Local Predictability in the Lexicon

Max Bane and Ed King
bane@uchicago.edu, etking@uchicago.edu

Department of Linguistics
University of Chicago

45th Annual Meeting of the Chicago Linguistic Society
April 16, 2009

Introduction

- Phonotactics: Languages exhibit significant, cross-linguistically variable preferences and dispreferences for segment sequences in their lexica.
- These tendencies can be. . .

Introduction

- Phonotactics: Languages exhibit significant, cross-linguistically variable preferences and dispreferences for segment sequences in their lexica.
- These tendencies can be...
  - Local generalizations about how often different segments appear adjacent to each other.
    - E.g., impossible or improbable consonant clusters, consonant-vowel sequences, etc.

Introduction

- Phonotactics: Languages exhibit significant, cross-linguistically variable preferences and dispreferences for segment sequences in their lexica.
- These tendencies can be...
  - Local generalizations about how often different segments appear adjacent to each other.
    - E.g., impossible or improbable consonant clusters, consonant-vowel sequences, etc.
  - Nonlocal generalizations about word structure spanning non-adjacent segments.
    - E.g., harmony—agreement between segments arbitrarily far from each other.

Local Predictability

- QUESTION:
  - To what extent is a language's lexical material predictable by purely local, statistical generalizations?

Local Predictability

- QUESTION:
  - To what extent is a language's lexical material predictable by purely local, statistical generalizations?
- A novel reframing of this question:
  - Out of all the logically possible words a language could have, how well "optimized" are the actual words for local, segement-to-segment predictability?

Local Predictability

- RESULTS:
  - Looking at the lexica of seven languages:
    - English, Dutch, French, Japanese, Finnish, Hungarian, Turkish
  - Among all the ways that words could be segmentally permuted or edited, the actual attested words are much closer to optimal than expected by chance.

Local Predictability

- RESULTS:
  - Looking at the lexica of seven languages:
    - English, Dutch, French, Japanese, Finnish, Hungarian, Turkish
  - Among all the ways that words could be segmentally permuted or edited, the actual attested words are much closer to optimal than expected by chance.
  - The languages with known, significant, non-local generalizations (vowel harmony—Finnish, Hungarian, Turkish) are detectably less well optimized for local predictability.

“Purely local, statistical generalizations”

Here, the bigram statistics of the lexicon.

- Probability (frequency) of bigram (immediately adjacent pair of segments) xy:

$$P(xy) = \frac{C(xy)}{\sum_{u,v \in \Sigma} C(uv)}$$

- Where:
  - $C(xy) =$ count of occurrences of xy in the lexicon,
  - $\Sigma =$ inventory of segments.

Pointwise mutual information

- Quantifies affinity between two segments as discrepancy between probability of their co-occurrence versus their individual, independent probabilities of occurrence.

Pointwise mutual information

- Quantifies affinity between two segments as discrepancy between probability of their co-occurrence versus their individual, independent probabilities of occurrence.
- $PMI(x, y) = \log_2 \frac{P(xy)}{P(x)P(y)} = \log_2 \frac{P(yx)}{P(y)P(x)}$

Pointwise mutual information

- Quantifies affinity between two segments as discrepancy between probability of their co-occurrence versus their individual, independent probabilities of occurrence.
- $PMI(x, y) = \log_2 \frac{P(xy)}{P(x)P(y)} = \log_2 \frac{P(yx)}{P(y)P(x)}$
  - $PMI(x, y) > 0 \Leftrightarrow x$ and y attract each other
  - $PMI(x, y) < 0 \Leftrightarrow x$ and y repel each other

Pointwise mutual information

- Quantifies affinity between two segments as discrepancy between probability of their co-occurrence versus their individual, independent probabilities of occurrence.
- $PMI(x, y) = \log_2 \frac{P(xy)}{P(x)P(y)} = \log_2 \frac{P(yx)}{P(y)P(x)}$
  - $PMI(x, y) > 0 \Leftrightarrow x$ and y attract each other
  - $PMI(x, y) < 0 \Leftrightarrow x$ and y repel each other
- Total PMI of word W: $\sum_{xy \in W} PMI(x, y)$
  - Quantifies how "good"/probable a word is according to purely local tendencies.
  - Units are bits.

Pointwise mutual information of “happy”

#	h	æ	p	i	#
	PMI(#, h)	*PMI*(h, æ)	*PMI*(æ, p)	*PMI*(p, i)	*PMI*(i, #)
	2.55	2.74	0.99	0.35	1.63

Total *PMI*(#hæpi#) = 8.26 bits.

A locally very probable word, comprising all “attractive” adjacent pairs of segments.

Pointwise mutual information of “zeus”

#	z	u	s	#
	PMI(#, z)	*PMI*(z, u)	*PMI*(u, s)	*PMI*(s, #)
	-4.27	-1.75	0.19	0.85

Total *PMI*(#zus#) = -4.98 bits.

A locally improbable word, comprising “repulsive” or only marginally “attractive” adjacent pairs of segments.

The question made more precise

- QUESTION:
  - To what extent are languages' lexica "optimal" according to purely local, statistical generalizations?
  - $\Rightarrow$ To what extent do they choose lexical material that maximizes *PMI* from among some space of alternative, possible lexical material?

The question made more precise

- QUESTION:
  - To what extent are languages' lexica "optimal" according to purely local, statistical generalizations?
  - $\Rightarrow$ To what extent do they choose lexical material that maximizes *PMI* from among some space of alternative, possible lexical material?
- If L_x is the lexicon of language x, is it the case that:

$$L_x = \underset{L \in \mathcal{L}_x}{\text{argmax}} \sum_{w \in L} PMI(w)$$

where $\mathcal{L}_x$ is some space of "possible" lexica for language x?

Preview of results

We've considered two possibilities for the space $\mathcal{L}_X$ of lexical alternatives:

Preview of results

We've considered two possibilities for the space $\mathcal{L}_x$ of lexical alternatives:

- If $\mathcal{L}_x$ is the space of lexica gotten by permuting the segments of words in any way, then yes, languages' actual lexica are very nearly ($\sim$95%) optimal in that space.

Preview of results

We've considered two possibilities for the space $\mathcal{L}_x$ of lexical alternatives:

- If $\mathcal{L}_x$ is the space of lexica gotten by permuting the segments of words in any way, then yes, languages' actual lexica are very nearly ($\sim$95%) optimal in that space.
- If $\mathcal{L}_x$ is the space of lexica in which words have been modified by any single edit (segmental insertion, deletion, substitution), then yes, languages are strongly ($\sim$80%) optimal in that space.

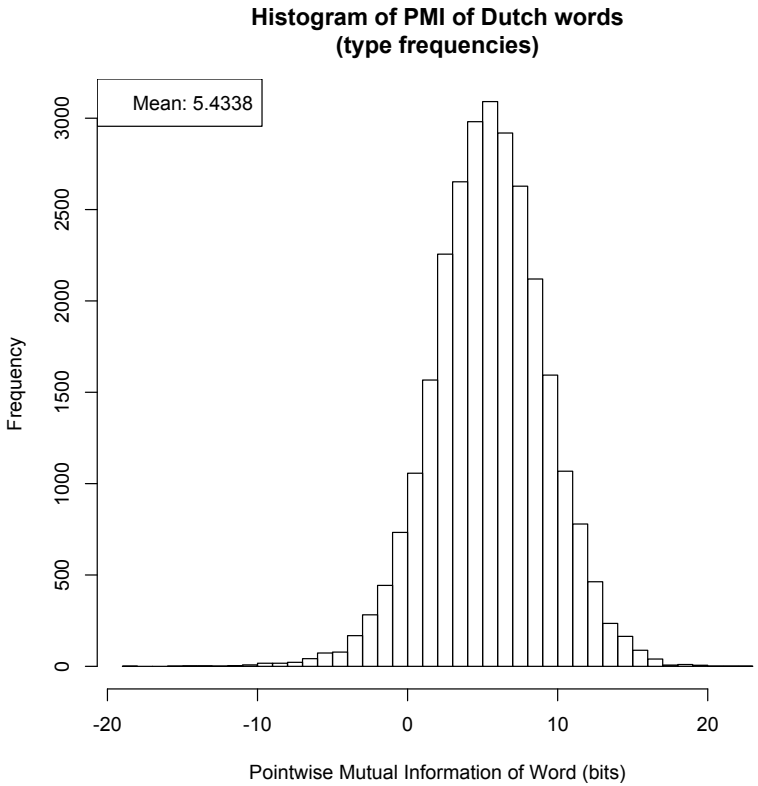
The Baseline: A Random Lexicon

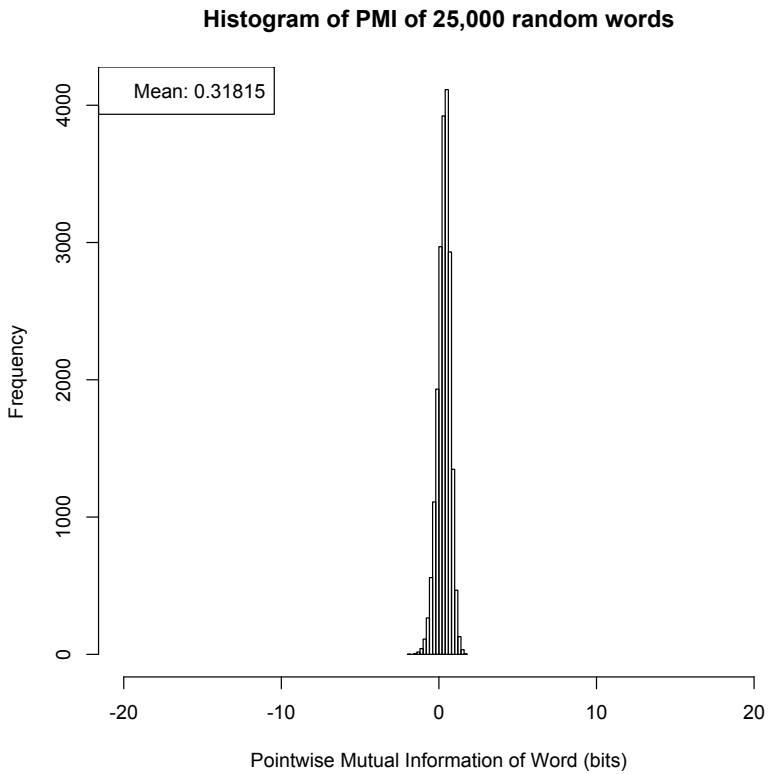
- For comparison: A randomly generated lexicon of 25k words over 39 segments (the Darpabet).
- Each word generated by:
  - Choosing a length (max 8) uniformly at random
  - Choosing that many segments uniformly at random (with replacement), in random order.
- $\Rightarrow$ a lexicon that follows no strictly local statistical generalizations.

The language lexica

- Phonemically transcribed lexica of English (125k words), French (22k), Dutch (57k), Japanese (58k), Finnish (188k), Turkish (117k), Hungarian (33k).
- Some extra-segmental information not included (e.g., stress), but most segment-level phonemic distinctions are represented (vowel length, etc.).
- Lexicon of English is from the CMU pronouncing dictionary, others from the Leipzig Corpora Collection.
- All statistics based on bigram frequencies in the lexicon, i.e., type frequencies.

The distribution of word-PMI: Dutch vs Random





The distribution of word-PMI

Language	Mean Word-PMI
English	6.2308 bits
Dutch	5.4338 bits
French	5.1956 bits
Japanese	7.8657 bits
Finnish	3.3068 bits
Hungarian	3.6657 bits
Turkish	3.6275 bits
Random	0.3182 bits

- Note difference between harmony and non-harmony languages.

Permutation neighborhoods

- The permutation neighborhood of a word w is the set of all $|w|!$ permutations of that word.
- Each permutation neighbor has some PMI.
- Are attested words the PMI-best words among their permutation neighbors?

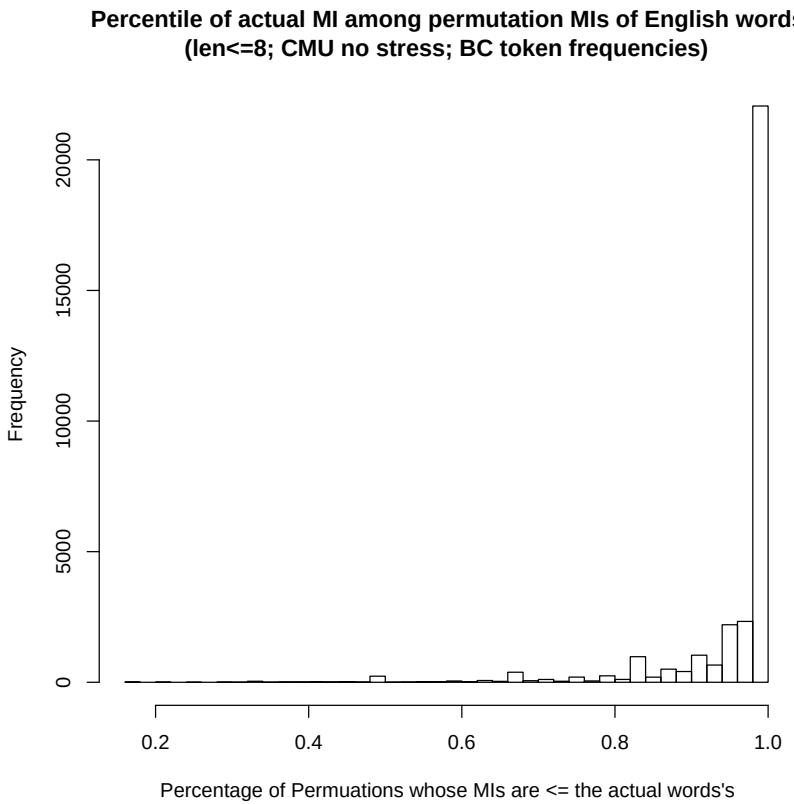
Permutation neighborhoods

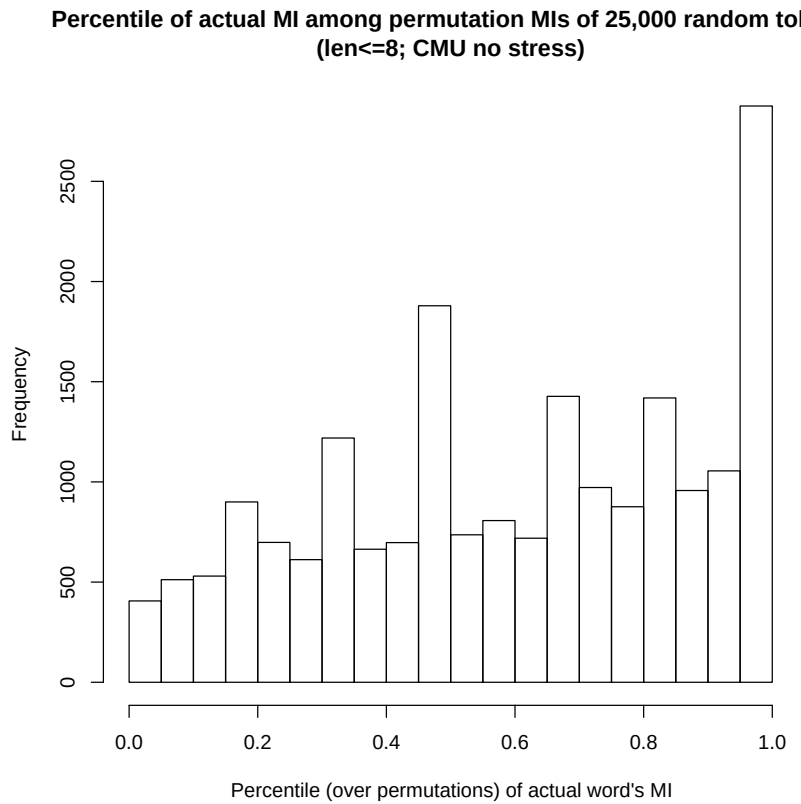
- The permutation neighborhood of a word w is the set of all $|w|!$ permutations of that word.
- Each permutation neighbor has some PMI.
- Are attested words the PMI-best words among their permutation neighbors?
  - No, only for a few hundred words, out of tens of thousands.
  - But, attested words have higher PMI than the vast majority of their permutation neighbors.

Permutation neighborhoods

- The permutation neighborhood of a word w is the set of all $|w|!$ permutations of that word.
- Each permutation neighbor has some PMI.
- Are attested words the PMI-best words among their permutation neighbors?
  - No, only for a few hundred words, out of tens of thousands.
  - But, attested words have higher PMI than the vast majority of their permutation neighbors.
- We calculate a percentile for each word w:
  - What percentage of w's permutations does w have higher PMI than?

Percentile of actual word PMI versus permutations: English versus Random



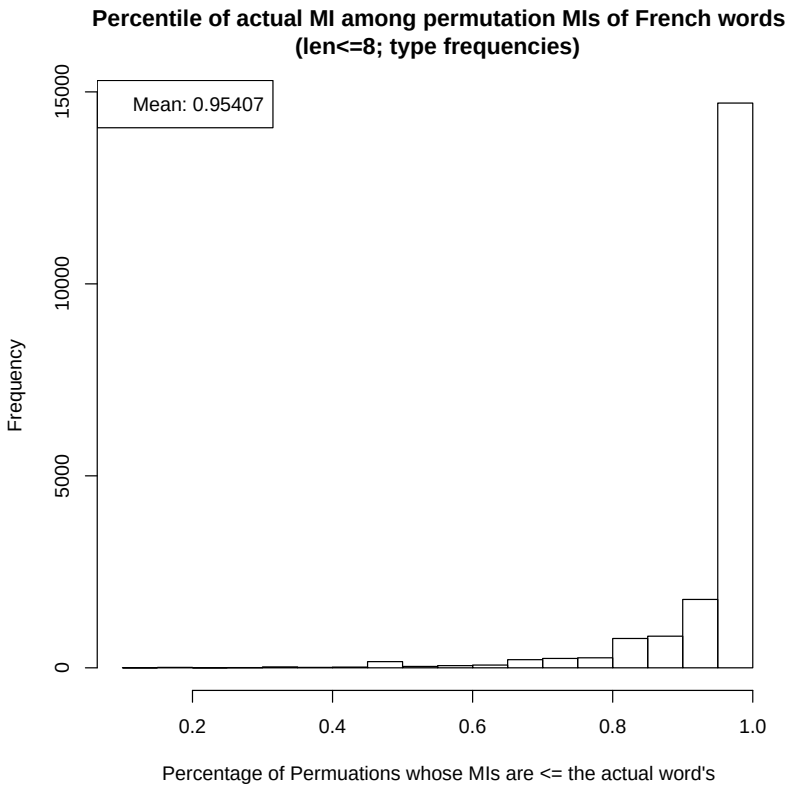


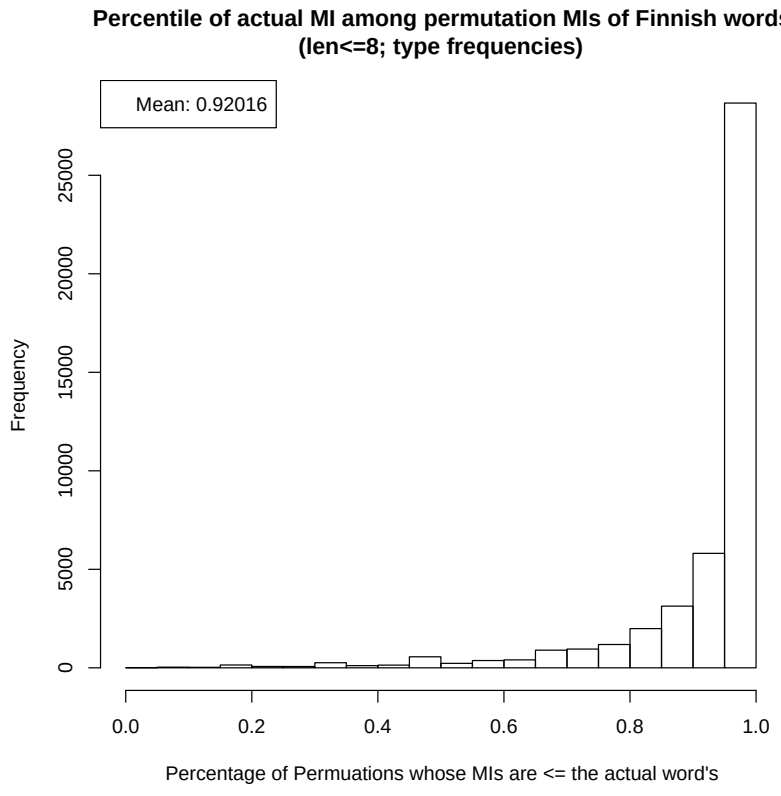
Percentile of actual word PMI versus permutations

Language	Mean Percentile
English	96.03
Dutch	96.34
French	95.41
Japanese	97.71
Finnish	92.02
Hungarian	92.55
Turkish	92.35
Random	59.95

- Significantly different from random (Mann-Whitney U-tests, $p < .01$)
- Again, difference between harmony and non-harmony languages.

Distribution of percentiles: Harmony vs Non-Harmony





- Harmony languages have notably less rightward skew than non-harmony languages.

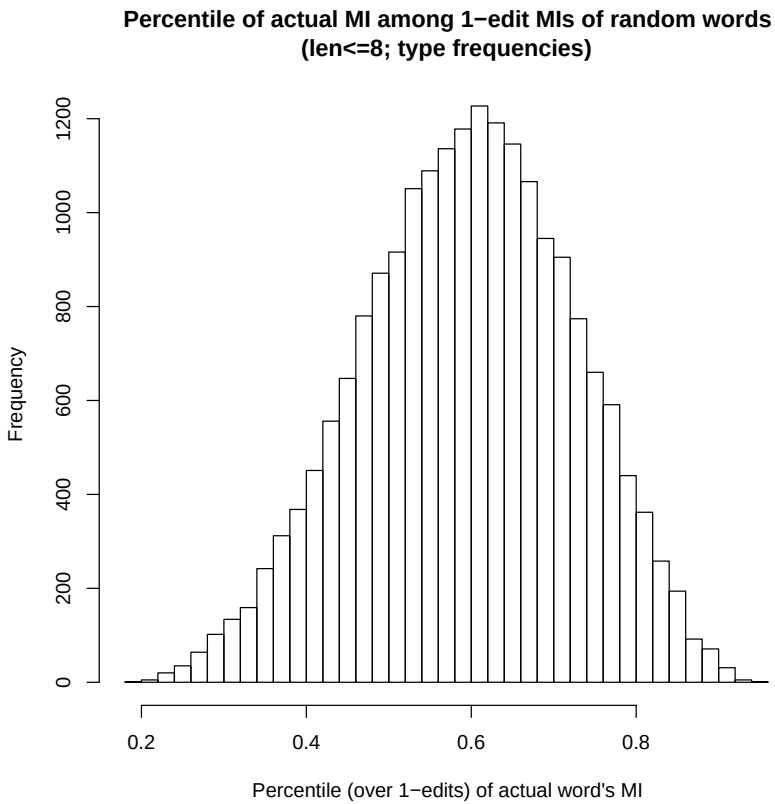
Edit neighborhoods

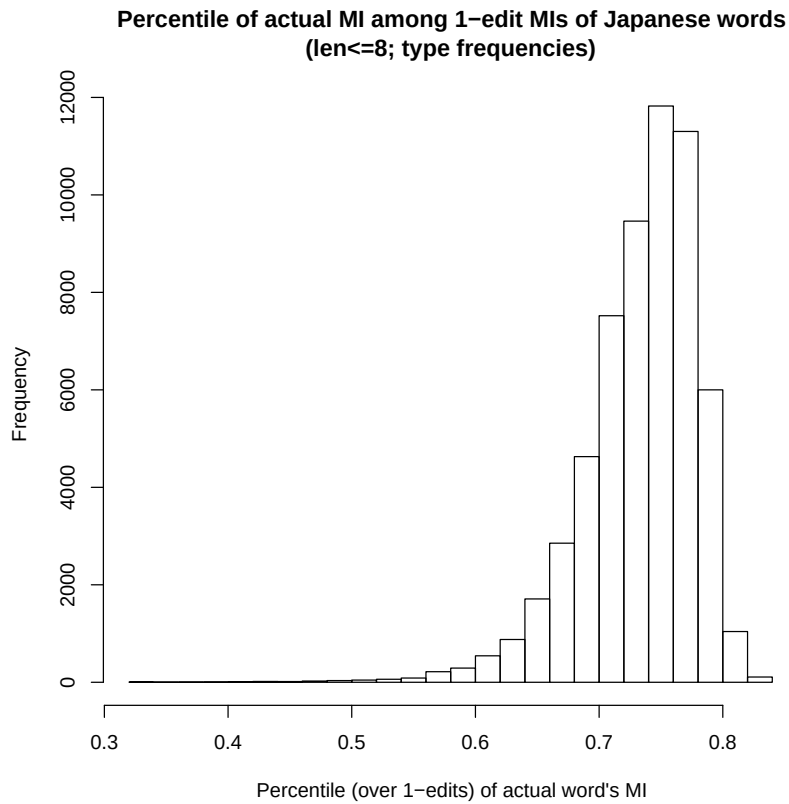
- The *k*-edit neighborhood of a word *w* is the set of all words that can result from making *k* or fewer edits (insertions, deletions, substitutions) to *w*.
- Perhaps linguistically more plausible space of possible realizations of a word than permutations.
- We've examined 1-edit neighborhoods.

Edit neighborhoods

- The *k*-edit neighborhood of a word *w* is the set of all words that can result from making *k* or fewer edits (insertions, deletions, substitutions) to *w*.
- Perhaps linguistically more plausible space of possible realizations of a word than permutations.
- We've examined 1-edit neighborhoods.
- Result: Attested words have higher PMI than large majority of their 1-edit neighbors.
  - Mean attested word percentile: ∼73–81% across languages.
  - Versus random: mean word percentile 59%.

Percentile of actual word PMI versus 1-edits: Random versus Japanese





Percentile of actual word PMI versus 1-edits

Language	Mean Percentile
English	80.82
Dutch	80.28
French	76.28
Japanese	73.34
Finnish	79.38
Hungarian	78.01
Turkish	82.35
Random	59.48

- Significantly different from random (Mann-Whitney U-tests, $p < .01$)
- Here harmony languages are less distinct from non-harmony.

Summary

- A language's phonotactics chooses words that tend to maximize the predictability of the lexicon by strictly local, statistical generalizations.
- A novel way of quantifying the strength of this effect:
  - How close the lexicon comes to

$$\operatorname*{argmax}_{L \in \mathcal{L}_x} \sum_{w \in L} PMI(w)$$

within some set of possible lexica $\mathcal{L}_x$.

Summary

- In the languages surveyed, actual words have much higher PMI than the great majority of alternatives
  - Where alternatives include permutations of the words, or single segmental edits.
- True even of languages with important non-local phonotactic effects.

Summary

- In the languages surveyed, actual words have much higher PMI than the great majority of alternatives
  - Where alternatives include permutations of the words, or single segmental edits.
- True even of languages with important non-local phonotactic effects.
- A substantial portion of the “shape” of languages’ lexical material can be explained by local, segment-to-segment preferences.

Further Work

- Might expect other phonological phenomena to be driven by similar, strictly local preferences.
  - Phonological alternations, phonological change.
- Ongoing research:
  - Attested changes in English phonology result in increased local predictability of the lexicon:
    - Mergers: “caught”∼“cot,” “pin”∼“pen,” “pull”∼“pool”

Thank You!

Special thanks to John Goldsmith and Jason Riggle.