

Phonology as Compression

Capturing Vowel Harmony

Adam C. Baker
adamc@uchicago.edu

Department of Linguistics
University of Chicago

Linguistics Society of America, 2009

Outline

Overview: Phonology as Compression

The Phonological Framework

Models

Bigram Model

Long Distance V-to-V Model

Results

Turkish

Finnish

Other Languages

Vowel Harmony Statistics

Why do most of the vowel to vowel interactions in Finnish resemble Italian more than they resemble Turkish?

Vowel Harmony Statistics

Why do most of the vowel to vowel interactions in Finnish resemble Italian more than they resemble Turkish?

1. Answer (Kirchner (1993)): common vowels [a, i, u, e, o] are more free in their distribution than cross linguistically less common vowels.

Vowel Harmony Statistics

Why do most of the vowel to vowel interactions in Finnish resemble Italian more than they resemble Turkish?

1. Answer (Kirchner (1993)): common vowels [a, i, u, e, o] are more free in their distribution than cross linguistically less common vowels.
2. Model of vowel harmony (Goldsmith and Riggle (2007)): phonology as compression.

Vowel Harmony Statistics

Why do most of the vowel to vowel interactions in Finnish resemble Italian more than they resemble Turkish?

1. Answer (Kirchner (1993)): common vowels [a, i, u, e, o] are more free in their distribution than cross linguistically less common vowels.
2. Model of vowel harmony (Goldsmith and Riggle (2007)): phonology as compression.
3. Comparing V-to-V interactions and compression gains by modeling vowel interactions, Finnish is closer to Italian than Turkish.

Outline

Overview: Phonology as Compression The Phonological Framework

Models

Bigram Model

Long Distance V-to-V Model

Results

Turkish

Finnish

Other Languages

The Minimal Description Length Principle

Minimal Description Length (Rissanen (1978); de Marken (1996); Ellison (1994); Goldsmith (2001))

- ▶ Grammar assigns probability to a corpus which is used to compute a compressed length. (Shannon and Weaver (1949))

The Minimal Description Length Principle

Minimal Description Length (Rissanen (1978); de Marken (1996); Ellison (1994); Goldsmith (2001))

- ▶ Grammar assigns probability to a corpus which is used to compute a compressed length. (Shannon and Weaver (1949))
- ▶ Grammar is assigned an encoding.
 - ▶ More complex grammars have longer encodings. See appendix for grammar code length formula.

The Minimal Description Length Principle

Minimal Description Length (Rissanen (1978); de Marken (1996); Ellison (1994); Goldsmith (2001))

- ▶ Grammar assigns probability to a corpus which is used to compute a compressed length. (Shannon and Weaver (1949))
- ▶ Grammar is assigned an encoding.
 - ▶ More complex grammars have longer encodings. See appendix for grammar code length formula.
- ▶ Add code length of corpus and grammar.
 - ▶ Smaller total code length = better grammar.

Gibbs models

Probabilities assigned using a Gibbs distribution (Geman and Johnson (2001))

Gibbs models

Probabilities assigned using a Gibbs distribution (Geman and Johnson (2001))

- ▶ Phonological representations are evaluated by a set of scoring functions.

Gibbs models

Probabilities assigned using a Gibbs distribution (Geman and Johnson (2001))

- ▶ Phonological representations are evaluated by a set of scoring functions.
- ▶ The results of the scoring functions are totaled.

Gibbs models

Probabilities assigned using a Gibbs distribution (Geman and Johnson (2001))

- ▶ Phonological representations are evaluated by a set of scoring functions.
- ▶ The results of the scoring functions are totaled.
- ▶ Higher total score = lower probability.

Outline

Overview: Phonology as Compression

The Phonological Framework

Models

Bigram Model

Long Distance V-to-V Model

Results

Turkish

Finnish

Other Languages

Bigram Model

- ▶ The score is the bigram encoding length.

Bigram Model

- ▶ The score is the bigram encoding length.
- ▶ Captures all effects between adjacent segments.

Bigram Model

- ▶ The score is the bigram encoding length.
- ▶ Captures all effects between adjacent segments.
- ▶ Serves as a baseline of comparison for the vowel models.

Bigram Model

- ▶ The score is the bigram encoding length.
- ▶ Captures all effects between adjacent segments.
- ▶ Serves as a baseline of comparison for the vowel models.
- ▶ Allows us to define Mutual Information (pointwise mutual information):

Bigram Model

- ▶ The score is the bigram encoding length.
- ▶ Captures all effects between adjacent segments.
- ▶ Serves as a baseline of comparison for the vowel models.
- ▶ Allows us to define Mutual Information (pointwise mutual information):
 - ▶ $MI(ab) = \log(p(ab)) - \log(p(a)) - \log(p(b))$

Bigram Model

- ▶ The score is the bigram encoding length.
- ▶ Captures all effects between adjacent segments.
- ▶ Serves as a baseline of comparison for the vowel models.
- ▶ Allows us to define Mutual Information (pointwise mutual information):
 - ▶ $MI(ab) = \log(p(ab)) - \log(p(a)) - \log(p(b))$
 - ▶ $MI(ab)$ is the number of bits saved encoding ab together, compared to encoding a and b independently based on their frequencies.

Bigram Model

- ▶ The score is the bigram encoding length.
- ▶ Captures all effects between adjacent segments.
- ▶ Serves as a baseline of comparison for the vowel models.
- ▶ Allows us to define Mutual Information (pointwise mutual information):
 - ▶ $MI(ab) = \log(p(ab)) - \log(p(a)) - \log(p(b))$
 - ▶ $MI(ab)$ is the number of bits saved encoding ab together, compared to encoding a and b independently based on their frequencies.
 - ▶ Positive MI means ab attract
 - ▶ Negative MI means ab repel

Bigram Example

Bigram model example: Turkish “sonucu”

$\frac{3.52}{s}$ $\frac{4.93}{o}$ $\frac{2.43}{n}$ $\frac{4.98}{u}$ $\frac{6.06}{c}$ $\frac{4.00}{u}$ $\frac{3.44}{\#}$

Bigram Example

Bigram model example: Turkish “sonucu”

$\# \xrightarrow{3.52} s \xrightarrow{4.93} o \xrightarrow{2.43} n \xrightarrow{4.98} u \xrightarrow{6.06} c \xrightarrow{4.00} u \xrightarrow{3.44} \#$

$$s(\#sonucu\#) = 29.36$$

$$p(\#sonucu\#) = 1.45 \times 10^{-9}$$

Outline

Overview: Phonology as Compression

The Phonological Framework

Models

Bigram Model

Long Distance V-to-V Model

Results

Turkish

Finnish

Other Languages

V-to-V Model

- ▶ Add long-distance V-to-V information to the bigram model.

V-to-V Model

- ▶ Add long-distance V-to-V information to the bigram model.
- ▶ Compute $MI(V_1, V_2)$ for vowels in $V_1C^+V_2$ configuration.

V-to-V Model

- ▶ Add long-distance V-to-V information to the bigram model.
- ▶ Compute $MI(V_1, V_2)$ for vowels in $V_1C^+V_2$ configuration.
- ▶ Subtract those MIs from bigram score.

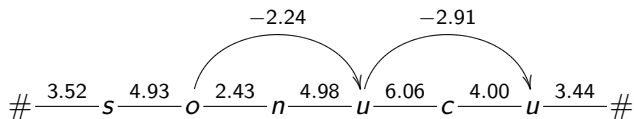
V-to-V Example

Start with the Bigram model

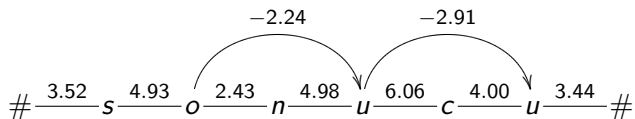
$\frac{3.52}{s}$ $\frac{4.93}{o}$ $\frac{2.43}{n}$ $\frac{4.98}{u}$ $\frac{6.06}{c}$ $\frac{4.00}{u}$ $\frac{3.44}{\#}$

V-to-V Example

Add another scoring function that subtracts distant V-to-V MI



V-to-V Example

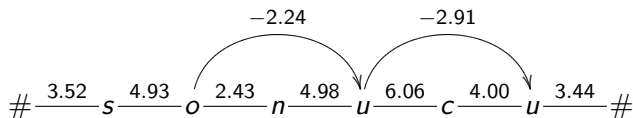


$$S_{\text{bigram}}(\#sonucu\#) = 29.36$$

$$s_V(\#sonucu\#) = -5.15$$

$$S_{\text{total}} = 24.21$$

V-to-V Example



$$S_{\text{bigram}}(\#sonucu\#) = 29.36$$

$$s_V(\#sonucu\#) = -5.15$$

$$S_{\text{total}} = 24.21$$

$$p_{\text{bigram}}(\#sonucu\#) = 1.45 \times 10^{-9}$$

$$p_V(\#sonucu\#) = 5.12 \times 10^{-8}$$

Predictions for Vowel Harmony

- ▶ In vowel harmony languages:
 - ▶ if both Vs are in the **same** class, they should **attract**
 - ▶ if they are in **different** classes, they should **repel**
 - ▶ Adding vowel-to-vowel MI to a bigram model should improve description length more for vowel harmony languages than non-vowel harmony languages.
- ▶ Finnish data replicates Goldsmith and Riggle (2007)

Outline

Overview: Phonology as Compression

The Phonological Framework

Models

Bigram Model

Long Distance V-to-V Model

Results

Turkish

Finnish

Other Languages

Turkish Vowel Harmony

- ▶ Two dimensional: front/backness and roundness.
- ▶ Front/back harmony primary.
- ▶ Roundness harmony in high vowels.
- ▶ ü,ö, and ı do not occur in front/back disharmonic roots or fixed suffixes. (Kirchner, 1993)
- ▶ ı does not occur in round/unround disharmonic roots.

	front		back	
	unround	round	unround	round
high	i	ü	ɪ	u
low	e	ö	a	o

Turkish Results

- ▶ Vowel model 95.29% of the size of the bigram model.
- ▶ V-to-V MI captures Turkish vowel harmony generalizations.

Turkish Vowel MI

	i	e	ü	ö	ı	a	u	o
i	0.69	0.71	-2.34	-0.73	-6.81	-0.79	-3.32	0.22
e	0.89	0.79	-2.06	0.26	-4.42	-1.86	-2.52	-0.32
ü	-2.51	0.99	3.77	0.54	-5.31	-1.88	-2.11	-0.82
ö	-4.20	1.33	3.65	1.57	-6.66	-3.74	-3.64	-1.12
ı	-4.26	-3.17	-4.83	-0.99	1.89	0.84	-4.94	-1.29
a	-0.52	-1.61	-2.30	0.34	1.16	0.59	-1.54	-0.04
u	-2.84	-2.04	-2.02	-2.19	-5.98	0.77	2.90	-0.72
o	-0.91	-0.98	-0.88	-0.08	-5.17	0.43	2.24	1.44

Turkish Vowel MI

	i	e	ü	ö	ı	a	u	o
i	0.69	0.71	-2.34	-0.73	-6.81	-0.79	-3.32	0.22
e	0.89	0.79	-2.06	0.26	-4.42	-1.86	-2.52	-0.32
ü	-2.51	0.99	3.77	0.54	-5.31	-1.88	-2.11	-0.82
ö	-4.20	1.33	3.65	1.57	-6.66	-3.74	-3.64	-1.12
ı	-4.26	-3.17	-4.83	-0.99	1.89	0.84	-4.94	-1.29
a	-0.52	-1.61	-2.30	0.34	1.16	0.59	-1.54	-0.04
u	-2.84	-2.04	-2.02	-2.19	-5.98	0.77	2.90	-0.72
o	-0.91	-0.98	-0.88	-0.08	-5.17	0.43	2.24	1.44

Turkish Vowel MI

	i	e	ü	ö	ı	a	u	o
i	0.69	0.71	-2.34	-0.73	-6.81	-0.79	-3.32	0.22
e	0.89	0.79	-2.06	0.26	-4.42	-1.86	-2.52	-0.32
ü	-2.51	0.99	3.77	0.54	-5.31	-1.88	-2.11	-0.82
ö	-4.20	1.33	3.65	1.57	-6.66	-3.74	-3.64	-1.12
ı	-4.26	-3.17	-4.83	-0.99	1.89	0.84	-4.94	-1.29
a	-0.52	-1.61	-2.30	0.34	1.16	0.59	-1.54	-0.04
u	-2.84	-2.04	-2.02	-2.19	-5.98	0.77	2.90	-0.72
o	-0.91	-0.98	-0.88	-0.08	-5.17	0.43	2.24	1.44

Adding Word Boundaries to the Turkish Vowel Model

	i	e	ü	ö	ı	a	u	o	#
i	0.7	0.6	-2.7	-2.7	-6.6	-0.9	-3.5	-0.4	0.6
e	1.0	0.7	-2.4	-1.6	-4.2	-1.9	-2.6	-0.9	0.2
ü	-2.3	1.0	3.5	-1.3	-5.0	-1.8	-2.1	-1.3	-0.4
ö	-3.9	1.4	3.4	-0.2	-6.2	-3.6	-3.6	-1.5	-1.4
ı	-4.2	-3.2	-5.3	-3.0	1.9	0.6	-5.2	-1.9	0.7
a	-0.3	-1.6	-2.6	-1.5	1.4	0.5	-1.6	-0.6	0.0
u	-2.7	-2.0	-2.3	-4.1	-5.7	0.6	2.7	-1.3	0.1
o	-0.7	-0.9	-1.1	-1.9	-4.9	0.3	2.1	0.9	-0.2
#	-0.6	0.0	0.9	2.0	-1.5	0.3	0.5	1.2	-3.7

Outline

Overview: Phonology as Compression

The Phonological Framework

Models

Bigram Model

Long Distance V-to-V Model

Results

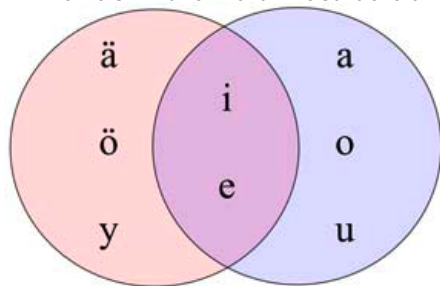
Turkish

Finnish

Other Languages

Finnish Vowel Harmony

All vowels in the word must be either front or back.



See Ringen and Heinämäki (1999) for data on disharmony.

Finnish Vowel MI

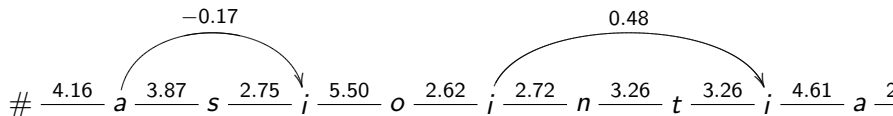
	y	ä	ö	e	i	o	u	a
y	2.00	1.74	1.75	0.26	0.17	-1.73	-2.94	-2.42
ä	1.16	1.85	2.68	0.03	0.33	-2.25	-2.29	-2.39
ö	1.57	1.63	2.03	-0.20	-0.43	-0.45	-0.43	-0.95
e	0.31	0.64	-0.87	-0.40	0.24	-0.02	0.20	-0.31
i	0.21	0.05	0.30	0.29	-0.47	0.01	-0.04	0.04
o	-1.58	-2.63	-4.12	0.18	0.18	0.25	-0.30	0.20
u	-1.45	-2.42	-3.42	0.12	-0.13	-0.01	0.58	0.23
a	-2.03	-2.73	-4.21	-0.46	0.16	0.33	0.17	0.42

Finnish Results

- ▶ Total vowel model length 99.01% of total bigram length.
- ▶ Better gains adding V-to-V MI than non-vowel harmony languages.
- ▶ Model fails to capture the transparency of the neutral vowels.

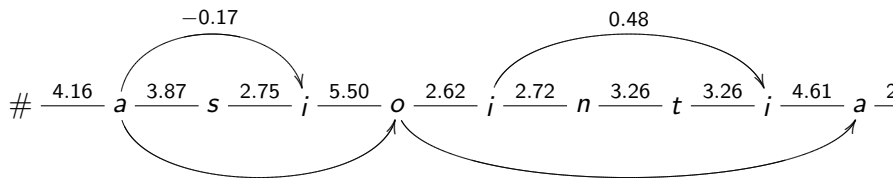
The Problem with Finnish Vowels

TOP:Goldsmith and Riggle (2007) model of "asiointia":



The Problem with Finnish Vowels

TOP: Goldsmith and Riggle (2007) model of "asiointia":



BOTTOM: V-to-V connections we want to model.

Alternate Finnish Model

- ▶ Leaving out neutral vowels improves corpus probability and simplifies grammar.
- ▶ Vowel model with no neutral vowels is 98.70% encoding length of bigram model.
- ▶ Recall with neutral models vowel model was 99.01% of the length of the bigram model.

Outline

Overview: Phonology as Compression

The Phonological Framework

Models

Bigram Model

Long Distance V-to-V Model

Results

Turkish

Finnish

Other Languages

Italian V to V MI

	e	i	a	o	u
e	0.08	-0.07	0.10	-0.16	0.05
i	-0.09	-0.12	0.20	-0.03	0.25
a	-0.05	0.08	-0.31	0.29	-0.12
o	0.19	0.01	-0.00	-0.25	-0.19
u	-0.38	0.22	0.21	-0.29	-0.04

English and Italian Language Models

- ▶ English vowel model saves 99.82% over bigram model.
- ▶ Italian vowel model saves 99.96% over bigram model.
- ▶ For English and Italian V-to-V MI is low magnitude.
- ▶ Hungarian vowel model saves 97.52% over bigram model.

Summary

The proposed model:

- ▶ captures Turkish vowel harmony well.
 - ▶ does not deal with transparency of neutral vowels in Finnish.
 - ▶ shows much smaller gains when applied to non-vowel harmony languages.
-
- ▶ Further research
 - ▶ More languages: Hungarian, Korean, Washo, Japanese, Arabic
 - ▶ Methods for excluding neutral vowels.
 - ▶ Incorporate syllable phonotactics and stress.

Selected References

- Ellison, T Mark. 1994. The iterative learning of phonological constraints. *Computational Linguistics* .
- Geman, Stuart, and Mark Johnson. 2001. Probability and statistics in computational linguistics, a brief review. URL <http://www.cog.brown.edu/mj/papers/Review.pdf>, manuscript.
- Goldsmith, John. 2001. The unsupervised learning of natural language morphology. *Computational Linguistics* .
- Goldsmith, John, and Jason Riggle. 2007. Information theoretical approaches to phonological structure: The case of vowel harmony. Under review.
- Kirchner, Robert. 1993. Turkish vowel harmony and disharmony: An optimality theoretic account. In *Rutgers Optimality Workshop I (ROW-I)*.
- de Marken, Carl. 1996. Unsupervised language acquisition. Doctoral Dissertation, MIT.
- Ringen, Catherine O., and Orvokki Heinämäki. 1999. Variation in finnish vowel harmony: An ot account. *Natural Language and*

Computing Grammar Length

- ▶ All scoring functions are finite state string to weight transducers.
- ▶ First state is assumed as the only start state (start weight 0). All states final.
 - ▶ $s = \#$ of states,
 - ▶ $a = \#$ of arcs,
 - ▶ $l = \#$ of letters in the alphabet
- ▶ $L = a(2\log_2(s) + \log_2(l) + 64) + \log_2(s)$
- ▶ Small added overhead to mark where one transducer ends and the next begins:
 $8^{\lceil \log_{127}(L) \rceil}$