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

These tendencies can be...
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, segment-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.

\[ \text{PMI}(x, y) = \log_2 \frac{P(xy)}{P(x)P(y)} = \log_2 \frac{P(yx)}{P(y)P(x)} \]

- \( \text{PMI}(x, y) > 0 \iff x \text{ and } y \text{ attract each other} \)
- \( \text{PMI}(x, y) < 0 \iff x \text{ and } y \text{ repel each other} \)
Pointwise mutual information

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

\[ PMI(x, y) = \log_2 \frac{P(xy)}{P(x)P(y)} = \log_2 \frac{P(yx)}{P(y)P(x)} \]

- \( PMI(x, y) > 0 \iff x \text{ and } y \text{ attract each other} \)
- \( PMI(x, y) < 0 \iff x \text{ and } y \text{ repel each other} \)

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”

<table>
<thead>
<tr>
<th></th>
<th>#</th>
<th>h</th>
<th>æ</th>
<th>p</th>
<th>i</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>PMI(#, h)</td>
<td>2.55</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PMI(h, æ)</td>
<td></td>
<td>2.74</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PMI(æ, p)</td>
<td></td>
<td></td>
<td>0.99</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PMI(p, i)</td>
<td></td>
<td></td>
<td></td>
<td>0.35</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PMI(i, #)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.63</td>
<td></td>
</tr>
</tbody>
</table>

Total $PMI(\#æpi\#) = 8.26$ bits.

A locally very probable word, comprising all “attractive” adjacent pairs of segments.
### Pointwise mutual information of “zeus”

<table>
<thead>
<tr>
<th></th>
<th>#</th>
<th>z</th>
<th>u</th>
<th>s</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PMI(#, z)</strong></td>
<td>-4.27</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>PMI(z, u)</strong></td>
<td>-1.75</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>PMI(u, s)</strong></td>
<td></td>
<td>0.19</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>PMI(s, #)</strong></td>
<td></td>
<td></td>
<td>0.85</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Total $\text{PMI}(\#zus\#) = -4.98$ 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?
- ⇒ 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:

$$L_x = \arg\max_{L \in \mathcal{L}_x} \sum_{w \in L} PMI(w)$$

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

Histogram of PMI of Dutch words
(type frequencies)

Mean: 5.4338

Histogram of PMI of 25,000 random words

Mean: 0.31815
## The distribution of word-PMI

<table>
<thead>
<tr>
<th>LANGUAGE</th>
<th>MEAN WORD-PMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>6.2308 bits</td>
</tr>
<tr>
<td>Dutch</td>
<td>5.4338 bits</td>
</tr>
<tr>
<td>French</td>
<td>5.1956 bits</td>
</tr>
<tr>
<td>Japanese</td>
<td>7.8657 bits</td>
</tr>
<tr>
<td>Finnish</td>
<td>3.3068 bits</td>
</tr>
<tr>
<td>Hungarian</td>
<td>3.6657 bits</td>
</tr>
<tr>
<td>Turkish</td>
<td>3.6275 bits</td>
</tr>
<tr>
<td>Random</td>
<td>0.3182 bits</td>
</tr>
</tbody>
</table>

- 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.
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 MI among permutation MIs of English words
(len<=8; CMU no stress; BC token frequencies)

Percentile of actual MI among permutation MIs of 25,000 random tokens
(len<=8; CMU no stress)
Percentile of actual word PMI versus permutations

<table>
<thead>
<tr>
<th>LANGUAGE</th>
<th>MEAN</th>
<th>PERCENTILE</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>96.03</td>
<td>96.03</td>
</tr>
<tr>
<td>Dutch</td>
<td>96.34</td>
<td>96.34</td>
</tr>
<tr>
<td>French</td>
<td>95.41</td>
<td>95.41</td>
</tr>
<tr>
<td>Japanese</td>
<td>97.71</td>
<td>97.71</td>
</tr>
<tr>
<td>Finnish</td>
<td>92.02</td>
<td>92.02</td>
</tr>
<tr>
<td>Hungarian</td>
<td>92.55</td>
<td>92.55</td>
</tr>
<tr>
<td>Turkish</td>
<td>92.35</td>
<td>92.35</td>
</tr>
<tr>
<td>Random</td>
<td>59.95</td>
<td>59.95</td>
</tr>
</tbody>
</table>

- Significantly different from random (Mann-Whitney U-tests, $p < .01$)
- Again, difference between harmony and non-harmony languages.
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.
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: $\sim 73$–$81\%$ across languages.
- Versus random: mean word percentile $59\%$. 
Percentile of actual word PMI versus 1-edits: Random versus Japanese

Percentile of actual MI among 1−edit MIs of random words
(len<=8; type frequencies)

Percentile of actual MI among 1−edit MIs of Japanese words
(len<=8; type frequencies)
**Percentile of actual word PMI versus 1-edits**

<table>
<thead>
<tr>
<th>LANGUAGE</th>
<th>MEAN</th>
<th>PERCENTILE</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>80.82</td>
<td></td>
</tr>
<tr>
<td>Dutch</td>
<td>80.28</td>
<td></td>
</tr>
<tr>
<td>French</td>
<td>76.28</td>
<td></td>
</tr>
<tr>
<td>Japanese</td>
<td>73.34</td>
<td></td>
</tr>
<tr>
<td>Finnish</td>
<td>79.38</td>
<td></td>
</tr>
<tr>
<td>Hungarian</td>
<td>78.01</td>
<td></td>
</tr>
<tr>
<td>Turkish</td>
<td>82.35</td>
<td></td>
</tr>
<tr>
<td>Random</td>
<td>59.48</td>
<td></td>
</tr>
</tbody>
</table>

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

$$\arg\max_{L \in \mathcal{L}_x} \sum_{w \in L} \text{PMI}(w)$$

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.