

Goldsmith, John (2002). Probabilistic  
Models of Grammar: Phonology as  
Information Minimization.  
*Phonological Studies* 5, 21-46.

Presented by Sung-Hoon Hong  
(HUFS)

# Organization

- Information Theory and the two fundamental measures, *positive log probability* and *entropy*
- Other information measures:
  - Phonological Complexity (PC)
  - Mutual Information
- Probability models: unigram vs. bigram models
- Maximize the probability of the observations
- Bayes' rule: Identifying the language from which a word is drawn
- Relation to other approaches (conjecture)
- Application to other phonological phenomena

# What is Information Theory?

- Information Theory is concerned with representing mathematically how much information is needed to convey a message given the constraints imposed on a communication system. (Hume & Malihot, in press:4)

# Two fundamental information measures

- We can measure the amount of information for a set of elements (or a system) or a particular element within a system.
  - Information of a particular element (개별 요소의 정보량)
    - ***positive log probability*** ('plog' (Goldsmith 2006, 2011), 'surprisal' (Levy 2008, Hume & Malihot, in press), or 'information content' (Hume 2004, 2006, 2008, Hume & Bromberg 2005))
  - Information of a system of elements (시스템의 정보량)
    - ***entropy***
  - An 'element' here is usually a phone/phoneme.
  - A 'system' here refers to any set of phones/phonemes, such as the set of coronal consonants, the set of front vowels, the set of sonorants, the set of all vowels, etc.

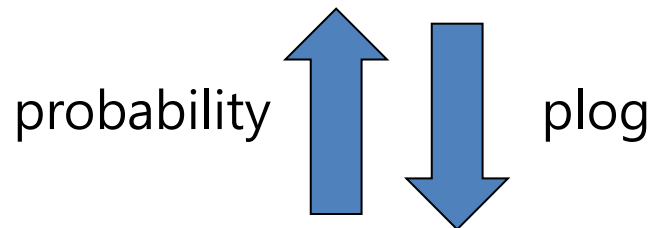
# How to measure information?

## (cont)

- Information is measured by the **predictability** or **probability** of an element.
- Predictability/probability is inversely related to the amount of information.
  - The more predictable and probable, the less the amount of information;
  - The less predictable and probable, the greater the amount of information.

# Information of an element: *plog*

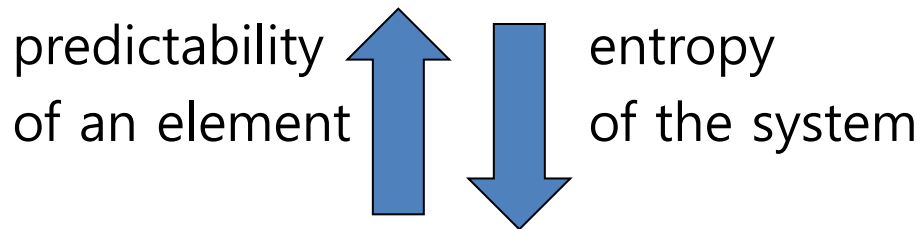
- The higher the probability of an element (and the higher the frequency of the element), the lower the **plog**.



- Assume a corpus that consists of 1,000,000 phonemes, of which /t/ appears 1000 and /h/ 10 times. Which phoneme has more probability, and hence is lower in plog?

# Information of a system: *entropy*

- The more variation and difference there are in a system, (the less predictable the elements of a system), the higher the **entropy** of the system.



- Compare the systems of two coronal consonants and eight coronal consonants. In which system are the occurrences of coronal consonants more predictable? In other words, which system of coronal consonants is lower in entropy?

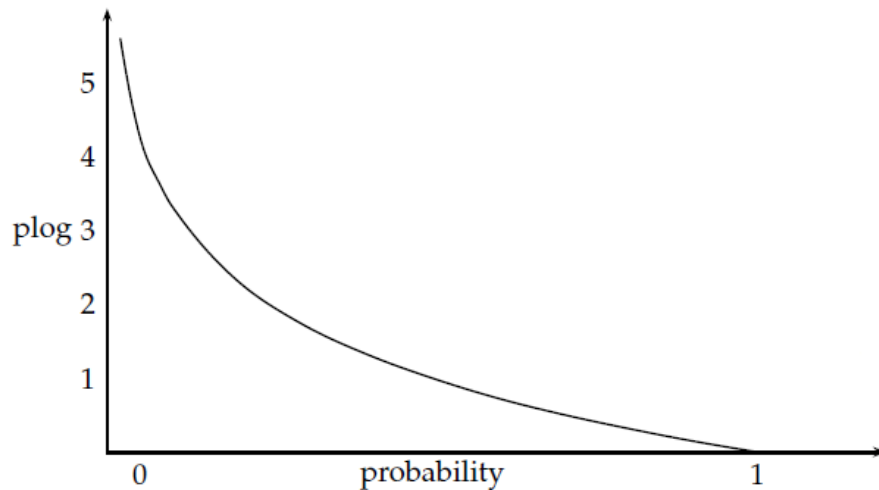
# How to measure information

- A simple probability value is not used to represent the amount of information.
- The values of probability are typically very small, and thus logarithm is used to avoid such small numbers and make the numbers seem more meaningful. Also, by using logarithm, multiplication can be replaced with simple addition (cf.  $\log xy = \log x + \log y$ )
- But  $\log(\text{probability})$  is negative because probability values are less than 1, and to make the values positive, it is multiplied by -1.
- The amount of information is usually represented by 'bits', so the base of the log is 2.



# Two fundamental measures: plog

$$plog(x) = -\log_2 Prob(x)$$



| $P(x)$                            | $plog(x)$ |
|-----------------------------------|-----------|
| 1                                 | 0         |
| $\frac{1}{2}$ (=0.5)              | 1         |
| $\frac{1}{4}$ (=0.25)             | 2         |
| $\frac{1}{16}$ (=0.125)           | 3         |
| $\frac{1}{1024}$ (=0.000987)      | 10        |
| $\frac{1}{1048576}$ (=0.00000095) | 20        |

$$-\log_2(1) = -\log_2 2^0 = 0$$

$$-\log_2\left(\frac{1}{2}\right) = -\log_2 2^{-1} = 1$$

$$-\log_2\left(\frac{1}{4}\right) = -\log_2 2^{-2} = 2$$

$$-\log_2\left(\frac{1}{1024}\right) = -\log_2 2^{-10} = 10$$

$$-\log_2\left(\frac{1}{1048576}\right) = -\log_2 2^{-20} = 20$$

# Two fundamental measures: entropy

- Information of a system of elements  
entropy = average *plog* per element in a **system**

$$\text{entropy}(\{x_1 \dots x_N\})$$

$$\begin{aligned} &= \frac{1}{N} \left( \sum_{i=1}^N (-\log_2 \text{Prob}(x_i)) \right) = \frac{1}{N} \left( \sum_{j=1}^V \text{count}(x_j) \times (-\log_2 \text{Prob}(x_j)) \right) \\ &= \sum_{j=1}^V \left( \frac{\text{count}(x_j)}{N} \times (-\log_2 \text{Prob}(x_j)) \right) = \sum_{j=1}^V \left( \text{Prob}(x_j) \times (-\log_2 \text{Prob}(x_j)) \right) \end{aligned}$$

(N: the total number of elements or tokens in the corpus)

(V: the total number of *distinct* elements or types in the corpus)

# Illustration

- Suppose we are interested in measuring the information of the three systems of coronal consonants as follows:
  - coronal system *A* where there are only two coronal consonants, /t/ and /d/
  - coronal system *B* where there are four coronal consonants /t, d, s, n/, which occur with the equal probability of 0.25%.
  - coronal system *C* where there are four coronal consonants /t, d, s, n/, which occur with different probabilities such that /t/ appears 0.5%, /d/ 0.25%, and /s/ and /n/ 0.125% each.

# Illustration: system A

$$Prob(t) = Prob(d) = 0.5$$

$$S(t) = -\log_2 Prob(t) = -\log_2 \frac{1}{2} = -\log_2 2^{-1} = 1$$

$$S(d) = S(t) = -\log_2 \frac{1}{2} = 1$$

$$entropy(\{t, d\}) = \sum_{i=1}^2 Prob(cor_i) \times (-\log_2 Prob(cor_i)) == \frac{1}{2} \times 1 + \frac{1}{2} \times 1 = 1$$

$$entropy(\{t, d\}) = \frac{1}{2} \left( \sum_{i=1}^2 (-\log_2 Prob(cor_i)) \right) == \frac{1}{2} (1 + 1) = 1$$

# Illustration: system B

$$Prob(t) = Prob(d) = Prob(s) = Prob(n) = 0.25$$

$$S(t) = -\log_2 Prob(t) = -\log_2 \frac{1}{4} = -\log_2 2^{-2} = 2$$

$$S(t) = S(d) = S(s) = S(n)$$

$$entropy(\{t, d, s, n\}) = \sum_{i=1}^4 Prob(cor_i) \times (-\log_2 Prob(cor_i)) = \frac{1}{4} \times 2 + \frac{1}{4} \times 2 + \frac{1}{4} \times 2 + \frac{1}{4} \times 2 = 2$$

$$entropy(\{t, d, s, n\}) = \frac{1}{4} \left( \sum_{i=1}^4 (-\log_2 Prob(cor_i)) \right) = \frac{1}{4} (2 + 2 + 2 + 2) = 2$$

# Illustration: system C

$$Prob(t) = 0.5 \quad Prob(d) = 0.25 \quad Prob(s) = 0.125 \quad Prob(n) = 0.125$$

$$S(t) = -\log_2 \frac{1}{2} = -\log_2 2^{-1} = 1$$

$$S(d) = -\log_2 \frac{1}{4} = -\log_2 2^{-2} = 2$$

$$S(s) = S(n) = -\log_2 \frac{1}{8} = -\log_2 2^{-3} = 3$$

$$entropy(\{t, d, s, n\}) = \sum_{i=1}^4 Prob(cor_i) \times (-\log_2 Prob(cor_i)) = \frac{1}{2} \times 1 + \frac{1}{4} \times 2 + \frac{1}{8} \times 3 + \frac{1}{8} \times 3 = 1.75$$

$$entropy(\{t, d, s, n\}) = \frac{1}{8} \sum_{i=1}^8 (-\log_2 Prob(cor_i)) = \frac{1}{8} (1 + 1 + 1 + 1 + 2 + 2 + 3 + 3) = 1.75$$

# Phonological Complexity

- Phonological Complexity (PC) is the average *plog* per element in a phonological string (such as a word, a syllable, etc).

$$PC(w = x_1x_2\dots x_n) = \frac{1}{n} \sum_{i=1}^{n=length} (-\log_2 P(x_i))$$

- PC reflects well-formedness (or how well any given individual word fits into the phonotactic patterns of the language).

# Probability models

- Prior probability (or relative frequency) → unigram

$$P(x) = \frac{C(x)}{N} \quad (\text{C: count})$$

- Conditional probability given previous → bigram

$$P(x_i | x_{i-1}) = \frac{C(x_{i-1}x_i)}{C(x_{i-1})}$$

- Conditional probability given next → bigram

$$P(x_i | x_{i+1}) = \frac{C(x_i x_{i+1})}{C(x_{i+1})}$$

- Conditional probability given surrounding → trigram

$$P(x_i | x_{i-1} \cdots x_{i+1}) = \frac{C(x_{i-1}x_i x_{i+1})}{C(x_{i-1} \cdots x_{i+1})}$$



# How to obtain PC: unigram & bigram models

- Calculate the PC of "man" based on the Brown corpus (Francis & Kucera 1964)

- N=4,763,448
- C(m)=120,380
- C(a)=381,030
- C(n)=336,338
- C(#)=1,013,057
- C(#m)=39,929
- C(ma)=20,824
- C(an)=73,073
- C(n#)=87,994

$$\begin{aligned}
 PC(man) &= \frac{1}{3} ((-\log_2 P(m)) + (-\log_2 P(a)) + (-\log_2 P(n))) \\
 &= \frac{1}{3} ((-\log_2 \frac{120380}{4763448}) + (-\log_2 \frac{381030}{4763448}) + (-\log_2 \frac{336338}{4763448})) \\
 &= 4.258
 \end{aligned}$$

$$\begin{aligned}
 PC(\#man\#) &= -\frac{1}{5} (\log_2 P(\#) + \log_2 P(m | \#) + \log_2 P(a | m) + \log_2 P(n | a) + \log_2 P(\# | n)) \\
 &= -\frac{1}{5} (\log_2 \frac{C(\#)}{N} + \log_2 \frac{C(\#m)}{C(\#)} + \log_2 \frac{C(ma)}{C(m)} + \log_2 \frac{C(an)}{C(a)} + \log_2 \frac{C(n\#)}{C(n)}) \\
 &= -\frac{1}{5} (\log_2 \frac{1013057}{4763448} + \log_2 \frac{39929}{1013057} + \log_2 \frac{20824}{120380} \\
 &\quad + \log_2 \frac{73073}{381030} + \log_2 \frac{87994}{336338}) = 2.749
 \end{aligned}$$

# Unigram vs. bigram models

- The shift from the unigram model to the bigram model leads to an significant increase in the probability assigned to the corpus.
- The positive log probability of the entire English corpus (the sum of the positive log probabilities of the individual words): 1,883,085 bits in unigram model and 1,559,194 bits in bigram model.

# Unigram vs. bigram models (cont)

- Maximize the probability of the observations:
  - The goal is to develop a model of phonology which assigns probabilities, and in particular to find the phonological model which assigns the highest probability to the set of observed data.
  - The discovery of any significant regularity will always lead to an increase in the probability assigned to the observations (i.e. a decrease in complexity).
- The probability of the data under the bigram model is greater than it is under the unigram model, and therefore we must prefer the bigram model.

# Complexity Sorter

- [Complexity Sorter](#)
- calculates "Phonological complexity" or the average *information content*
- the lowest average probability = the highest average complexity

- Goldsmith (2011)

| rank   | orthography | phonemes | <i>avg. plog</i> |
|--------|-------------|----------|------------------|
| 1      | a           | ə        | 3.11             |
| 2      | an          | ən       | 3.44             |
| 3      | to          | tə       | 3.47             |
| 4      | and         | ənd      | 3.80             |
| 5      | eh          | é        | 3.88             |
| 6      | the         | ðə       | 3.88             |
| 7      | can         | kən      | 3.90             |
| 8      | an          | æn       | 3.91             |
| 9      | Ann         | æn       | 3.91             |
| 10     | in          | ín       | 3.91             |
| 63,195 | bourgeois   | bǎržwá   | 7.21             |
| 63,196 | Ceausescu   | čǎčěskǔ  | 7.21             |
| 63,197 | Peugeot     | pyǔžó    | 7.22             |
| 63,198 | Giraud      | žǎyró    | 7.24             |
| 63,199 | Godoy       | gádoŷ    | 7.27             |
| 63,200 | geoid       | ǵíǔyd    | 7.40             |
| 63,201 | Cesare      | čězárě   | 7.40             |
| 63,202 | Thurgood    | θǎgǎd    | 7.47             |
| 63,203 | Chenoweth   | čénǔwěθ  | 7.49             |
| 63,204 | Qureshey    | kǎréšě   | 7.54             |

# PC and well-formedness

- The “bad” ends of the lists contain primarily borrowings into the language, compounds, and onomatopoeia, while the “good” ends of the lists contain words whose phonological patterns are the most central in the language.
- PC expresses quite well the phonologist's intuition regarding the phonological well-formedness of words (or how well any given individual word fits into the phonotactic patterns of the language).
- Nativization of borrowings decreases the average log probability.

# Mutual Information

- Mutual information (MI) is the ratio of the number of occurrences observed to the number of occurrences expected.

$$\begin{aligned} MI(x, y) &= \log_2 \frac{prob(xy)}{prob(x) \times prob(y)} \\ &= \log_2 prob(xy) - \log_2 prob(x) - \log_2 prob(y) \end{aligned}$$

- MI is a measure of the interdependence of two elements with respect to one another. Adjacent elements with a high value of MI are highly interdependent.

# Mutual Information (cont)

- Mutual information is the difference of log probability between bigram and unigram models.

$$MI(x_1, x_2) = \log_2 P(x_1 x_2) - \log_2 P(x_1) - \log_2 P(x_2)$$

$$= \log_2 \frac{P(x_1 x_2)}{P(x_1) P(x_2)}$$

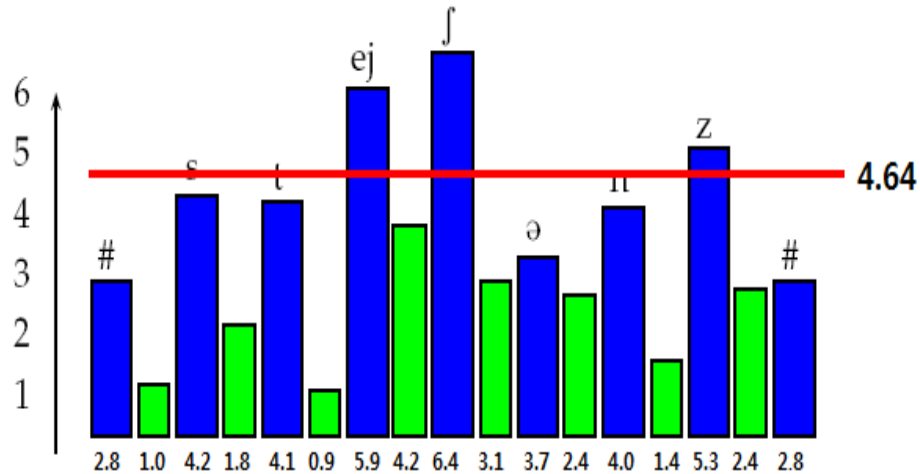
$$= \log_2 \frac{P(x_2 | x_1)}{P(x_2)}$$

$$= \log_2 P(x_2 | x_1) - \log_2 P(x_2)$$

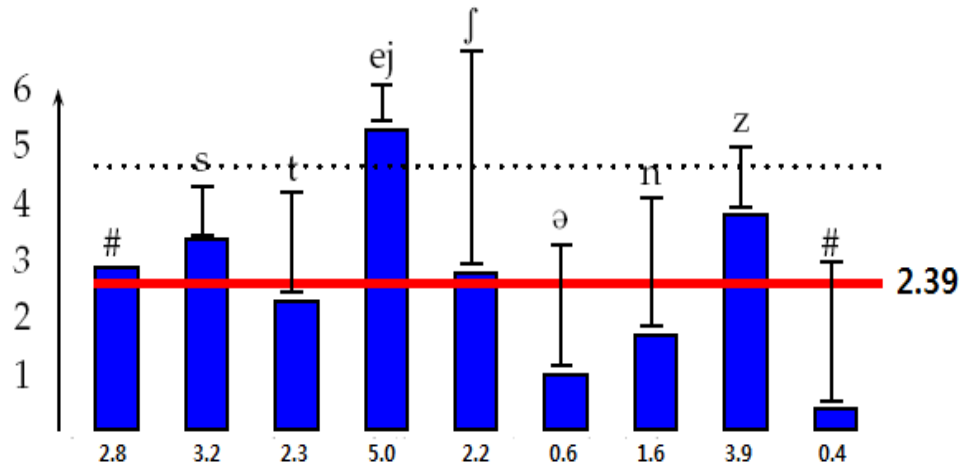
$$= (-\log_2 P(x_2)) - ((-\log_2 P(x_2 | x_1)))$$



# Unigram/bigram models & MI



Unigram model



Bigram model

# Bayes' rule: Identifying the language from which a word is drawn

- $$P(A|B) = \frac{\text{prob}(B|A)\text{prob}(A)}{\text{prob}(B)}$$

5.31

$$P(W \text{ comes from English} \mid W=\text{mitsubishi}) = \frac{P(W=\text{mitsubish} \mid W \text{ comes from English}) * P(W \text{ comes from English})}{P(W=\text{mitsubishi})}$$

3.36

$$P(W \text{ comes from Japan} \mid W=\text{mitsubishi}) = \frac{P(W=\text{mitsubish} \mid W \text{ comes from Japan}) * P(W \text{ comes from Japan})}{P(W=\text{mitsubishi})}$$

# Relating to morphophonology

- Instead of computing the complexity of the string #klab#, I can make a variable out of the third position (let us indicate this as #k?ab#; we can call that a *representation schema*), and then what I have is a function from all of the phonemes to the real numbers: for each phoneme P, I can replace ? in #k?ab# by P, and compute the complexity. We may then ask, which value of ? gives us the smallest value for complexity. In that way, we can compute the optimal log probability of a representation-schema.  
(p.20)

# Application to other phenomena: epenthetic vowel

- Hume & Bromberg (2005): Epenthetic vowel is lowest in 'context-sensitive' *plog*.

$$CS\text{-}plog(v) = plog(v) + MI(pC, v) + MI(v, fC)$$

- Hong (2011a)

| Vowel | Freq      | $plog(v)$ | $MI(pC, v)$ | $MI(v, fC)$ | $CS\text{-}plog(v)$ |
|-------|-----------|-----------|-------------|-------------|---------------------|
| ɪ     | 5,559,893 | 4.3423    | 14.7001     | -21.2083    | -2.1659             |
| e     | 2,210,627 | 5.6729    | 4.6581      | -4.1774     | 6.1536              |
| i     | 5,831,331 | 4.2735    | 4.9605      | 10.6195     | 19.8535             |
| o     | 3,931,431 | 4.8423    | 18.4288     | -2.0648     | 21.2063             |
| ə     | 5,979,700 | 4.2372    | 13.8785     | 4.1897      | 22.3054             |
| u     | 2,674,162 | 5.3982    | 16.7618     | 0.4044      | 22.5644             |
| a     | 9,077,880 | 3.6350    | 22.0551     | 3.6478      | 29.3378             |
| æ     | 1,761,357 | 6.0006    | 19.4430     | 15.1942     | 40.6377             |

# Application to other phenomena: feature distinctiveness

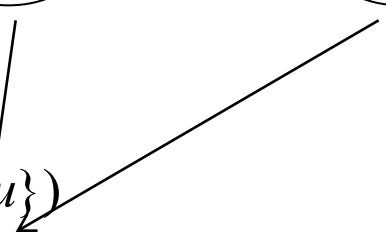
|   | [son] | [round] | [high] | [back] | [cons] | Prob |
|---|-------|---------|--------|--------|--------|------|
| i | 1     | 0       | 1      | 0      | 0      | 0.25 |
| e | 1     | 0       | 0      | 0      | 0      | 0.05 |
| a | 1     | 0       | 0      | 1      | 0      | 0.35 |
| o | 1     | 1       | 0      | 1      | 0      | 0.20 |
| u | 1     | 1       | 1      | 1      | 0      | 0.15 |

- The distinctiveness of a feature can be measured by its Entropic Contribution (Hume & Malihot 2011).
- EC of a vowel feature [F] is measured by the drop in entropy of the vowel system when [F] is eliminated.
- Eliminating [back] collapses the distinction between [e] and [a].

$$EC([\text{back}]) = \text{entropy}(\{i, e, a, o, u\}) - \text{entropy}(\{i, (e + a), o, u\})$$

# Application to other phenomena: feature distinctiveness(cont)

$$\text{entropy}(\{i, e, a, o, u\})$$

$$= 0.25 \times \log_2 0.25 + 0.05 \times \log_2 0.05 + 0.35 \times \log_2 0.35 + 0.2 \times \log_2 0.2 + 0.15 \times \log_2 0.15$$
$$= 2.12$$


$$\text{entropy}(\{i, (e + a), o, u\})$$

$$= 0.25 \times \log_2 0.25 + 0.4 \times \log_2 0.4 + 0.2 \times \log_2 0.2 + 0.15 \times \log_2 0.15$$
$$= 1.90$$

$$EC([\text{back}])$$

$$= \text{entropy}(\{i, e, a, o, u\}) - \text{entropy}(\{i, (e + a), o, u\})$$

$$= 2.12 - 1.90 = 0.22$$

# Application to other phenomena: feature distinctiveness (cont)

Korean vowel features (Hong 2011b)

|   | [high] | [low] | [back] | [round] | Prob   |
|---|--------|-------|--------|---------|--------|
| i | 1      | 0     | 0      | 0       | 0,1575 |
| e | 0      | 0     | 0      | 0       | 0,0597 |
| æ | 0      | 1     | 0      | 0       | 0,0476 |
| ɪ | 1      | 0     | 1      | 0       | 0,1502 |
| ə | 0      | 0     | 1      | 0       | 0,1615 |
| a | 0      | 1     | 1      | 0       | 0,2452 |
| u | 1      | 0     | 1      | 1       | 0,0722 |
| o | 0      | 0     | 1      | 1       | 0,1062 |

$$E\alpha(\text{high}) = \text{entropy}(\{i, e, \text{æ}, \text{ɪ}, \text{ə}, a, u, o\}) - \text{entropy}(\{(i+e), \text{æ}, (i+\text{ə}), a, (u+o)\}) = 0.6693$$

$$E\alpha(\text{low}) = \text{entropy}(\{i, e, \text{æ}, \text{ɪ}, \text{ə}, a, u, o\}) - \text{entropy}(\{i, (e+\text{æ}), \text{ɪ}, (\text{ə}+a), u, o\}) = 0.5004$$

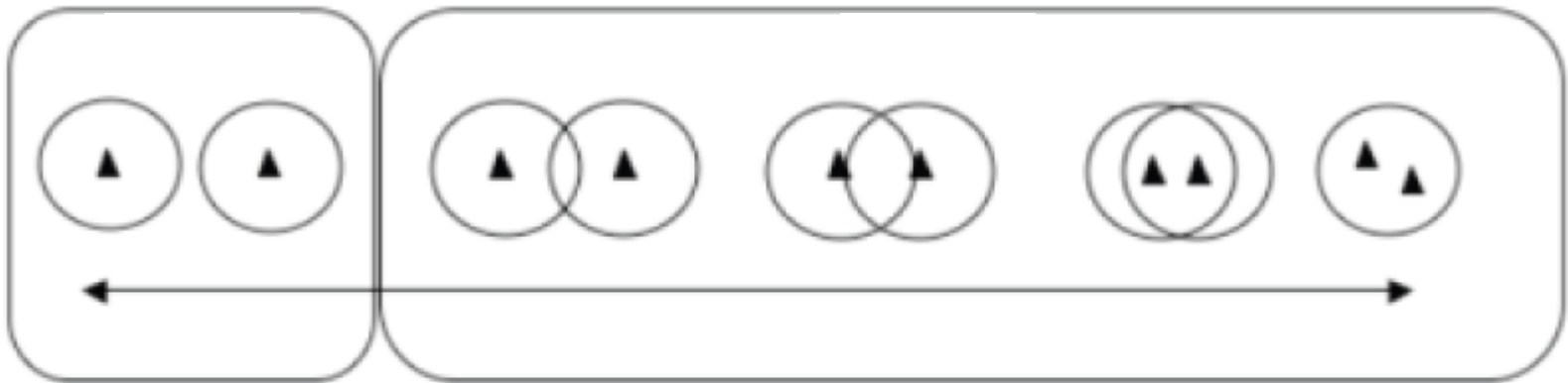
$$E\alpha(\text{back}) = \text{entropy}(\{i, e, \text{æ}, \text{ɪ}, \text{ə}, a, u, o\}) - \text{entropy}(\{(i+\text{ɪ}), (e+\text{ə}), (\text{æ}+a), u, o\}) = 0.6811$$

$$E\alpha(\text{round}) = \text{entropy}(\{i, e, \text{æ}, \text{ɪ}, \text{ə}, a, u, o\}) - \text{entropy}(\{i, e, \text{æ}, (\text{ɪ}+u), (\text{ə}+o), a\}) = 0.6850$$

[round] > [back] > [high] > [low]

# Application to other phenomena: sound distribution

- Hall (2009)





# Application to other phenomena: sound distribution (cont)

- Toy grammar (Hall 2009:137)

|     | <u># a</u>    | <u>a #</u> | <u>a a</u>               | <u>i i</u> |
|-----|---------------|------------|--------------------------|------------|
| [t] | ta, tara, tat | at, tat    | *                        | iti        |
| [d] | da, dara      | ad         | *                        | *          |
| [r] | *             | *          | ara, tara, dara,<br>sara | iri        |
| [s] | sa, sara      | as         | *                        | *          |

$$\text{entropy}([t],[d])_{[\#\_a]} = - \left( \frac{3}{5} \times \log_2 \frac{3}{5} + \frac{2}{5} \times \log_2 \frac{2}{5} \right) = 0.97$$

$$\text{entropy}([t],[d])_{[a\_ \#]} = - \left( \frac{2}{3} \times \log_2 \frac{2}{3} + \frac{1}{3} \times \log_2 \frac{1}{3} \right) = 0.92$$

$$\text{entropy}([t],[d])_{[i\_i]} = - (1 \times \log_2 1 + 0 \times \log_2 0) = 0$$

# Application to other phenomena: sound distribution (cont)

$$\text{entropy}([t],[d])$$

$$= \text{entropy}([t],[d])_{[\#\_a]} \times P([t],[d])_{[\#\_a]}$$

$$+ \text{entropy}([t],[d])_{[a\_ \#]} \times P([t],[d])_{[a\_ \#]}$$

$$+ \text{entropy}([t],[d])_{[i\_j]} \times P([t],[d])_{[i\_j]}$$

$$= 0.97 \times 5/9 + 0.91 \times 3/9 + 0 \times 1/9 = 0.85$$

Probability-weighted

$$\text{entropy}([d],[r]) = 0$$

$$\text{entropy}([t],[r]) = 0.18$$

$$\text{entropy}([d],[s]) = 1$$

$$\text{entropy}([d],[s]) > \text{entropy}([t],[d]) > \text{entropy}([t],[r]) > \text{entropy}([d],[r])$$

← unpredictable

→ predictable