Informatics 1 Cognitive Science

Lecture 11: Communication and Efficiency

Frank Keller

8 February 2024

School of Informatics
University of Edinburgh
keller@inf.ed.ac.uk

Slide credits: Frank Mollica, Chris Lucas, Mirella Lapata

Overview

The Mathematical Theory of Communication

Information Theory Fundamentals

Principle of Least Effort

Examples: Reading and Speech

Recap

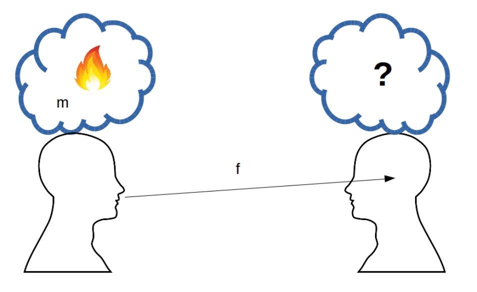
So far, we have focused on aspects of language acquisition:

- Transitional probabilities are useful for modeling word segmentation
- Bayes rule models how different probabilities can be combined: prior vs. likelihood
- We've applied this to model the acquisition of number concepts:
  - simplicity prior: prefer hypotheses with fewer primitives
  - likelihood: prefer the smallest hypothesis that's compatible with the data: size principle

Today, we will look at language as a communication system. Notions from probability and information theory will again be useful.

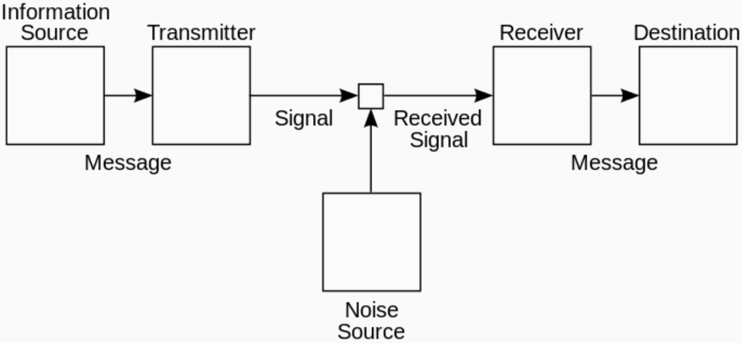
The Mathematical Theory of Communication

The Mathematical Theory of Communication (Shannon, 1948)



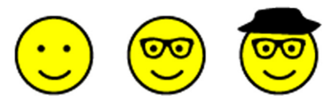
- **Communication:** the transfer of information from a source to a destination.
- Communication is often not *the goal* of language use but a necessary subgoal.
- Shannon's theory (now called Information Theory) deals with the efficient encoding and decoding of messages.
- Widely used in cognitive science, neuroscience, computer science.

The Mathematical Theory of Communication (Shannon, 1948)

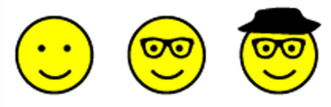


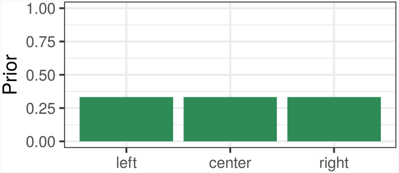
- A **source** has a probability distribution over meanings $P(m)$.
- A **transmitter** (or encoder) is a function that maps meanings to forms with distribution $P(f|m)$.
- A **receiver** (or decoder) is a function that maps forms to meanings with $P(m|f)$.

Reference Resolution: A demo



Reference Resolution: A demo

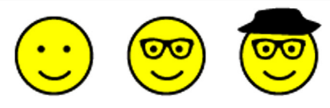


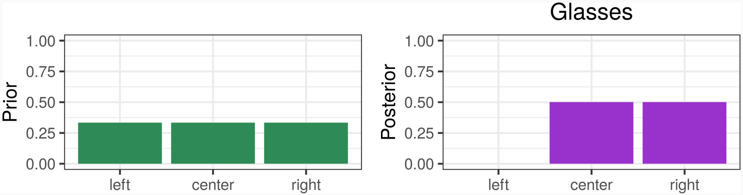


Bayes Rule for Decoding

$$P(m|f) = \frac{P(f|m)P(m)}{\sum_i P(f|m_i)P(m_i)}$$

Reference Resolution: A demo

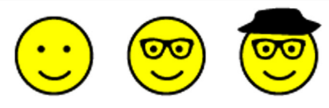


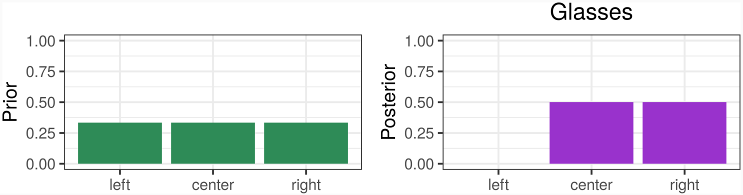


Bayes Rule for Decoding

$$P(m|f) = \frac{P(f|m)P(m)}{\sum_i P(f|m_i)P(m_i)} = \frac{0 \cdot \frac{1}{3}}{0 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3}} = 0$$

Reference Resolution: A demo

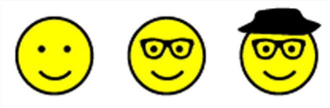


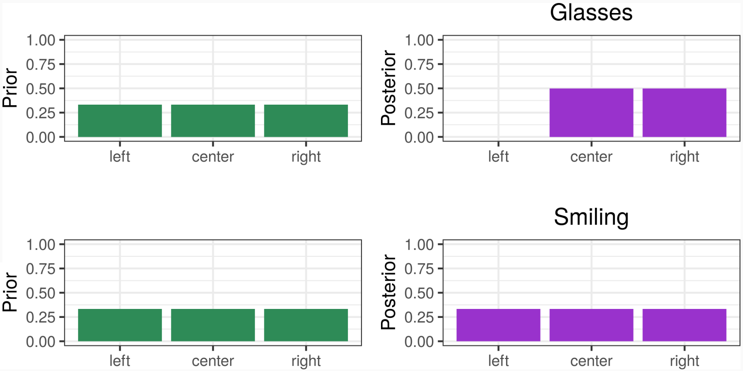


Bayes Rule for Decoding

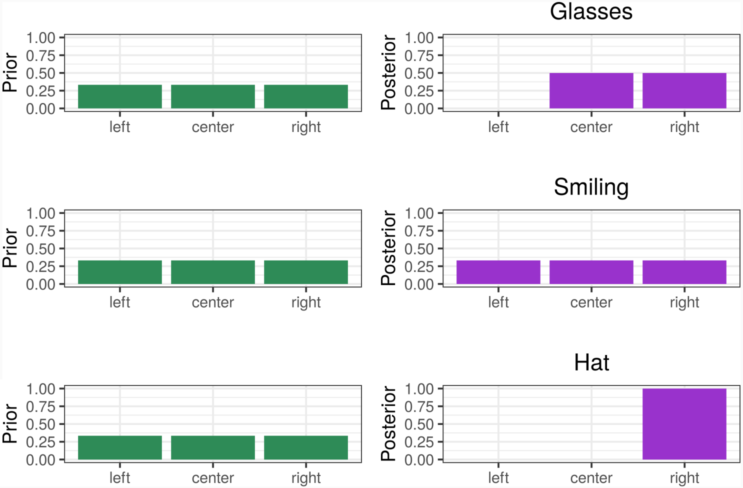
$$P(m|f) = \frac{P(f|m)P(m)}{\sum_i P(f|m_i)P(m_i)} = \frac{1 \cdot \frac{1}{3}}{0 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3}} = \frac{1}{2}$$

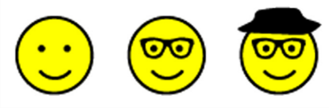
Reference Resolution: A demo





Reference Resolution: A demo





Information Theory Fundamentals

Information Theory Fundamentals

Information content (surprisal): a measure how unexpected an event is:

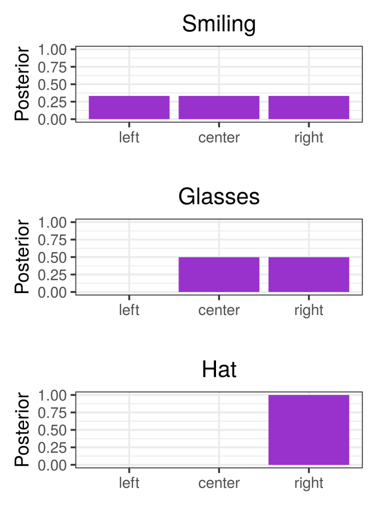
$$I(x) = -\log_2 P(x)$$

If an event is highly predicted (unsurprising), it has low information content.
If an event is very unpredicted (surprising), it has high information content.

Entropy: a measure of the uncertainty about an outcome:

$$H(X) = \sum_x P(x)I(x) = -\sum_x P(x)\log_2 P(x)$$

Entropy



Smiling:

$$H(M) = -(\frac{1}{3}\log_2\frac{1}{3} + \frac{1}{3}\log_2\frac{1}{3} + \frac{1}{3}\log_2\frac{1}{3}) = 1.58$$

Glasses:

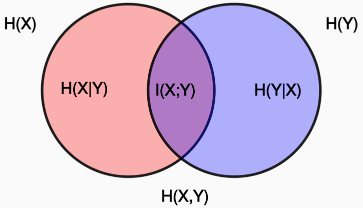
$$H(M) = -(0\log_2 0 + \frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}) = 1$$

Hat:

$$H(M) = -(0\log_2 0 + 0\log_2 0 + 1\log_2 1) = 0$$

By convention, we define $\log_2 0 = 0$.

Information Theory Fundamentals

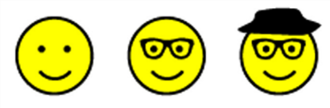


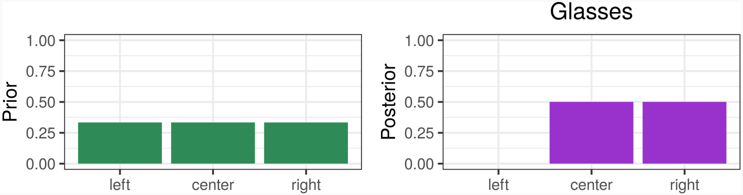
- **Mutual information:** The information shared between two variables.

$$I(M; F) = H(M) - H(M|F)$$

- **Conditional entropy:** $H(M|F)$: how deterministic is the mapping between from forms to meanings.
- If M and F are independent, $H(M|F) = H(M)$.
- If F deterministically maps to M, $H(M|F) = 0$.

Information Theory Fundamentals





Kullback-Leibler (KL) Divergence

The information gained by using distribution P compared to distribution Q.

$$KL(P||Q) = -\sum_{x} P(x) \log \frac{Q(x)}{P(x)}$$

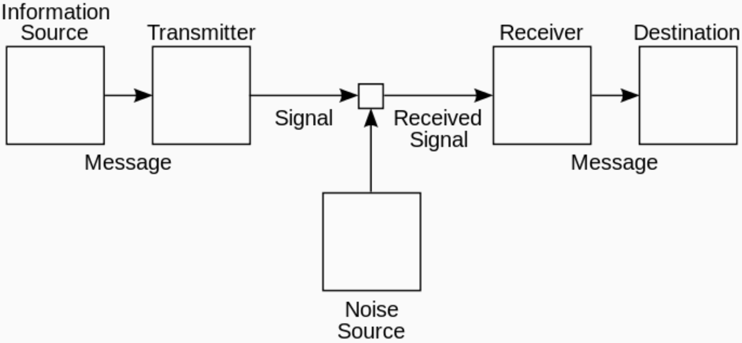
Information Theory Fundamentals

Time for a short quiz on Wooclap!

https://app.wooclap.com/FUZTIE

Principle of Least Effort

Big Questions



1. Are languages optimized for communicative efficiency?
2. Can we find efficiency effects in language processing?

Principle of Least Effort (Zipf, 1949)

Law of Abbreviation:

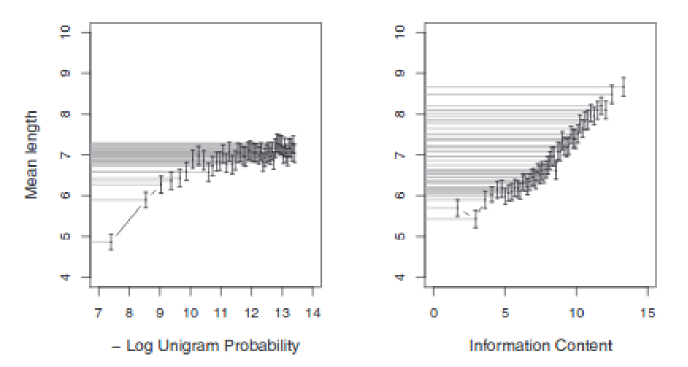
- Tools that are frequently used should be easier to use.
- Tools that are frequently used should be closer to you.

Law of Diminishing Returns:

- Frequently used tools should be versatile, i.e., have multiple uses.
- Frequently used tools should be used in concert rather than construct a specialized tool.
- If a permutation of tools becomes too complicated, make a specialized tool.

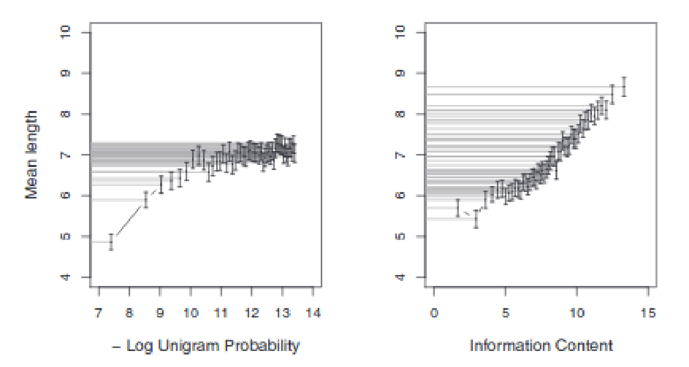
Are Frequent Words Easier to Use?

A shorter word is easier to produce. Are more frequent words shorter (Piantadosi et al., 2011)?



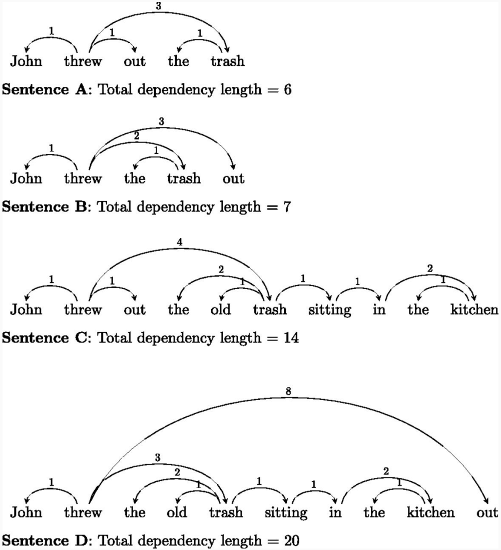
Are Frequent Words Easier to Use?

A shorter word is easier to produce. Are more frequent words shorter (Piantadosi et al., 2011)?



- Frequent words are shorter than infrequent words.
- The relationship is much stronger when we calculate it in terms of information content (surprisal).

Are Frequent Words Closer Together?


3
1 1 1
John threw out the trash
3
2
1 1
John threw the trash out
4
2
1 1 1 1 1 2 1
John threw out the old trash sitting in the kitchen
8
3
2
1 1 1 1 2 1
John threw the old trash sitting in the kitchen out

Sentence A: Total dependency length = 6

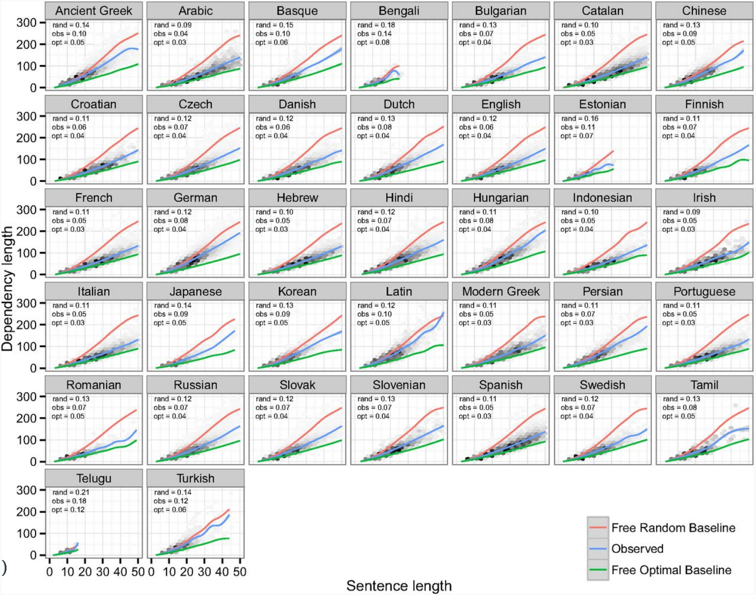
Sentence B: Total dependency length = 7

Sentence C: Total dependency length = 14

Sentence D: Total dependency length = 20

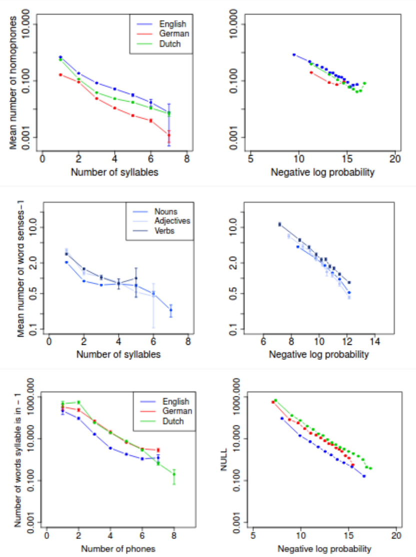
(Futrell et al., 2015)

Are Frequent Words Closer Together?



(Futrell et al., 2015)

Do Frequent Words Have More Functions?



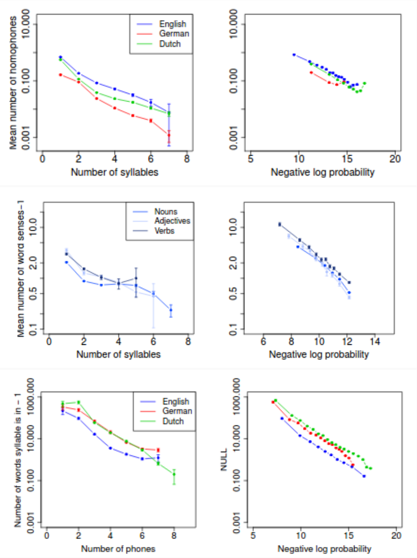
Homophones

Same sounds but different meanings

Word Sense

A single meaning of a word

Do Frequent Words Have More Functions?



Homophones

Same sounds but different meanings

Word Sense

A single meaning of a word

- Shorter words have more meanings.
- More probable (frequent) words have more meanings.

(Piantadosi et al., 2012)

Principle of Least Effort

Time for a short quiz on Wooclap!

https://app.wooclap.com/FUZTIE

Interim Summary

- Language as a communication system seems optimized for efficiency.
- Frequent words are:
    - shorter in length
    - closer to each other in sentences
    - versatile in how they can be used
- Can we find efficiency effects also language processing, i.e., when we understand language in real time?

Examples: Reading and Speech

Processing Difficulty in Reading

Some sentences are harder to process than others, even though the words are the same:

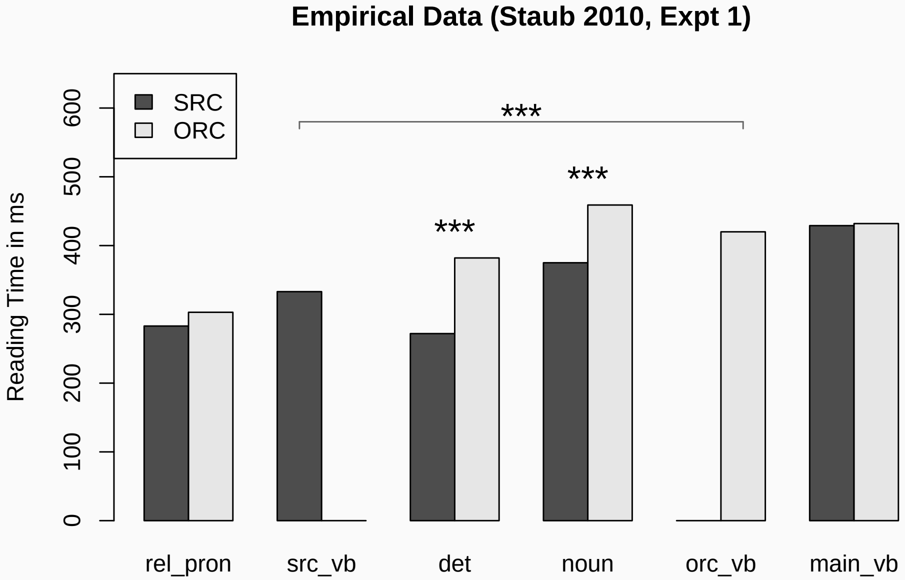
(1) The reporter that attacked the senator admitted the error.

(2) The reporter that the senator attacked admitted the error.

The object relative clause (ORC) in (2) is more difficult to process than the subject relative clause (SRC) in (1).

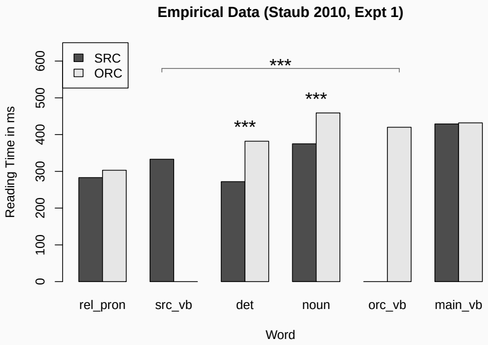
To be modeled: reading time differences on the relative clause verb and noun phrase (Staub, 2010).

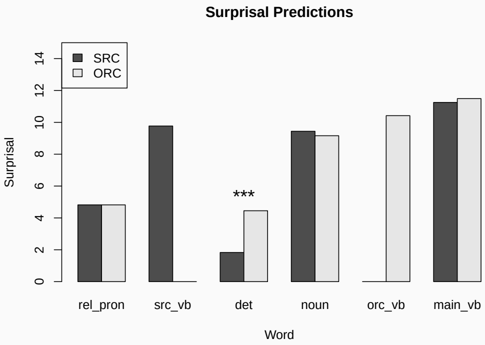
Processing Difficulty in Reading



Processing Difficulty in Reading

Compare reading times for relative clauses against surprisal values:

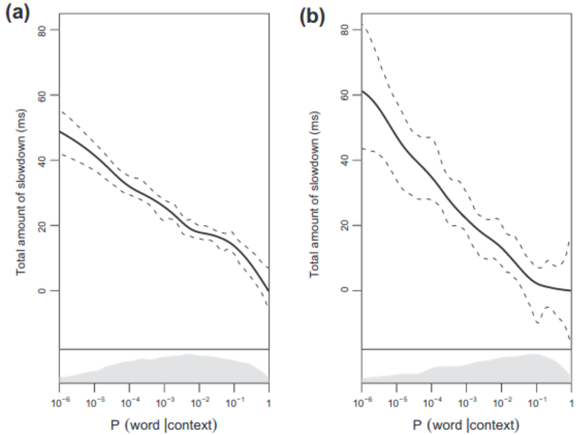




Surprisal successfully models only the difference at the NP. To model the difference at the verb, we need to add a distance-based component (Demberg et al., 2013).

General Effect of Surprisal on Reading Times

This effect generalizes when we look at reading times in a large text corpus: More surprising words take longer to read.

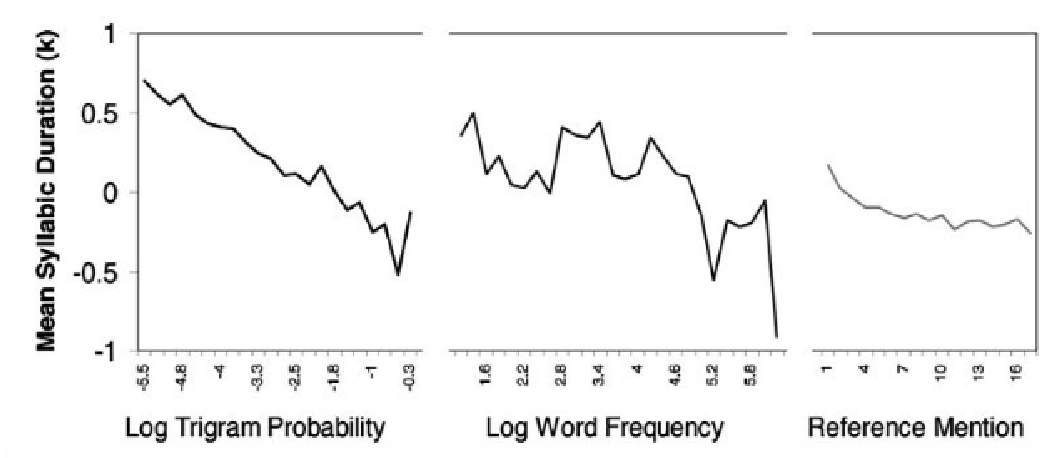


(Smith & Levy, 2013)

Fig. 4. By summing the curves in Fig. 3, we can estimate the total slowdown caused by an unpredictable word, regardless of where in the spillover region this slowdown occurs. (a) First-pass gaze durations. (b) Self-paced reading times. Lower panels show the proportion of data available at each level of probability.

General Effect of Surprisal on Speech

The effect also generalizes to speech: More surprising syllables have a longer duration.



(Aylett & Turk, 2004)

Surprisal Effects

Time for a short quiz on Wooclap!

https://app.wooclap.com/FUZTIE

Summary

- Communicative efficiency can be formalized using Information Theory.
- The forms of languages appear optimized for efficient communication.
- Efficiency also has observable effects on language processing.