EXTRACTING MORPHOPHONOLOGY FROM SMALL CORPORA

THE UNIVERSITY OF CHICAGO

Marina Ermolaeva
mermolaeva@uchicago.edu

Introduction

- What can be learned from a **small sample of glossed words**?
  - alternations behind allomorphy
  - distribution of allomorphs
  - missing morphs
  - instances of mislabeling
- Which of these tasks can be accomplished **automatically**...
  ... and how much data is required?

- **Case study**:
  Mishar dialect of Tatar, Turkic/Kipchak (LYUTIKOVA et al. 2007)

	Tokens
Total words	3090
Polymorphemic words	1736
Gloss-tagged morphemes	2850

(Morpho)phonological phenomena

- Vowel harmony:

	[−BK −RND]	[−BK +RND]	[+BK −RND]	[+BK +RND]
[+HI −LO]	i	ü	ɤj	u
[−HI −LO]	e		ɤ	
[−HI +LO]	ä		a	

(1) a. bala-lar-ɤbɤz-ga
child-PL-P1PL-DAT
b. täräz-lär-ebez-gä
window-PL-P1PL-DAT

- Non-canonical harmony:

(2) a. tarix-ɤ
history-P3
b. činovnig-ɤ
official-P3

- Local processes:

(3) a. kibet-tän shop-ABL b. kɤz-dan girl-ABL c. urɤn-nan place-ABL

(4) a. jɤrt-ta yard-LOC b. kɤz-da girl-LOC c. ten-dä night-LOC

(5) a. at-lar horse-PL b. kɤz-lar girl-PL c. ujɤn-nar game-PL

- Interacting processes:

(6) a. bala-m child-P1SG b. set-em milk-P1SG c. kɤz-ɤm girl-P1SG

- Non-canonical voicing:

(7) a. i-kän AUX-PFCT b. di-gän speak-PFCT

- Allomorphy for morphosyntactic features:

(8) a. bašl-a-r begin-ST-POT b. bul-m-a-s be-NEG-ST-POT

String differences

- GOLDSMITH 2011: **difference** and **commonality** operators over strings and **paradigms** (sets of strings)

Right difference: $\frac{\text{jumps}}{\text{jumping}}R = \frac{\text{s}}{\text{ing}}$ Left difference: $\frac{\text{like}}{\text{dislike}}L = \frac{\emptyset}{\text{dis}}$

- A paradigm is **regular** iff its self-difference array has a single common numerator in each row and all its commonalities are identical

	jump	jumps	jumped	jumping
jump		$\frac{\emptyset}{\text{s}}$	$\frac{\emptyset}{\text{ed}}$	$\frac{\emptyset}{\text{ing}}$
jumps	$\frac{\text{s}}{\emptyset}$		$\frac{\text{s}}{\text{ed}}$	$\frac{\text{s}}{\text{ing}}$
jumped	$\frac{\text{ed}}{\emptyset}$	$\frac{\text{ed}}{\text{s}}$		$\frac{\text{ed}}{\text{ing}}$
jumping	$\frac{\text{ing}}{\emptyset}$	$\frac{\text{ing}}{\text{s}}$	$\frac{\text{ing}}{\text{ed}}$	

	jump	jumps	jumped	jumping
jump		jump	jump	jump
jumps	jump		jump	jump
jumped	jump	jump		jump
jumping	jump	jump	jump	

	try	tries	tried	trying
try		$\frac{\text{y}}{\text{ies}}$	$\frac{\text{y}}{\text{ied}}$	$\frac{\emptyset}{\text{ing}}$
tries	$\frac{\text{ies}}{\text{y}}$		$\frac{\text{s}}{\text{d}}$	$\frac{\text{ies}}{\text{ying}}$
tried	$\frac{\text{ied}}{\text{y}}$	$\frac{\text{d}}{\text{s}}$		$\frac{\text{ied}}{\text{ying}}$
trying	$\frac{\text{ing}}{\emptyset}$	$\frac{\text{ying}}{\text{ies}}$	$\frac{\text{ying}}{\text{ied}}$	

	try	tries	tried	trying
try		tr	tr	try
tries	tr		trie	tr
tried	tr	trie		tr
trying	try	tr	tr	

Extracting alternations

- An **alternation** is the set of numerators in a self-difference array with the following properties:
  - regular (ignoring any known alternations)
  - short differences (up to one character)
  - nonzero commonalities

- **Preprocessing**: arrange **morphs** (instances of morphemes) into sets by gloss, each in its own **group**

(9) Q: {[m ɤ], [m e]}
PST: {[d ɤ], [d e], [t ɤ], [t e]}
P1PL: {[b e z], [e b e z], [ɤ b ɤ z]}

- **Extraction step**: find alternations between groups within sets
- **Reduction step**: collapse groups that are identical up to known alternations
- Repeat until the number of groups stops decreasing

Input	Extraction	Reduction	Extraction	Reduction
[m ɤ], [m e]	{e, ɤ}	[m ɤ, m e]	–	[m ɤ, m e]
[d ɤ], [d e], [t ɤ], [t e]	–	[d ɤ, d e], [t ɤ, t e]	{d, t}	[d ɤ, d e, t ɤ, t e]
[b e z], [e b e z], [ɤ b ɤ z]	–	[b e z], [e b e z, ɤ b ɤ z]	{∅, e, ɤ}	[∅ b e z, e b e z, ɤ b ɤ z]

References:
BOUMA, G. 2009. Normalized (pointwise) mutual information in collocation extraction. • GOLDSMITH, J. 2011. A group structure for strings: Towards a learning algorithm for morphophonology. • LYUTIKOVA, E., KAZENIN, K., SOLOVYEV, V. & TATEVOSOV, S., eds. 2007. Mishar Dialect of Tatar Language: Essays on Syntax and Semantics

Intermediate results

- Started with 55 gloss tags and 160 distinct morphs
- Converged in **3 iterations**
- All morphs collapsed into 85 groups
- Discovered **12 alternations**

Correct	{d, n, t}, {d, t}, {∅, k, g}, {l, n}, {k, g}, {∅, e, ɤ}, {a, ä}, {e, ɤ}
Incomplete	{∅, ɤ}, {∅, ä}
Incorrect	{s, g}, {n, ŋ}

(10)
```
---- ATR
---- Group 0
[['s' 'ɤ' 'z']
 ['s' 'e' 'z']]
---- Group 1
[['l' 'e']
 ['l' 'ɤ']]
---- Group 2
[['g' 'e']
 ['g' 'ɤ']
 ['k' 'e']]
```

(11)
```
---- PL
---- Group 0
[['n' 'ä' 'r']
 ['n' 'a' 'r']
 ['l' 'ä' 'r']
 ['l' 'a' 'r']]
```

(12)
```
---- ORD
---- Group 0
[['e' 'n' 'č' 'e']
 [' ' 'n' 'č' 'e']
 ['ɤ' 'n' 'č' 'ɤ']]
```

(13)
```
---- COMP
---- Group 0
[['d' 'i' 'p']]
---- Group 1
[['r' 'ä' 'k']
 ['r' 'a' 'k']]
```

(14)
```
---- CMPR
---- Group 0
[['r' 'a' 'k']]
```

Contexts and PMI

- A **rule** consists of three components:
  - a set of triggers for each alternant
  - a set of transparent characters
  - directionality: **left** or **right**

- Bigrams for {e, ɤ} in #bašɤn#:

	Left	Right
Local	šɤ	ɤŋ
Nonlocal	#ɤ, bɤ, aɤ, šɤ	ɤŋ, ɤ#

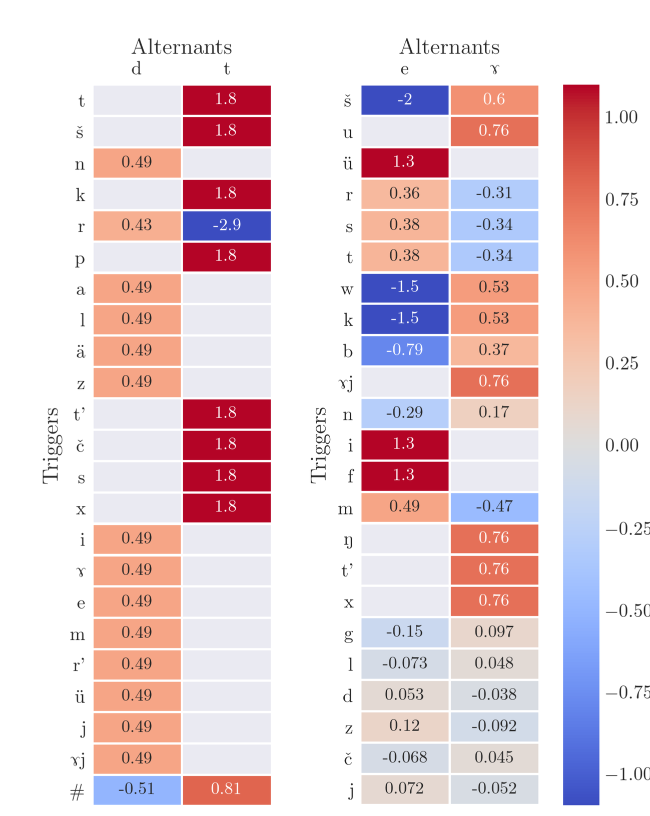
- **Pointwise mutual information**: measure of attraction between a pair of events (BOUMA 2009)

$$PMI_A(s;a) = \log_2 \frac{p(a,s)}{p(a)p(s)}$$

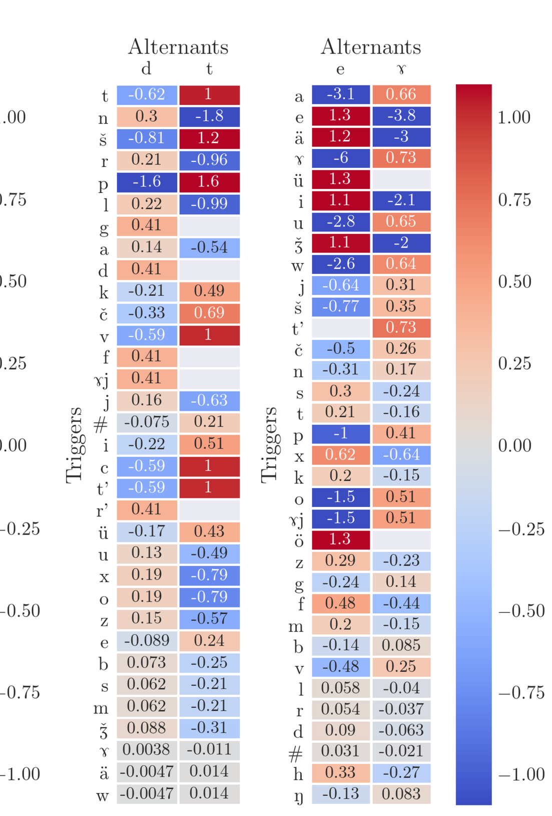
- PMI values correctly pair alternants with their trigger characters...

  ... if the directionality is correct

  ... and unless the character is transparent!



{d, t} {e, ɤ}
Local left bigrams

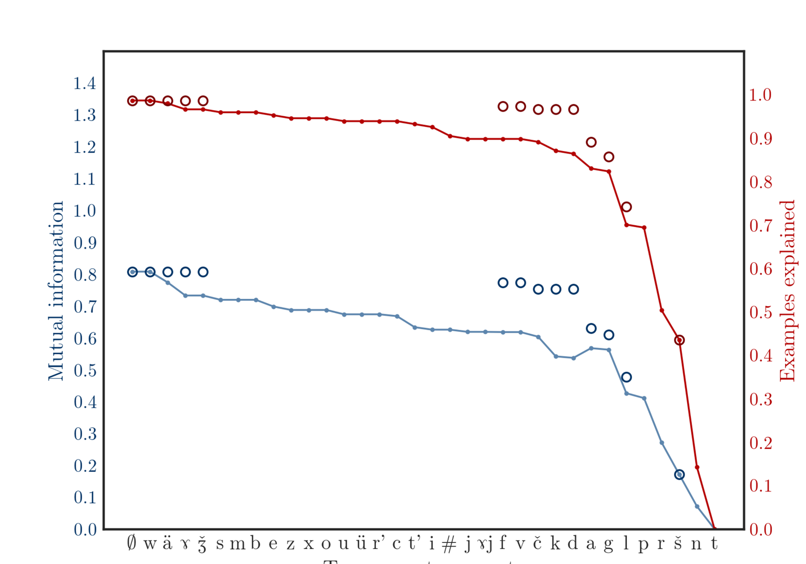


{d, t} {e, ɤ}
Nonlocal left bigrams

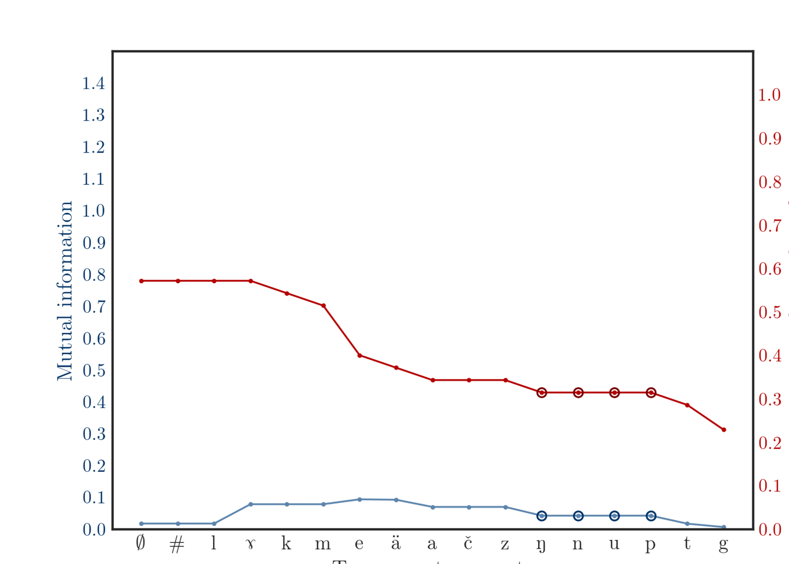
Learning and evaluating rules

- Start with **nonlocal PMIs**
- Marginalize over alternants to **rank** context segments
- Expand the list of transparent segments, recalculating **local PMIs**
- Track metrics to select the best rule:
  - average PMI over all bigrams
  - portion of examples explained
- Use the notion of natural class to find **phonologically viable** configurations

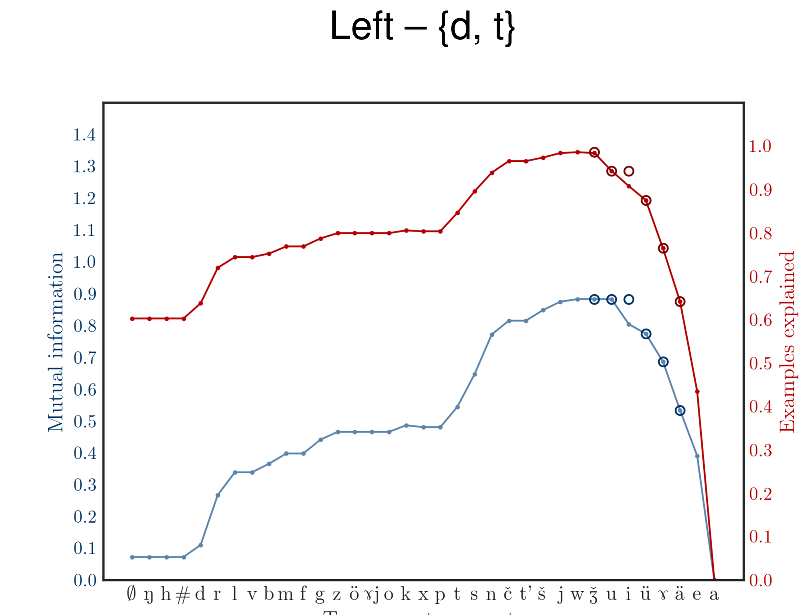
Alternation	MI (left)	EE (left)	MI (right)	EE (right)
{d, n, t}	1.4277	**1.0000**	0.0000	0.0000
{d, t}	0.8079	**0.9864**	0.0421	0.3143
{∅, k, g}	0.4951	**0.4909**	0.0000	0.0000
{l, n}	0.5547	**0.9738**	0.0514	0.1154
{k, g}	0.0814	**0.1653**	0.0000	0.0000
{∅, e, ɤ}	0.8431	**0.6220**	0.6205	0.5357
{a, ä}	0.8319	**0.9733**	0.8331	0.8387
{e, ɤ}	0.8825	**0.9857**	0.7148	0.7912
{∅, ɤ}	0.8113	1.0000	1.0000	1.0000
{∅, ä}	0.1651	0.4688	0.5586	**1.0000**



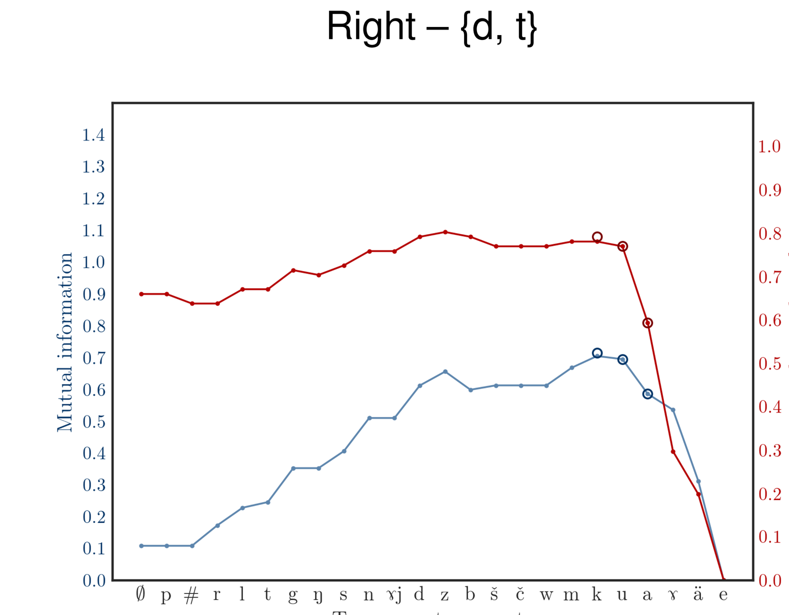
Left – {d, t}



Right – {d, t}



Left – {e, ɤ}



Right – {e, ɤ}