Computational Psycholinguistics

Lecture 10: **Information Theoretic Approaches 2**

Matthew W. Crocker
crocker@coli.uni-sb.de

Surprisal & Psycholinguistics

- In addition to measuring the average information for a language, we can of course measure the **information conveyed by any given linguistic unit** (e.g. phoneme, word, utterance) in context. This is often called *surprisal*:

$$Surprisal(x) = \log_2 \frac{1}{P(x \mid context)}$$

- **Surprisal will be high**, when *x* has a low conditional probability, and **low**, when *x* has a high probability.

- Claim: **Cognitive effort** required to process a word is **proportional** to its **surprisal** (Hale, 2001).

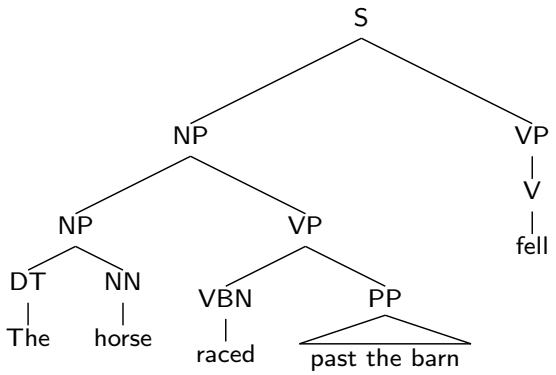
Computing Surprisal

$$\text{Surprisal}_{k+1} = -\log P(w_{k+1} \mid w_1 \ldots w_k)$$

- There are various ways we can compute surprisal from different kinds of underlying probabilistic language models
- N-gram surprisal:

$$\text{Surprisal}(w_{k+1}) = -\log_2 p(w_{k+1} \mid w_{k-2}, w_{k-1}, w_k)$$

Lexical vs. structural surprisal



$$S_{w_n} = -\log \sum_{T \in Trees} \frac{p(T|w_1 \ldots w_n)}{p(T|w_1 \ldots w_{n-1})}$$

$$structS_{w_n} = -\log \sum_{\mathrm{POS}_n \in POS} \sum_{T \in Trees} \frac{p(T|w_1 \ldots \mathrm{POS}_n)}{p(T|w_1 \ldots w_{n-1})}$$

$$lexS_{w_n} = -\log \sum_{\mathrm{POS}_n \in POS} \sum_{T \in Trees} \frac{p(T|w_1 \ldots w_n)}{p(T|w_1 \ldots \mathrm{POS}_n)}$$

Cloze Probabilities and Predictability

- Ask participants to fill in the blanks (Taylor, 1953)

  I went to the _____ and bought some milk and eggs. I knew it was going to rain, but I forgot to take my _____, and ended up getting wet on the way _____.

- Cloze probability is the likelihood of a particular word occuring in a particular context:

  (a) My brother came inside to _____.
  (b) The children went outside to _____.

- "play" is plausible in both sentences, but is 1st choice 90% of the time in (b) never the first choice for (a).

Cloze and Reading

- But **cloze** is an off-line production task:
  - many low probability words are never produced
  - participants have more time to determine likely words
  - may also reflect knowledge, not just linguistic experience
- Cloze indexes predictability, but may not tell us much about how comprehenders might actually predict upcoming words on-line

Cloze and Reading

- Rayner & Well (1996) directly investigated the influence of contextual constraints on reading

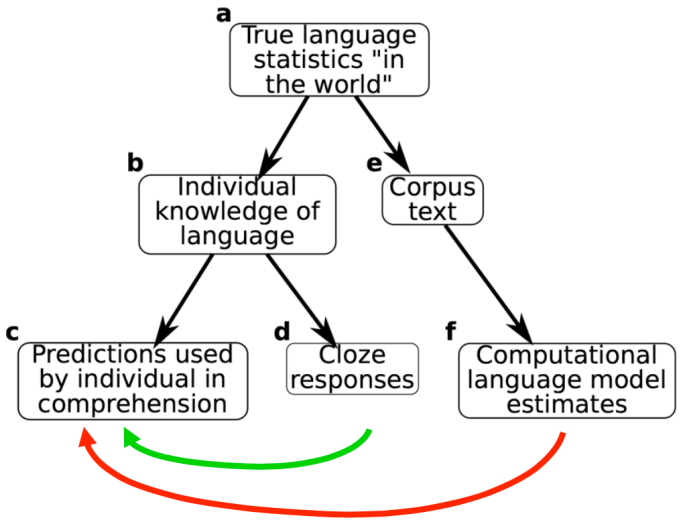
(a) The woman took the warm cake out of the oven. (high – 93%)
(b) The woman took the warm cake out of the stove. (med – 33%)
(c) The woman took the warm cake out of the pantry. (low – 3%)

- Low-constraint (3-8%) words were fixated longer than high(>73%) and medium (13-68%).

- High-constraint words were skipped more often than low and medium.

Close verse Corpora

- What is the better predictor of reading times?

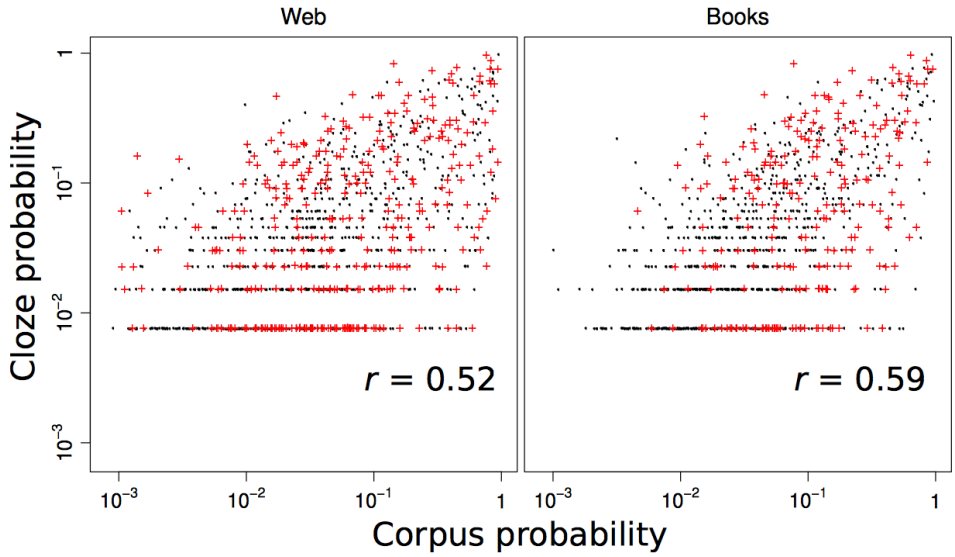


He played a key...			*After a cup of...*		
role	94%	42%	coffee	39.6%	28%
part	2.3%	3.8%	tea	39.1%	61%
			hot	3.0%	—
When she began to...			*The time needed to...*		
speak	5.2%	9.0%	complete	41%	6.7%
cry	2.5%	19%	reply	3.2%	—
work	2.2%	1.5%	finish	0.3%	10%
It usually takes the...			*In the winter and...*		
form	34%	1.5%	spring	66%	40%
shape	23%	—	early	13%	—
following	2.7%	—	summer	4.2%	32%
cake	—	8.3%	fall	2.3%	19%

Table 1: Sample continuation distributions from Experiment 1. In each case, the left column is corpus probability, and the right column is measured cloze probability. '—' denotes continuations that were never observed.

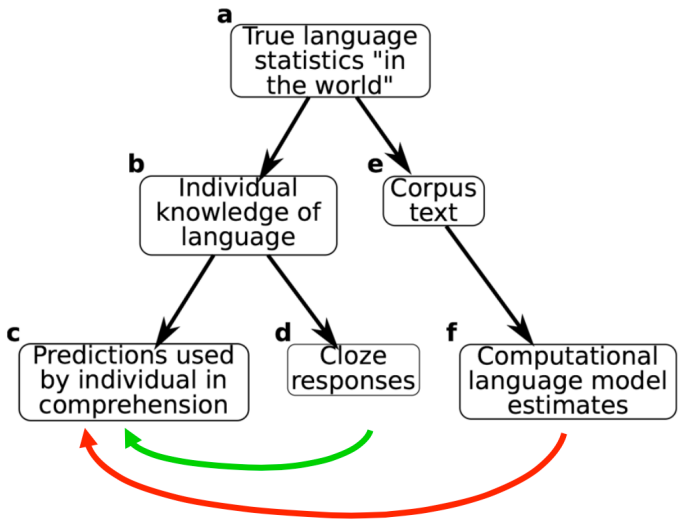
Cloze vs. Corpora

- Smith & Levy (2011) determined corpus & cloze probabilities for a set of 4 word contexts:



What predicts reading times?

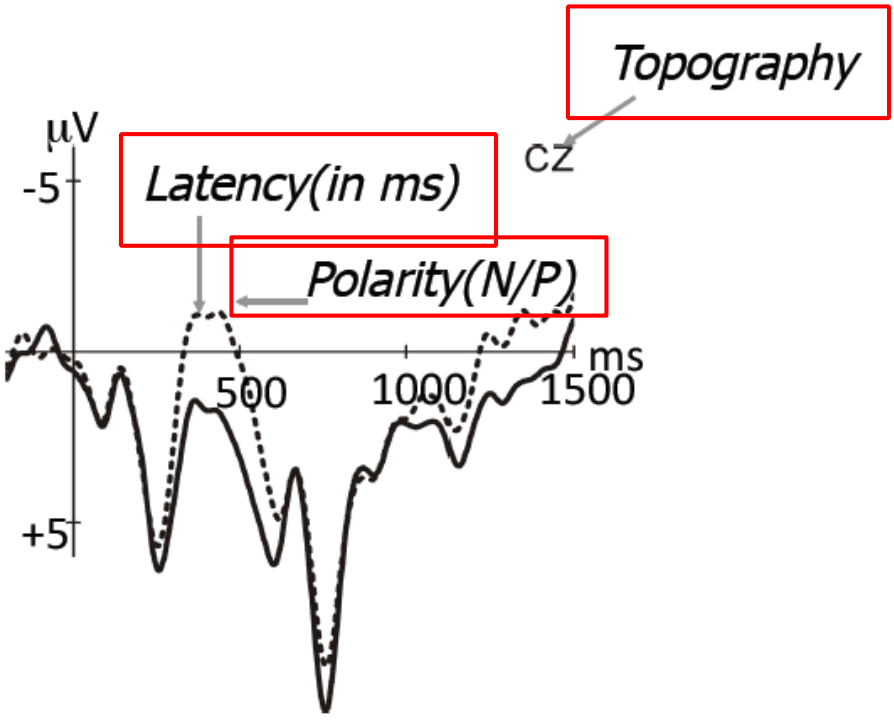
- Cloze significantly predicted reading times, after controlling for corpus probability
- Corpus-based probability estimates did not predict reading times, after controlling for Cloze
- How probabilities contribute to human predictions and reading times is not yet clear



On-line Measures

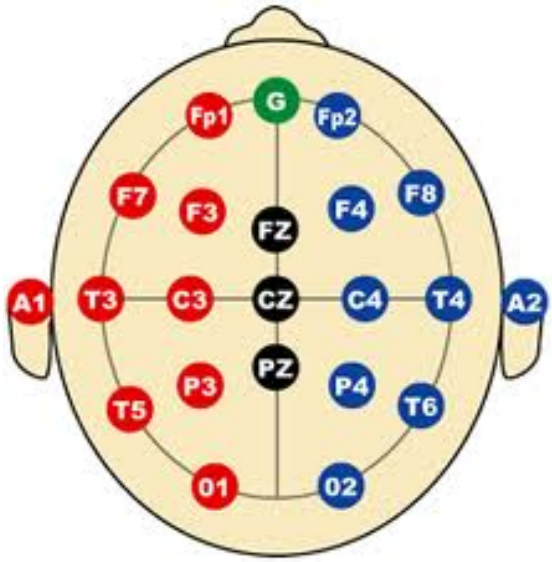
- Reading times are known to reflect processing difficulty due to lexical, syntactic and semantic factors … more on this later.
- Event-related potentials are a neurophysiological measure that indexes processes of lexical retrieval (N400) and integration (P600)
- The visual world paradigm.

ERP Components



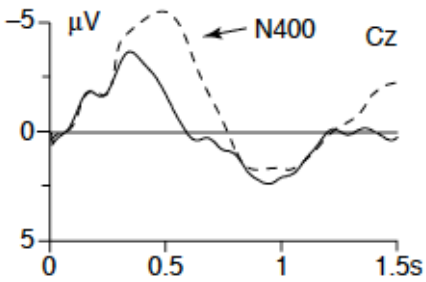
Topographical distribution

- **Where is the ERP found on the scalp?**
- ERP components may have a broad/ frontal/central/posterior/ lateralized distribution
- NB: Topography is not informative about the brain areas generating the signal
- However, different topographical distributions suggest different neural generators



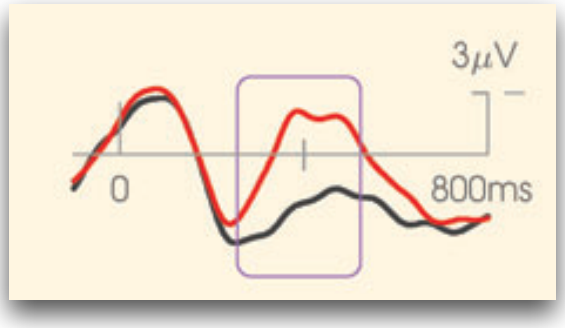
The N400

- Negative deflection peaking around 400ms after stimulus onset



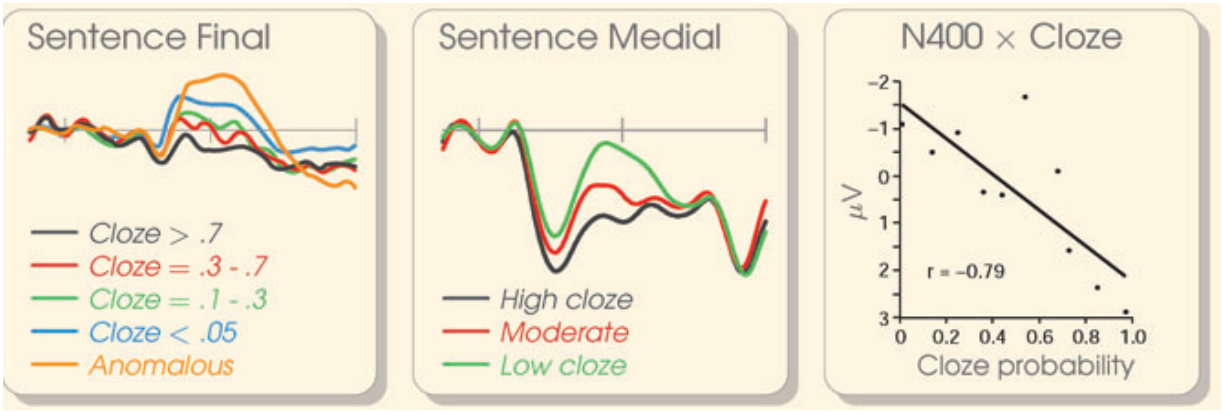
- Maximal over centro-posterior sites, bilateral
- Discovered by Kutas and Hillyard in the early 80s

Some factors influencing N400 amplitudes



- Frequency (LF>HF)
- Repetition (New>Repeated)
- Sentence position (Initial words > Medial > Final)
- Lexical association (priming): Unrelated > Associated
- Semantic congruency: Incongruent > Congruent
- Off-line expectancy (cloze probability)
  - Unexpected > Expected

N400 and cloze probability



Kutas & Federmeier (2010)

The N400 is inversely correlated with the cloze probability of a word

N400 and cloze probability

- The N400 sensitivity to word predictability is consistent with either of two views:

1) Words are actively predicted and reduced N400 amplitudes reflect the benefits of facilitated **lexical retrieval**

2) Predictable words fit better with the wider context and reduced N400 amplitudes reflect **easier semantic integration** (regardless of prediction)

Federmeier and Kutas (1999)

- Examined the relationship between word predictability and semantic memory

- *They wanted to make the hotel look more like a tropical resort. So along the driveway they planted rows of* palms./pines./tulips.

Manipulation

Cloze probability

palms / pines / tulips
0.74 / < 0.05 / < 0.05

Category membership

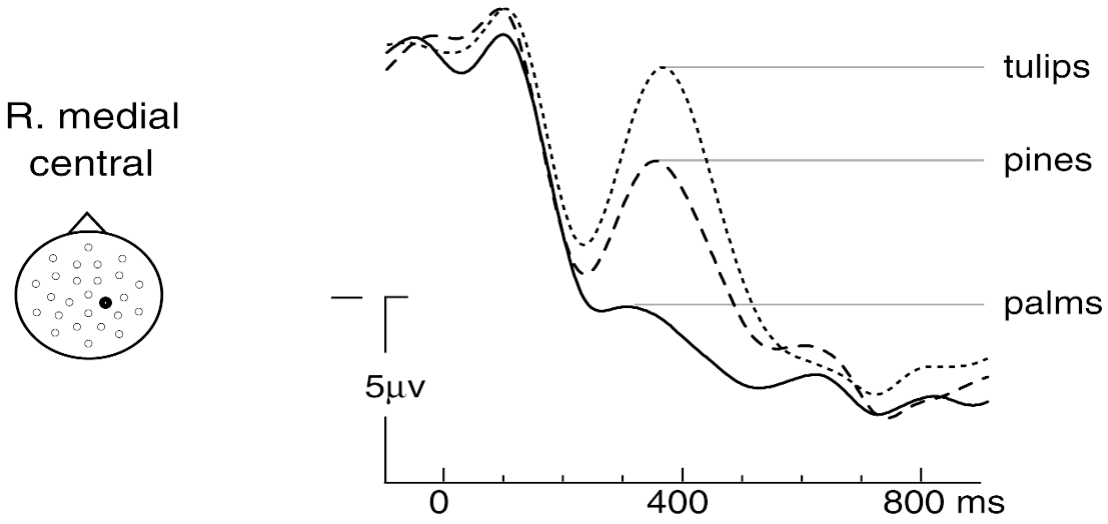
palms / pines / tulips
[tree] / [tree] / [flower]

Unexpected within-category violation
Unexpected between-category violation

Federmeier & Kutas (1999)

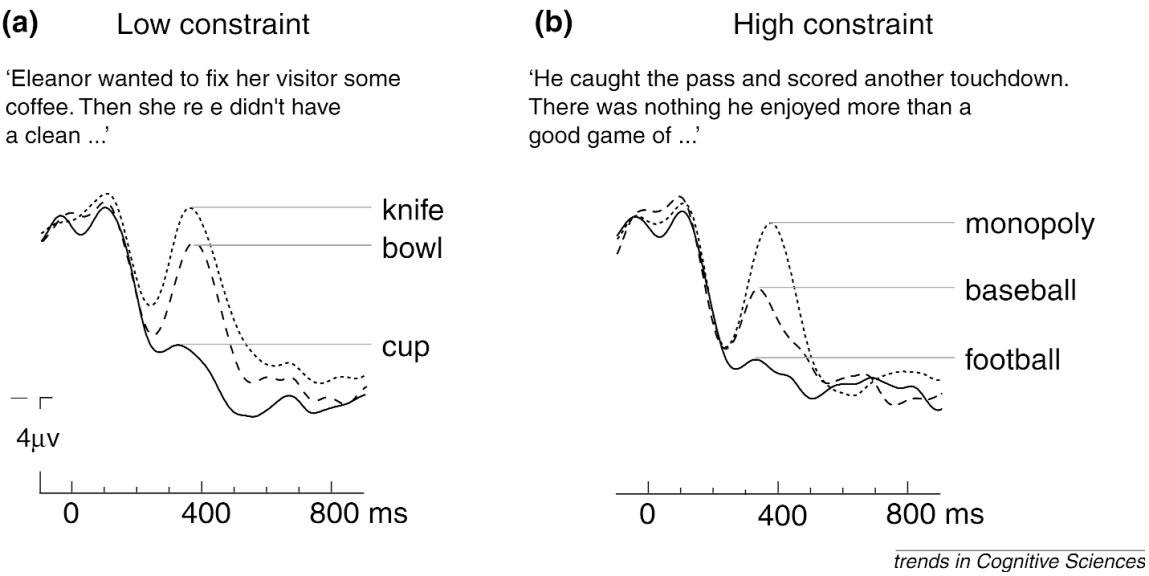
Results

'They wanted to make the hotel look more like a tropical resort.
So along the driveway they planted rows of ...'



Federmeier & Kutas (1999)

Results



Federmeier & Kutas (1999)

Discussion

- The incremental language processor generates expectations for the semantic features of the upcoming words
- Words that are almost never produced off-line – but are more congruent with the brain's predictions – are easier to process
- But do people ever predict specific words?

Word Pre-activation

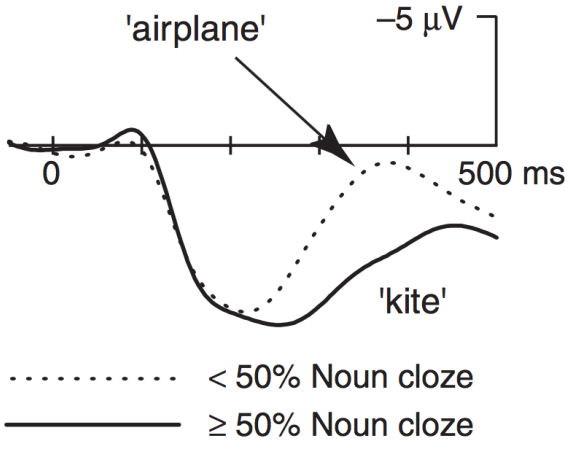
- Consider the sentence:
  - *The day was breezy so the boy went outside to fly _____*
  - *… **a kite** / an airplane*
- We would predict an increased N400 for *airplane*
- But what about for the determiner “a” versus “an”

Delong, Urbach & Kutas, *Nature Neuroscience*, 2005

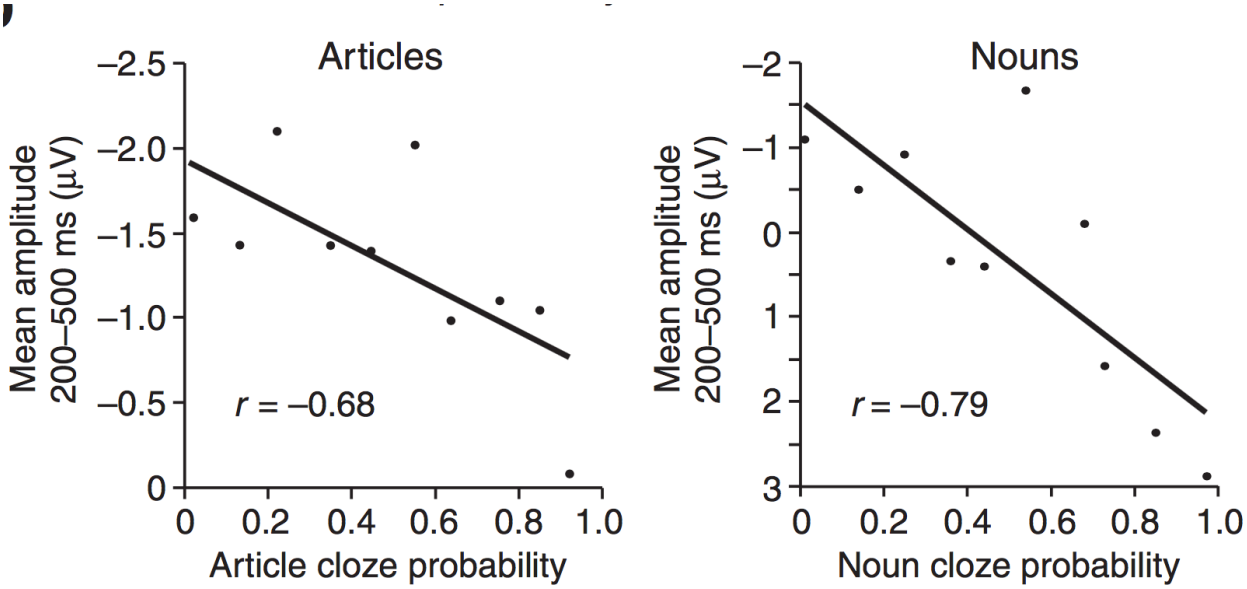
Lexical Prediction?

e.g., 'The day was breezy so the boy went outside to fly ...'

Nouns



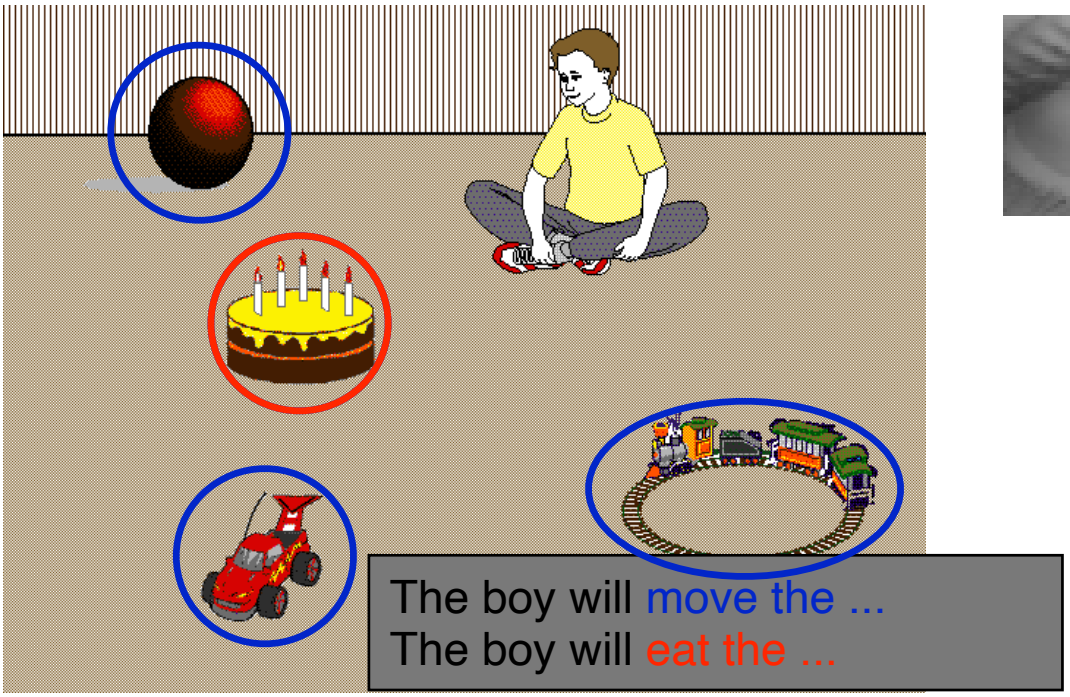
Lexical Prediction?

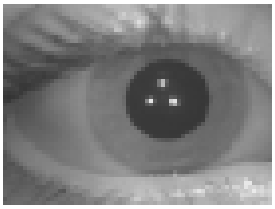


Evidence for On-line Prediction

- Many reading studies demonstrate how different aspects of syntactic and semantic context influence the reading times or ERPs for words.
  - But these are measured *on* the word of interest.
  - Mostly only offering indirect evidence of prediction.
- Is there some way to determine what people might be predicting, before they encounter a word?
  - YES! The visual world paradigm.

Parsing as Prediction

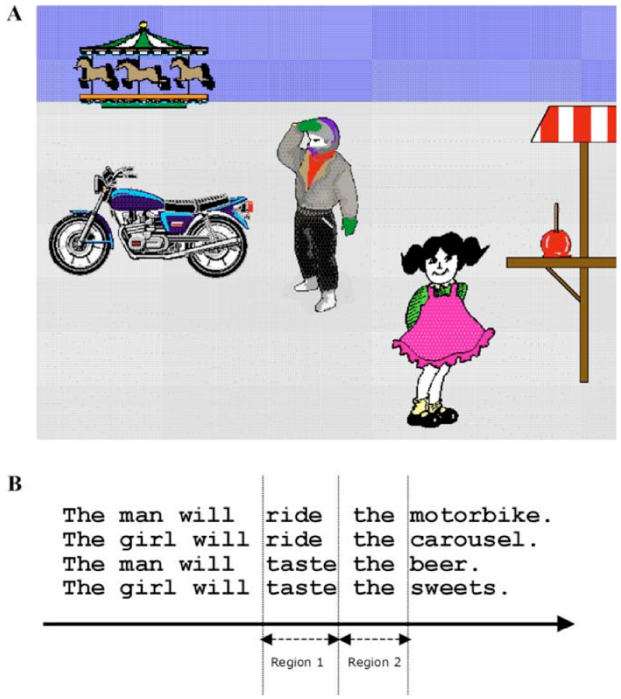


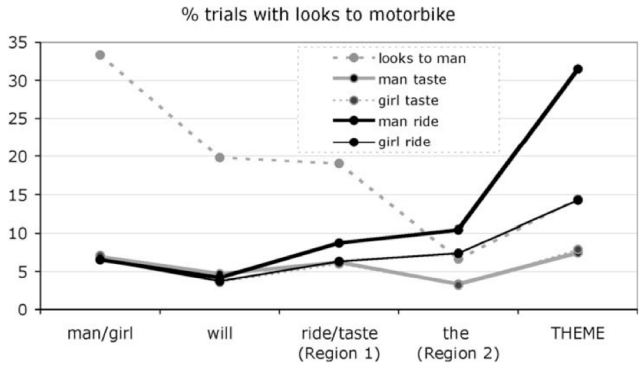


But hang on a second ..

- Is this really “prediction”?
- What kind of experiments might be more convincing to address these doubts?
- Can we use the paradigm to investigate other kinds of prediction?
- Even if it is prediction, is it limited to, or even determined by the visual context?

Compositional Prediction



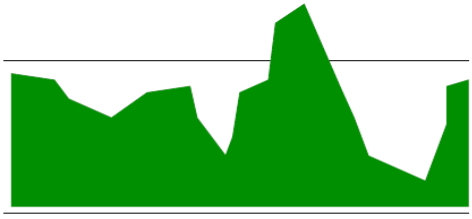


Rational Communication

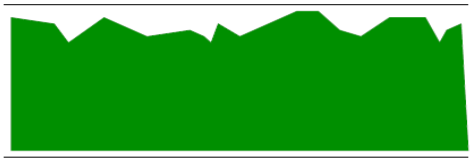
- Linguistic forms are being reduced/expanded at all linguistic levels
- Variation enables speakers to modulate the rate and linearization of message transmission
  - Evidence: Word length, speech, reading times
- Rational communication systems:
  - How is information communicated optimally?
  - Are speakers adapted to listeners constraints?

Hypotheses

- Rational language use is shaped by general information theoretic principles
  - There is an upper bound on the amount of information: Channel Capacity
  - Language users prefer to distribute information uniformly over a message
  - Variation in encoding serves to modulate information density
- Production choices are modulated by predictability:
  - Expand of high surprisal expressions, reduce predictable ones



bad use of channel
ID very variable



good use of channel
ID uniformly distributed

Information Density

$$Information(event) = \log_2 \frac{1}{P(event)}$$

$$= \log_2 \frac{1}{P(w_1)} + \log_2 \frac{1}{P(w_2 \mid w_1)} + \ldots + \log_2 \frac{1}{P(w_n \mid w_1 \ldots w_{n-1})}$$

- Uniform Information Density:
  - Maximizes information transmission
  - Avoids comprehender difficulty

Example: that-omission

- The complementizer “that” is optional in English:

  My boss confirmed (that) I am absolutely crazy.

- Uniform Information Density: Use of overt “that” increases with ID at onset of the CC “I ...”

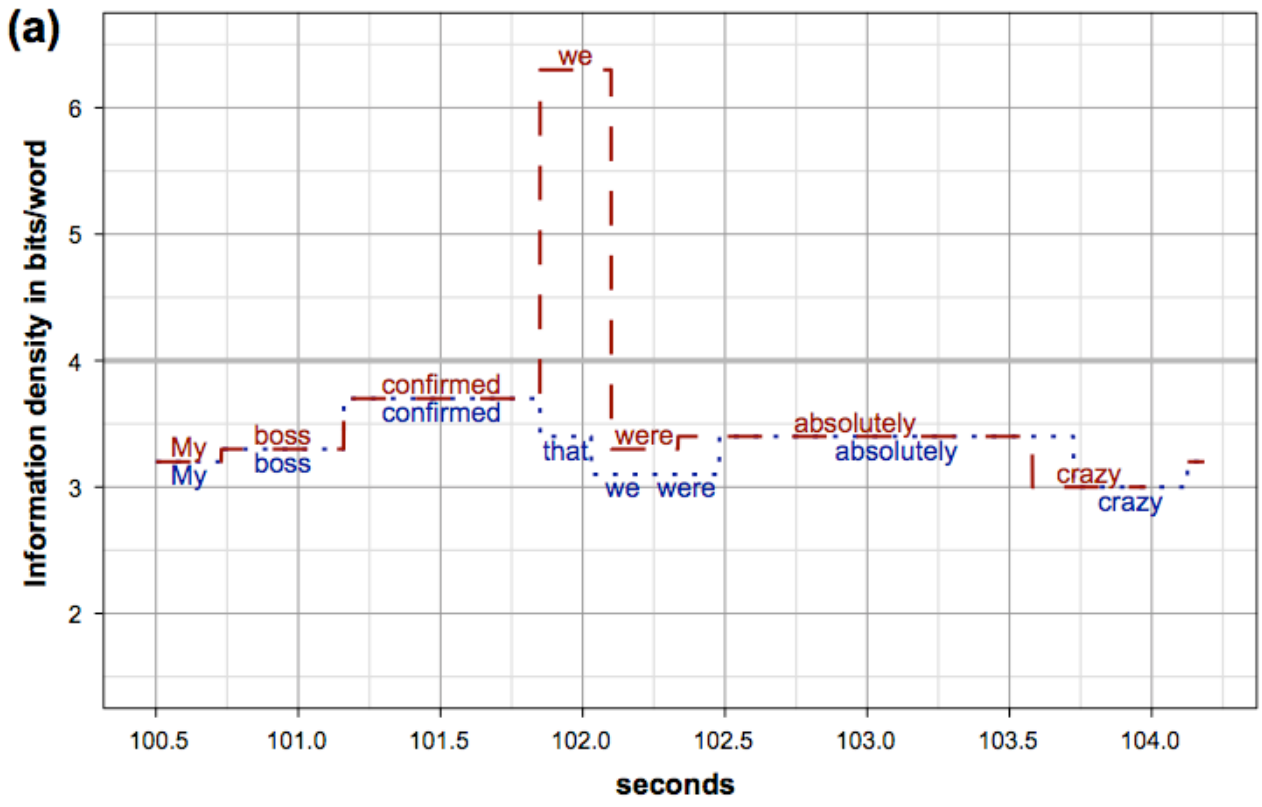
Overt *that*

$$= \log_2 \frac{1}{P(w_1 \mid CC, that, w_{\ldots -1})}$$

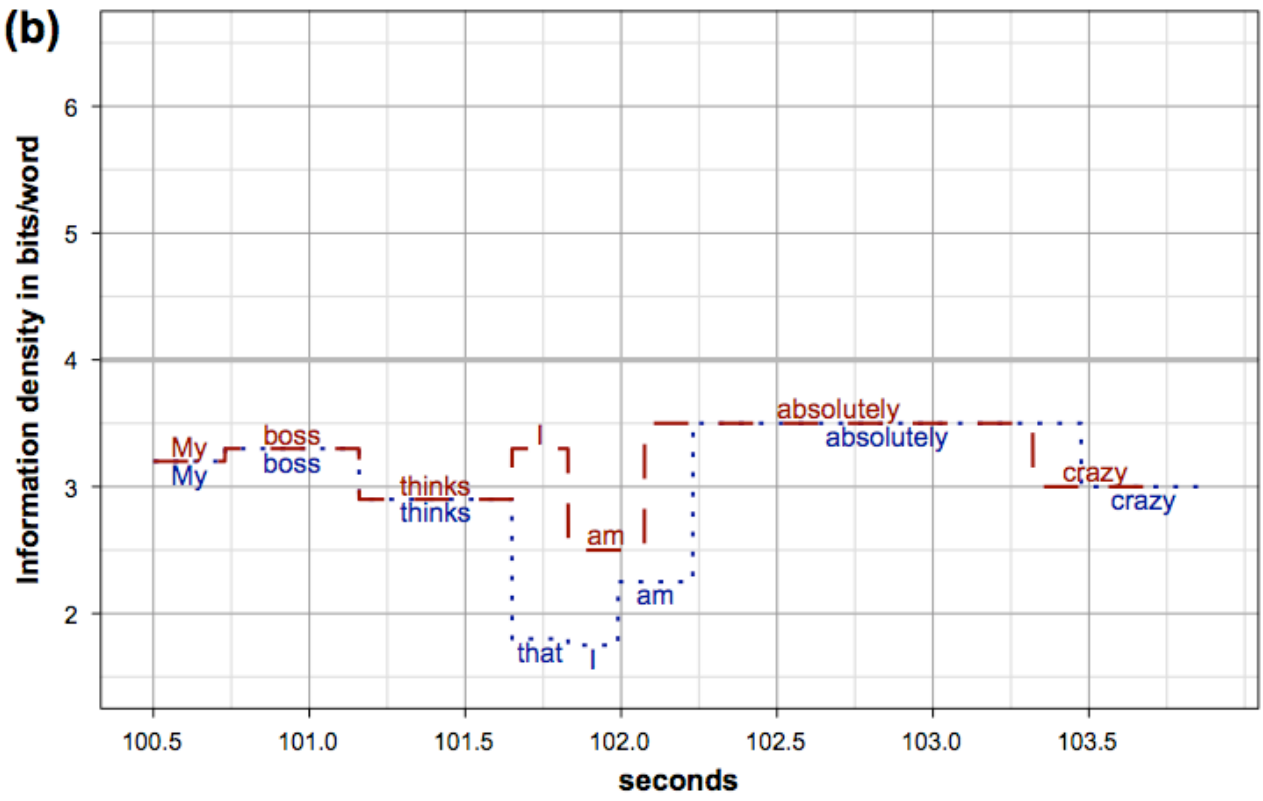
Omitted *that*

$$= \log_2 \frac{1}{P(CC \mid w_{\ldots -1})} + \log_2 \frac{1}{P(w_1 \mid CC, w_{\ldots -1})}$$

Jaeger, 2010



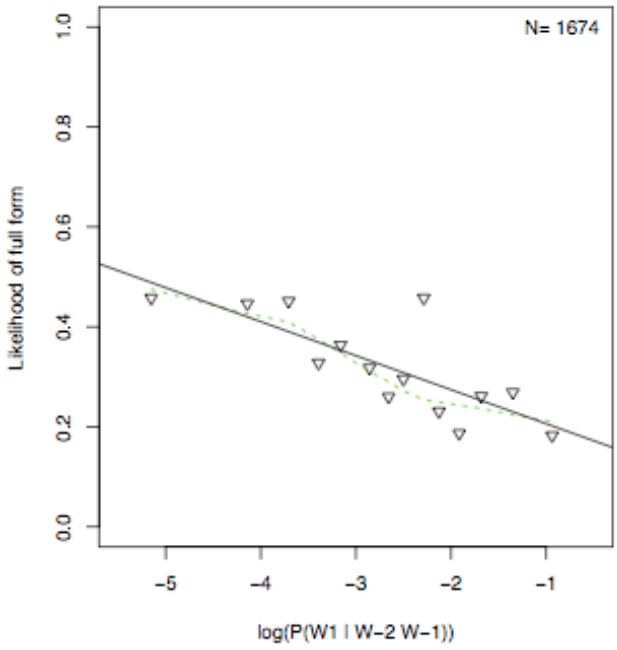
Jaeger, 2010



Jaeger, 2010

Example: *that*-omission

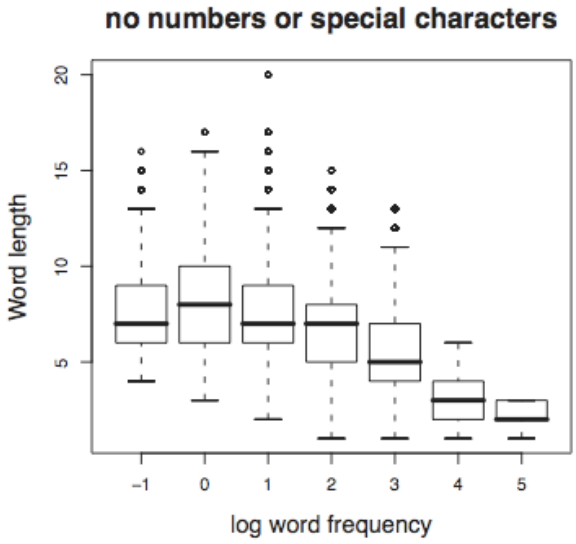
- N-gram estimates of ID predicted use of “that”
- Additionally, evidence that purely structural ID also predicts use of “that”



Levy & Jaeger, 2007

Natural language lexica

- It's long been known that word length correlates with word frequency: frequent words are generally shorter (Zipf)



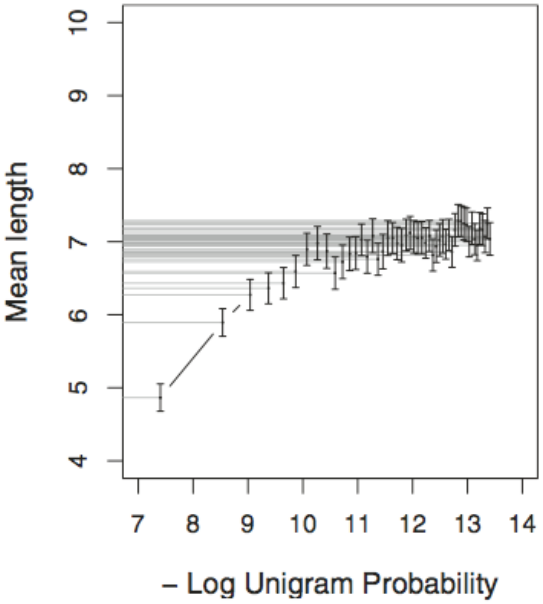
- If lexica are optimized to take into account the **likelihood of words in context**, then average predictability should be a better predictor of word length:

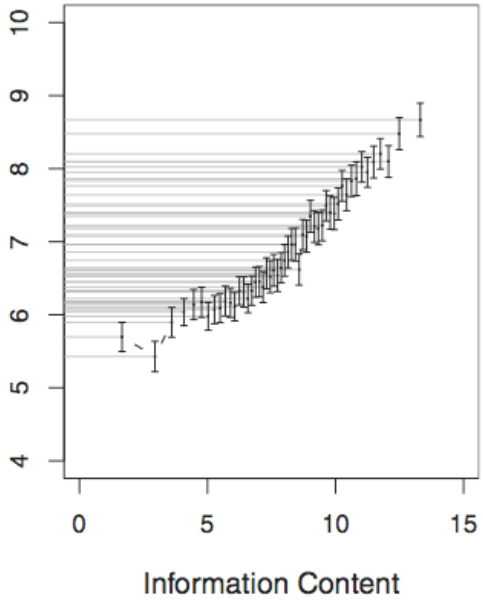
$$-\frac{1}{N}\sum_{i=1}^{N}\log P(W = w \mid C = c_i)$$

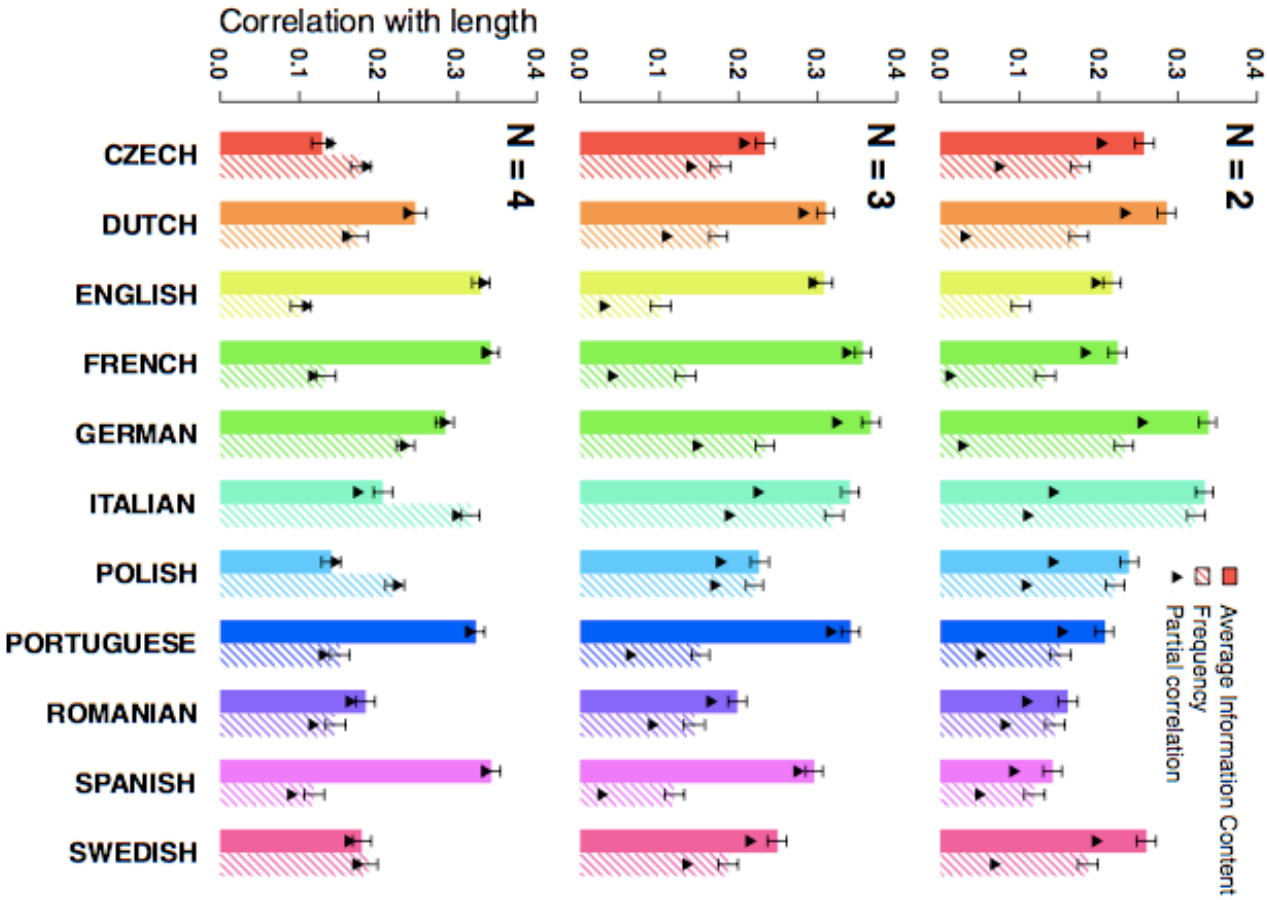
- Piantadosi, et al. *PNAS*, 2011.

Correlating Information with Length

- Higher average information content corresponds to greater length
  - And is better predictor of length then unigram

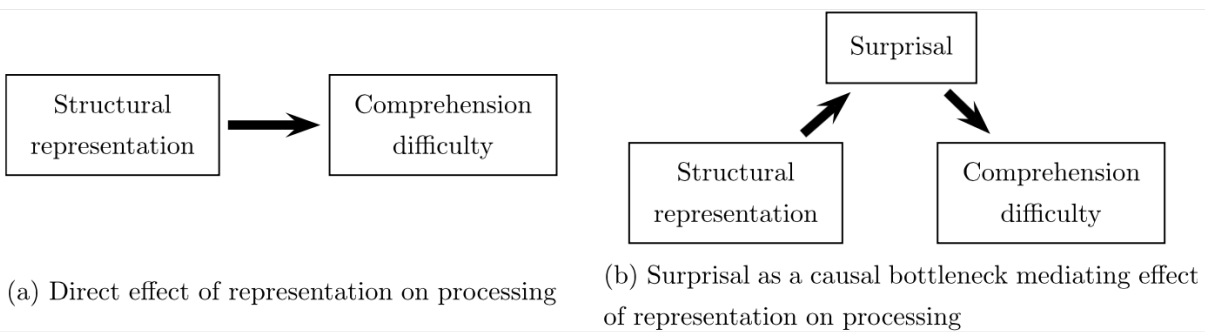






Causal Bottleneck

- Surprisal Theory assumes difficulty is determined by a word's predictability
  - Abstracts away from detailed representational or mechanistic accounts
  - Only depends on the quality of the conditional word probabilities
- If true, evidence regarding processing difficulty will shed little light on the nature of mental grammar



(a) Direct effect of representation on processing

(b) Surprisal as a causal bottleneck mediating effect of representation on processing

Information Theoretic Approaches

- Information theory offers a (linguistic) theory neutral measure of the information conveyed by linguistic events: **Surprisal**
- Surprisal also offers a good index of on-line lexical and syntactic ***comprehension*** effort, both for ambiguous and unambiguous constructions (Hale, 2001; Levy, 2008).
- Finally, evidence suggests ***speakers*** may modulate surprisal to avoid peaks (and troughs) of information (UID: Levy & Jaeger, 2007).
- The average surprisal of a word has been shown to correlate with word length, suggesting ***lexica*** have "evolved" towards an optimised encoding
  - predictable words are shorter

No lecture on January 6th, but read:

YOU CAN'T PLAY 20 QUESTIONS WITH NATURE AND WIN:
PROJECTIVE COMMENTS ON THE PAPERS OF THIS SYMPOSIUM

Allen Newell
Carnegie-Mellon University

I am a man who is half and half. Half of me is half distressed and half confused. Half of me is quite content and clear on where we are going.

-> Surprisal Tutorial on January 8th

$$y = \frac{\log_e\left(\frac{x}{m} - sa\right)}{r^2}$$

$$yr^2 = \log_e\left(\frac{x}{m} - sa\right)$$

$$e^{yr^2} = \frac{x}{m} - sa$$

$$me^{yr^2} = x - msa$$

$$me^{rry} = x - mas$$