An Entropy-based Assessment of the Unicode Encoding for Tibetan

Paul G. Hackett

Columbia University

<ph2046@columbia.edu>

Unicode

- Represent Languages with minimum overhead
- Characters vs. Glyphs
- Goal: Efficiency with Comprehensive Coverage

History:

Tibetan Unicode 1.0beta proposed in May, 1993 using *virāma*-model but not accepted

Entropy

- a measure of the uncertainty or amount of disorder in a system
- a measure of the lower bound of the compression ratio of any compression algorithm for linguistic data
- the lower bound to the average number of bits per symbol needed to encode a message

Entropy & Information Theory:

Claude Shannon, A Mathematical Theory of Communication, *Bell System Technical Journal*, 27 [1948]: 379-423, 623-656.

Calculating Entropy I: Theoretical Estimates

- **Contingent on Language Model**
  - —Traditional Tibetan morphology, UNICODE
  - **and the Transition Probability function**
  - —Independent graphemes, conditional bi-grams, tri-grams, etc.
- **Entropy, $\mathcal{H}$** where Probability, P_i

$$\mathcal{H} = -\sum_{i=1}^{n} P_i \log_2 P_i \qquad P_i = \sum_{j=1}^{n} P(x_i \mid x_j)$$

Traditional Tibetan Orthography

- **44 discrete graphemes:**

  - Thirty letters, four explicit vowels, three super-scripts, four subscripts, *tsheg*, *shad*, whitespace

- **Independent Grapheme model:**

  - Theoretical upper bound for the grapheme Entropy of Tibetan, $\mathcal{H}$ = 5.46 bits per symbol (bps)

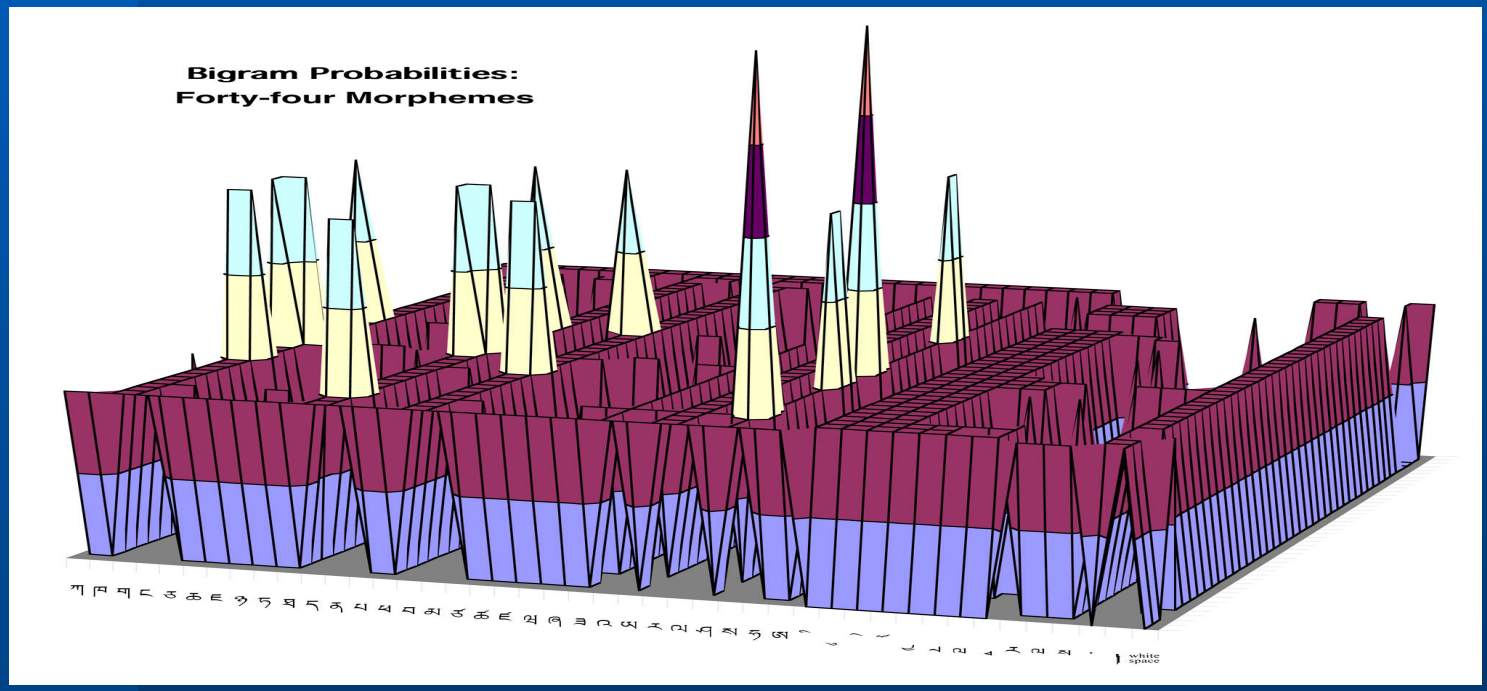
For Comparison ...

Shannon estimated the Entropy of English at 1.3 bits per symbol

— Claude Shannon, Prediction and Entropy of Printed English, *Bell System Technical Journal*, 30 [1951]: 50-64.

Traditional Tibetan Orthography

- **Morphologically Constrained Bi-grams:**



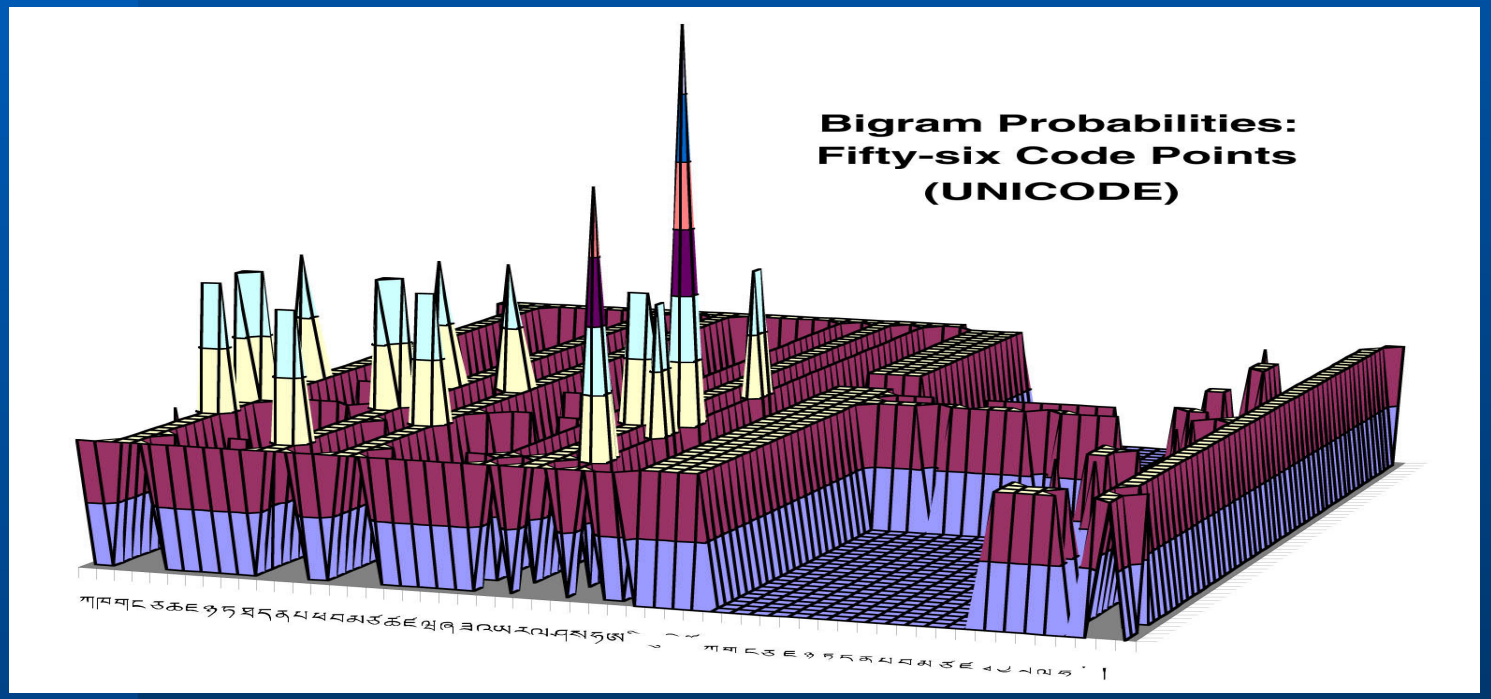
Theoretical upper bound Entropy, $\mathcal{H}$ = 4.82 bps

Tibetan Orthography in Unicode

- **56 discrete graphemes:**
  - **Thirty letters, nineteen subjoined characters, four explicit vowels, *tsheg*, *shad*, whitespace**
- **Independent Grapheme model:**
  - **Theoretical upper bound for the grapheme Entropy of Tibetan, $\mathcal{H}$ = 5.81 bits per symbol (bps)**

Tibetan Orthography in Unicode

- **Morphologically Constrained Bi-grams:**



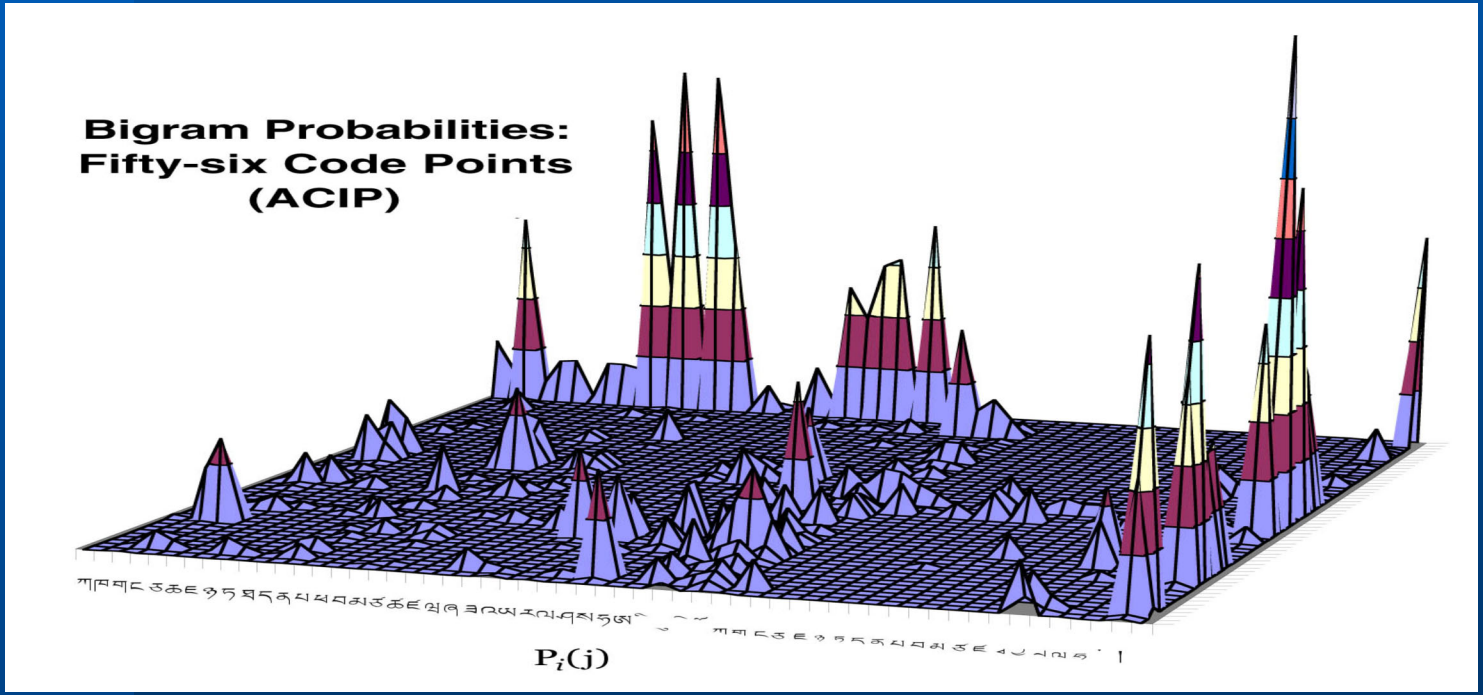
Theoretical upper bound Entropy, $\mathcal{H}$ = 4.79 bps

Calculating Entropy II: Statistical Estimates

- **Contingent on Statistically Broad Sample**
  - —ACIP corpus of religious literature
  - —41 million syllables
  - —81,000 possible valid syllables
  - —17,000 unique syllables in ACIP corpus
  - —3,136 possible bi-grams
  - —1,970 morphologically constrained bi-grams
  - —2,237 bi-grams attested in ACIP (including Sanskrit)

Statistically-derived Probabilities

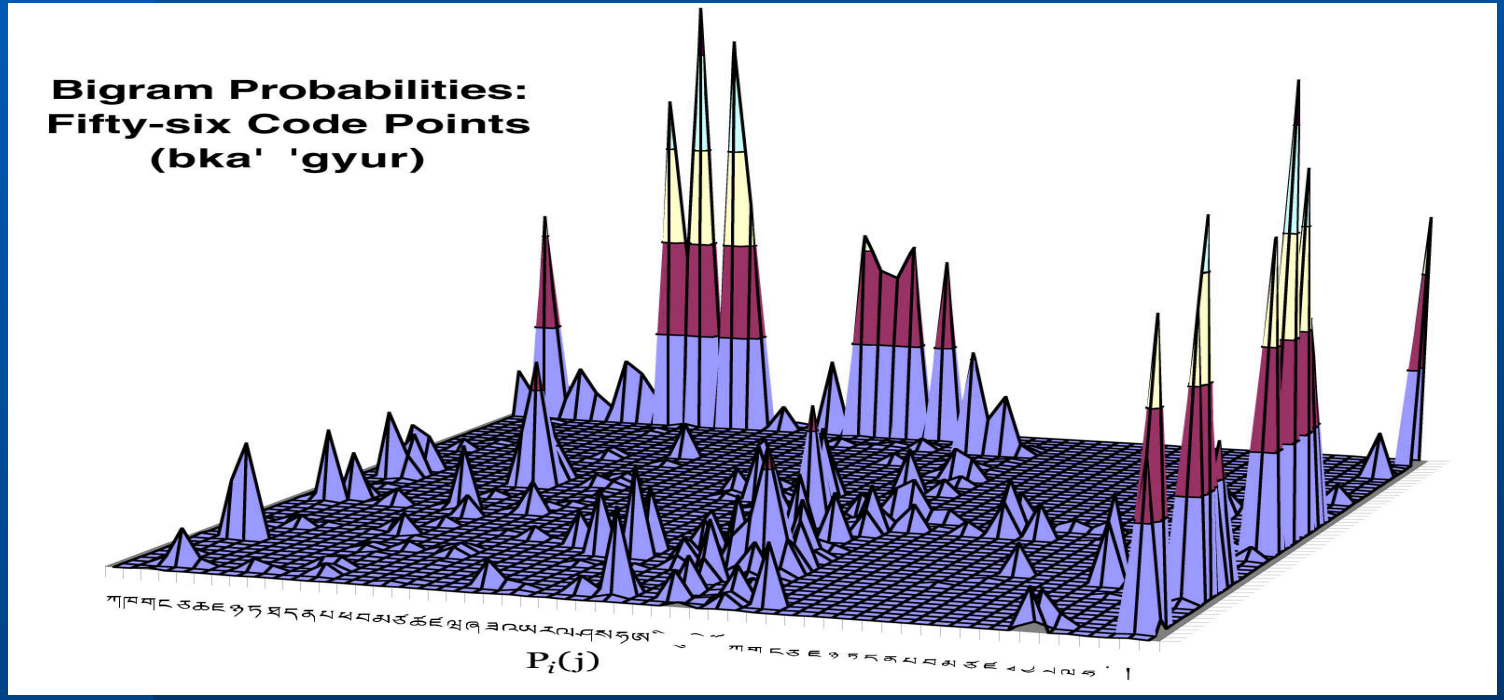
- **Contingent Bigrams in ACIP corpus:**



Entropy, $\mathcal{H}$ = 4.35 bps

Statistically-derived Probabilities

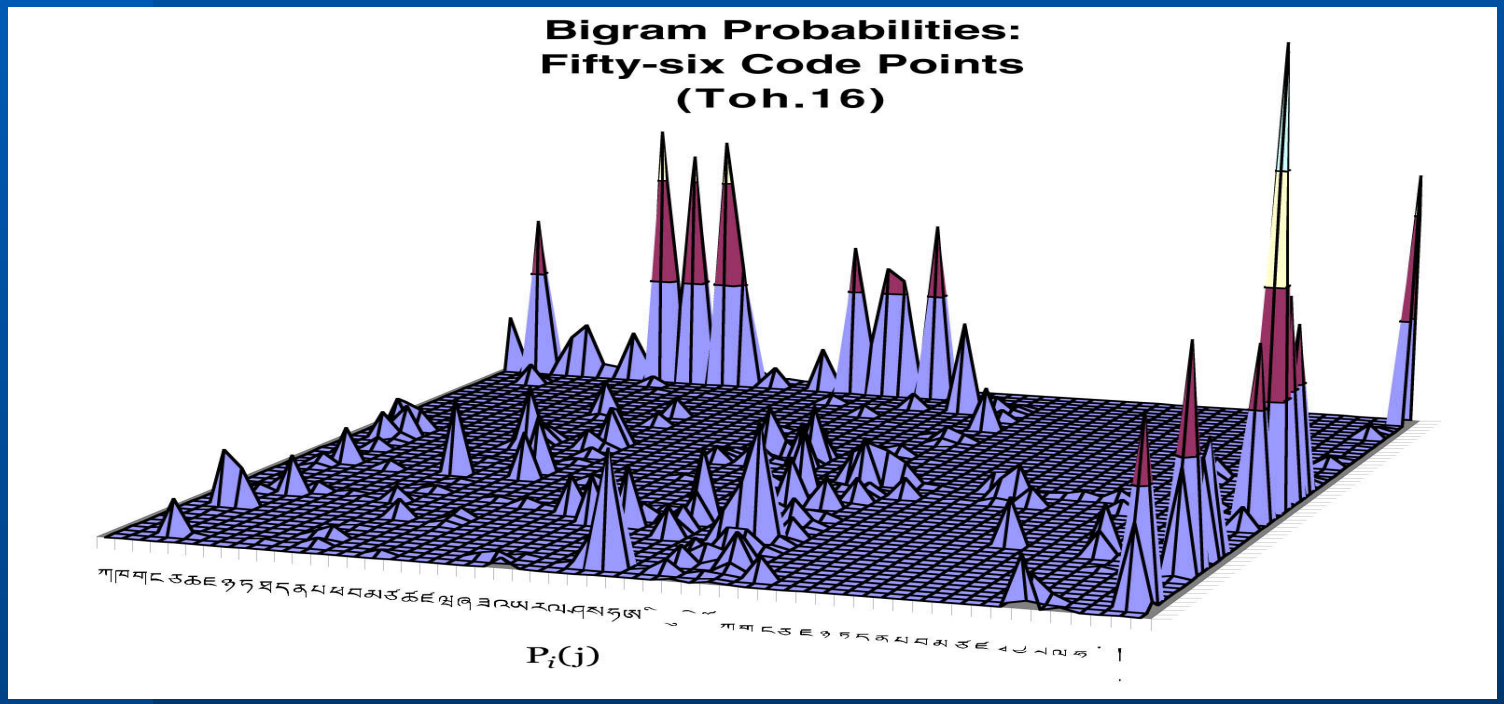
- **Contingent Bigrams in 20 བཀའ་འགྱུར་ texts:**



Entropy, $\mathcal{H}$ = 4.29 bps

Statistically-derived Probabilities

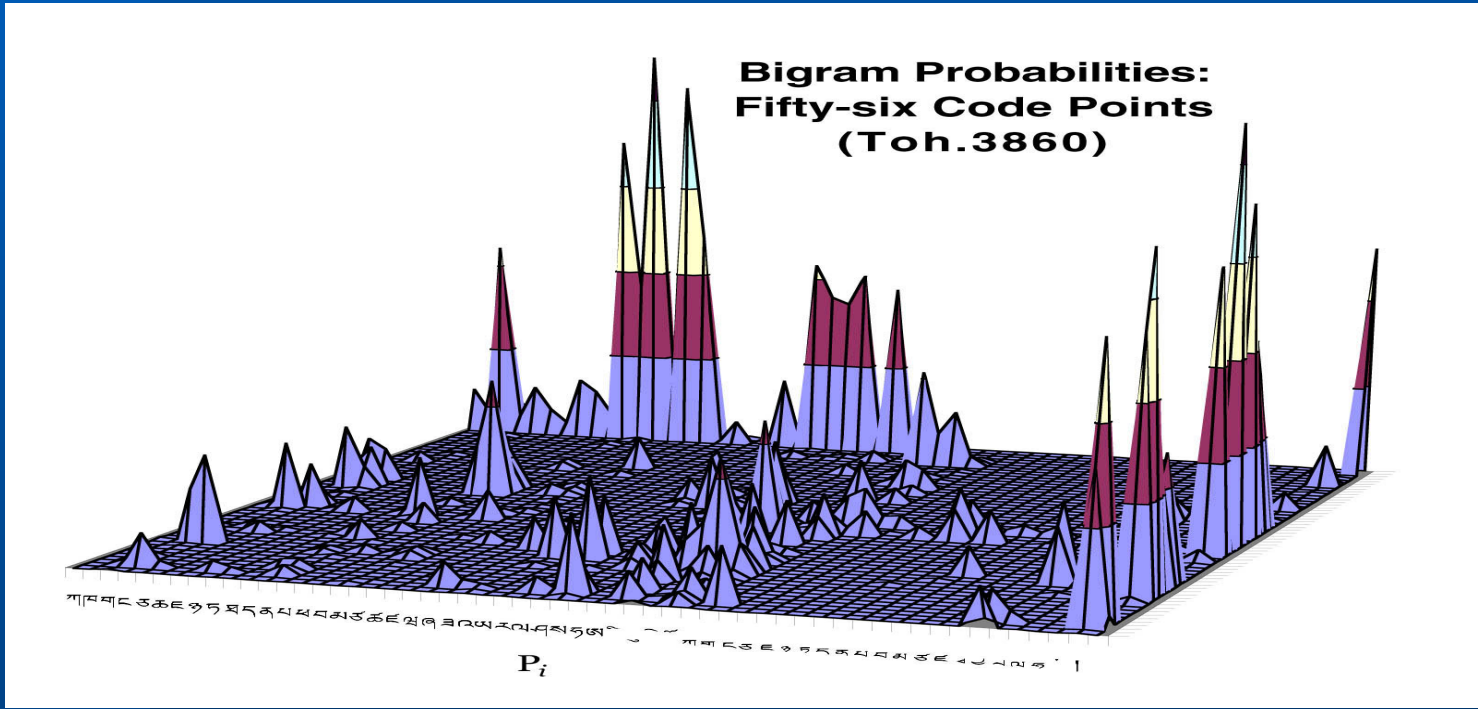
- **Contingent Bigrams in Toh.16:**



Entropy, $\mathcal{H}$ = 4.17 bps

Statistically-derived Probabilities

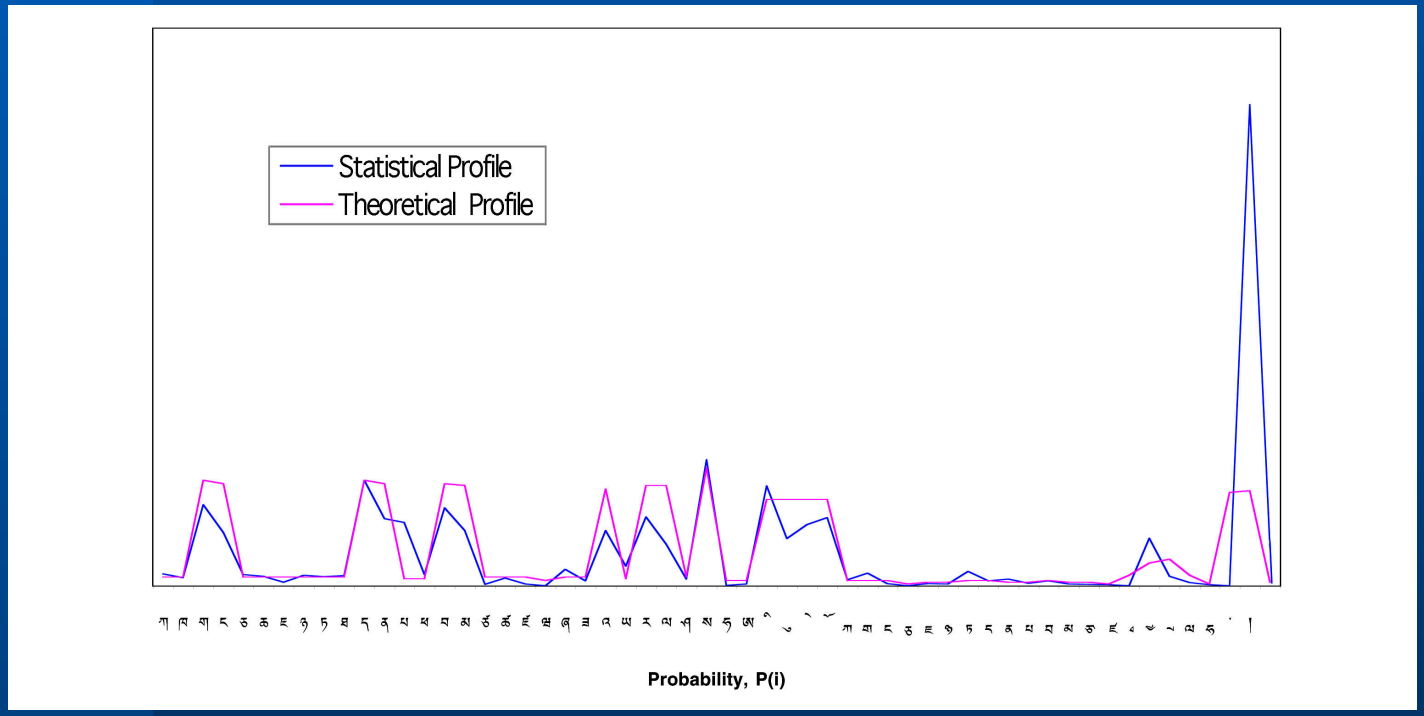
- **Contingent Bigrams in Toh.3860:**



Entropy, $\mathcal{H}$ = 3.78 bps

Summary

- Statistical probabilities differ slightly from theory
- Dominated by syllable boundary effects



Implications

- **UNICODE**
  - —Not statistically different from traditional representation; possibly better
- **Entropy Language-model**
  - —Theoretical upper bound insufficient (~4.8 bps)
  - —Statistical bound lower (~4.3 bps), and is heavily dominated by syllable boundaries
  - —Compression limit can be lower still for individual texts

Applications

- **Statistical Entropy Language-model**
  - —**OCR error detection**
  - —**Efficient Generic Compression Scheme**
  - —**Calculated quickly for Custom Compression**