

Experiments in Morphology Induction

Bryan Jurish

moocow@bbaw.de

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

June 20, 2006



Overview

The Big Picture

- Desiderata
- Evaluation

Experiments

- Phonetization
- Goldsmith
- Creutz & Lagus

Next Steps

- Corpus Manipulation
- Alternate Learning Procedures
- Applications



The Big Picture



Desiderata

Input

- Raw electronic text corpus (tokenized)
- Historical orthography
- (Later): phonetized

Output

- Morphological Model
 - Surface segmentation $M_{seg} : \Sigma^* \rightarrow (\Sigma \cup \{.\})^*$
 - and/or Conflation pairs $M_{con} \subseteq \Sigma^* \times \Sigma^*$
- (?) Morphological paradigms (“signatures”)
- (?) Analyzer for arbitrary text (“stemmer”)



Evaluation: Segmentation

Segmentation Precision & Recall

$$\text{Precision} = \frac{\# \text{ correctly induced boundaries}}{\# \text{ induced boundaries}}$$
$$\text{Recall} = \frac{\# \text{ correctly induced boundaries}}{\# \text{ true boundaries}}$$

Segmentation “Pyrite” Standard

- Derived from TAGH lemmata for NEGRA
- Segment extraction FST + string alignment
- min edit distance \ll min # morphs \ll TAGH cost
- Inherently erroneous:

prüfung(en) \rightsquigarrow *prüfung.en* \neq *prüf.ung.en*

verwesung \rightsquigarrow *verwes.ung* \neq *ver.wes.ung*

tänzerinnen \rightsquigarrow *tänz.e.rinn.en* \neq *tänz.er.in.nen*



Evaluation: Conflation

Conflation Precision & Recall

$$\text{Precision} \quad ?= \frac{\# \text{ correctly induced pairs}}{\# \text{ induced pairs}}$$

$$\text{Recall} \quad ?= \frac{\# \text{ correctly induced pairs}}{\# \text{ true pairs}}$$

Conflation “Pyrite” Standard

(TODO!)

- Idea: derive from TAGH lemmata:

$$\begin{aligned} Conf(w) &= Lemma^{-1}(Lemma(w)) \\ &= \{v \mid Lemma(v) = Lemma(w)\} \end{aligned}$$

- Problem: Ambiguity resolution \Rightarrow moot + ???



Experiments



Phonetization

Motivation

- Mitigate sparse data due to historical orthography

Procedure

- Adapted Letter-to-Sound (LTS) ruleset from IMS German festival TTS system
- (Re-)implemented as a deterministic FST

Results (Grimm / “R” / headwords & verse)

- Speed improvement of 555.77% vs. festival

	Orthographic	Phonetic	Improvement
Tokens	220,967	220,967	0.00 %
Types	44,644	40,774	8.67 %
Hapax legomena	30,820	27,694	10.14 %



Goldsmith: Overview

- Minimum Description Length (MDL): seeks the model θ that maximimally compresses the analysis system (model & corpus):

$$\text{len}(\text{Analysis}) = \text{len}(\theta) + \text{len}(\text{Corpus}|\theta)$$

- Concatenative Language Model:

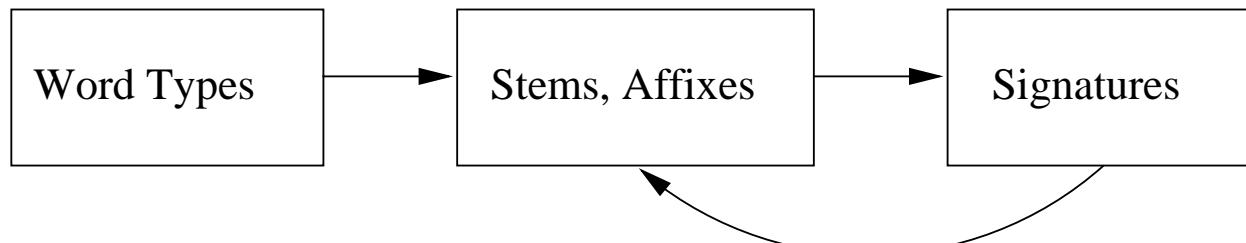
$$\mathcal{W} \subseteq (\text{Prefix?}) \text{ Stem } (\text{Suffix?})$$

- Paradigm Inclusion (“Signatures”):

$$\text{sig} : \text{Stem} \rightarrow \mathcal{P}(\text{Suffix}), \text{ where}$$

$$t \in \text{Stem}, f \in \text{sig}(t) \Rightarrow tf \in \mathcal{W}$$

- Heuristic Bootstrapping:

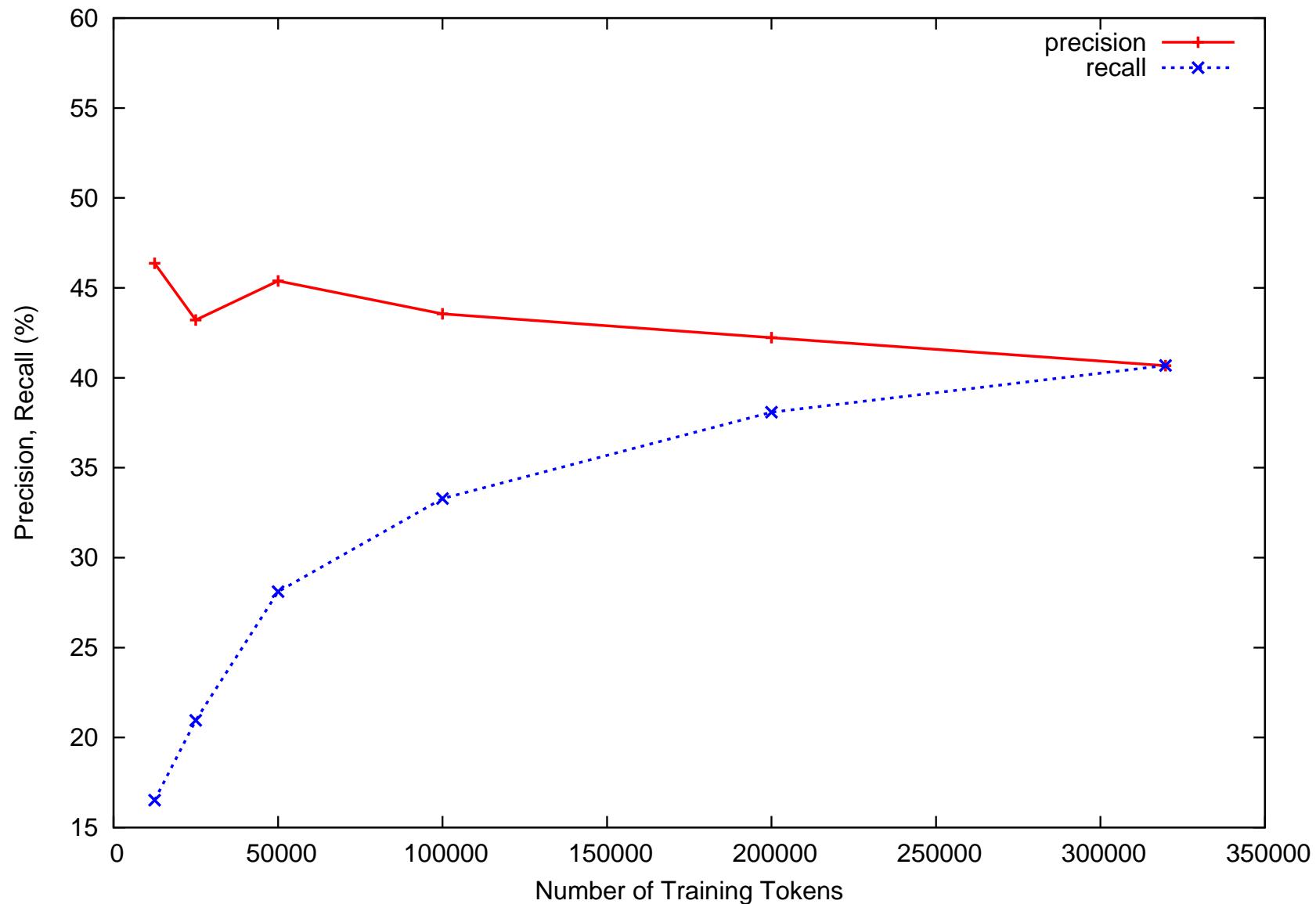


Goldsmith: Heuristics

- Successor “Frequency” (cardinality)
 - $SF(w, i) = \text{outDegree}(\text{ptaNode}(w_1 \cdots w_i))$
 - **Harris**: SF peaks indicate morpheme boundaries
 - **Goldsmith**: prefer short suffixes & impose SF maximum for boundary-adjacent states
- Minimum stem length
- Maximum affix length
- Minimum number of stems per suffix
- Minimum number of short-suffix stems
- Maximum entropy of stem-final n -grams
- ... *and many, many more!*

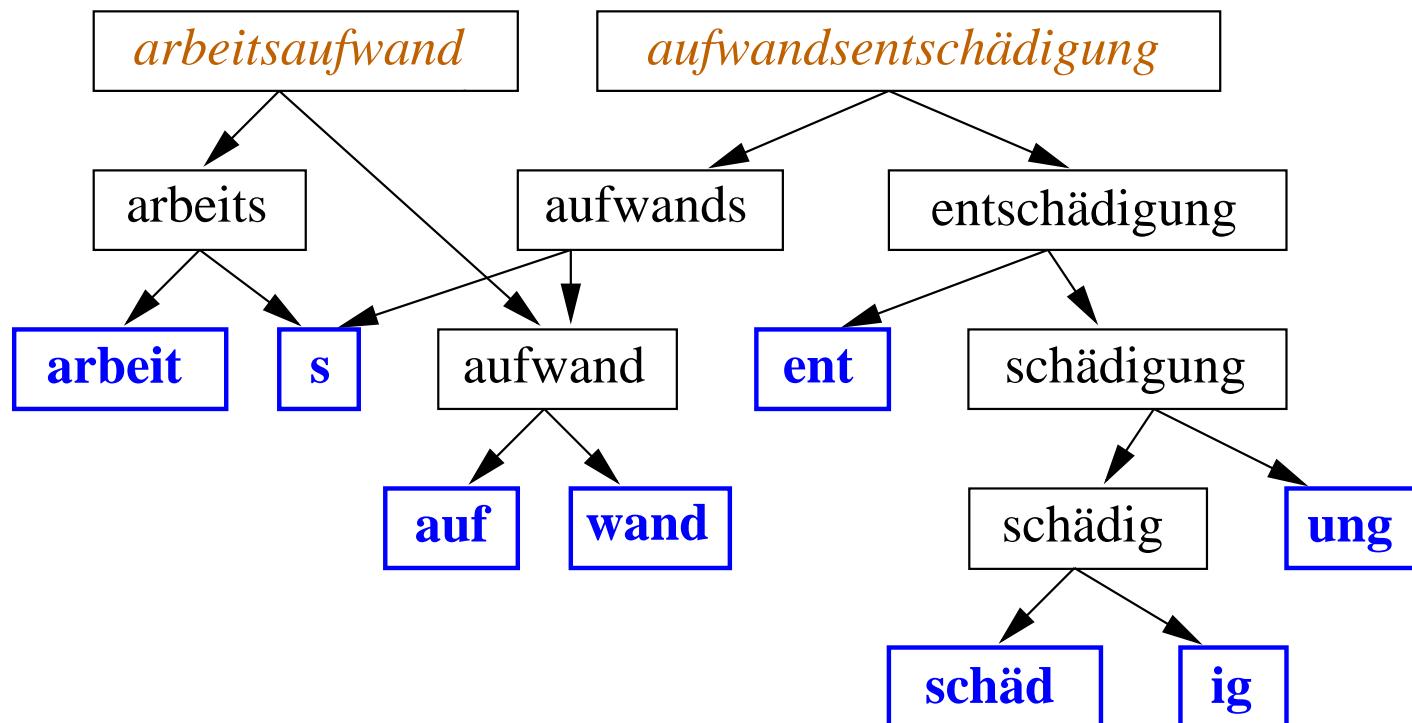


Goldsmith: Results (NEGRA)



Creutz & Lagus: Overview

- Incremental MDL procedure
- Designed for **highly inflective languages** (Finnish):
$$\mathcal{W} \subseteq ((\text{Prefix}^*) \text{ Stem } (\text{Suffix}^*))^+$$
- Recursive segmentation using hierarchical lexicon:



Creutz & Lagus: Procedure

Incremental Search

- Each **type** resegmented for each **token** instance
- “**Dreaming**”: random batch resegmentation

Recursive Binary Segmentation

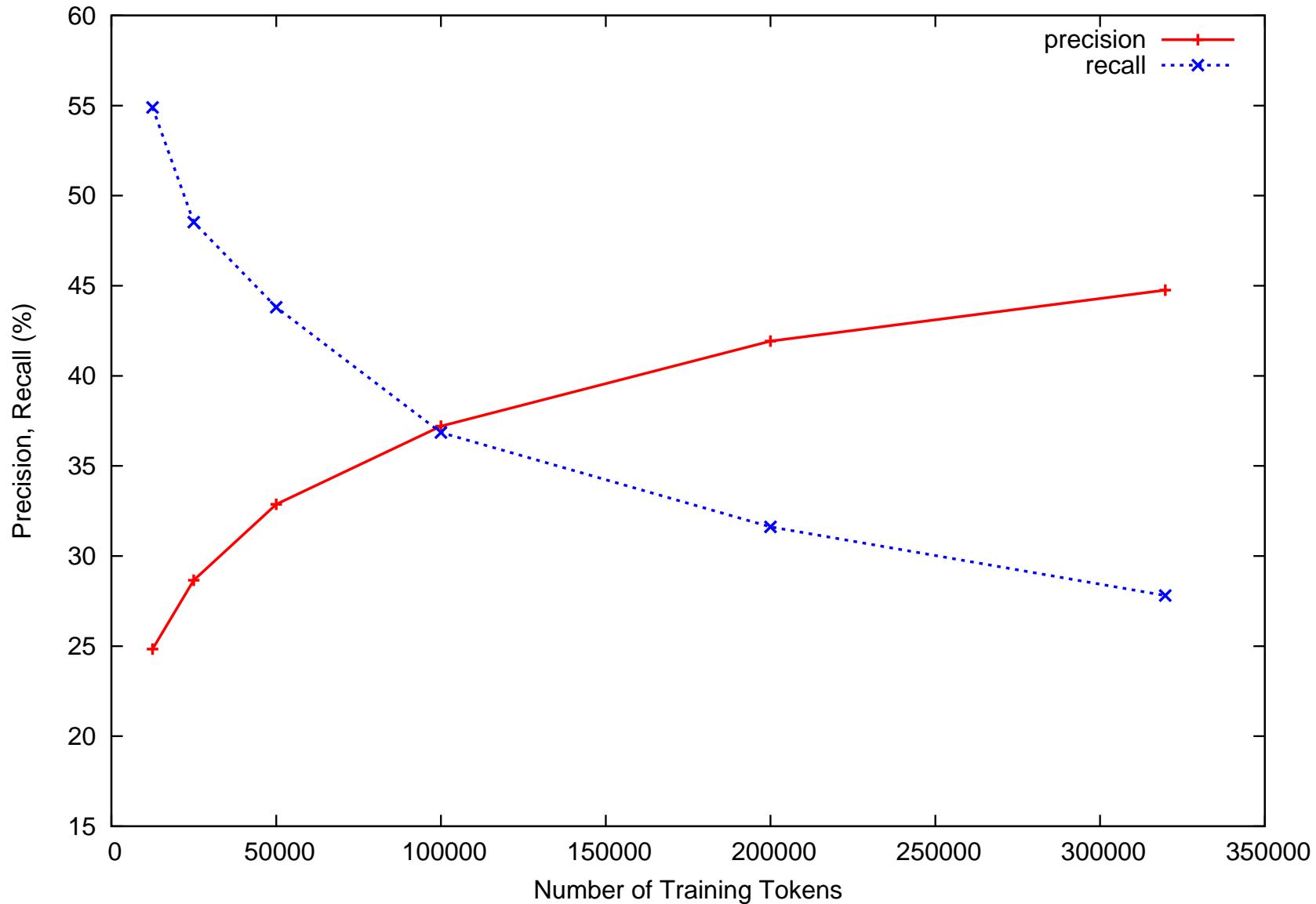
- Consider **every split** of each (sub)string
- **Greedy MDL**: choose **maximally compressing** split

Extensions

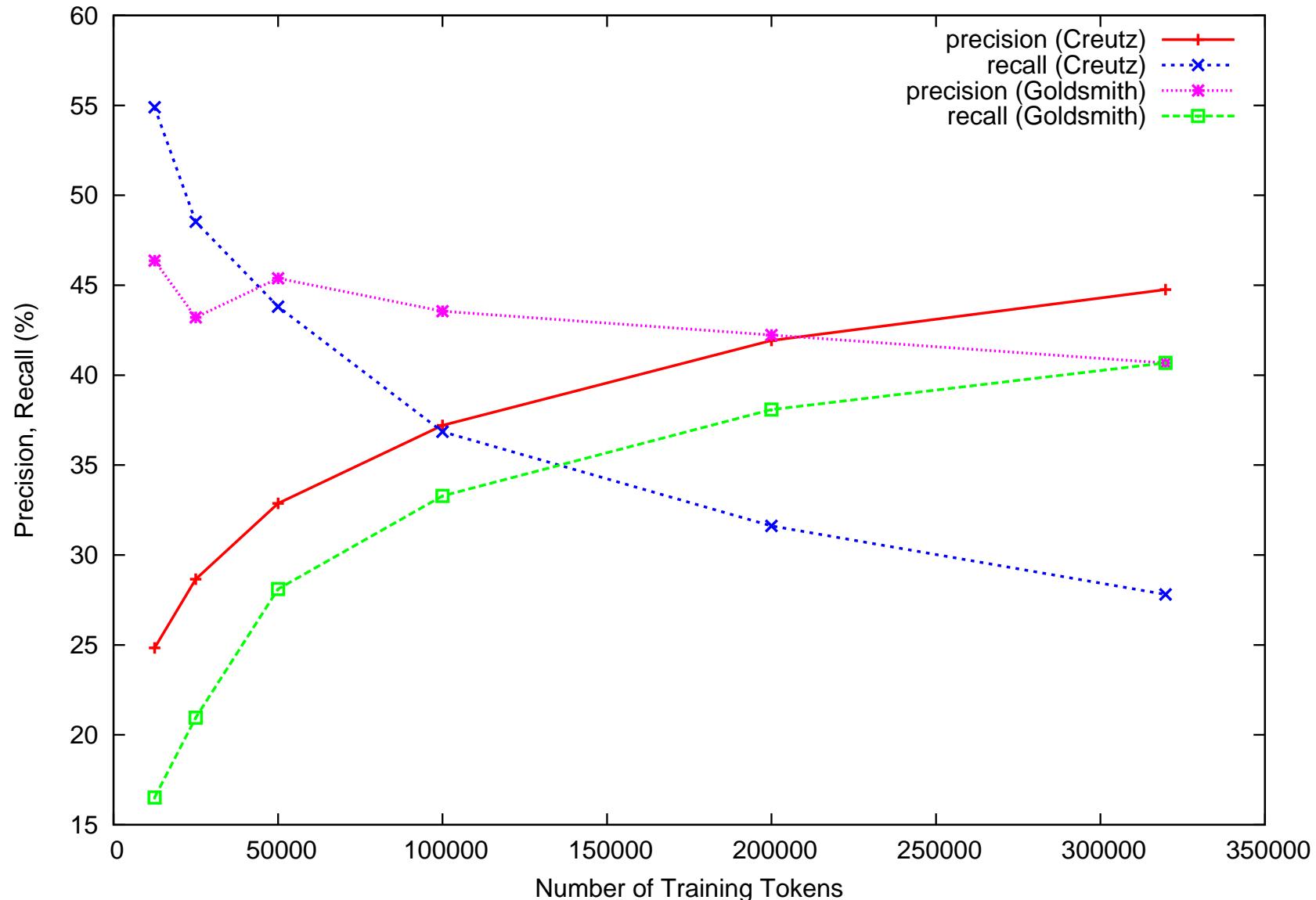
- **Bayesian priors** on morph **frequency & length**
- **Viterbi search** of morph grammar FSA



Creutz & Lagus: Results (NEGRA)



Creutz vs. Goldsmith: Results



Error Analysis

Goldsmith

- Tends to under-segment words (low recall):

abgebaut \rightsquigarrow *ab.gebaut* \neq *ab.ge.bau.t*

kaputtgemacht \rightsquigarrow *kaputt.gemacht* \neq *kaputt.ge.mach.t*

straßenbahnlinie \rightsquigarrow *straßenbahnlini.e* \neq *straße.n.bahn.linie*

Creutz & Lagus

- Tends to over-segment words (low precision):

abbremsen \rightsquigarrow *ab.b.r.em.s.en* \neq *ab.brems.en*

nein \rightsquigarrow *n.ein* \neq *nein*

wandel \rightsquigarrow *wand.el* \neq *wandel*

- Some putative false positives result directly from the erroneous pyrite standard



Next Steps

or:

What now?



Next Steps: Overview

Corpus Manipulation

- Phonetization
- Gold Standard

Alternate Learning Procedures

- (Hierarchical) Segmentation Methods
- Conflation Methods
- (Partial) Supervision

Applications

- Corpus Indexing
- “Open” Lexicon



Corpus Manipulation

Grimm Corpus

- Conversion to **UTF-8**
- Extraction of **prose quotation evidence**

Phonetization

- Adaptation of **IMS LTS Rules**
 - diacritics (e.g. ò, ó, ô, õ, ø, ...)
 - non-latin characters (e.g. cyrillic, greek, ...)
 - non-printing characters (e.g. [\202], ...)
- Fine-tuning for **historical orthography** (?)

umb ~> /?ümp/ ≠ /?üm/ ↪ um

sym ~> /zÿm/ ≠ /tsüm/ ↪ zum

het ~> /hët/ ≠ /hät/ ↪ hat



Gold Standard

Motivation

- Auto-generated “pyrite” standard is **erroneous**
- Enable **justifiable interpretation** of data

Goal(s)

- 10k–30k token *testing gold standard*
- Hand-segmented
- **Optional Annotations:**
 - phonetic form, POS, syntactic features, ...

Method

- **Correct pyrite standard**



Gold Standard: GUI

MUDL::Morph::Editor::Gtk2

File Corpus Analyses

Word Types

Word	Freq	An?
einziehen	1	<input checked="" type="checkbox"/>
einzige	3	<input checked="" type="checkbox"/>
einzig	1	<input checked="" type="checkbox"/>
einzustellen	2	<input type="checkbox"/>
eis	1	<input type="checkbox"/>
eisenbergs	1	<input type="checkbox"/>
eisensteins	1	<input type="checkbox"/>
e-jugend	1	<input type="checkbox"/>
ekd-kirchenamtes	1	<input type="checkbox"/>
elektro-unternehmen	1	<input type="checkbox"/>

Contexts

i-4	i-3	i-2	i-1	i	i+1	i+2	i+3	i+4	
der	woche	auf	praktikanten	praktikanten	einzustellen	,	weil	sie	befürchteten
sich	auf	die	siebener-formation	siebener-formation	einzustellen	.			
								

Edit

Segments: ein.zu.stell.en

Analyses Morphs

Segments	Distance	Analysis	Weight
ein.zu.stell.en	1	ein.zu.stell.e	0.000000
ein.zu.stell.en	1	ein.zu.stell.en	0.000000
ein.zu.stell.en	2	ein.zu.stell.t	0.000000
ein.zu.stell.en	2	ein.zu.stell.te	0.000000
ein.zu.stell.en	2	ein.zu.stell.ten	0.000000
einzu.stell.en	2	ein.stell.en	0.000000
ein.zu.stell.en	3	ein.zu.stell.st	0.000000
ein.zu.stell.en	3	ein.zu.stell.tet	0.000000

Info Filters

Words: 3622

Word: einzustellen

ID: 2746

Freq: 2

Analyzed: no



Alternate Learning Procedures

(Hierarchical) Segmentation Methods

- Alignment-Based Learning (ABL) [*van Zaanen (2000)*]
- Hierarchical Lexicon Compression
 - [*de Marcken (1995, 1996)*]
- **Problem(s):** termination threshhold

Conflation Methods

- Semantic estimators (WPMI, LSA)
 - [*Schone & Jurafsky (2000), Baroni et al. (2002)*]
- Lexical class estimators (inflection vs. derivation)
 - [*Yarowsky & Wicentowski (2000)*]
- **Problem(s):** corpus structure, rule extraction



(Partial) Supervision #1

Phone-Based Morphology

- **Idea:** Compose TAGH (M) and LTS (L) FSTs:

$$M' = (M^{-1} \circ L^{-1})^{-1}$$

- **Advantages:** uses existing tools (TAGH, moot, ...)
- **Disadvantages:** insufficient coverage

Nondeterministic Analysis

- **Idea:** extend LTS FST with phonetic distances
- **Advantages:** compatibility, linear-time search
- **Disadvantages:** thresholds (beam width), large search space



(Partial) Supervision #2

Learning from Paradigms

- **Idea:** encode *morphological paradigms* as *declarative constraints* [Forsberg et. al (2006)]
- **Advantages:** efficient use of supervisory effort
- **Disadvantages:** sparse data \Rightarrow fine tuning

Exploiting Dictionary Structure

- **Idea:** use **headwords** as **cues** to **conflation targets**
- **Advantages:** efficient use of corpus structure
- **Disadvantages:** limited applicability,
conflation-based approach



... and finally:

Applications

Corpus Index

- Segment-level search
- Conflation set search (period-independent)
- sounds-like queries

“Open” Lexicon

- Greedy affixation [Allen et al. (1987), Dutoit (1997)]
 - Can be implemented with (weighted) FSTs
 - Also useful for words unknown to TAGH
- Can be extended to accommodate compounds



The End



Thanks for listening!

