pairs?
- rewrite editor
- HMM smoothing *PoS, semantics, morphs, ...*

þe Olde Last Shde ("The End")

===== Gory Details =====

Gory Details

- Related work
- LTS FST construction
- Dictionary lemma instantiation
- Online k -best lookup algorithm
- HMM disambiguator
- Evaluation Measures

Online Demos

- <http://kaskade.dwds.de/demo/cab> canonicalization server
- <http://kaskade.dwds.de/demo/caberr> error database
- <http://kaskade.dwds.de/demo/dtaec> evaluation corpus editor
- <http://kaskade.dwds.de/dta-ecp> evaluation corpus pruning
- <http://kaskade.dwds.de/dtaos> corpus search
- <http://kaskade.dwds.de/dtaq/web> top-level web site

— Related Work —

Spelling Correction

Damerau (1964); Levenshtein (1966)

- string edit distance (“DL distance”)

Kernighan et al. (1990); Church & Gale (1991)

- explicit error models
- local k -gram context disambiguation

Mays, Damerau & Mercer (1991)

- word trigram model
- full sentential context

Brill & Moore (2000)

- generic k -local string-to-string error operations

Diachronic Conflation: English

Hartlib Papers, Sheffield

(Rogers & Willett, 1991; Robertson & Willett, 1992; 1993)

(16th–18th C.)

- phonetic digest codes
 - ▶ SOUNDEX (Russel, 1918)
 - ▶ PHONIX (Gadd, 1988; 1990)
- approximate (inverse) matching
 - ▶ string edit distance (Levenshtein, 1966)
 - ▶ letter n -gram distance (Pollock & Zamora, 1984)
- improved recall ($rc_{typ} \approx 96\%$) for $k_{Lev^{-1}} \in \{10, 20\}$

UCREL/UCLan VARD Project, Lancaster (16th–19th C.)

(Rayson, Archer & Smith, 2005; Baron & Rayson, 2008; 2011)

- manual or trained letter-replacement heuristics
 - ▶ unweighted k -local rules (Rayson & Baron, p.c.)
- improved accuracy vs. conventional spell-checkers
 - ▶ $\approx 50\%$ token-wise error reduction (rw $\approx 46\%$)

Diachronic Conflation: German

RSNSR Project, Duisburg-Essen

(14th–19th C.)

"Regelbasierte Suche in Textdatenbanken mit nichtstandardisierter Rechtschreibung"

(Kempken, 2005; Pilz et al., 2006, 2008; Ernst-Gerlach & Fuhr, 2006, 2007, 2010; ...)

- inverse canonicalization for information retrieval
- trained edit distance (*Cendrowska, 1987; Ristad & Yianilos, 1998*)
- improved type-wise precision & recall vs. DL distance
 - ▶ $\Delta(\text{pr}_{\text{typ}}) \approx 39\%$; $\Delta(\text{rc}_{\text{typ}}) \approx 31\%$ ($\text{rw} \approx 56\%; 45\%$)

IMPACT Project (OCR), CIS München

(16th–19th C.)

(Gotscharek, Neumann, Reffle, Ringlstetter & Schulz, 2009a; b; c)

- “hypothetical dictionary” for canonicalization
 - ▶ unweighted k -local rules (*Ringlstetter, p.c.*)
 - ▶ 19th C. tok: pr = 97.7%; rc = 95.0% ($\text{hmm} = 99.7\%; 99.1\%$)
- “channel error profiles” for precision recovery

— LTS FST Construction —

Step 1: Rule Match Detection

- Aho-Corasick Pattern Matchers (ACPMs)
- Distinguish **left** vs. **right** of current I/O position
- Intermediate **alphabets**: matched rule subsets
- Composition of (reversed) partial ACPMs models
LTS **configurations**

Step 2: Output Filter

- Composed with match-detection FST
- Selects **applicable** rules
- Simulates **position increment**

Aho-Corasick Pattern Matching

Given

- A set $P = \{p_1, \dots, p_{n_P}\} \subseteq \Sigma^*$ of **patterns**

Task

- **Identify all occurrences** of any pattern $p \in P$ in an input string $w \in \Sigma^*$

Aho-Corasick Algorithm

(Aho & Corasick, 1975)

- Constructs a **pattern-matching FST**

$$\text{AC}(P) : \mathcal{A}^* \rightarrow \wp(P)^*$$

- **Linear** runtime complexity:

$$\mathcal{O} = \mathcal{O}(|w|)$$

- ▶ More like $\mathcal{O}(|w| + \sum_{i=1}^{n_P} |p_i| + \sum_{j=1}^{|w|} |P \sim_j w|)$
... but who's counting?

ACPM Construction Sketch

Input

- Prefix Tree Acceptor (trie) for P

Output

- FST $\text{AC}(P) : \mathcal{A}^* \rightarrow \wp(P)^*$ such that

$$[\text{AC}(P)](w) = \text{cat}_{i=1}^{|w|} \left\{ p \in P \mid p = w_{i-|p|+1}^i \right\}$$

Method

- Transition function
- Failure function
- Output function

goto : $Q \times \mathcal{A} \rightarrow Q$

fail : $Q \rightarrow Q$

out : $Q \rightarrow \wp(P)$

Together, **goto**, **fail**, and **out** define a conventional arc-dependent input-deterministic transition function

$$\delta : Q \times \mathcal{A} \rightarrow Q \times \wp(P)$$

ACPM Limitations & Workarounds

Delayed Output

- ACPM output occurs at pattern terminus
- **Problem** for LTS rule lookahead / -behind
- **Solution:** construct 2 independent ACPMs

Information Loss

- ACPM outputs only pattern sets, not input symbols
- **Problem** for dual-ACPM construction (need both)
- **Solution:** preserve input in lookbehind ACPM

Fixed Increment

- ACPM outputs a pattern set for each input symbol
- **Problem** for LTS rule-dependent increment
- **Solution:** filter out unneeded outputs

LTS FST: Match Detection

Left-Context Matching

(Lookbehind)

$$\begin{aligned}
 M_L &\cong \text{AC}\left(\{\alpha \mid (\alpha[\beta]\gamma \rightarrow \pi) \in R\}\right) \\
 &: \mathcal{A}_w^* \rightarrow (\mathcal{A}_w \times \wp(R))^* \\
 &: w \mapsto \text{cat}_{i=1}^{|w|} \left\langle w_i, \left\{ (\alpha[\beta]\gamma \rightarrow \pi) \in R \mid w_{i-|\alpha|}^{i-1} = \alpha \right\} \right\rangle
 \end{aligned}$$

Target & Right-Context Matching

(Lookahead)

$$\begin{aligned}
 M_R &\cong \text{rev}\left(\text{AC}\left(\{\text{rev}(\beta\gamma) \mid (\alpha[\beta]\gamma \rightarrow \pi) \in R\}\right)\right) \\
 &: (\mathcal{A}_w \times \wp(R))^* \rightarrow \wp(R)^* \\
 &: \langle w_i, S_i \rangle_{i=1}^{|w|} \mapsto \\
 &\quad \text{cat}_{i=1}^{|w|} \left(S_i \cap \left\{ (\alpha[\beta]\gamma \rightarrow \pi) \in R \mid w_i^{i+|\beta\gamma|-1} = \beta\gamma \right\} \right)
 \end{aligned}$$

Match Detection FST

(Current Match Subset)

$$M_{LR} = M_L \circ M_R : \mathcal{A}_w^* \rightarrow \wp(R)^* : w \mapsto \text{cat}_{i=1}^{|w|} R \sim_i w$$

LTS FST: Output Filter

Notation

- Let $\mathcal{A}_{LR} \subseteq \wp(R)$ be the output alphabet of M_{LR}
- For each $S \in \mathcal{A}_{LR}$, let $(\alpha_S[\beta_S]\gamma_S \rightarrow \pi_S) = \min_{\prec} S$

Output Filter Construction

- Define output filter $M_O : \mathcal{A}_{LR}^* \rightarrow \mathcal{A}_p^*$ as:

$$M_O = \left(\bigcup_{S \in \mathcal{A}_{LR}} \left(\underbrace{(S : \pi_S)}_{\text{apply}} \underbrace{(\mathcal{A}_{LR} \times \{\varepsilon\})^{|\beta_S|-1}}_{\text{increment}} \right) \right)^*$$

LTS Transducer

- Then, the desired LTS FST can be defined as:

$$M_{\text{pho}} = (M_{LR} \circ M_O) = (M_L \circ M_R \circ M_O) : \mathcal{A}_w^* \rightarrow \mathcal{A}_p^*$$

— Lemma Instantiation —

Lemma Instantiation: Sketch

Idea

- Exploit dictionary-corpus structure
- Assume each quote $q \in \mathcal{Q}$ contains an instance $i \in \mathcal{W}$ of the associated dictionary lemma $\ell \in \mathcal{L}$

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: Method

Empirical Probability Estimates

$$P(\ell, i) = \frac{\sum_{w_i \in \text{pho}^{-1}(i)} \sum_{w_\ell \in \text{pho}^{-1}(\ell)} f(W=w_i, L=w_\ell)}{|Corpus|}$$

$$P(\ell) = \sum_i P(\ell, i)$$

$$P(i) = \sum_\ell P(\ell, i)$$

Pointwise Mutual Information

$$I(\ell, i) = \log_2 \frac{P(\ell, i)}{P(\ell)P(i)}$$

$$\tilde{I}(i|\ell) = \frac{I(\ell, i) - \min I(\ell, \mathcal{W})}{\max I(\ell, \mathcal{W}) - \min I(\ell, \mathcal{W})}$$

$$\tilde{I}(\ell|i) = \frac{I(\ell, i) - \min I(\mathcal{L}, i)}{\max I(\mathcal{L}, i) - \min I(\mathcal{L}, i)}$$

Edit-Distance Threshold

$$d_{\max}(\ell, i) = \min\{|\ell|, |i|\} - 1$$

Lemma Instantiation: Method (cont.)

Phonetic Similarity

$$\text{sim}(\ell, i) = \begin{cases} \frac{d_{\max}(\ell, i) - d_{\text{Lev}}(\ell, i)}{d_{\max}(\ell, i)} & \text{if } d_{\text{Lev}}(\ell, i) \leq d_{\max}(\ell, i) \\ 0 & \text{otherwise} \end{cases}$$

Instantiation Likelihood Function L

$$L(i, \ell) = \frac{\text{sim}(\ell, i) \times (\tilde{I}(\ell|i) + \tilde{I}(i|\ell))}{2}$$

Instantiation Heuristic

$$\mathcal{I}(\cdot) : \mathcal{Q} \rightarrow \mathcal{W}$$

$$: q \mapsto \arg \max_{w \in q} L(\text{pho}(w), \text{pho}(\text{lemma}(q)))$$

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

$$\left(\text{lemma}(\mathcal{I}^{-1}(w)) \cap \text{lemma}(\mathcal{I}^{-1}(v)) \neq \emptyset \right)$$

— Online k -Best Lookup —

Modified Dijkstra (1959) Algorithm

```
1: function DIJKSTRA-KBEST-STRINGS( $\vec{w}, C, k, c_{\max}, x_{\max}$ )
2:    $S := \{\langle q_0, \varepsilon \rangle\}$                                      /* Initialize */
3:    $d[\cdot] := \{\langle q_0, \varepsilon \rangle \mapsto \bar{1}\}$            /* Sparse partial minimum-cost map */
4:    $d_{\Pi, k}[\cdot] := \emptyset$                                 /* Track  $k$ -best final configurations */
5:   while  $S \neq \emptyset$                                          /* Main loop */
6:      $\langle q, s \rangle := \arg \min_{\langle q, s \rangle \in S, <_{\mathcal{K}}} d[q, s]$       /* Best-first search */
7:      $S := S \setminus \{\langle q, s \rangle\}$ 
8:     if  $x_{\max} = 0$  then break                                 /* Too many extractions */
9:      $x_{\max} := x_{\max} - 1$ 
10:    if  $d[q, s] >_{\mathcal{K}} c_{\max}$  then break                  /* Cost upper-bound */
11:    if  $q \in F$  then                                         /* Finality check */
12:       $d_{\Pi, k}[q, s] := d[q, s]$ 
13:      if  $|d_{\Pi, k}| = k$  then break                         /* Greedy termination */
14:    foreach  $e \in \text{ARCS}(\vec{w}, C, q)$                  /* Outgoing arcs (online expansion) */
15:       $d' := d[q, s] \otimes c[e]$                                /* Accumulate */
16:       $s' := s \circ o[e]$                                        /* Append */
17:      if  $d' <_{\mathcal{K}} \text{COST}(d[\cdot], \langle n[e], s' \rangle)$  then /* Relax */
18:         $d[n[e], s'] := d'$ 
19:       $S := S \cup \{\langle n[e], s' \rangle\}$ 
20:    return  $d_{\Pi, k}[\cdot]$                                      /* Final output */
```

k -Best Lookup: Subroutines

```
21: function ARCS( $\vec{w}, C, q$ )                                /* Online cascade composition */
22:     return EXPAND-ARCS( $C, q, 1, \vec{w}_1$ )  $\cup$  EXPAND-ARCS( $C, q, 1, \varepsilon$ )
23: function EXPAND-ARCS( $C, q, i, a$ )                         /* Recursive guts for ARCS() */
24:     local  $A := \emptyset$ 
25:     foreach  $e \in E_i \cup \{(q[i], q[i], \varepsilon, \varepsilon, \bar{1})\} : p[e] = q[i] \& i[e] = a$ 
26:         if  $i = |C|$  then
27:              $A := A \cup \{e\}$ 
28:         else
29:             foreach  $e' \in \text{EXPAND-ARCS}(C, q, i + 1, o[e])$ 
30:                  $A := A \cup \left\{ (\langle p[e], p[e'] \rangle, \langle n[e], n[e'] \rangle, i[e], o[e'], c[e] \otimes c[e']) \right\}$ 
31:     return  $A$ 
32: function COST( $d[\cdot], q$ )                                /* Minimum cost tracking for sparse partial map */
33:     if  $d[q]$  defined then return  $d[q]$ 
34:     return  $\bar{0}$ 
```

Complexity

$$\mathcal{O}(\text{DIJKSTRA}) = \mathcal{O}(|E| + |V| \log |V|)$$

$$\mathcal{O}(\text{KBEST}[-x_{\max}]) = \mathcal{O}\left(|E_{\Pi,k}| + \mathcal{O}(\text{ARCS}) |V_{\Pi,k}| \log |V_{\Pi,k}|\right)$$

$$\mathcal{O}(\text{KBEST}[+x_{\max}]) = \mathcal{O}\left(x_{\max} \times \mathcal{O}(\text{ARCS}) \log x_{\max}\right)$$

$$\mathcal{O}(\text{ARCS}) = \mathcal{O}\left(\prod_{i=1}^{n_C} \deg(q \in Q_i)\right)$$

where:

$$V_{\Pi,k} = \left\{ \langle q, s \rangle \in Q \times \mathcal{B}^* \mid d[q, s] \leq_k \max(\text{rng}(d_{\Pi,k})) \right\}$$

$$E_{\Pi,k} = \left\{ \langle \langle q, s \rangle, e \rangle \in V_{\Pi,k} \times E \mid p[e] = q \right\}$$

- Fibonacci Heap *(Fredman & Tarjan, 1987)*
- Linear sorted arc expansion

Degenerate Cycles

A cycle $\pi \in E^*$ is “**degenerate**” if $\exists \pi' \in E^*, \forall n \in \mathbb{N}$:

- $\pi'\pi \in \Pi(q_0, Q)$ /* accessible */
- $p[\pi] = n[\pi]$ /* cyclic */
- $i[\pi] = \varepsilon$ /* empty input */
- $o[\pi] \neq \varepsilon$ /* non-empty output */
- $c[\pi^n] \leq_{\mathcal{K}} c_{\max}$ /* cost-free */

Consequence: Infinite Loop

- **Accessibility** \rightsquigarrow enqueue-ability
 - **Empty input** \rightsquigarrow word-length independence
 - **Non-empty output** \rightsquigarrow new queue items
 - **Cost freedom** \rightsquigarrow iterated RELAX, EXTRACT
- \rightsquigarrow **KABOOM!**

— HMM Disambiguator —

Basic Model

Common Variables

- $\mathcal{W} \subset \tilde{\mathcal{A}}^*$ known extant words
- $\mathbf{u} \notin \mathcal{W}$ designated “unknown word” symbol
- $S = \langle w_1, \dots, w_{n_S} \rangle$ current input sentence
- $R = \{r_1, \dots, r_{n_R}\}$ (opaque) type-wise conflators
- $m \in \mathbb{N}$ model order (m -grams)

Disambiguator $D = \langle \mathcal{Q}, \mathcal{O}_S, \Pi, A, B_S \rangle$

- $\mathcal{Q} = (\mathcal{W} \cup \{\mathbf{u}\}) \times R$
- $\mathcal{O}_S = \bigcup_{i=1}^{n_S} \{w_i\}$
- $\Pi : \mathcal{Q} \rightarrow [0, 1] : q \mapsto p(Q_1 = q)$
- $A : \mathcal{Q}^m \rightarrow [0, 1] : q_1^m \mapsto p(Q_i = q_m | Q_{i-m+1}^{i-1} = q_1^{m-1})$
- $B_S : \mathcal{Q} \times \mathcal{O}_S \rightarrow [0, 1] : \langle q, o \rangle \mapsto p(O = o | Q = q)$

Probability Computation

- Sentence probability
 - ▶ $p(S = w_1^{n_S}) = \sum_{q_1^{n_S} \in \mathcal{Q}^{n_S}} p(S = w_1^{n_S}, Q = q_1^{n_S})$
- Path probability
 - ▶ $p(S = w_1^{n_S}, Q = q_1^{n_S}) = \prod_{i=1}^{n_S} p(q_i | q_{i-m+1}^{i-1}) p(w_i | q_i)$

Markov Assumptions

- m -local state dependencies
 - ▶ $p(q_i | q_1^{i-1}, w_1^{i-1}) = p(q_i | q_{i-m+1}^{i-1})$
- Context-independent observations (surface forms)
 - ▶ $p(w_i | q_1^i, w_1^{i-1}) = p(w_i | q_i)$

Transition Probabilities

Assumptions

- $p(Q = \langle \tilde{w}_q, r_q \rangle) = p(W = \tilde{w}_q)p(R = r_q)$
: independence of extant form from conlator
- $p(R = r) = \frac{1}{n_R}$: *conflators are uniformly distributed*

Transition Probability Estimation

- $\Pi(q) := p(W_1 = \tilde{w}_q) / n_R$
- $A(q_1^m) := p(W_i = \tilde{w}_{q_m} | W_{i-m+1}^{i-1} = \tilde{w}_{q_1}^{q_{m-1}}) / n_R$
- \equiv word m -gram model for contemporary forms

Smoothing

- Unknown words: $f(u) := \frac{1}{2}$ *(Lidstone, 1920)*
- Linear interpolation of m -grams *(Jelinek & Mercer, 1980; 1985)*

Lexical Probabilities

Maxwell-Boltzmann Approximation

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

Parameters

$$\begin{aligned} b &= 2 \\ \beta &= -1 \\ R &= \{\text{xlit}, \text{pho}, \text{rw}\} \\ \downarrow[w]_r &= [w]_r \cap \mathcal{W} \\ d_{\text{xlit}}(w, \tilde{w}) &= 2/|w| && \text{if } \tilde{w} = \text{xlit}^*(w) \\ d_{\text{pho}}(w, \tilde{w}) &= 1/|w| && \text{if } \tilde{w} \in \downarrow[w]_{\text{pho}} \\ d_{\text{rw}}(w, \tilde{w}) &= \|M_{\text{rw}} \circ A_{\text{Lex}}\|(w, \tilde{w})/|w| && \text{if } \tilde{w} = \text{best}_{\text{rw}}(w) \end{aligned}$$

Runtime Disambiguation

Viterbi Algorithm

(Viterbi, 1967)

$$\hat{q}_1^{n_S} = \text{VITERBI}(S, D) = \arg \max_{q_1^{n_S} \in Q^{n_S}} p(q_1^{n_S}, S | D)$$

Disambiguator Output

$$\hat{S} = \langle \hat{w}_1, \dots, \hat{w}_{n_S} \rangle$$

where:

$$\hat{w}_i := \begin{cases} \text{witness}(\downarrow[w]_{r_{\hat{q}_i}}) & \text{if } \tilde{w}_{\hat{q}_i} = \mathbf{u} \\ \tilde{w}_{\hat{q}_i} & \text{otherwise} \end{cases}$$

— Evaluation Measures —

Evaluation Measures: Basic

Common Definitions

- canonicalizers

$$C = \{c_1, \dots, c_{n_C}\}$$

- test corpus

$$G = \langle g_1, \dots, g_{n_G} \rangle$$

$$\triangleright g_i = \langle w_i, \tilde{w}_i, [w_i]_{c_1}, \dots, [w_i]_{c_{n_C}} \rangle \in \mathcal{A}^* \times \mathcal{A}^* \times \wp(\mathcal{A}^*)^{n_C}$$

- queries

$$Q = \bigcup_{i=1}^{n_G} \{\tilde{w}_i\}$$

Relevance & Retrieval

for each canonicalizer $c \in C$ and query $q \in Q$, define:

- by token

$$\text{relevant}_{\text{tok}}(q) = \{i \in \mathbb{N} \mid q = \tilde{w}_i\}$$

$$\text{retrieved}_{\text{tok},c}(q) = \{i \in \mathbb{N} \mid q \in [w_i]_c\}$$

- by type

$$\text{relevant}_{\text{typ}}(q) = \bigcup_{i \in \text{relevant}_{\text{tok}}(q)} \{w_i\}$$

$$\text{retrieved}_{\text{typ},c}(q) = \bigcup_{i \in \text{retrieved}_{\text{tok},c}(q)} \{w_i\}$$

Precision & Recall

For each evaluation mode $m \in \{\text{tok}, \text{typ}\}$, define:

Precision

$$\text{pr}_{m,c} = \frac{|\bigcup_{q \in Q} \text{retrieved}_{m,c}(q) \cap \text{relevant}_m(q)|}{|\bigcup_{q \in Q} \text{retrieved}_{m,c}(q)|}$$

- Likelihood of **relevance**, given **retrieval**

Recall

$$\text{rc}_{m,c} = \frac{|\bigcup_{q \in Q} \text{retrieved}_{m,c}(q) \cap \text{relevant}_m(q)|}{|\bigcup_{q \in Q} \text{relevant}_m(q)|}$$

- Likelihood of **retrieval**, given **relevance**

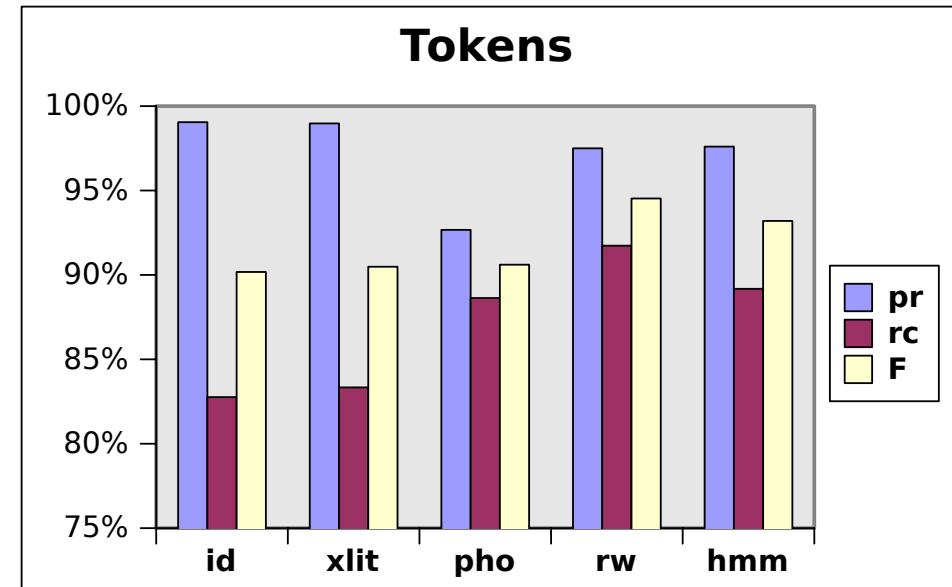
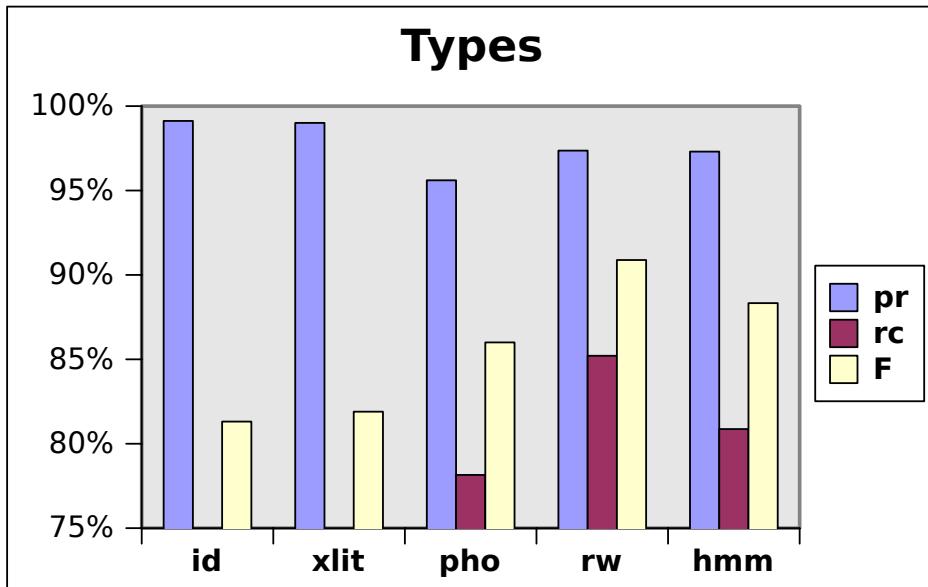
Precision-Recall Composite F

$$F_{m,c} = \frac{2 \times \text{pr}_{m,c} \times \text{rc}_{m,c}}{\text{pr}_{m,c} + \text{rc}_{m,c}}$$

- Unweighted harmonic average

(van Rijsbergen, 1979: "E")

Evaluation: DWB-1 Verse



	% Types			% Tokens		
	pr	rc	F	pr	rc	F
id	99.1	68.9	81.3	99.0	82.8	90.2
xlit	99.0	69.8	81.9	99.0	83.3	90.5
pho	95.6	78.2	86.0	92.7	88.6	90.6
rw	97.4	85.2	90.9	97.5	91.7	94.5
hmm	97.3	80.9	88.3	97.6	89.2	93.2