Language as communication 2

Ted Gibson
9.59

Language as communication

- Information theory
- Words
- Sentences
- Communication-based models of language evolution and processing

Optimally designing a language

What features of a language might make it *optimal*?

What do we mean by *optimal*?

- optimal for use?
- optimal for comprehension?
- optimal for production?
- optimal for acquisition?

Words: Optimized for Communication?
Designing a language: **Ithkuil**

(thanks to Kyle Mahowald)

Foer, from the New Yorker, 2012:

Languages are something of a mess. They evolve over centuries through an unplanned, democratic process that leaves them teeming with irregularities, quirks, and words like "knight." No one who set out to design a form of communication would ever end up with anything like English, Mandarin, or any of the more than six thousand languages spoken today.

Portrait of John Quijada removed due to copyright restrictions.

Hence: **Ithkuil**, developed by John Quijada, a 53-year-old former employee of the California State Department of Motor Vehicles

John Quijada

Goals of Ithkuil:

- no ambiguity
- concision of expression
- broad coverage of ideas

Words: Optimized for Communication?
Designing a language: **Ithkuil**

(thanks to Kyle Mahowald)

from Wikipedia

Ithkuil words can be divided into just two parts of speech, *formatives* and *adjuncts*. Formatives can function both as nouns and as verbs, depending on the morpho-semantic context.[8] Both nominal and verbal formatives are inflected to one of the possible 3 *stems*, 3 *patterns*, 2 *designations* (*formal* or *informal*), 9 *configurations*, 4 *affiliations*, 4 *perspectives*, 6 *extensions*, 4 *contexts*, 2 essences, and 96 *cases*; formatives also can take on some of the 153 affixes, which are further qualified into one of 9 *degrees*. Verbal formatives are additionally inflected for 7 *illocutions* and 7 *conflations*.

Verbal adjuncts work in conjunction with adjacent formatives to provide additional grammatical information.[9] Verbal adjuncts are inflected to indicate 14 *valencies*, 6 *versions*, 8 *formats*, 37 *derivations*, 30 *modalities*, 4 *levels*, 14 *validations*, 9 *phases*, 9 *sanctions*, 32 *aspects*, 8 *moods*, and 24 *biases*.

Ithkuil would not be mistaken for a natural language

Foer: *Ideas that could be expressed only as a clunky circumlocution in English can be collapsed into a single word in Ithkuil. A sentence like "On the contrary, I think it may turn out that this rugged mountain range trails off at some point" becomes simply "Tram-mļöi hhâsmařpţuktôx."*

Portrait of John Quijada removed due to copyright restrictions.

John Quijada

Words: Optimized for Communication?
Designing a language: **Ithkuil**

(thanks to Kyle Mahowald)

- Ithkuil is not like a human language. (how so?)
- What is the baseline? Should we expect a language to look like Ithkuil?
- Important idea from the New Yorker article: How should a language be designed for optimal communication?

Information theory

- Claude Shannon:

  A Mathematical Theory of Communication (1948)

Portrait of Claude Shannon removed due to copyright restrictions.

Information theory / communication:

(1) Minimize code length;

(2) Noisy channel, so we need extra bits of information for robustness, especially for low frequency events

Diverging paths

Portrait of Claude Shannon removed due to copyright restrictions.

Portrait of Noam Chomsky removed due to copyright restrictions.

The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. ... The significant aspect is that the actual message is one *selected from a set* of possible messages. The system must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design.

But it must be recognized that the notion of "probability of a sentence" is an entirely useless one, under any known interpretation of this term.

Information

- The information of an event relates to the *probability* that the event occurs
- The more surprised you are by the event, the greater its *surprisal*: the more information in it
- An event with 0 information is already known (P = 1)
- An event that is infinitely unknowable should be infinitely informative (P = 0)
- Units of information: bits = coin flips = -log2(P(event)) = *surprisal* of event

Guess a word

- Suppose that there are 10,000 words in the lexicon
- -log2(10,000) = 13.3
  - 13.3 bits of information
- The optimal number of Yes-no questions that you might need to guess this word.
- "Is it *pizza*?" is not a good question: how many bits in the answer to that q?
  - -log2(10,000) - -log2(9999) = .0001 bits

Goal: 13.3 bits

It's a noun

- 5,000 nouns
- How much information did we gain?
- 1 bit

Current: 1 bit Goal: 13.3 bits

It's an animal

- There are 200 animals. How much information did we gain?
- log2(5000/200) = log2(25) = 4.64 bits
- Total: 5.64 bits

Current: 5.64 bit Goal: 13.3 bits

What if it were a planet?

- There are 8 planets. How much information would we have gained?
- log2(5000/8) = 9.3 bits
- Or: 9.1 bits (Pluto)

Current: 10.3 bit Goal: 13.3 bits

But it's really an animal

- Total information thus far: 5.64 bits
- Needed: 13.3

Current: 5.64 bit Goal: 13.3 bits

But it's really an animal

- 200 animals
- It starts with a 'b'
- 20 out of 200 start with b
- 3.3 bits from "starts with b"
- Total bits: 8.94

Current: 8.94 bit Goal: 13.3 bits

What if it starts with z?

- 200 animals
- It starts with a ‘z’
- 1 out of 200 start with z
- 7.64 bits from “starts with z”
- Total bits: 13.3: uniquely identified

Goal: 13.3 bits

Mackay 2003

Brainteaser

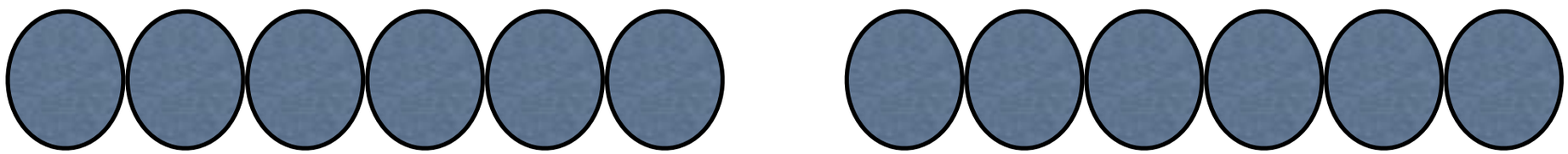
- You are given 12 balls and a scale. Of the 12 balls, 11 are identical and 1 weighs either slightly more or slightly less. How do you find the special ball (and whether it is heavier or lighter) using the scale only three times?
- The scale can only tell you which side is heavier.

Mackay 2003

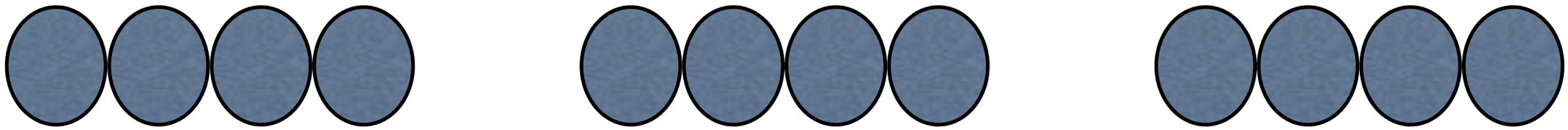
Brainteaser

- How many bits of information do you need to get?
  - 12 balls, each with 2 possibilities (normal, special) = 24 possibilities, giving -log2(24) = 4.58 bits
- How much information can you get (at most) from each weighing?
  - You get three possible answers: left heavier, right heavier, same = -log2(3) = 1.58 bits
- If you can divide the groups of balls into smaller groups of 3 with each weighing, you might be able to get the needed information after 3 weighings

Mackay 2003

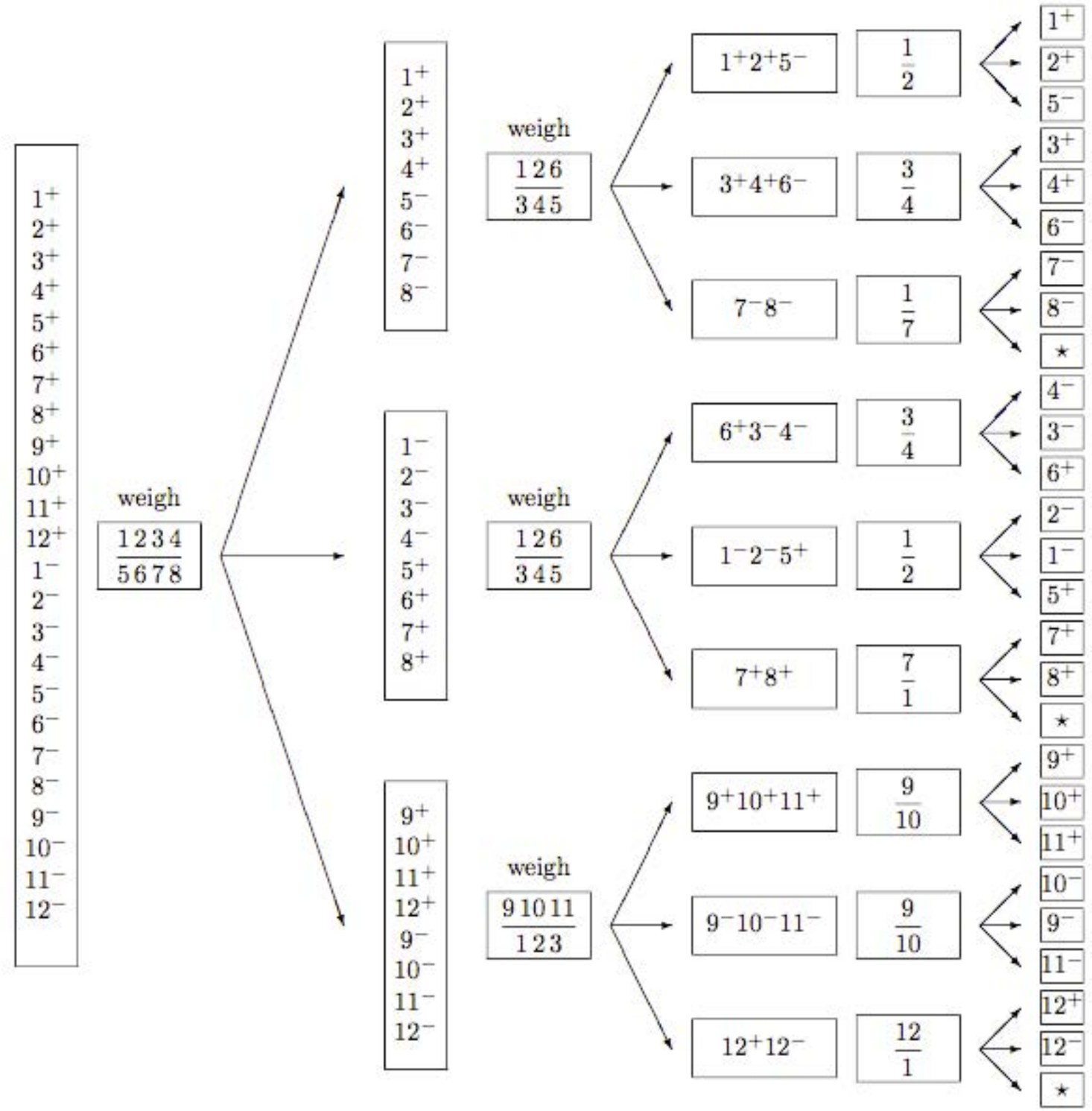


1 bit



1.58 bits

Mackay 2003



Coding

Suppose I want to transmit information: *communicate*

I need a **code**

Simplified code: just parts of speech (POS):

00 - NOUN
01 - VERB
10 - ADJECTIVE
11 - OTHER

The ugly man ran quickly to the rhinoceros

11 10 00 01 11 11 11 00 [16 bits]

Coding

The ugly man ran quickly to the rhinoceros

11 10 00 01 11 11 11 00 [16 bits]

00 - NOUN (2 times)
01 - VERB (1 time)
10 - ADJECTIVE (1 time)
11 - OTHER (4 times)

Different POS tags occur with different frequencies / probabilities. We can therefore use this to build a more efficient code: *to minimize the expected code length (optimize efficiency)*

Coding

The	ugly	man	ran	quickly	to	the	rhinoceros	
1	001	01	000	1	1	1	01	[14 bits]

1 - OTHER (4/8)
01 - NOUN (2/8)
000 - VERB (1/8)
001 - ADJECTIVE (1/8)

Information content

Suppose we have a distribution P on events (words, part of speech tags, weather conditions, notes in a song, etc.)

- The amount of information it takes to specify which event occurred is the <u>average</u> number of bits the <u>best</u> code must send

The	ugly	man	ran	quickly	to	the	rhinoceros	
1	001	01	000	1	1	1	01	[14 bits]

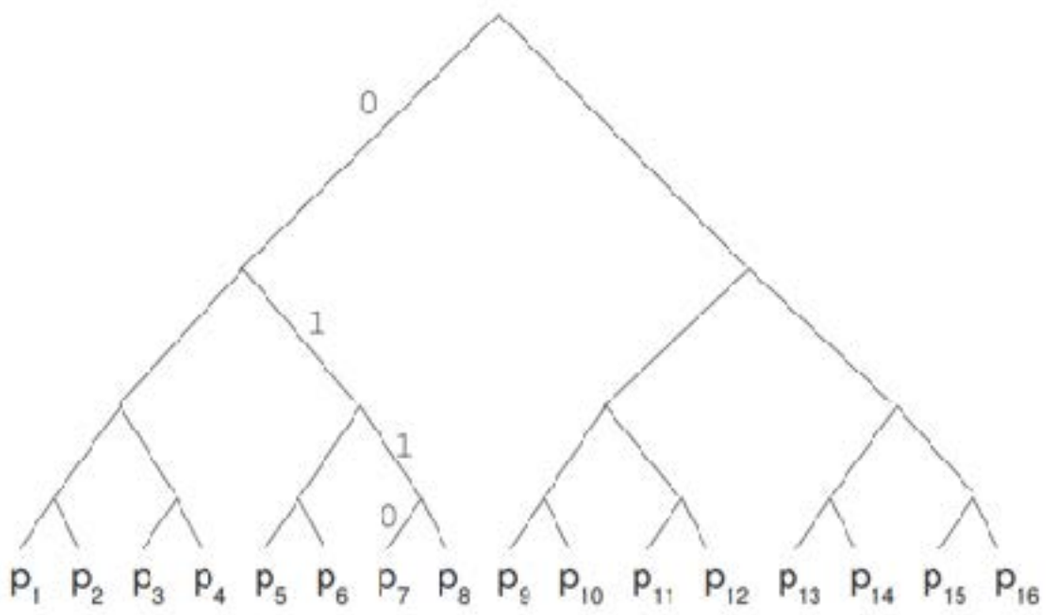
Result from Information theory (Shannon, 1948)

The best code will assign an event of probability p a code word of length ***-log(p)*** (roughly, in the limit): the *surprisal* of that event

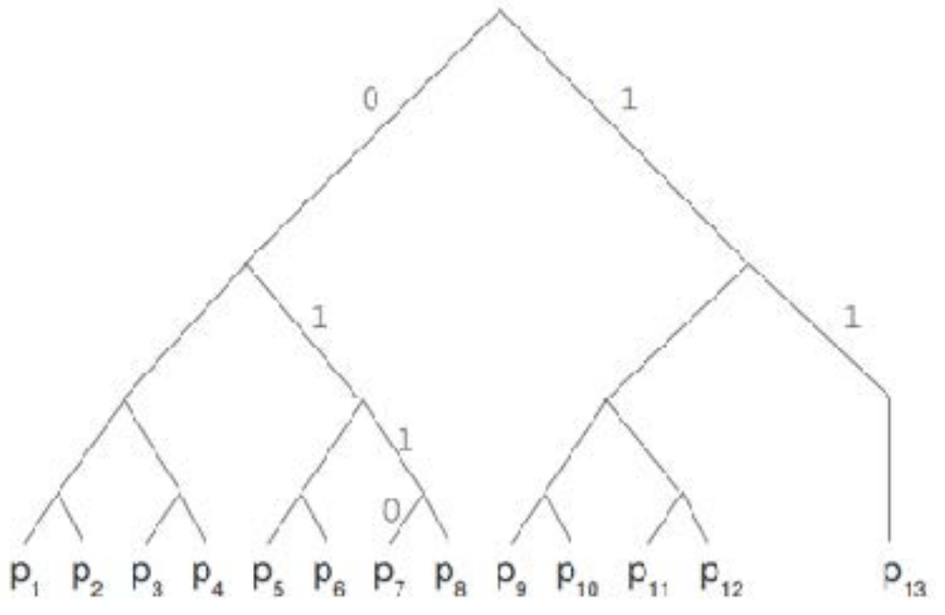
Likely events have low surprisal: few bits of information
Unlikely events have high surprisal: many bits of information:

the depth of the binary tree is the negative log probability

Intuitive relationship:
log probability and code length



Intuitive relationship:
log probability and code length



Entropy

- **What is the average (expected) number of bits required to specify which event happened?**
- **This measure is the entropy.**

$$H[X] = -\sum_x p(x) \log p(x)$$

Probability of having to communicate x

Bits required to communicate x

- **-Log probability (surprisal)** – measures of the amount of information it takes to specify that a specific event occurred (measured on *events*)
- **Entropy** – measures the average number of bits it takes to specify which event will occur (measured on *distributions*)

Uniform distribution

- A uniform distribution maximizes entropy

```
calc.entropy <- function(x) {return (sum(-x*log2(x)))}
```

- [.97, .01, .01, .01] : Entropy =

```
-(.03 * log2(.01) + .97 * log2(.97)) = 0.24 bits
```

- [.25, .25, .25, .25] : Entropy =

```
-4 * (1/4 * log2(.25)) = 2 bits
```

Distributions

- ABABABABCABABACABABA
- ABABCACBACCBABCBCBCA
- eiqtyp2q3450761Q[WR8Y[82Qdsiew92

Conditional surprisal

We are typically are in situations where events are not independent:

- The …
- The silly…
- The silly grasshopper…
- The silly grasshopper wanted to find his friend the …

Conditional surprisal

- **To estimate the probability of w in context c, see how often w is observed in c in a corpus**

$$p(w \mid c) \approx \frac{cnt(c, w)}{cnt(c)}$$

The man <u>ate</u> → high probability

- log p(ate | the man) = 4.4

The man <u>galloped</u> → low probability

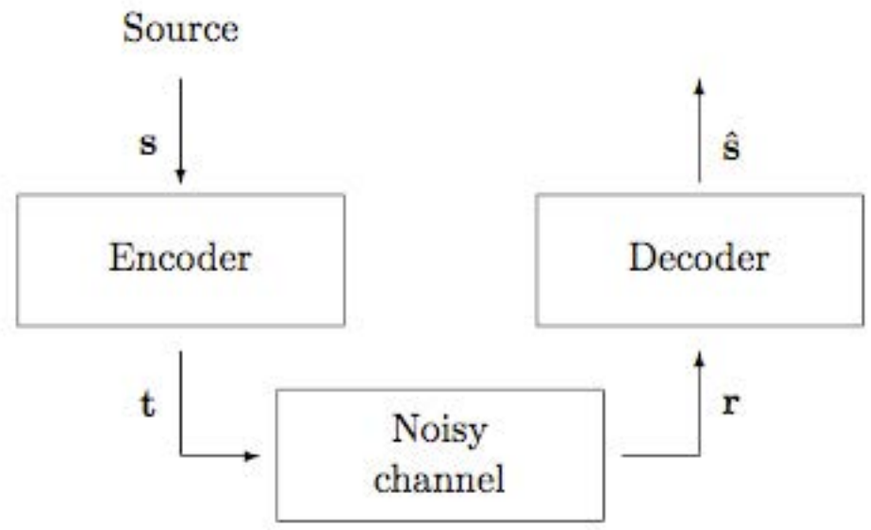
- log p(galloped | the man) = 13.2

The man <u>scissors</u> → very low probability

- log p(scissors | the man) = 25.1

Efficient communication

- Hypothesis: Natural language is a largely efficient code
  - Concise while still being robust to noise
- Longer words are more robust to noise



s	0	0	1	0	1	1	0
t	000	000	111	000	111	111	000
n	000	001	000	000	101	000	000
r	000	001	111	000	010	111	000

redundant code

Language / Communication: Words
Piantadosi, Tily & Gibson (2011)

Zipf (1949): more frequent words are shorter:

- "Principle of least effort"

Extension: more *predictable* words should be shorter.

- e.g., to maintain Uniform Information Density (Aylett & Turk, 2004; Jaeger, 2006; Levy & Jaeger, 2007)
- Estimate of predictability: n-grams (3-grams) over large corpora

phrase	count	freq	information in last word (bits)
"to be or not to **be**"	86/87	0.99	$-\log_2(86/87) = 0.01$
"to be or not to **bop**"	1/87	0.01	$-\log_2(1/87) = 6.44$

from corpus of contemporary American English (COCA)

Average information

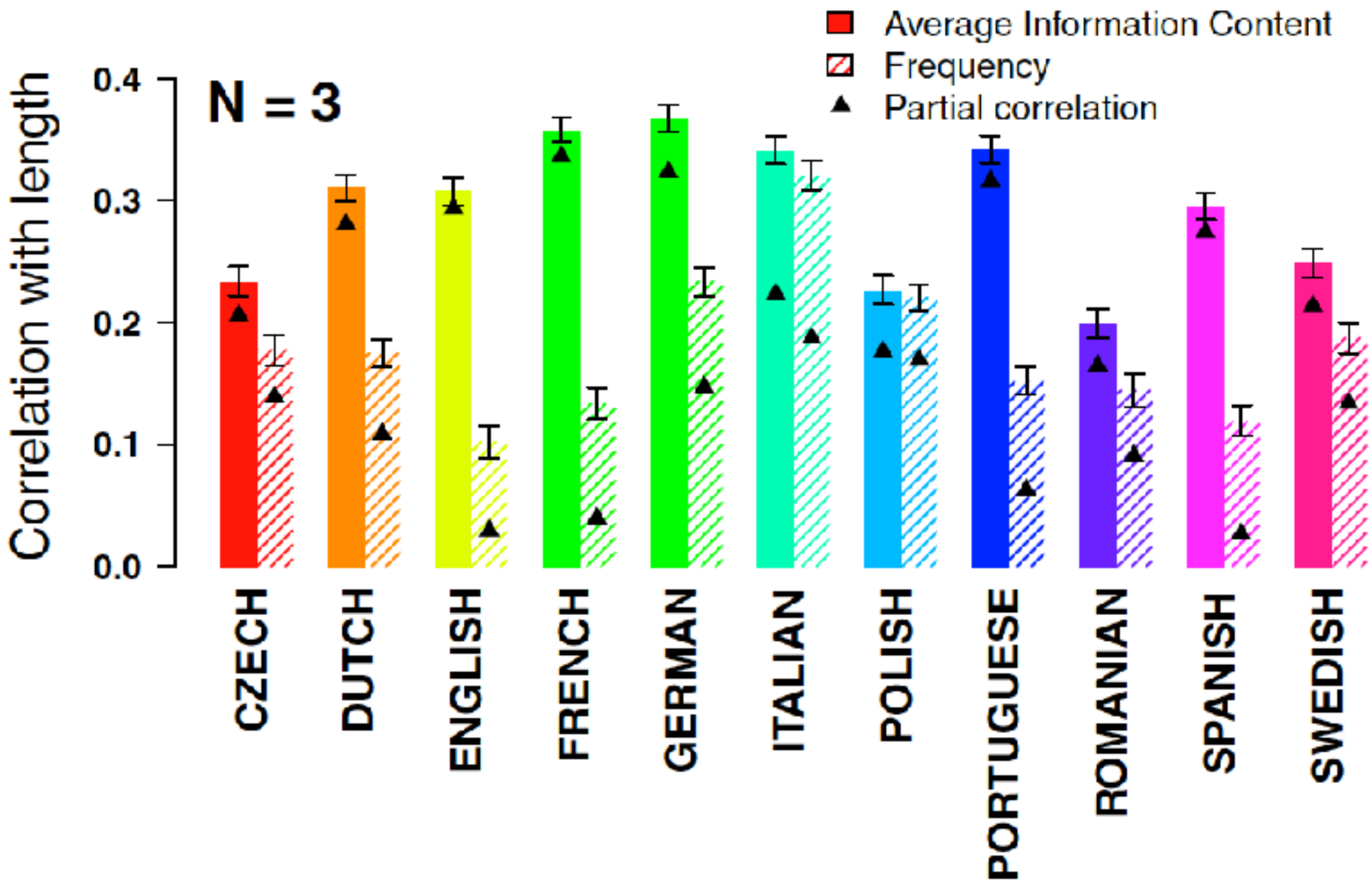
- average information of a word *w* over all contexts in which it appears

$$-\frac{1}{N}\sum_{i=1}^{N}\log P(W = w \mid C = c_i)$$

if nothing else in vocabulary:

phrase	word	average surprisal
"to be or not to **be**"	be	0.01
"to be or not to **bop**"	bop	6.44

Language for communication: Words
Piantadosi, Tily & Gibson (2011)



More predictable words are shorter!

How does the effect arise?

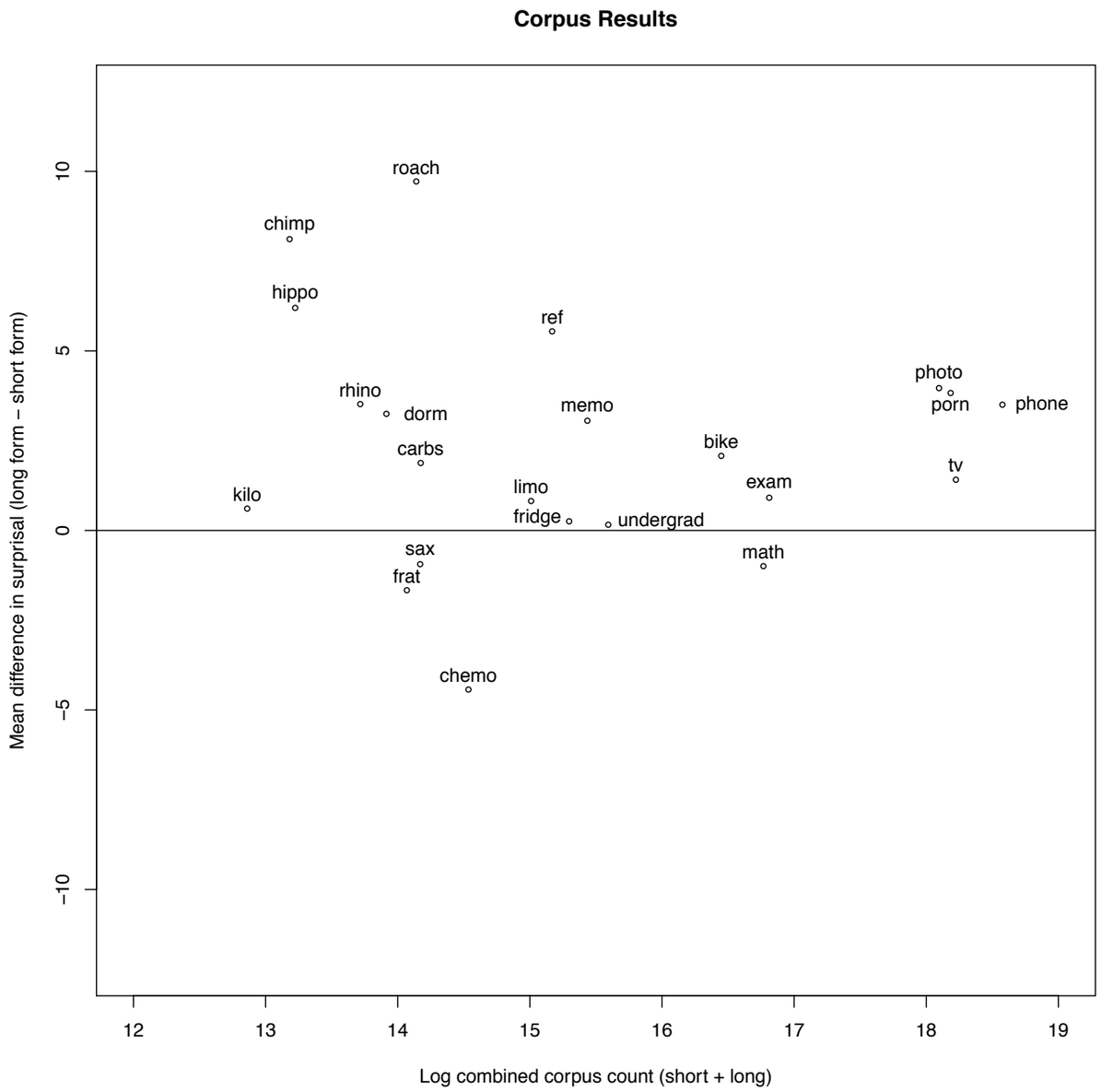
- Is it just differences among broad classes of words like content vs. function words? Or within class too?
- How does the effect come about in the lexicon? Long-term evolution?
- look at long/short pairs (chimpanzee → chimp), which differ in length but are controlled for meaning

Info/Information theory

Using Google trigrams, we looked at average surprisal for long forms vs. short forms.

Mean surprisal for long forms (9.21) is significantly higher than mean surprisal for short forms (6.90) (P = .004 by Wilcoxon signed rank test)

Linear regression shows significant effect of log frequency on surprisal (t = 2.76, P = .01) even when controlling for frequency.



Mahowald, Fedorenko, Piantadosi and Gibson (Cognition 2013)

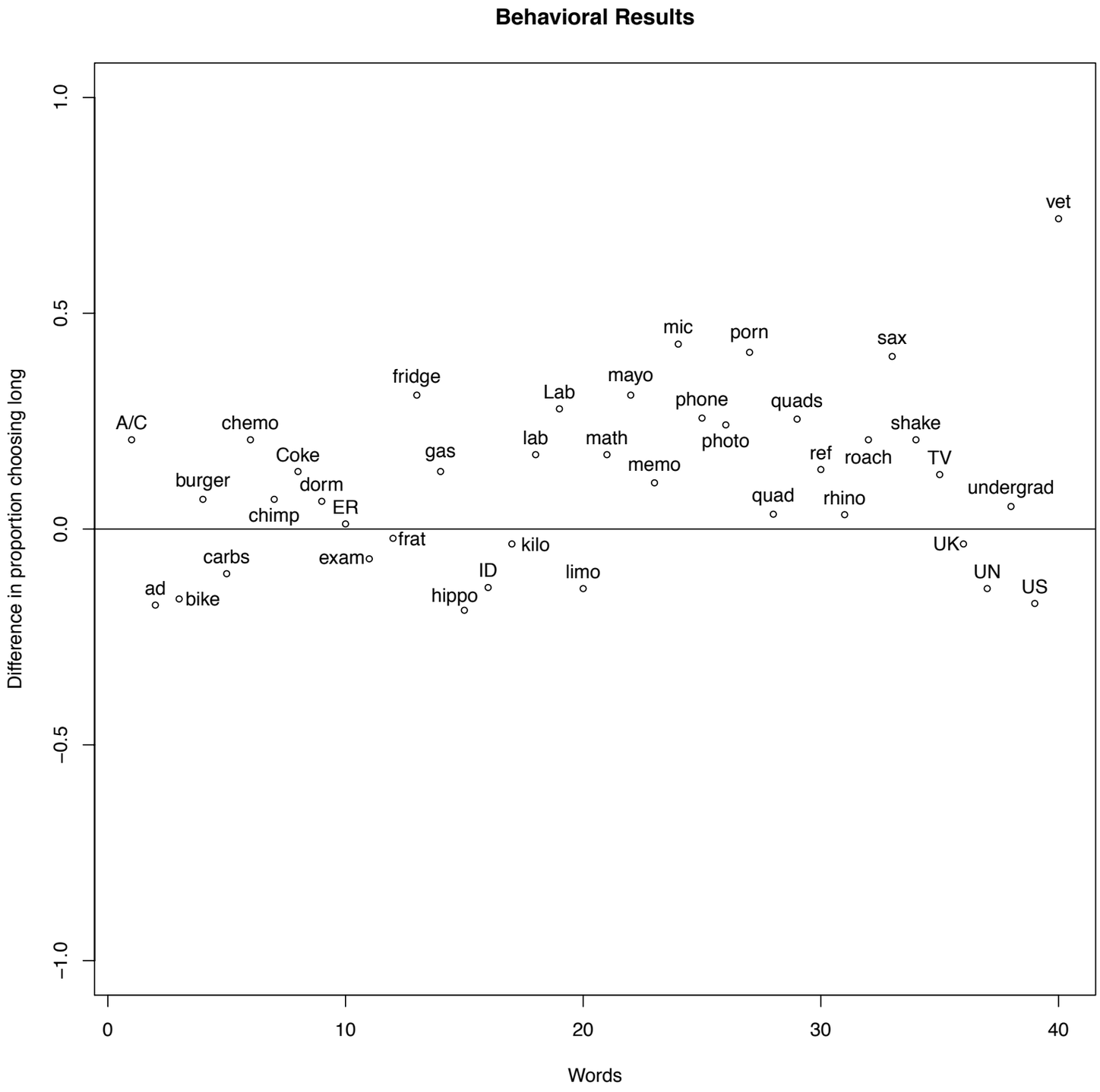
Forced-choice sentence completion in supportive and neutral contexts:

supportive-context: Bob was very bad at algebra, so he hated...
1. math 2. mathematics

neutral-context: Bob introduced himself to me as someone who loved...
1. math 2. mathematics

Short form is chosen 67% of the time in supportive-context sentences vs. just 56% of the time in neutral-context sentences.

Significant by maximal mixed effect logistic regression with both item and participant slopes and intercepts (β = .67, z = 2.59, P < .01).



Courtesy of Elsevier, Inc., http://www.sciencedirect.com. Used with permission. Source: Mahowald, Kyle, Evelina Fedorenko, Steven T. Piantadosi, and Edward Gibson. "Info/information theory: Speakers choose shorter words in predictive contexts." Cognition 126, no. 2 (2013): 313-318.

Mahowald, Fedorenko, Piantadosi and Gibson (Cognition 2013)

Examples from Mahowald et al. behavioral study

supportive: For commuting to work, John got a 10-speed...
neutral: Last week John finally bought himself a new…

bicycle / bike

supportive: Henry stayed up all night studying for his...
neutral: Henry was stressed because he had a major...

examination / exam

supportive: Jason moved off campus because he was tired of living in a...
neutral: After leaving Dan's office, Jason did not want to go to the...

dormitory / dorm

Audience design?

Clark (1996): Yes, for word choices

Asking for directions: Speakers use words that are appropriate to listeners background knowledge

Audience design?

Ferreira & Dell (2000): Exploring **syntactic optionality** in sentence production

Method: produce memorized sentences, either for a listener or not.

Materials contained optional "that"

No ambiguity:
Match: I knew (that) I had ...
No match: You knew (that) I had ...

Ambiguity:
No match: I knew (that) you had ...
Match: You knew (that) you had ...

Audience design?

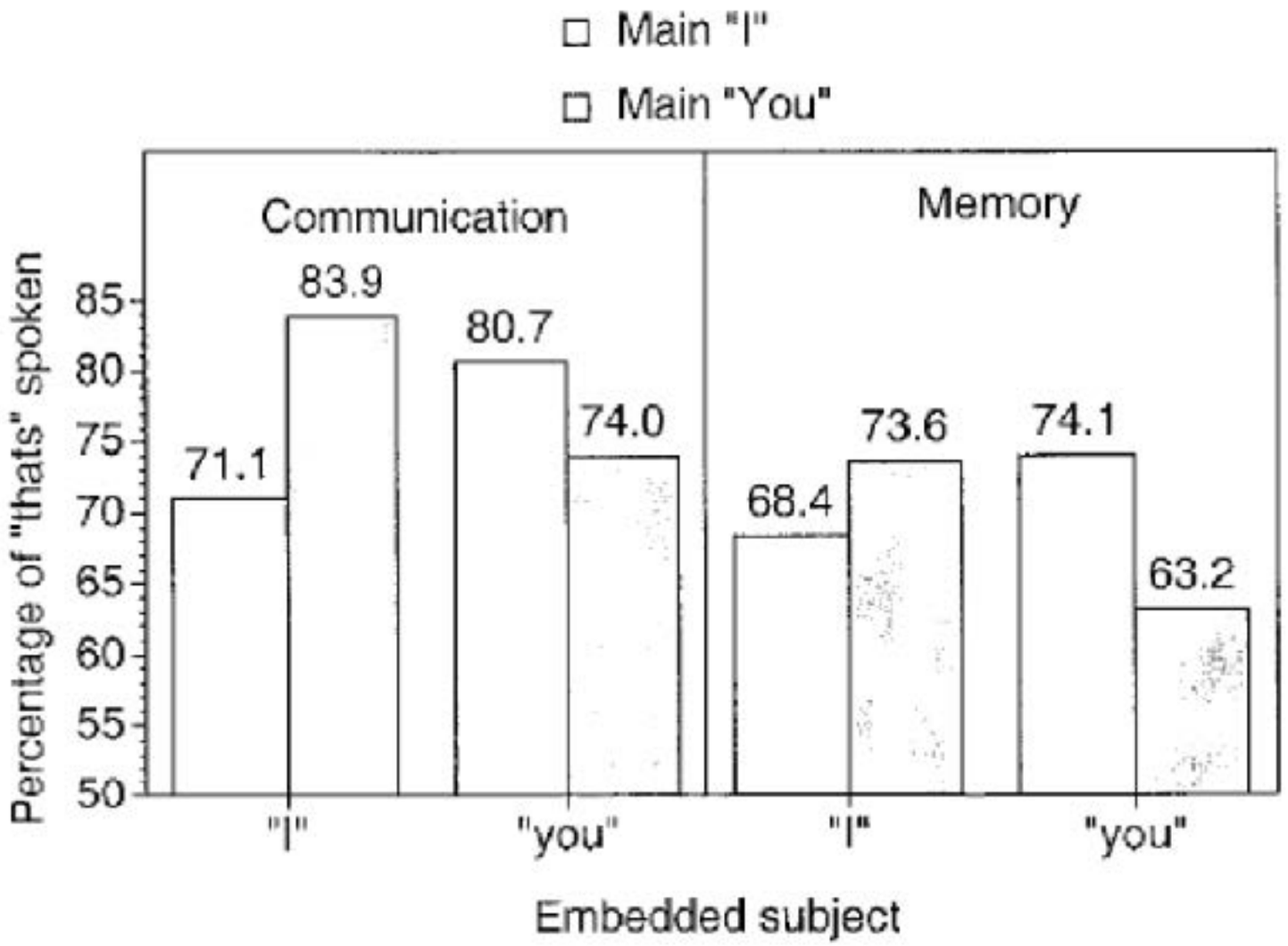


FIG. 3. Percentages of full sentences produced with different main and embedded subjects by speakers in a communication or memory task in Experiment 6.

The left bar in each pair is "I" in the main clause; the right is "you".
e.g., "I / you knew that I / you had missed practice"

What do you want to know more about?

I'm curious about the algebra adopted in the corpus study to calculate surprisal of each word.

I would like to know more concretely how the surprisal of each word was estimated in the corpus study. I would appreciate it very much if you could give a couple of actual examples in the class.

I would like elaboration on the details of the information theory involved in this and especially in other similar experiments.

I am curious to learn more on what Shannon information content/theory.

Does this pattern of information-theoretical optimization hold true for other languages as well as English? Is there any data on how language evolved through time to become more efficient at communicating information?

I want to learn more about the robustness of the negative log-probability surprisal measure and about how non-lexical, non-syntactic, context-based information interacts with this word-length, information correlation.

Can we talk more about the statistics behind how you perform a corpus study? How exactly does an n-gram model allow you to generate probabilities for not just word predictions, but also for how contextual a sentence is?

I would be interested in seeing other examples of questions from the behavioral study.

I'm interested in the difference between supportive contexts and neutral contexts; more broadly, I'd love a continued exploration of different areas of syntax.

1. what's the relationship of linguistic theories and information theory?
2. As is mentioned in the last part of the paper, this research might help account for the language change where words become shorter. I was wondering how we understand/explain the condition where the surprisal and word length increases.

The results of this paper are convincing, but I'm also curious about the selection of synonyms of different lengths in neutral context and very unsupportive context. Following the line of this research, we might predict that words of long length would be preferred in very unsupportive context.

I'm interested in learning more about other factors that cause people to chose longer words

I would like to discuss more about the difference between written and verbal communications because there may also be differences in word length between these two methods of communication. Additionally, I want to know what kinds of follow up studies would be/ have been conducted as a result of this study; what has it led to?

9.59J/24.905J Lab in Psycholinguistics
Spring 2017