Local predictability in the lexicon

Max Bane and Ed King
University of Chicago

1 Introduction

All languages exhibit phonotactic preferences and dispreferences for the occurrence of particular segment sequences in their lexical material. While some such phonotactic generalizations may take the form of categorical constraints (e.g., the impossibility of stop-obstruent onset clusters in English (Duanmu 2002), or the absence of any consonant clusters in Hawaiian (Lyovin 1997)), many others emerge as statistical tendencies of varying strength and significance. Among either kind of generalization, categorical or statistical, one can distinguish two types:

- *Local* generalizations about whether, or how often, segments may appear adjacent to each other or in contiguous sequences.
- *Nonlocal* generalizations that span discontiguous collections of segments, possibly separated by arbitrarily much intervening material.

The above mentioned categorical constraints of English and Hawaiian clusters thus constitute local generalizations, as does, for example, the statistical rarity of morpheme-initial /sf/ sequences in the English lexicon. Each of these restrictions can be represented as a ban or statistical tendency on sequences of adjacent segments in surface forms:

(1) $*[_{\mu}$[+stop][+obstruent] — no morpheme-initial stop-obstruent clusters.

(2) *[+consonant][+consonant] — no adjacent consonants.

(3) $P([_{\mu}\mathrm{sf}) \ll P([_{\mu})P(\mathrm{s})P(\mathrm{f})$ — the probability of morpheme-initial /sf/ is much less than independent chance.[1]

Note that by treating a morpheme boundary as a kind of symbol or segment (here $[_{\mu}$), environments like "morpheme-initial" and "morpheme-final" can be respresented in this local way.

The canonical example of a nonlocal phonotactic generalization is harmony: any requirement or tendency for segments of certain specified classes to look like each other in some respect throughout a morphophonological domain, in spite of (though possibly influenced by) any intervening segments. So, for instance, a tendency toward vowel harmony of the scheme in (4) exemplifies a nonlocal phonotactic generalization.

[1]The notation is as follows: $P(x)$ is the empirical probability (i.e., relative frequency) of the string x in some lexicon, where a lexicon is understood to be a long string containing all the words of a language, each separated from the next, and from the ends of the string, by any number of special boundary symbols. Where SPE-style (Chomsky and Halle 1968) regular expression notation is used, $P(x)$ is meant to be the sum/integral of the probabilities of each string in the set of strings x.

(4) $P([_{\mu}\mathrm{V}_{[\alpha\ \mathrm{back}]}\mathrm{C}^0\mathrm{V}_{[\alpha\ \mathrm{back}]}) \gg P([_{\mu})P(\mathrm{V}_{[\alpha\ \mathrm{back}]})P(\mathrm{C}^0)P(\mathrm{V}_{[\alpha\ \mathrm{back}]})$ — the probability of a morpheme-initial vowel agreeing in backness with the next vowel, zero or more consonants later, is much greater than independent chance.

In this paper, we investigate the extent to which the contents of a language's lexicon can be predicted solely on the basis of its local statistical tendencies. In particular, we consider what might be called the *strictly local* or *unigram* and *bigram* statistics of a lexicon—the frequencies of segment-sequences of length $n \leq 2$—and quantify each lexical item's predictability according to these frequencies through the *pointwise mutual information* (PMI) of its constituent pairs of adjacent segments.[2] If L_x is the lexicon of a language x, our aim, then, is to ascertain whether (or to what extent) it is the case that:

(5) $$L_x = \operatorname*{argmax}_{L \in \mathscr{L}_x} \sum_{w \in L} PMI(w)$$

where $\mathscr{L}_x$ is some space of "possible" lexica for language x, and $PMI(w)$ is the summed pointwise mutual information of all the adjacent pairs of segments in word $w \in L$. That is, we ask not just how predictable a language's lexicon is by its strictly local statistics, but how "optimal" it is according to those same statistics among the alternative lexica presented by $\mathscr{L}_x$.

The answer will of course depend on the hypothesized space $\mathscr{L}_x$ of alternative lexica for x. Here we consider two possible such spaces:

(6) $\mathscr{L}_{x!}$: all lexica gotten by permuting the segments of words of x in any way.

(7) $\mathscr{L}_{e_1(x)}$: all lexica in which the words of x have been modified by any single edit (segmental insertion, deletion, or replacement).

We survey the lexica of seven languages—English, Dutch, French, Japanese, Finnish, Hungarian, and Turkish—and find that in both spaces of possible lexica, the languages' attested lexica are very close to optimal, much closer than could be reasonably expected by chance. In other words, among the ways offered by $\mathscr{L}_{x!}$ and $\mathscr{L}_{e_1(x)}$ of "modifying" a language's lexicon, the actual attested lexicon is very nearly the best or most probable in terms of its adherence to strictly local phonotactic tendencies. This suggests that the preponderance of a language's lexical "shape" is determined by strictly local generalizations, with a comparatively minor (and quantifiably so, for a given space of supposed alternates) remainder being driven by other, presumably nonlocal (or at least less than *strictly* local) tendencies. Furthermore, of the languages surveyed, those with known significant nonlocal phonotactics (vowel harmonies of various sorts—Finnish, Hungarian, Turkish) are detectably further from optimal within at least the first of these lexical spaces.

In what follows, section 2 reviews the basic mathematical tools that we apply for measuring the local tendencies of a language's lexicon, as well as the nature of our data, and our construction of a random control lexicon. Section 3 then details

[2] Manning and Schütze (2000) offer a concise and linguistically-oriented review of n-grams and mutual information.

the results of our comparison of the actual attested lexica, in terms of the quantities described in section 2, to the spaces $\mathscr{L}_{x!}$ and $\mathscr{L}_{e_1(x)}$ of possible alternatives. Section 4 finally discusses the possible future progress of this work and offers some concluding remarks.

2 Quantifying and measuring local predictability

2.1 Bigram frequencies and mutual information

To measure the force of a language's strictly local phontactic preferences, we count the occurrences of immediately adjacent pairs of segments across a phonemic transcription of its lexicon. We will refer to each such pair as a bigram or biphone. We assume that each distinct lexical item is flanked on both ends by instances of a special word boundary symbol, #. Thus bigrams like #x and x# refer to word-initial and word-final occurrences of x, respectively, while the bigram xy ($x \neq$#, $y \neq$#) refers to an x immediately preceding a y at any location in some word.

Counting the occurrences of bigrams in the lexicon provides the basis of a joint probability model describing the relative frequencies of adjacent pairs of segments:

$$P(xy) = \frac{C(xy)}{\sum_{u,v\in\Sigma} C(uv)} \tag{8}$$

where $C(xy)$ is the count of bigram xy, and Σ is the inventory of segments for the language in question.

From this joint bigram probability distribution we can construct a conditional distribution $P(y|x)$, describing the probability of encountering a segment y conditioned on the fact that the immediately preceding segment is x, by normalizing $P(xy)$ over the implied marginal probability $P(x)$ of x:

$$P(y|x) = \frac{P(xy)}{P(x)} = \frac{P(xy)}{\sum_{u\in\Sigma} P(xu)} = \frac{P(xy)}{\sum_{u\in\Sigma} P(ux)} \tag{9}$$

We can then quantify an "affinity" or local "attractiveness" between the constituent segments of a bigram as the discrepancy between the bigram's probability of occurrence versus the segments' individual, independent probabilities of occurrence; or in other words, between the conditional probability of y given the preceding x, and the marginal probability of y:

$$PMI(xy) = \log_2 \frac{P(y|x)}{P(y)} = \log_2 \frac{P(xy)}{P(x)P(y)} \tag{10}$$

This quantity is called the pointwise mutual information of the bigram xy. Positive values indicate an "attraction" or tendency for x to be immediately followed by y, the strength of that tendency being proportional to the magnitude of the mutual information. Conversely, negative values correspond to a "repulsive" tendency between the segments.[3] Finally, we define the PMI of a lexical item w to be the sum

[3] It's worth noting that pointwise mutual information differs by only a $\log_2$ transformation from Frisch et al.'s (2004) and others' "observed-over-expected" ratios of co-occurrence. The advantage of PMI is that it can be added up meaningfully across all the segment pairs in a word.

of the PMIs of its constituent bigrams:

$$PMI(w) = \sum_{xy \in w} PMI(xy) \tag{11}$$

This quantity serves to indicate the degree of a word's overall adherence to the strictly local statistical generalizations of the lexicon; its distribution across words, lexica, and spaces of possible lexica will be of primary interest in this paper.

As an example, consider the English word *happy* /hæpi/. The calculation of its total PMI is shown in (12), with figures based on the frequencies of phoneme bigrams in a large sample of the vocabulary of English (as transcribed in the CMU pronouncing dictionary; see below). The relatively large value[4] of 8.26 bits indicates that *happy* is a locally very probable word, comprising all "attractive" adjacent pairs of segments.

(12)

#	h	æ	p	i	#
PMI(# h)	*PMI*(hæ)	*PMI*(æp)	*PMI*(pi)	*PMI*(i#)	
2.55	2.74	0.99	0.35	1.63	

Total *PMI*(#hæpi#) = 8.26 bits.

We may contrast this with the case of *zeus* /zus/, an overall improbable word according to the local statistics of English, as shown in (13). One can see that it is made up of mostly "repulsive" or only marginally attractive adjacent pairs of segments, which is reflected in its strongly negative total *PMI* of -4.99 bits.

(13)

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

Total *PMI*(#zus#) = -4.99 bits.

2.2 Language corpora

The work presented in this paper is based on a survey of the lexical material of English, Dutch, French, Japanese, Finnish, Hungarian, and Turkish. In each instance, our primary data consist of a large list of words in that language. That of English is due to the CMU pronouncing dictionary,[5] a public-domain database of over 125,000 English words phonemically transcribed in the ARPAbet for segmental and accentual content; for the results reported here, we ignored all notation of stress, focusing just on the segment level. The word-lists of Dutch, French, and Japanese were made available to us by John Goldsmith (p.c.), all phonemically transcribed for segmental content (no accent or tone) and containing approximately 57,000, 22,000, and 58,000 words each. The lists of Finnish, Hungarian, and Turkish words are due to the Leipzig Corpora Collection,[6] respectively containing 188,000, 33,000, and 117,000 words in standard orthography, which for all three languages we judge to be phonemically transparent.

[4] What constitutes "relatively large" becomes apparent upon looking at the distribution of *PMI* across all words of English. See below.

[5] `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`

[6] `http://corpora.uni-leipzig.de`

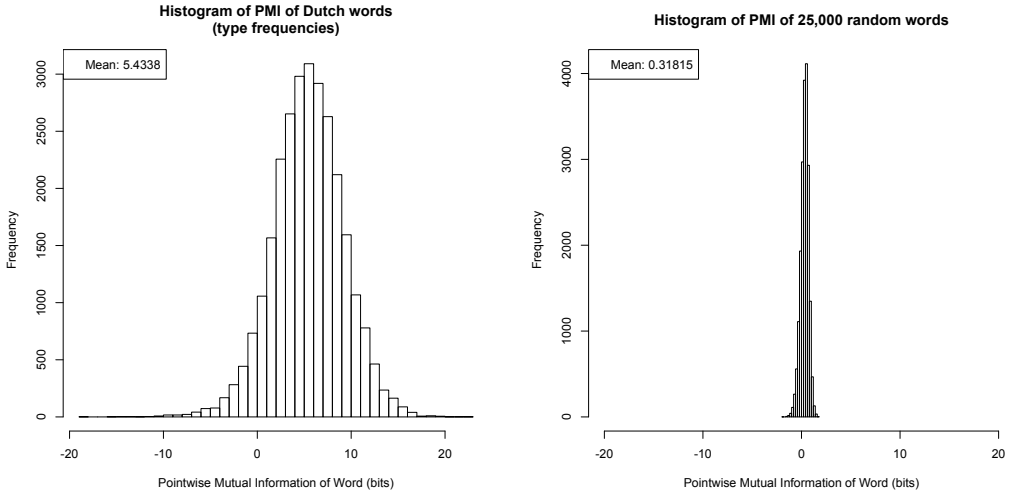

Figure 1: Histograms of the distribution of total word PMI in Dutch (left) and the random lexicon (right). The height of each bar indicates the number of words with total PMI in that bin.

Thus we remain fairly agnostic in this study about what exactly should be counted as lexical material for a language. That is, we include any word (phonemically transcribed, where possible) that may occur in written text surrounded by spaces. For our present purposes, we do not wish to espouse any particular linguistic theory about lexical specification, or the mental lexicon and its contents, simply viewing any utterable surface-string as lexically present in its entirety (modulo some notion of wordhood).

2.3 The random lexicon

The essence of our claim is that the shapes of words in human languages are much more strongly governed by strictly local statistical generalizations than may be expected by chance. To get a handle one what exactly can be expected to occur by chance, we've constructed a random lexicon of 25,000 words as a control against which to compare the properties of attested lexica. Each word of the random lexicon was generated by choosing a length uniformly at random between one and eight segments, and then choosing that many segments uniformly a random from the 39 phonemes of English (according to the CMU Pronouncing Dictionary) in random order. A maximum length of eight segments was used so that all words could viably have their permutations enumerated (see below). Since each segment of each word was chosen independently and uniformly at random, the random lexicon exhibits the minimal possible adherence to statistical bigram tendencies—in the limit of an infinite random lexicon, all bigrams become equally common.

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

Table 1: The mean total PMI of words in each lexicon.

3 Predictability among lexical neighborhoods

3.1 The distribution of word-PMI

We may begin by noting the distribution of PMI across the words of the attested lexica, and how this differs significantly from the random case. For each word in a lexicon, we can calculate its total PMI as described in section 2; figure 1 contrasts how these word-PMIs are distributed in Dutch, a representative human language, with their distribution in the random lexicon. Table 1 gives the average word-PMI in each of the lexica surveyed. Two facts bear noting: first, that the actual lexica have words of much higher average PMI than the random lexicon; and second, that the three vowel harmony languages (Finnish, Hungarian, Turkish) show markedly lower averages than the non-harmony languages. The actual languages' divergence from random is as expected, and quantifies to some extent the strength of the local phonotactics in each. The apparent difference between harmony and non-harmony languages indicates that less of the formers' lexica is captured by a strictly local, bigram probability model; whether this is significant and actually due to the prevalence of nonlocal harmony tendencies, as opposed to other factors (e.g., a less transparent orthography-phonemics mapping than we've supposed), is difficult to assess from this small sample. Interestingly, though, this difference continues to hold true when we compare lexica against the spaces $\mathscr{L}_{x!}$ and $\mathscr{L}_{e_1(x)}$ of alternatives.

3.2 Permutation neighborhoods

We examine the first of these spaces, $\mathscr{L}_{x!}$, as follows. For each word in a lexicon, we enumerate its "permutation neighborhood," the set of all words that can be made by rearranging the segments of the original word in any way. Since the size of a word w's permutation neighborhood grows with $|w|!$, the enumeration of this neighborhood quickly becomes intractable for large words. For this reason, we limit our examination to words of eight segments or fewer. Now for each word examined, we calculate two values with respect to its permutation neighborhood. The first is the percentile of the actual attested word's PMI among its neighborhood—that is, the percentage of words in w's permutation neighborhood whose PMIs the PMI of w is greater than. We take this as a measure of how "optimized" for PMI (and thus the adherence to strictly local phonotactics) the word is among the permutation-space

LANGUAGE	MEAN PERCENTILE	MEAN NEARNESS
English	96.03	92.46%
Dutch	96.34	91.11%
French	95.41	91.62%
Japanese	97.71	95.40%
Finnish	92.02	85.94%
Hungarian	92.55	86.73%
Turkish	92.35	86.91%
Random	59.95	61.71%

Table 2: Each lexicon's average percentile of attested word-PMI versus PMIs of permutations, and mean nearness of attested word-PMI to permutation-maximal PMI.

of alternatives. The second quantity is what we call the "nearness" of w's PMI to maximal, an indicator of how close to the best possible PMI in the permutation neighborhood the actual word's PMI is. Among the neighborhood, some permutation(s) will attain the maximal PMI for that neighborhood, and some other permutation(s) will attain the minimal PMI; call these values $\text{PMI}_{\max}(w!)$ and $\text{PMI}_{\min}(w!)$, respectively. The nearness of the actual word w's PMI to the maximum is then defined as:

$$\frac{|\text{PMI}(w) - \text{PMI}_{\min}(w!)|}{|\text{PMI}_{\max}(w!) - \text{PMI}_{\min}(w!)|} \tag{14}$$

which is the ratio of the distance between $\text{PMI}(w)$ and $\text{PMI}_{\min}(w!)$ to the distance between $\text{PMI}_{\max}(w!)$ and $\text{PMI}_{\min}(w!)$.

Table 2 shows the values of these two quantities for each of the seven languages, plus the random lexicon. In addition, figures 2–6 give histograms of the distribution of these quantities in a selection of the surveyed lexica. As with the distribution of word-PMI, we find that the mean percentiles and nearnesses of the attested lexica are much greater than those of the random lexicon, and figure 2 shows that their distribution is qualitatively different. Furthermore, the harmony languages show systematically lower mean percentiles and nearness values than the non-harmony languages, and the distribution of the latter, at least, is qualitatively different in the harmony languages (see figures 5, 6).

Interestingly, in all of the languages surveyed, it is extremely rare for an attested word's PMI to actually achieve the maximum of all its permutation neighbors (this only happens on the order of several hundred times per lexicon, out of tens of thousands of words). Thus, though we interpret the high percentiles and nearnesses found here as indicating that the attested lexica are indeed very nearly optimal within the space $\mathscr{L}_{x!}$, they very infrequently achieve actual optimality for any given word, which perhaps reflects the influence of nonlocal phonotactic generalizations.

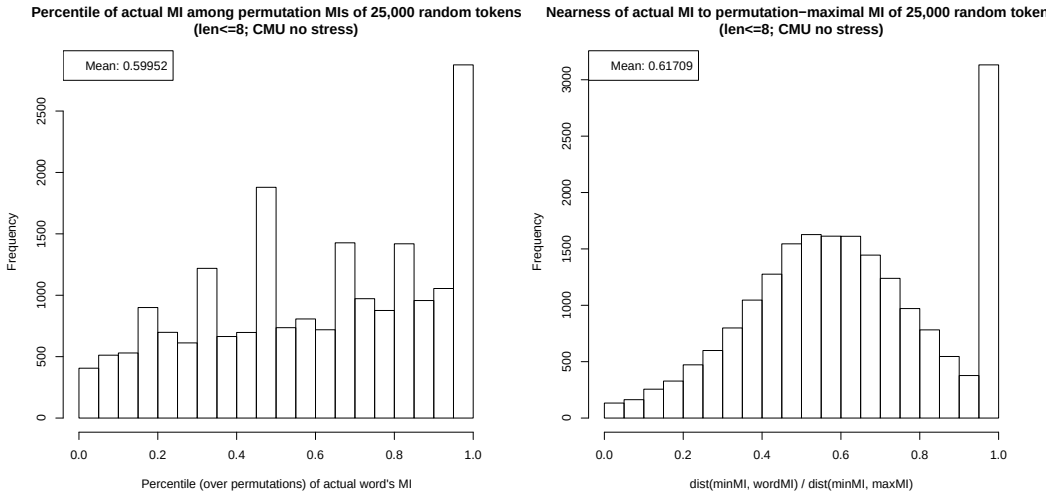

Figure 2: Histograms of the distribution of the percentile of a word's PMI with respect to its permutation neighbors' (left) and the nearness of a word's PMI to the maximum achieved by its permutation neighbors (right) in the random lexicon. Notice the significantly different character of these distributions from those of actual languages in figures 3–6.

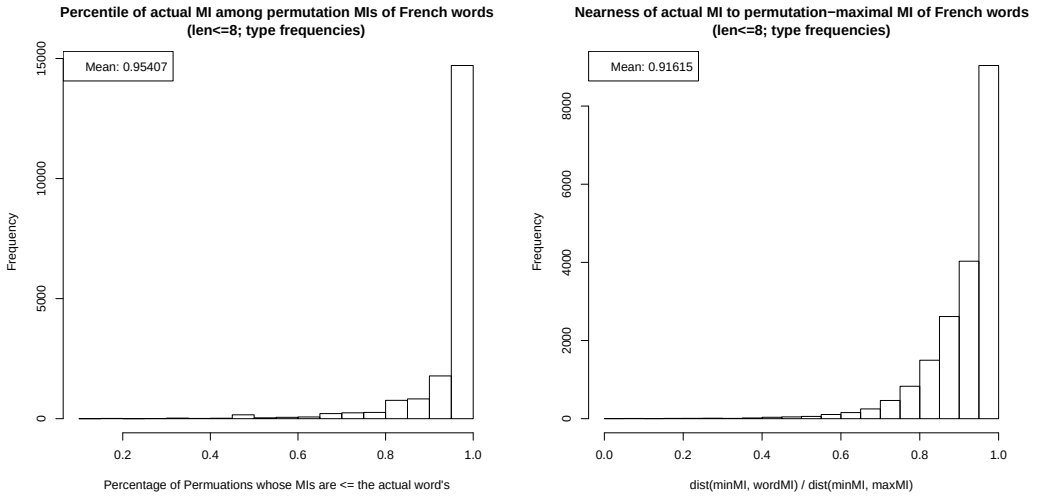

Figure 3: Histograms of the distribution of the percentile of a word's PMI with respect to its permutation neighbors' (left) and the nearness of a word's PMI to the maximum achieved by its permutation neighbors (right) in the French lexicon.

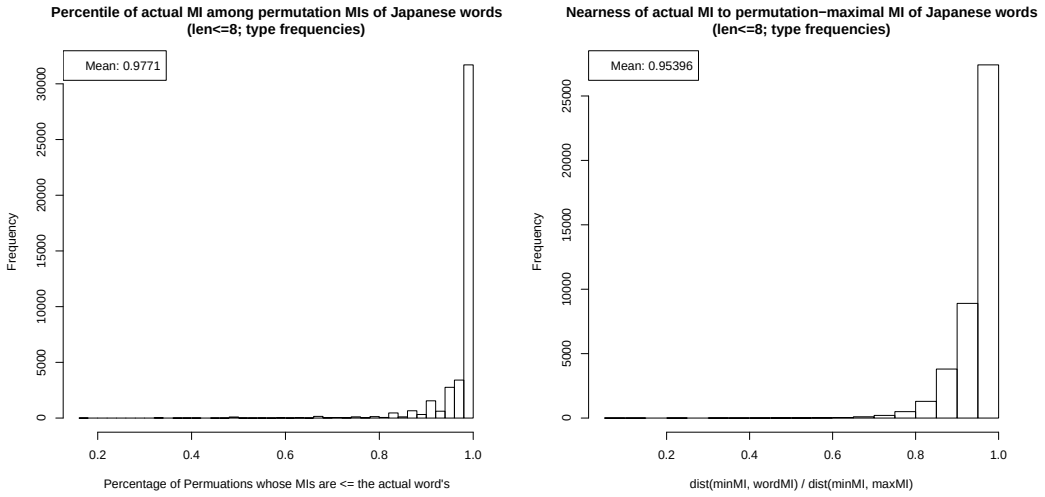

Figure 4: Histograms of the distribution of the percentile of a word's PMI with respect to its permutation neighbors' (left) and the nearness of a word's PMI to the maximum achieved by its permutation neighbors (right) in the Japanese lexicon.

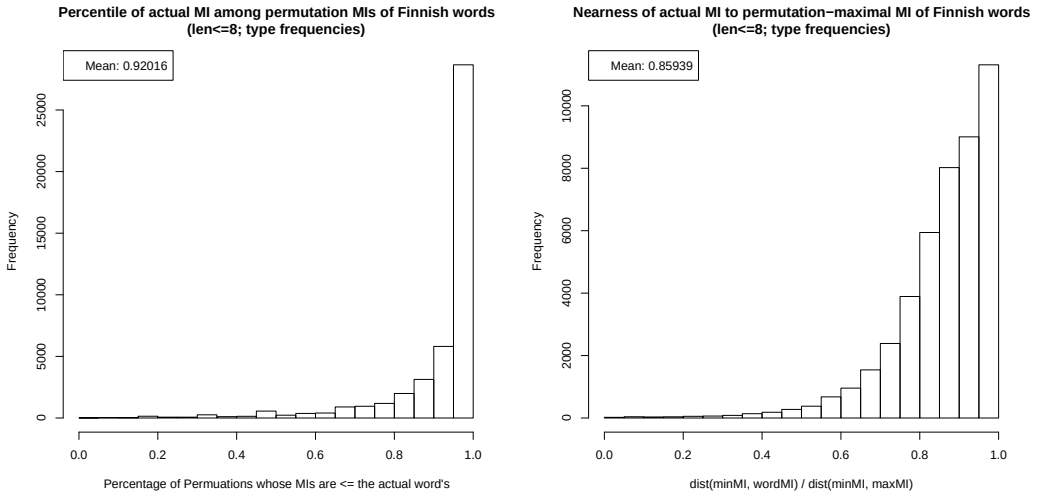

Figure 5: Histograms of the distribution of the percentile of a word's PMI with respect to its permutation neighbors' (left) and the nearness of a word's PMI to the maximum achieved by its permutation neighbors (right) in the Finnish lexicon. Note the "thicker" distribution of the latter compared to those of the non-harmony languages (figs. 3–4).

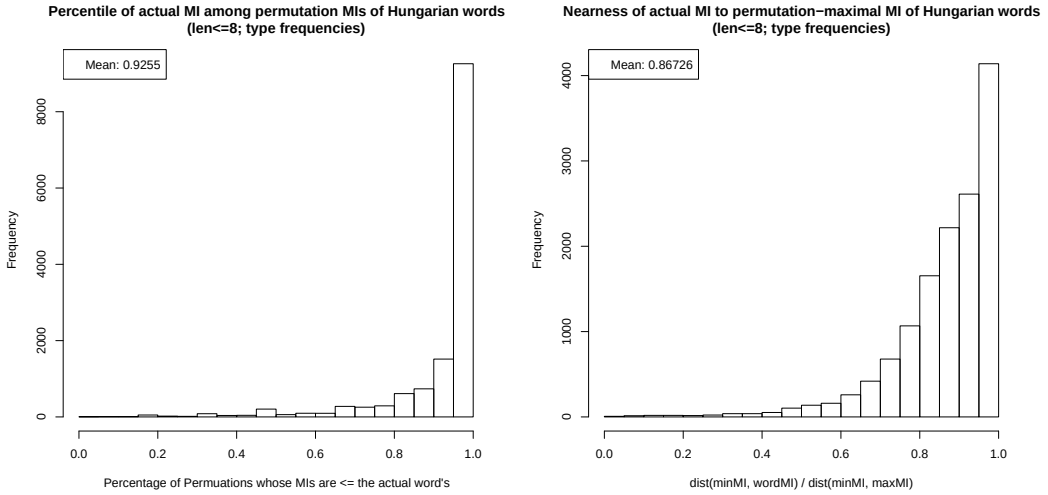

Figure 6: Histograms of the distribution of the percentile of a word's PMI with respect to its permutation neighbors' (left) and the nearness of a word's PMI to the maximum achieved by its permutation neighbors (right) in the Hungarian lexicon. Note the "thicker" distribution of the latter compared to those of the non-harmony languages (figs. 3–4).

3.3 Edit neighborhoods

The 1-edit space, $\mathscr{L}_{e_1(w)}$, is derived as follows. For each word w in a lexicon we can enumerate $e_1(w)$, the set of all words that result from making any single edit (or none) to w, that is, any single insertion of a segment from the language's phoneme inventory, any single deletion of a segment in w, or any single substitution of a segment in w by one in the inventory. Unlike in the case of the permutation neighborhood, the size of this set grows only linearly with $|w|$:

(15) $|e_1(w)| = |w| + |\Sigma|(2|w|+1) + 1$
($|w|$ possible deletions + $|w||\Sigma|$ possible substitutions + $(|w|+1)|\Sigma|$ possible insertions, plus w itself).

Thus it is tractable to enumerate $e_1(w)$ for any attested word, and so we do not constrain ourselves to words of eight segments or fewer as above.

As with the permutation neighborhoods, we calculate the percentile of attested words' PMI in these 1-edit neighborhoods, as well as their nearness to the 1-edit-maximal PMI. Table 3 gives the means of these quantities for each of the lexica, and figures 7–12 show their distribution in the random lexicon, in two of the non-harmony languages (English and Dutch), and in all three of the harmony languages. The first thing to note is that mean percentiles and nearnesses achieved in the 1-edit neighborhoods are all lower than was found for permutation neighborhoods, with some languages' mean nearness even dropping below that found in the random lexicon. Nonetheless, the mean percentiles remain robustly above random, and the distributions of both, as seen in the figures, are significantly more skewed than random across languages.

LANGUAGE	MEAN PERCENTILE	MEAN NEARNESS
English	80.82	72.43%
Dutch	80.28	47.99%
French	76.28	62.63%
Japanese	73.34	45.09%
Finnish	79.38	55.54%
Hungarian	78.01	49.38%
Turkish	82.35	62.88%
Random	59.48	57.74%

Table 3: Each lexicon's average percentile of attested word-PMI versus PMIs of 1-edits, and mean nearness of attested word-PMI to 1-edit-maximal PMI.

We interpret the greater-than-random mean percentiles and rightward skew of the distributions as again indicating that, among the space of possible 1-edits, the words of attested languages are near optimal in their adherence to strictly local bigram phonotactics, though less so than is the case in the permutation space. The lower percentiles reached here make some sense intuitively, insofar as most edits of a word will represent linguistically more plausible alternatives than do most permutations (many phonological processes correspond to segmental insertions, deletions, or substitutions, while comparatively few look like segmental permutations, much less arbitrary ones). Unlike in the permutation space, here it is not clear that a systematic difference between harmony and nonharmony languages emerges, though the distribution of Finnish's metric of nearness to 1-edit-maximal PMI surely looks anomalous in comparison to the other lexica.

4 Coclusions and further work

We have presented a general methodology of assessing how locally driven the lexical phonotactics of a language are, by comparing its attested lexical material to some hypothesis of what else that material "could have been" (here, permutations and edits), in terms of its local, bigram predictability. In other words, we ask how "optimized" the lexicon is for local predictability within some space of possible lexica. The method finds that languages exhibit interesting variation in the degree to which they are optimized in this way, but that every language is significantly more locally optimal than random; and furthermore, that the languages which appear least optimized in this sense are precisely those known to obey important nonlocal lexical tendencies (vowel harmony).

We can identify a number of deficiencies in the work presented here. First, the application of these methods to a broader swath of the world's languages would aid immensely in the interpretation of our results, as would a deeper analytic understanding of the distribution, in general, of pointwise mutual information among the kinds of string-spaces we've been considering. There are also immediate extensions and improvements to this work that suggest themselves naturally, such as the ap-

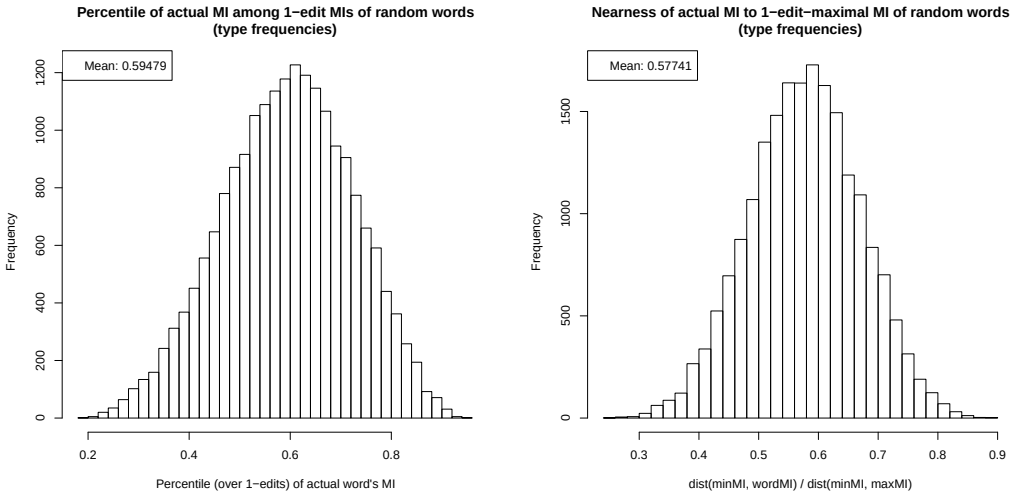

Figure 7: Histograms of the distribution of the percentile of a word's PMI with respect to its 1-edit neighbors' (left) and the nearness of a word's PMI to the maximum achieved by its 1-edit neighbors (right) in the random lexicon. Notice the significantly different character of these distributions from those of actual languages.

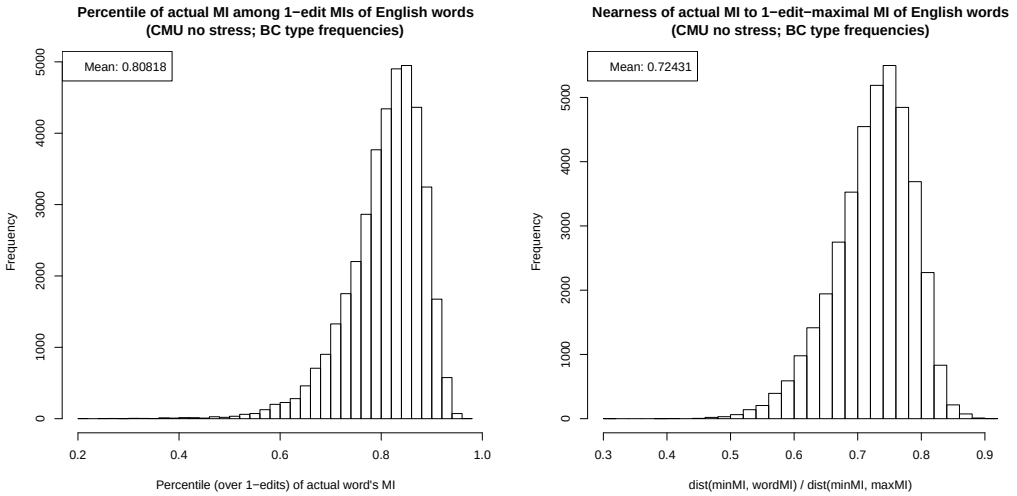

Figure 8: Histograms of the distribution of the percentile of a word's PMI with respect to its 1-edit neighbors' (left) and the nearness of a word's PMI to the maximum achieved by its 1-edit neighbors (right) in the English lexicon.

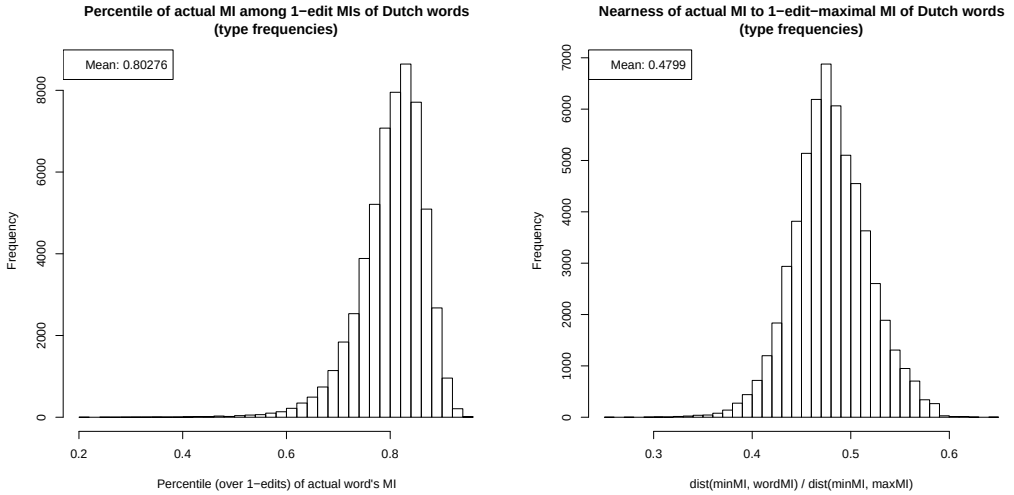

Figure 9: Histograms of the distribution of the percentile of a word's PMI with respect to its 1-edit neighbors' (left) and the nearness of a word's PMI to the maximum achieved by its 1-edit neighbors (right) in the Dutch lexicon.

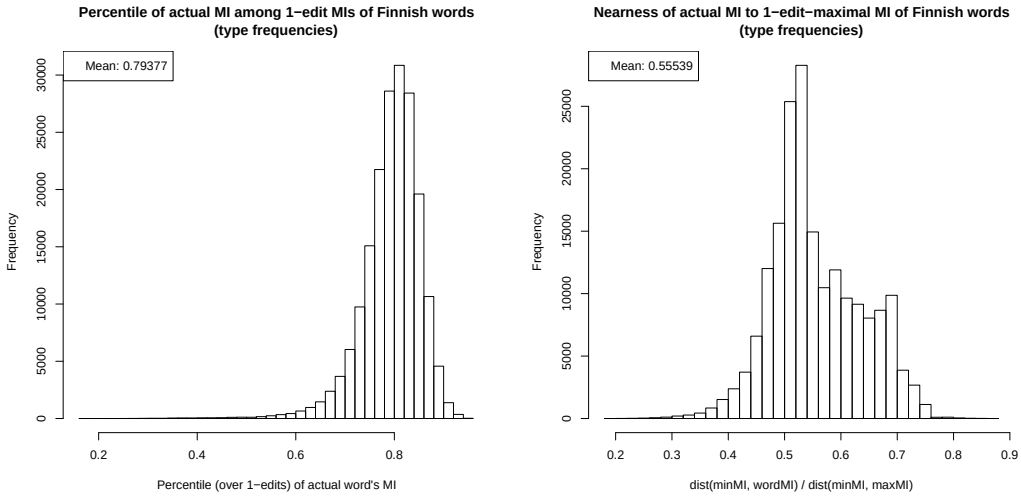

Figure 10: Histograms of the distribution of the percentile of a word's PMI with respect to its 1-edit neighbors' (left) and the nearness of a word's PMI to the maximum achieved by its 1-edit neighbors (right) in the Finnish lexicon.

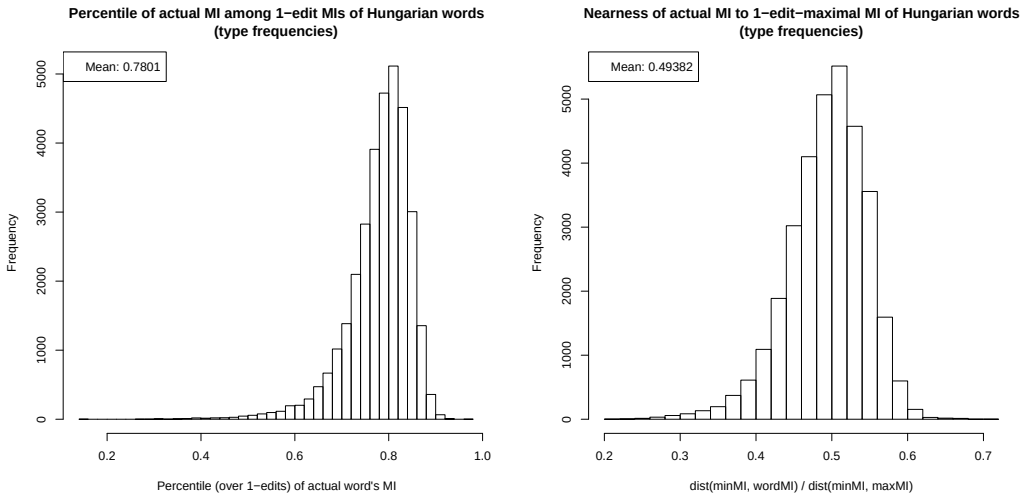

Figure 11: Histograms of the distribution of the percentile of a word's PMI with respect to its 1-edit neighbors' (left) and the nearness of a word's PMI to the maximum achieved by its 1-edit neighbors (right) in the Hungarian lexicon.

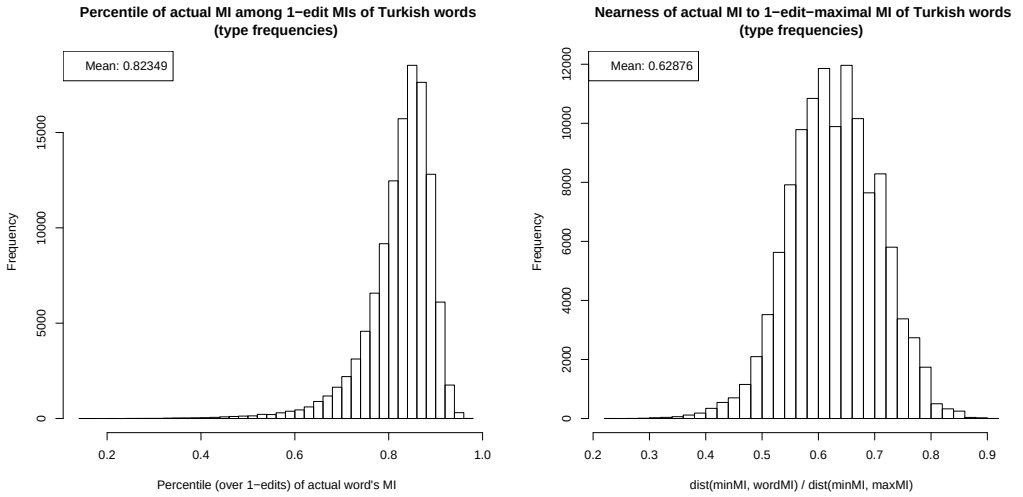

Figure 12: Histograms of the distribution of the percentile of a word's PMI with respect to its 1-edit neighbors' (left) and the nearness of a word's PMI to the maximum achieved by its 1-edit neighbors (right) in the Turkish lexicon.

plication of standard statistical methods to quantify the significance of the lexica's deviations from random and from each other, or the consideration of n-edit spaces for $n > 1$.

It is also true that our central thesis—that languages' lexica represent near-optima in some space of possibilities, with respect to strictly local phonotactic probability—has only been approached in a somewhat indirect fashion here, by merely performing word-by-word measurements of relative optimality and then reporting the averages of those measurements for a whole lexicon. Ideally, we would like to attack the question directly, by enumerating or sampling the space of alternative lexica themselves; that is, by actually generating each alternative lexicon *in toto*, taking some measure of its overall bigram-probability, and comparing that to the same measure as applied to the attested lexicon. So for the 1-edit space, this would amount to enumerating (or more likely, sampling) each whole lexicon that results from making 1 or 0 edits to any subset of the words in the attested lexicon, and comparing it against the attested lexicon. At the least, it would be useful to work out how the word-by-word figures and distributions we've collected here might be expected to diverge from this ideal methodology.

References

Chomsky, N., and M. Halle. 1968. *The Sound Pattern of English*. New York: Harper and Row.

Duanmu, S. 2002. Two theories of onset clusters. *Chinese Phonology* 11.97–120.

Frisch, S., J. Pierrehumbert, and M. Broe. 2004. Similarity avoidance and the OCP. *Natural Language and Linguistic Theory* 22.179–228.

Lyovin, A. V. 1997. *An Introduction to the Languages of the World*. New York: Oxford University Press.

Manning, C. D., and H. Schütze. 2000. *Foundations of Natural Language Processing*. Cambridge: MIT Press, corrected edition.