Efficient Communication I
Information theory, evolution, and the mental lexicon

LSA Summer Institute 2011,
*Computational Psycholinguistics*, Lecture 3

Florian Jaeger, Human Language Processing Lab
http://www.hlp.rochester.edu/
Readings

• **Required:**
  – Pinker (2000) – 2pp
  – Piantadosi et al. (2011) – 3pp
  – Watch: [http://www.youtube.com/watch?v=z2Whj_nL-x8](http://www.youtube.com/watch?v=z2Whj_nL-x8)

• **Technical reading:**
  – **Basic:** Jurafsky and Martin 2009, Chapter 4, pp. 83-95, 114-116, 118-119.
  – **Advanced:** pp. 95-106, 116-118
Plan

• Introduction to information theory  [Shannon, 1948]
  – Shannon information
  – Entropy (uncertainty)
  – Noisy channel
  – Noisy Channel theorem

• Two application:
  – Language use and language evolution  
  – Language use and language change: Entropy and the mental lexicon  
    [Ferrer i Cancho, XXX; Manin, 2006; Piantadosi et al., 2011; Plotkin and Novak, 2000; Zipf, 1935, 1949]
Part 1

Biological and Cultural Evolution and the Functionalist Hypothesis
Biological vs. cultural evolution

Biological evolution

Cultural evolution/language change

Figure 5 Two aspects of language evolution. a, There is a biological evolution of universal grammar (UG) vis-a-vis genetic modifications that affect the architecture of the human brain and the class of languages it can learn. UG can change as a consequence of (1) random variation (neutral evolution), (2) as a by-product of selection for other cognitive function or (3) under selection for language acquisition and communication. At some point in the evolutionary history of humans, a UG arose that allowed languages with infinite expressibility. b, On a faster timescale, there is cultural evolution of language constrained by a constant UG. Languages change by (1) random variation, (2) by contact with other languages (red arrow), (3) by hitch-hiking on other cultural inventions, or (4) by selection for increased learnability and communication. Although many language changes in historical linguistics might be neutral, a global picture of language evolution must include mutation.

[taken from Nowak et al. 2002-Nature]
The functionalist hypothesis

• **Functional pressures on cultural evolution:** Grammatical properties may be observed more often across languages because they improve a language’s ‘utility’.
  [e.g. BatesMacWhinney82,89; Bybee01,02; ChristiansonChater08; Croft04; Givon91,92,01; Hawkins94,01,02,04,07; Hocket60; Langacker91; Slobin73; Zipf49]

• This is an intriguing possibility, as it promises to reduce the number of cognitively arbitrary (=linguistic specific) properties of language that we need to explain.
Challenges

1. ‘Transmission problem’: Where do hypothesized pressures operate? That is, how would such pressures come to shape language over time?

   - Biases on language acquisition, changing the structures acquired by the next generation [e.g. Fedzechkina, Jaeger, Newport, 2011, in prep; Tily, Frank, Jaeger, 2011]

   - Biases operating throughout adult life that change the output provided to the next generation [cf. Lecture 7]
Challenges

1. ‘Transmission problem’: Where do hypothesized pressures operate? That is, how would such pressures come to shape language over time?

2. What is ‘utility’? What is good?
   - **Learnability** [cf. Deacon, 1998; Slobin, ]
   - **Ease of processing, e.g.: Minimization of memory cost**
     [cf. GildeaTemperley08; Hawkins94,01,02,04,07,09; Levy05]  
   - **Trade-off between production and comprehension effort**
   - **Efficient and robust communication**
     [cf. Aylett and Turk, 2004; FerreriCancho05,07,10; GenzelCharniak02,03; Jaeger06,10; LevyJaeger07; PiantadosiTilyGibson11; QianJaeger09,10,submitted]
On language ‘utility’: processing complexity and communicative efficiency
T. Florian Jaeger\(^1\)\(^*\) and Harry Tily\(^2\)


The Cross-linguistic Study of Sentence Production
T. Florian Jaeger\(^1\)\(^*\) and Elisabeth J. Norcliffe\(^2\)
\(^1\)Brain and Cognitive Sciences, Computer Science, University of Rochester, and
\(^2\)Linguistics, Stanford University

Biological evolution

• Utility has also been hypothesized to affect biological evolution of language:
  – Why do all human languages share certain properties (e.g. that they have structure above the sound level)?
    [e.g. Gasser, 2004; Hurford, 1989; Hurford et al., 1998; Plotkin and Nowak, 2000; Nowak and Krakauer, 1999; Nowak et al., 1999, 2000, 2002]
Part 2

Language Fitness and Evolution
Historical background

- Originating in work on evolutionary game theory, research in the late 80s began to define the 'fitness' of a language for communication.

- Originally, this work makes simple (and perhaps somewhat ad-hoc) assumptions, but in the early 2000s, this work is combined with information theoretic consideration about communication through a noisy channel.
Defining the ‘fitness’ of a language

- We’ll start with a model of simple language without much structure to understand its limitations and then move to models of more structured languages.

1. Languages that **map sounds to meaning**

2. **Adding noise to the equation:** Languages that map sounds to meaning with a certain error probability (sound confusability)

3. **Adding structure to the equation:** Languages that map words and sentence to meaning with a certain error probability (sound confusability)
1. From sound to meaning

[Hurford, 1989; Hurford, Studdert-Kennedy, and Knight, 1998; Nowak, Plotkin, and Krakauer, 1999]

- Language as a mapping between meaning (‘objects’) and sounds (‘signal’)
  - **Meaning:** $n$ objects $\leftrightarrow$ *could be semantic or social meaning, etc.*
  - **Sound:** $m$ signals

- **Production matrix** (‘active matrix’), $P$: matrix with probabilities $p_{ij}$ that meaning $n_i$ is mapped to sound $m_j$

- **Comprehension matrix** (‘passive matrix’), $Q$: matrix with probabilities $q_{ij}$ that sound $m_j$ is mapped to meaning $n_i$

- Together $P$ and $Q$ define a language, $L(P, Q)$. 
Example production matrix

\[
P = \begin{bmatrix}
0.3 & 0.2 & 0.5 \\
0 & 0.7 & 0.3 \\
0 & 0 & 0 \\
1 & 0 & 0
\end{bmatrix}
\]

'ambiguity/homonymy'

'sparseness'
Aside: Is sparseness common?

A sentence is like a conventionalized line drawing. You have to abstract away from everything you know or can picture about a situation, and present a schematic version which conveys the essentials. In terms of grammatical marking, there is not enough time in the speech situation for any language to allow for the formal encoding of everything which could possibly be significant to the message. Probably there is not enough interest either. Language *evokes* ideas; it does not represent them. Linguistic expression is thus *not* a straightforward map of consciousness or thought. It is a highly selective and conventionally schematic map.

[taken from Slobin, 1979 as quoted in Bates and MacWhinney, 1982]
Example comprehension matrix

\[ P = \begin{bmatrix}
  'a' & 'u' & 'o' \\
  0.3 & 0.2 & 0.5 \\
  0 & 0.7 & 0.3 \\
  0 & 0 & 0 \\
  1 & 0 & 0
\end{bmatrix} \]

\[ Q = \begin{bmatrix}
  0 & 0 & 0 & 1 \\
  0 & 1 & 0 & 0 \\
  1 & 0 & 0 & 0
\end{bmatrix} \]
Communicative Payoff (‘fitness’)  
[Nowak, Plotkin, and Krakauer, 1999]

- (Ignoring frequency of objects, ignoring confusability of forms)
- Payoff for communication between users of two languages $L_1$ and $L_2$:  
  \[
  F(L_1, L_2) = F(L_2, L_1) = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{m} (p_{ij}^{(1)} q_{ji}^{(2)} + p_{ij}^{(2)} q_{ji}^{(1)})
  \]
- Payoff for communication between users of same language:  
  \[
  F(L_1, L_1) = \sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij}^{(1)} q_{ji}^{(1)}
  \]
\[ \begin{align*} Q &= \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \\ P &= \begin{bmatrix} .3 & .2 & .5 \\ 0 & .7 & .3 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} \end{align*} \]

\[ Q \cdot P = \begin{bmatrix} .5 \\ 1 \end{bmatrix} = .3 \cdot 0 + .2 \cdot 0 + .5 \cdot 1 
\]

\[ F(L,L) = 2.2 \]

- **Q1**: How could you improve the first row of P to increase F(L,L)?
- **Q2**: What’s the P that maximizes F(L,L)?
- **Q3**: Could we further improve Q? What is \( F_{\text{max}} \)?
What makes a language maximally fit?

• If \( m \) (number of sounds) = \( n \) (number of objects): The best \( P \) has exactly one 1 per row and column and zeros elsewhere. \( Q = P^T \)

• If \( n > m \): \( P \) should have at least one 1 in every column. I.e. each sound is used and for each sound there is at least one meaning that is always mapped to it (\( \rightarrow \) at least \( m \) meanings are mapped, though not necessarily unambiguously).

• If \( m > n \): \( P \) should have at least one 1 in every row. Etc.

\[ F_{\text{max}} = \min(n, m) \]
(Batch) Learning

[Nowak, Plotkin, and Krakauer, 1999]

- (Assume a learner with perfect and unlimited memory)
- An association matrix $A$ stores all observed mappings from meaning to form:
  
  \[
  A = \begin{pmatrix}
  33 & 0 & 0 \\
  0 & 11 & 25 \\
  0 & 0 & 0 \\
  1 & 407 & 0 
  \end{pmatrix}
  \]

  - From which $P$ (and $Q$) can be estimated:

  \[
  p_{ij} = \frac{a_{ij}}{\left( \sum_{l=1}^{m} a_{il} \right)}
  \]

- Language evolution:

  \[
  P_0 \to A_1 \to P_1 \to A_2 \to P_2 \to \ldots
  \]
Consequences

• Under various learning assumptions, this model converges against languages
  – with no synonymy
  – Homonymy is evolutionary stable

→ Compatible with what’s observed cross-linguistically

... but ...
Limits of this model

- No noise in transmission (no error)
- Objects (meanings) are assumed to occur equally often OR we assess fitness not in terms of the number of successful communication instances but in terms of successful *types* of communication.

[the first problem is recognized by e.g. Nowak, Plotkin, and Krakauer, 1999 and addressed in more detail in Plotkin and Nowak, 2000]
2. The noisy nature of the signal

[Nowak, Plotkin, and Krakauer, 1999; Nowak and Krakauer, 1999]

• Payoff for communication between users of two languages $L_1$ and $L_2$ without considering that sounds can be confused:

$$F(L_1, L_2) = F(L_2, L_1) = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{m} \left( p_{ij}^{(1)} q_{ji}^{(2)} + p_{ij}^{(2)} q_{ji}^{(1)} \right)$$

• Payoff for communication between users of two languages $L_1$ and $L_2$ considering that forms can be confused:

$$F(L_1, L_2) = F(L_2, L_1) = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{m} \left( u_{jk} q_{ki}^{(2)} + p_{ij}^{(2)} \sum_{k=1}^{m} u_{jk} q_{ki}^{(1)} \right)$$
Probability of correct form recognition

- If the confusability of two forms is determined by their similarity,

\[ u_{ij} = \frac{s_{ij}}{\sum_{k=1}^{m} s_{ik}} \]

and we assume that self-similarity is 1 and all sounds are equally similar to all other sounds (for all \( i \neq j \), \( s_{ij} = \epsilon \)), ...
• Let $\varepsilon = .05$, then the sound-to-sound confusion matrix is:

$\begin{bmatrix}
.91 & .045 & .045 \\
0.045 & .91 & .045 \\
0.045 & 0.045 & .91
\end{bmatrix}$

$P = \begin{bmatrix}
.3 & .2 & .5 \\
0 & .7 & .3 \\
0 & 0 & 0 \\
1 & 0 & 0
\end{bmatrix}$

$Q = \begin{bmatrix}
0 & 0 & 0 & 1 \\
0 & 1 & 0 & 0 \\
1 & 0 & 0 & 0
\end{bmatrix}$

• Q: Has $F(L, L)$ gone down or up?
Probability of correct form recognition

If the confusability of two forms is determined by their similarity,

$$u_{ij} = s_{ij} / \sum_{k=1}^{m} s_{ik}$$

and we assume that self-similarity is 1 and all sounds are equally similar to all other sounds (for all $i \neq j$, $s_{ij} = \varepsilon$),

then the probability of correct form recognition $u_{ii}$ for a language with $m$ forms is:

$$u_{ii} = \frac{1}{1 + (m - 1)\varepsilon}$$

Self-similarity

Joint confusability due to all other $m-1$ sounds
Maximum fitness

• In this case, the maximum fitness of a language for two individuals speaking the same language with \( m \) sounds is \( mu_{ii} \)

\[
\sum_{i=1}^{m} u_{ii} = m \frac{1}{1 + (m - 1)\varepsilon}
\]

• As the number of sounds increases, the fitness \( F \) converges against (cf. without noise: \( F_{\text{max}} = \min(m,n) \)):

\[
\lim_{m \to \infty} \frac{m}{1 + (m - 1)\varepsilon} = \frac{1}{\varepsilon}
\]

• E.g. for our example, of \( \varepsilon = .05 \), \( F_{\text{max}} \) can never exceed 20

[28]
Consequences

• *If a language has no structure beyond sounds*, the noisiness of the signal *imposes a hard limit on* the maximum fitness that can be achieved.
  – *If sounds are distinguished along a pre-compact metric space and similarity decreases monotonically with increasing distances in this space*, the *maximum fitness of a language is bounded by the properties that determine sound similarity*, but not by the number of sound.  [Nowak, Krakauer, and Dress, 1999]

*Intuition*: Often adding more sounds creates more *in principle* distinguishable forms, but their *actual* distinguishability might decrease because sounds become to confusible.
How confusable are sounds really?

- Highest recognition rates for isolated CVC syllables reported in Woods et al (2010) are 90.6%.

- Confusion rates are, of course, *not* phoneme independent. Neither are they position independent.
E.g. initial consonant confusion

|     | b | d | g | r | l | n | m | v | ð | z | ð | f | f | s | θ | f | p | t | k | h |
| b   | 1001 | 12 | 24 | 17 | 50 | 2 | 32 | 266 | 16 | 4 | 4 | 2 | 0 | 5 | 15 | 69 | 58 | 7 | 16 | 128 |
| d   | 32 | 1059 | 110 | 28 | 73 | 26 | 28 | 74 | 36 | 27 | 32 | 6 | 4 | 13 | 18 | 15 | 18 | 32 | 30 | 67 |
| g   | 36 | 76 | 1097 | 26 | 62 | 10 | 26 | 78 | 20 | 16 | 22 | 2 | 3 | 5 | 11 | 17 | 41 | 6 | 33 | 141 |
| r   | 39 | 35 | 85 | 924 | 206 | 27 | 62 | 114 | 17 | 29 | 20 | 16 | 2 | 12 | 4 | 15 | 27 | 12 | 20 | 62 |
| l   | 28 | 9 | 11 | 50 | 1334 | 19 | 48 | 150 | 28 | 2 | 1 | 1 | 3 | 2 | 0 | 6 | 10 | 1 | 1 | 24 |
| n   | 17 | 28 | 33 | 45 | 303 | 917 | 170 | 77 | 18 | 11 | 12 | 3 | 0 | 4 | 7 | 5 | 8 | 4 | 10 | 56 |
| m   | 39 | 5 | 18 | 43 | 203 | 106 | 1073 | 127 | 9 | 6 | 3 | 1 | 1 | 2 | 4 | 7 | 13 | 1 | 10 | 57 |
| y   | 57 | 10 | 12 | 37 | 61 | 6 | 15 | 1360 | 102 | 11 | 0 | 1 | 0 | 0 | 11 | 17 | 8 | 2 | 3 | 15 |
| ð   | 17 | 36 | 6 | 193 | 2 | 1 | 499 | 861 | 41 | 1 | 1 | 1 | 0 | 54 | 8 | 0 | 2 | 0 | 0 | 3 |
| z   | 14 | 51 | 26 | 12 | 38 | 10 | 20 | 54 | 30 | 1155 | 66 | 21 | 7 | 53 | 27 | 27 | 16 | 32 | 25 | 44 |
| ðδ | 12 | 85 | 78 | 24 | 53 | 9 | 23 | 23 | 9 | 45 | 1056 | 138 | 15 | 7 | 9 | 12 | 12 | 24 | 49 | 45 |
| f   | 3 | 7 | 13 | 2 | 7 | 2 | 1 | 4 | 2 | 15 | 253 | 1222 | 67 | 6 | 3 | 5 | 15 | 42 | 43 | 16 |
| f   | 5 | 15 | 9 | 3 | 12 | 2 | 4 | 3 | 1 | 23 | 278 | 451 | 813 | 23 | 5 | 4 | 10 | 12 | 17 | 38 |
| s   | 11 | 24 | 15 | 10 | 24 | 6 | 14 | 39 | 10 | 424 | 47 | 52 | 21 | 734 | 42 | 64 | 25 | 58 | 39 | 69 |
| θ   | 1 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 | 1224 | 455 | 3 | 6 | 0 | 2 |
| f   | 25 | 2 | 7 | 5 | 6 | 1 | 7 | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 15 | 1264 | 16 | 58 | 328 |
| p   | 9 | 3 | 7 | 5 | 6 | 1 | 7 | 1 | 0 | 0 | 0 | 1 | 15 | 1264 | 16 | 58 | 328 |
| t   | 22 | 47 | 27 | 10 | 42 | 8 | 24 | 34 | 10 | 39 | 36 | 41 | 7 | 21 | 23 | 25 | 92 | 880 | 165 | 175 |
| k   | 12 | 9 | 30 | 3 | 7 | 4 | 7 | 12 | 2 | 9 | 6 | 1 | 1 | 4 | 8 | 26 | 116 | 41 | 1253 | 177 |
| h   | 11 | 0 | 4 | 0 | 7 | 0 | 4 | 13 | 2 | 0 | 2 | 2 | 0 | 1 | 3 | 20 | 176 | 7 | 51 | 1425 |

[Table V, taken from Woods et al. 2000]
3. Adding structure
[Nowak and Krakauer, 1999; Plotkin and Nowak, 2000]

- However, languages which use **structure** can overcome this limit.

- For example, for a language with words (all of the maximum length l), the maximum fitness of a language increases exponentially with l (F_{max} = \epsilon^{-l})
  - For our example, if l = 4 and \epsilon = .05, F_{max} = 160000 instead of 20!)

- **Information theory** provides a proof for this.
(Re)defining languages

• Let a language be described by
  – a lexicon: a *subset* of all possible *l*-long phoneme sequences:

\[ \mathcal{L} \subseteq \Phi^{l} \]

• Let \( n \) be the size of the lexicon

  – an active (production) matrix \( P \)
  – a passive (comprehension) matrix \( Q \)
  – a phoneme error-matrix (confusion matrix) \( V \).
Word confusability

• So for words $\alpha$ and $\beta$, we can calculate their confusability under the Markov assumption:

$$U_{\alpha\beta} = \prod_{k=1}^{l} V_{\alpha^{(k)}\beta^{(k)}}$$

• (i.e. word confusability is the production of all phoneme confusabilities, assuming that phonological context does not matter → for example, the existence of e.g. OCP effects [cf. Graff and Jaeger, submitted, in prep] could limit the conclusions about maximal fitness presented below)
Fitness

- Language fitness is then defined absolutely parallel to the simpler cases discussed above:

\[
F(L, L') = \sum_{i=1}^{n} \sum_{\alpha \in \Phi^l} \sum_{\beta \in \Phi^l} P_{w_i\alpha} U_{\alpha\beta} Q_{\beta w_i}
\]

\[
= \sum_{i=1}^{n} \sum_{\alpha \in \Phi^l} P_{w_i\alpha} \sum_{\beta \in \Phi^l} Q_{\beta w_i} \prod_{k=1}^{l} V^{(k)}_{\alpha(k)} \beta^{(k)}
\]
Information theory

[Shannon, 1948]
Communication through a noisy channel

Relating Shannon’s noisy channel model to Plotkin and Nowak (2000): P and Q must be deterministic

Effects of environmental noise & biological noise aggregated into on noise model

[Figure 1 from Shannon 1949]
Shannon’s noisy channel theorem
(for mere mortals like us)
[Shannon, 1948; Wolfowitz, 1961]

• For any noisy digital channel with capacity $C > 0$ and any rate of information transmission $0 < R < C$, there is a finite sequence of $n$ codes (a finite language with $1 \leq n < \infty$ words) so that
  – The number of possible words increases exponentially with the maximum length $l < n$
  – The error probability decreases exponentially with $n$
  $\Rightarrow$ If $R < C$, it is possible to communicate at an arbitrarily low error rate

• The converse holds, too: for any $R > C$, the error probability will converge against 1 the larger the codeword vocabulary.
Maximum fitness in information theoretic terms

• Let’s think of the set of code words as a language. Again assuming that we talk about each object equally often (i.e. use each codeword $w_i$ equally often), we get the expected error probability:

$$ e(L) = \frac{1}{n} \sum_{i=1}^{n} p(\text{error} | w_i \text{ sent}) $$

• We can define the fitness of a language in information theoretic terms, depending on the number of code words and the expected error probability:

$$ \tilde{F}(L) = n[1 - e(L)] $$
• Plotkin and Nowak (2000) show that this is the same as their original definition of fitness, which means that the information theoretic proof about error probability translates into a proof that a language’s fitness improves exponentially with the maximum length of words.

\[
\hat{F}(L) = n[1 - e(L)] \\
= n - \sum_{i=1}^{n} p(\text{error} | w_i, \text{sent}) \\
= n - \sum_{i=1}^{n} 1 - p(\text{no error} | w_i, \text{sent}) \\
= \sum_{i=1}^{n} p(\text{no error} | w_i, \text{sent}) \\
= \ldots
\]
Summary

• Information theory can be used to derive – based on general assumptions about communication through a noisy channel— to motivate why languages have structure beyond phones (e.g. words, syntax, etc.): because this makes it possible to overcome the error limit (1/ ε).
Limits

- Limits of the model presented in Plotkin and Nowak (2000):
  - Objects (meanings) are assumed to occur equally often.
  - Exponentially increasing fitness with linearly increasing maximum word length only is shown to hold under assumption that word errors are the consequences of independent confusion probabilities of the sounds that form the word
  - **Word boundaries are assumed to be known with certainty** (cf. digital channel)!
• Next, let’s see how the same information theoretic ideas can be applied to study possible effects of a pressure for efficient information transfer on cultural language evolution (language change).
Part 3

Zipf
Why is the mental lexicon structured the way it is?

- Zipf first worked on sound change and noticed that frequency of use (in at least one part of the language community) often seemed to be correlated with shorter form.

- This observation was a driving factor for the development of the principle of least effort.
The Principle of Least Effort

[Zipf 1949:1]

In simple terms, the Principle of Least Efforts means, for example, that a person in solving his immediate problems will view these against the background of his probable future problems as estimated by himself. [...] The person will strive to minimize the probable average rate of his work-expenditure (over time).

[emphasis in original; Zipf attributes the roots of similar ideas to Maupertuis in the 18th century]

- NB: minimization of probable average rate of work
Two opposing forces

- Prior to considering that language use might (among other things) serve to communicate:
  - **Speaker economy** (force of unification): map all meanings onto the same (short) word /Ə/
  - **Hearer economy** (force of diversification): map each meaning onto a different word

- These two forces together are assumed to affect (through diachronic change) the structure of the mental lexicon.

[NB: Zipf did not consider confusability of words]
The frequency ~ length correlation

<table>
<thead>
<tr>
<th>Number of Syllables in Word</th>
<th>Number of Occurrences (Including Repetitions) of Words</th>
<th>Percentage of the Whole</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5,426,326</td>
<td>49.76%</td>
</tr>
<tr>
<td>2</td>
<td>3,156,448</td>
<td>28.94%</td>
</tr>
<tr>
<td>3</td>
<td>1,410,494</td>
<td>12.93%</td>
</tr>
<tr>
<td>4</td>
<td>646,971</td>
<td>5.93%</td>
</tr>
<tr>
<td>5</td>
<td>187,738</td>
<td>1.72%</td>
</tr>
<tr>
<td>6</td>
<td>54,436</td>
<td>.50%</td>
</tr>
<tr>
<td>7</td>
<td>16,993</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>5,038</td>
<td>.22%</td>
</tr>
<tr>
<td>9</td>
<td>1,225</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>461</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>59</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

[taken from Zipf 1935:23; based on Kaeding 1928]
## American Newspaper English

(According to R. C. Eldridge)

<table>
<thead>
<tr>
<th>Number of Occurrences</th>
<th>Number of Words</th>
<th>Average Number of Phonemes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2976</td>
<td>(6.656)</td>
</tr>
<tr>
<td>2</td>
<td>1079</td>
<td>(6.151)</td>
</tr>
<tr>
<td>3</td>
<td>516</td>
<td>(6.015)</td>
</tr>
<tr>
<td>4</td>
<td>294</td>
<td>(6.081)</td>
</tr>
<tr>
<td>5</td>
<td>212</td>
<td>(5.589)</td>
</tr>
<tr>
<td>6</td>
<td>151</td>
<td>(5.768)</td>
</tr>
<tr>
<td>7</td>
<td>105</td>
<td>(5.333)</td>
</tr>
<tr>
<td>8</td>
<td>84</td>
<td>(5.654)</td>
</tr>
<tr>
<td>9</td>
<td>86</td>
<td>(5.174)</td>
</tr>
<tr>
<td>10</td>
<td>45</td>
<td>(5.377)</td>
</tr>
<tr>
<td>11</td>
<td>40</td>
<td>(4.825)</td>
</tr>
<tr>
<td>12</td>
<td>37</td>
<td>(5.459)</td>
</tr>
<tr>
<td>13</td>
<td>25</td>
<td>(5.560)</td>
</tr>
<tr>
<td>14</td>
<td>28</td>
<td>(5.000)</td>
</tr>
<tr>
<td>15</td>
<td>26</td>
<td>(4.807)</td>
</tr>
<tr>
<td>16</td>
<td>17</td>
<td>(5.058)</td>
</tr>
<tr>
<td>17</td>
<td>18</td>
<td>(4.166)</td>
</tr>
<tr>
<td>18</td>
<td>10</td>
<td>(6.100)</td>
</tr>
<tr>
<td>19</td>
<td>15</td>
<td>(4.733)</td>
</tr>
<tr>
<td>20</td>
<td>16</td>
<td>(4.687)</td>
</tr>
<tr>
<td>21</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>10</td>
<td>(3.455)</td>
</tr>
<tr>
<td>27</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of Occurrences</th>
<th>Number of Words</th>
<th>Average Number of Phonemes</th>
</tr>
</thead>
<tbody>
<tr>
<td>31</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>1</td>
<td>(3.903)</td>
</tr>
<tr>
<td>42</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>43</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>44</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>45</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>46</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>47</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>48</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>49</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>51</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>52</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>53</td>
<td>1</td>
<td>(3.333)</td>
</tr>
<tr>
<td>54</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>55</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>56</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>58</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>1</td>
<td>(2.666)</td>
</tr>
<tr>
<td>61-4290</td>
<td>71</td>
<td></td>
</tr>
</tbody>
</table>

[taken from Zipf 1935:28]
Part 4

Taking Zipf to the age of information theory
Going back to the Noisy Channel Theorem

• Recall that the channel capacity defines the maximal rate of information per time step/sent signal that allows communication at an arbitrarily low error rate.

• An optimal code then transmits information at an average rate close to, but not exceeding the channel capacity.
  – Constant Entropy Rate [Genzel and Charniak, 2002]
  – Smooth Signal Redundancy [Aylett and Turk, 2004]
  – Uniform information density [Jaeger, 2006; Levy and Jaeger, 2007]
Information

• Shannon information, for example, of a word
  \[ I(w) = \log \left[ \frac{1}{p(w)} \right] = -\log p(w) \]

  – Log often taken to base 2 → units of information is “bits”

• Intuitive properties:
  – 0 bits new information, if something is perfectly predictable (cf. surprisal)
  – More new information, the less predictable something is
True or False?

• **Q1:** \( I(w_i \mid w_{i-1}) = I(w_i, w_{i-1}) - I(w_{i-1}) \)

• **Q2:** \( I(w_{i-1} w_i \mid w_1 \ldots w_{i-2}) = I(w_i \mid w_1 \ldots w_{i-1}) + I(w_{i-1} \mid w_1 \ldots w_{i-2}) \)
Entropy

• **Entropy is the expected information**, for example, for words:

\[ H(w) = \sum p(w) \log p(w) \]

  - So, for example, we can think of entropy as measuring the uncertainty about which word we will see before we see it; Shannon information is then the actual information we received once the word is seen.

• **Mutual information:**

\[ I(X; Y) = \sum_{y \in Y} \sum_{x \in X} p(x, y) \log \left( \frac{p(x, y)}{p_1(x) p_2(y)} \right), \]

[from wikipedia.org]
Let’s, for now, consider the number of phones in a word as the amount of signal provided.

So, we want words that on average carry more information to have more phones.

The average information a word $w$ carries in its different contexts $C$ is:

$$-\sum_c P(C = c \mid W = w) \log P(W = w \mid C = c).$$

which can be estimated from a corpus:

$$-\frac{1}{N} \sum_{i=1}^{N} \log P(W = w \mid C = c_i)$$
How to estimate the average information of a word?

- Piantadosi et al (2011) use ngram models with smoothing (n = 2,3,4) based on the Google ngram corpus available for 11 languages.
  - The large size of the Google ngram corpus is important to obtain reliable estimates of the average information a word carries in context.

- These estimates are then regressed against word length.
How is information estimated from a corpus?

- As Shannon information is defined with reference to probability, we need to *estimate the probability of words* in order to estimate their information.

- So called ngram models provide a simple way that is frequently employed to derive probability estimates from a collection of speech or writing.

- So, let’s assume we have such a corpus.
A very small corpus:

Over the last two decades, cognitive science has undergone a paradigm shift towards probabilistic models of the brain and cognition. Many aspects of human cognition are now understood in terms of rational use of available information in the light of uncertainty (e.g. models in memory, categorization, generalization and concept learning, visual inference, motor planning). Building on a long traditional of computational models for language, such rational models have also been proposed for language processing and acquisition. This class provides an overview to the newly emerging field of computational psycholinguistics, which combines insights and methods from linguistic theory, natural language processing, machine learning, psycholinguistics, and cognitive science into the study of how we understand and produce language. There has been a surge in work in this area, which is attracting scholars from many disciplines. The goal of this class is to provide students with enough background to start their own research in computational psycholinguistics.

• Now let’s extract the bigrams from this text. That’s really just the list of all two-word sequences in the above text, followed by how often they occur
From bigrams to Shannon information

this class 2
over the 1
the last 1
last two 1
this area 1
...

• (for a neat tool that let’s you estimate bigrams based on the Brown corpus, see But we can be lazy and use a tool, e.g. http://word.snu.ac.kr/ngram/)

• E.g. the word this is followed 2 out of 3 times by the word class. Hence, our best (maximum likelihood) estimate of $p(\text{class} \mid \text{this}) = -\log \frac{2}{3}$
Getting the *average* information of a word in context

- In our sample text, the word *class* only occurred twice, each time preceded by *this*. Recall that the average information of a word in context based on a corpus is calculated as:

\[-\frac{1}{N} \sum_{i=1}^{N} \log P(W = w | C = c_i)\]

where C is the context (here simply the preceding word) and N is the number of different context (here 2 since *class* occurs twice in the corpus).

- Hence, the average information of *class* given the preceding word in our sample is: 

\[-1/2(\log 2/3 + \log 2/3) = -1/2 \times \log 4/9 = -0.5849625\] bits of average information that *class* carries in its context.
Result for English

Fig. 2. Relationship between frequency (negative log unigram probability) and length, and information content and length. Error bars represent SEs and each bin represents 2% of the lexicon.

[Figure 2 from Piantadosi et al., 2011]
Bigram results for all 11 languages

[p]art of Figure 1 from Piantadosi et al., 2011
• Comparing the results for different $n$ in the ngram model
One more example: Russian

Manin uses human judgments from a cloze-like task to estimate information (or unpredictability)
Some open questions (1)

- How does usage (here: the average information content of a word in context) create these correlations over time?

- Under what condition does an innovation (e.g. shortening of a frequent word) survive? What’s the role of social networks?

  [e.g. Zipf hypothesized that it was sufficient for words to be frequent in at least one sub-community (e.g. automobile → auto, voltage → volt)]
Some open questions (2)

- What are contributions of phonetic reduction vs. phonological simplification? [cf. Johnson’s 2004 work on massive reduction; see also Zipf, 1949 and Kuperman et al., 2008-JASA arguing against Zipf’s interpretation]

Figure 2: The four syllable word *apparently* is realized [pʰeɾɪ] in this instance of the phrase *apparently not*. [Figure 2 from Johnson, 2004]
Take home points (1)

- Information theory provides several important results for language:
  - The fitness of the best language, as assessed in terms of its error probability, increases exponentially with word length. I.e., words overcome the error-limit found in earlier work for languages without words.
  - There is an upper bound of the information per signal that can be transferred (lower bound to error-probability). This bound is defined by the noisy channel’s capacity $C(V)$, depending only on the confusion matrix $V$. 

[67]
Take home points (2)

- Given that language has the properties expected if humans and languages evolved to transfer information efficiently, and given that these properties are unlikely to be due to chance (see e.g. Ferrer i Cancho and Sole, 2003 on Zipf’s law), this provides tentative support for the idea that functional pressures shape language over time.

- The noisy channel model provides a first approximation of at least one of these functional pressures.
Take home points (3) - Caveats

[from http://roosterteeth.com/comics/strip.php; thx to Nurit Melnik]
Take home points (3) - Caveats

• The models we have seen today make many simplifying assumptions that are problematic

• The idea that a uniform distribution of information is optimal is based on the noisy channel theorem, but
  – What does it mean to assume a digital channel?
  – What does it mean that we assumed that encoding and decoding is noise-free?
  – Even if an optimal encoding system exists, encoding and decoding might be too computationally expensive.
Readings

• **Required:** Jaeger, 2010 (20+ pp); Aylett & Turk 2006 (11pp)

• **Suggested:** Genzel & Charniak, 2002; Levy & Jaeger, 2007; Moscoso del Prado Martin, submitted; Qian and Jaeger, 2010, 2011

• **Technical reading (optional):** Shannon (1948)