Surprisal Theory and Empirical Evidence

Matthew W. Crocker
Vera Demberg

Block Course – Summer Semester 2015

Comprehension

- People understand language incrementally, integrating each word into an unfolding interpretation
- Ambiguity means there may be multiple interpretation for a the current initial substring
- Traditionally, psycholinguistics has tried to identify the parsing mechanisms by looking are relative processing difficulty in cases of ambiguity

Processing Difficulty

- Working memory:
  - *The mouse the cat the dog bit chased died.*
- Parse ambiguity and reanalysis:
  - *The horse raced past the barn fell*
- These interact with lexical and semantic aspects:
  - Word frequency, sentence plausibility, subcategorization preferences

Surprisal Theory

- We can measure the information conveyed by any given linguistic event (e.g. phoneme, word, utterance) encountered 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)

The Claim

- Surprisal is intended as high-level theory
    - a linking hypothesis that relates parsing to observed processing behaviour
- Subsumes many of the individual explanations of processing difficulty
- Is grounded in the principles of information theory, providing a possible *explanation* for difficulty

Surprisal Theory

$$Effort \propto \text{Surprisal} = \log_2 \frac{1}{p(w_i | w_{1 \ldots i-1})}$$

- Different kinds of probabilistic language models:

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

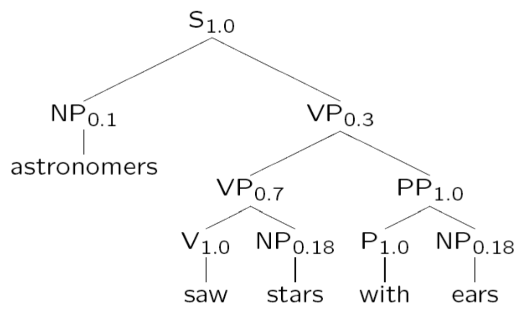
- N-gram surprisal:

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

- But n-grams don't model comprehension!

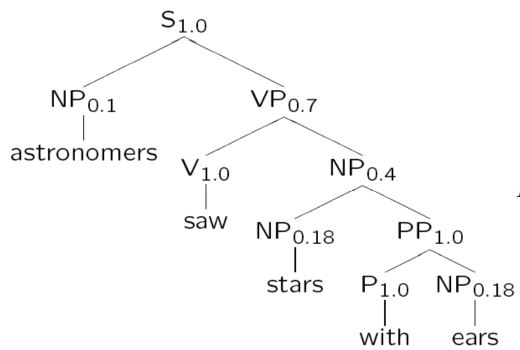
S → NP VP	1.0	NP → NP PP	0.4
PP → P NP	1.0	NP → astronomers	0.1
VP → VP NP	0.7	NP → ears	0.18
VP → VP PP	0.3	NP → saw	0.04
P → with	1.0	NP → stars	0.18
V → saw	1.0	NP → telescopes	0.1

t_2:



$$P(t_1) = 1.0 \times 0.1 \times 0.3 \times 0.7 \times 1.0 \times 0.18 \times 1.0 \times 1.0 \times 0.18 = 0.0006804$$

t_1:



$$P(t_1) = 1.0 \times 0.1 \times 0.7 \times 1.0 \times 0.4 \times 0.18 \times 1.0 \times 1.0 \times 0.18 = 0.0009072$$

Parser Surprisal

- We can also compute surprisal using probabilities recovered by a probabilistic grammar/parser:

$$\begin{aligned}
\text{Surprisal}_n &= -\log_2 P(w_n | w_1 \cdots w_{n-1}) \\
&= -\log_2 \frac{P(w_1 \cdots w_n)}{P(w_1 \cdots w_{n-1})} = log_2 \frac{P(w_1 \cdots w_{n-1})}{P(w_1 \cdots w_n)} \\
&= \log_2 P(w_1 \cdots w_{n-1}) - log_2 P(w_1 \cdots w_n) \\
&= log_2 \sum_T P(T, w_1 \cdots w_{n-1}) - log_2 \sum_T P(T, w_1 \cdots w_n) \\
&= prefprob_{n-1} - prefprob_n
\end{aligned}$$

Hale 2001

- Hale proposed that surprisal measures be determined by an incremental probabilistic Earley parser (Stolcke)

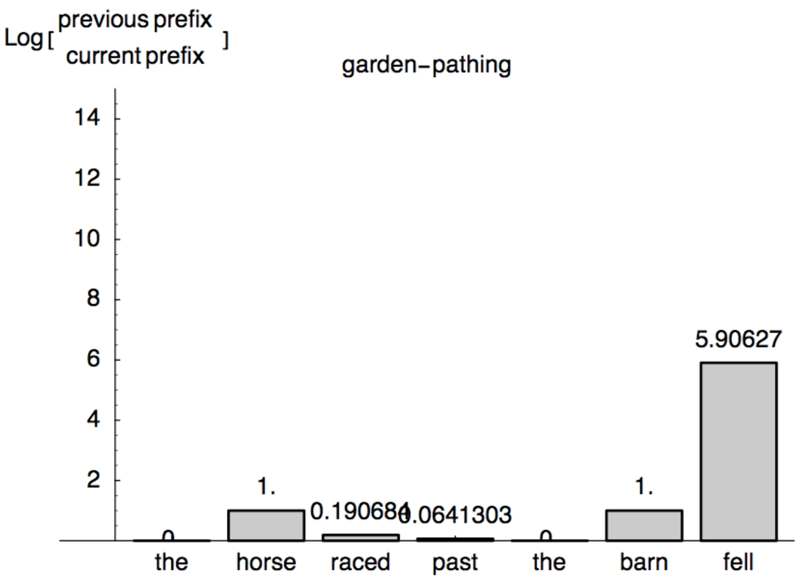
$$prefprob_n = log_2 \sum_T P(T, w_1 \cdots w_n)$$

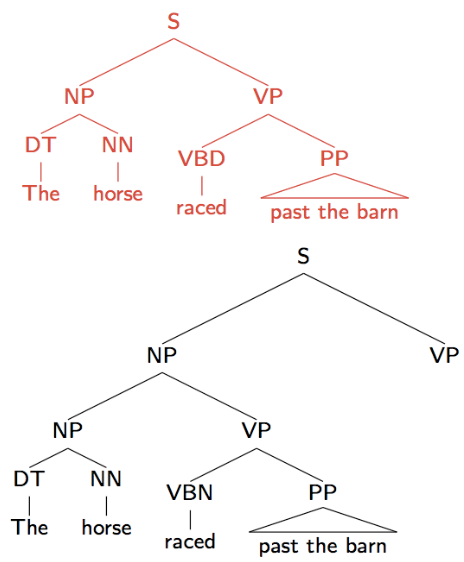
$$Surprisal_n = prefprob_{n-1} - prefprob_n$$

1.0	S	→	NP VP .
0.876404494831	NP	→	DT NN
0.123595505169	NP	→	NP VP
1.0	PP	→	IN NP
0.171428571172	VP	→	VBD PP
0.752380952552	VP	→	VBN PP
0.0761904762759	VP	→	VBD
1.0	DT	→	*the*
0.5	NN	→	*horse*
0.5	NN	→	*barn*
0.5	VBD	→	*fell*
0.5	VBD	→	*raced*
1.0	VBN	→	*raced*
1.0	IN	→	*past*

Reduced Relatives

The horse raced past the barn fell.





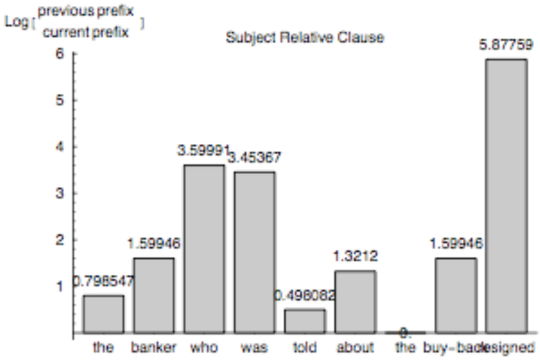


Figure 4: Mean 10.5

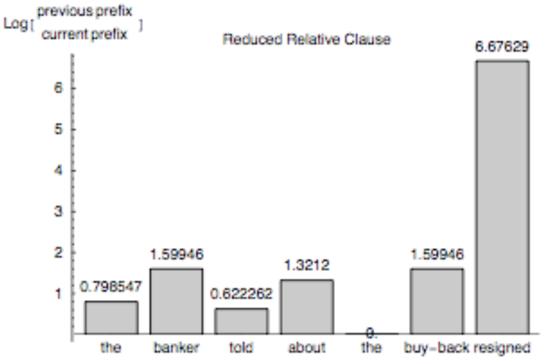
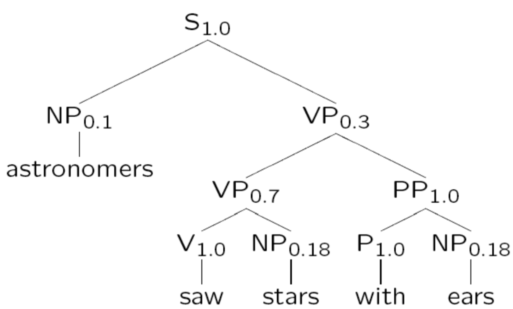


Figure 5: Mean: 16.44

0.574927953937	S	→	NP VP
0.425072046063	S	→	VP
1.0	SBAR	→	WHNP S
0.80412371161	NP	→	DT NN
0.082474226966	NP	→	NP SBAR
0.113402061424	NP	→	NP VP
0.11043	VP	→	VBD PP
0.141104	VP	→	VBD NP PP
0.214724	VP	→	AUX VP
0.484663	VP	→	VBN PP
0.0490798	VP	→	VBD
1.0	PP	→	IN NP
1.0	WHNP	→	*who*
1.0	DT	→	*the*
0.33	NN	→	*boss*
0.33	NN	→	*banker*
0.33	NN	→	*buy-back*
0.5	IN	→	*about*
0.5	IN	→	*by*
1.0	AUX	→	*was*
0.74309393	VBD	→	*told*
0.25690607	VBD	→	*resigned*
1.0	VBN	→	*told*

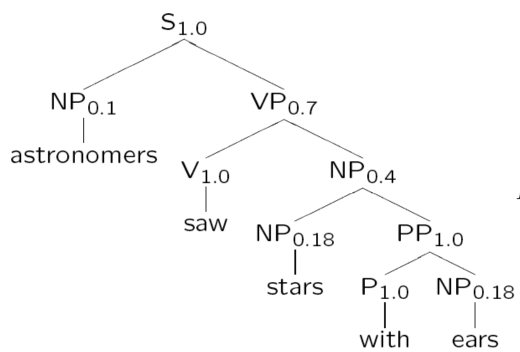
S → NP VP	1.0	NP → NP PP	0.4
PP → P NP	1.0	NP → astronomers	0.1
VP → VP NP	0.7	NP → ears	0.18
VP → VP PP	0.3	NP → saw	0.04
P → with	1.0	NP → stars	0.18
V → saw	1.0	NP → telescopes	0.1

t_2:



$P(t_1) = 1.0 \times 0.1 \times 0.3 \times 0.7 \times 1.0 \times 0.18 \times 1.0 \times 1.0 \times 0.18 = 0.0006804$

t_1:



$P(t_1) = 1.0 \times 0.1 \times 0.7 \times 1.0 \times 0.4 \times 0.18 \times 1.0 \times 1.0 \times 0.18 = 0.0009072$

Unambiguous example

- It is well known that subject relative clauses are processed more easily than object relatives:

The reporter who attacked the senator $<^{easier}$
The reporter who the senator attacked

Refining Surprisal

- Levy (2008) further develops Surprisal Theory, and proves that:

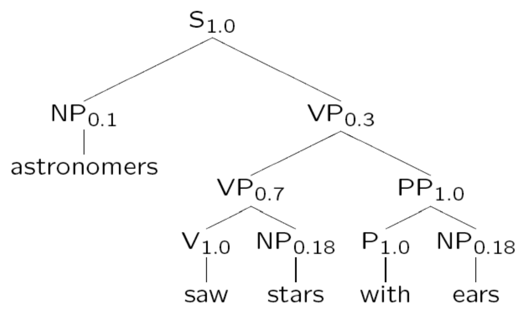
$$\text{Surprisal}_{k+1} = D(P_{k+1} || P_k) = -\log_2 p(w_{k+1} | w_{1 \ldots k})$$

- Conceptually: Surprisal reflects the change in the probability distribution over the possible parses of the input.

- Thus Surprisal simultaneously explains the cost of revising beliefs about the preferred parse, as well as difficulty due to a words expectancy

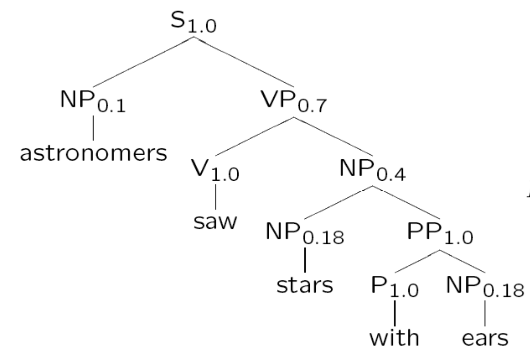
S ➜ NP VP	1.0	NP ➜ NP PP	0.4
PP ➜ P NP	1.0	NP ➜ astronomers	0.1
VP ➜ VP NP	0.7	NP ➜ ears	0.18
VP ➜ VP PP	0.3	NP ➜ saw	0.04
P ➜ with	1.0	NP ➜ stars	0.18
V ➜ saw	1.0	NP ➜ telescopes	0.1

t_2:



$P(t_1) = 1.0 \times 0.1 \times 0.3 \times 0.7 \times 1.0 \times 0.18 \times 1.0 \times 1.0 \times 0.18 = 0.0006804$

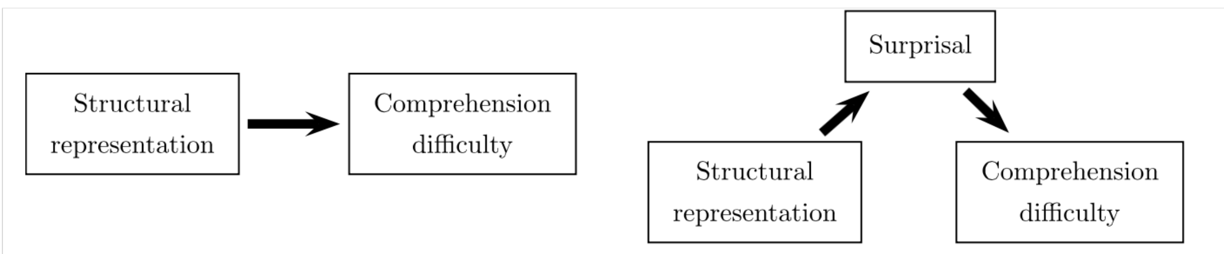
t_1:



$P(t_1) = 1.0 \times 0.1 \times 0.7 \times 1.0 \times 0.4 \times 0.18 \times 1.0 \times 1.0 \times 0.18 = 0.0009072$

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

Core Phenomena

- Predictability: predictable words are easier

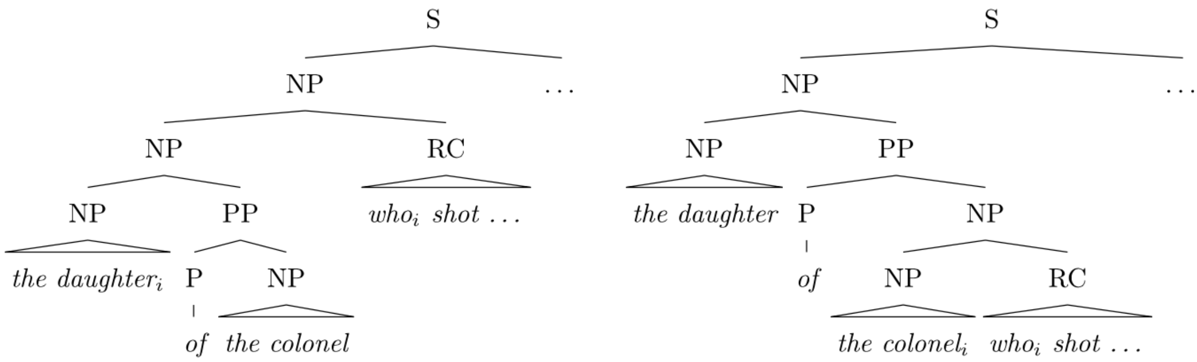
  a. He mailed the letter without a ***stamp***.
  b. There was nothing wrong with the ***car***.

- Locality: local extractions are more likely

  a. The reporter *who* attacked the senator admitted the error.
  b. The reporter *who* the senator attacked admitted the error

- Ambiguity advantage: the input doesn't lead to changes in relative entropy

  a. The daughter$_i$ of the colonel$_j$ who shot herself$_{i/*j}$ on the balcony had been very depressed.
  b. The daughter$_i$ of the colonel$_j$ who shot himself$_{*i/j}$ on the balcony had been very depressed.
  c. The son$_i$ of the colonel$_j$ who shot himself$_{i/j}$ on the balcony had been very depressed.



(a) High-attached relative clause ($\mathrm{RC_{high}}$) (b) Low-attached relative clause ($\mathrm{RC_{low}}$)

$$
\begin{aligned}
P_i(himself) &= P_i(\mathrm{RC}_{low})P_i(himself|\mathrm{RC}_{low}) \\
&+ P_i(\mathrm{RC}_{high})P_i(himself|\mathrm{RC}_{high})
\end{aligned}
$$

Why is himself easier for:
The son of the colonel who shot …

$$\text{Effort} \propto \text{Surprisal} = \log_2 \frac{1}{p(w_i | w_{1\ldots i-1})}$$

- Surprisal theory claims that predictable words will be easier to process, due to either of the following mechanisms:
  - Prediction: comprehenders actively predict what comes next
  - Integration: is it easier to integrate incoming words that fit the preceding context
- What aspects of the context – lexical, syntactic, semantic, conceptual – are used for prediction, or to facilitate integration?
- What kinds of experimental measure index these processes?

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
- Cloze indexes predictability, but may not tell us much about how readers might actually predict upcoming words on-line

Cloze and Reading

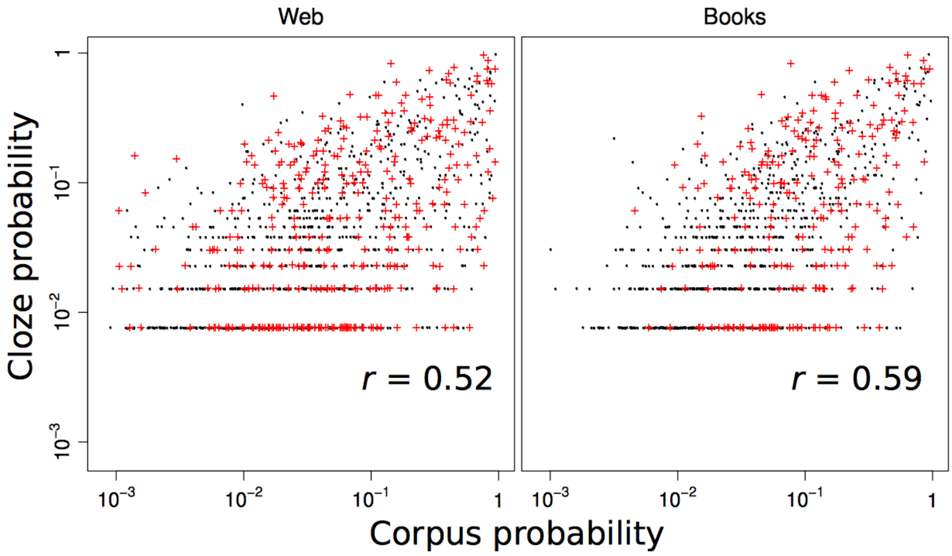
- Rayner & Well (1996) directly investigated the influence of contextual constraints on reading

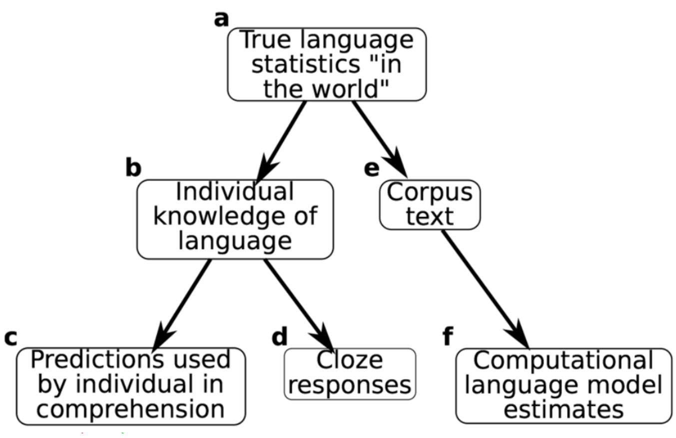
(a) The woman took the warm cake out of the <u>oven</u>. (high – 93%)
(b) The woman took the warm cake out of the <u>stove</u>. (med – 33%)
(c) The woman took the warm cake out of the <u>pantry</u>. (low – 3%)

- 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.

Cloze vs. Corpora

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





- Cloze significantly predicted reading times

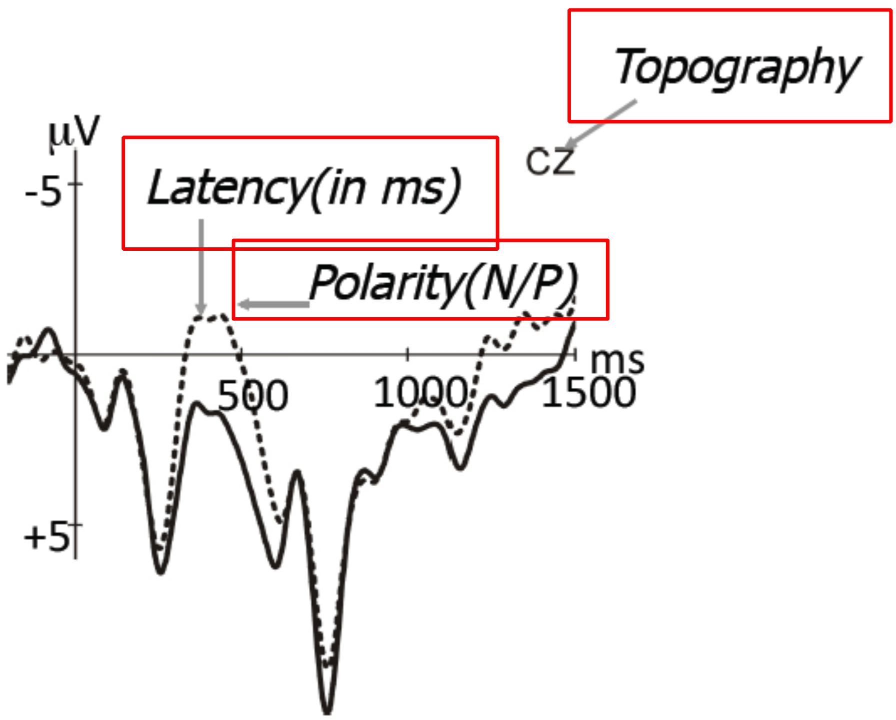
- Corpus-based probability estimates did not

- How probabilities contribute to human predictions and reading times is not yet clear

On-line Measures

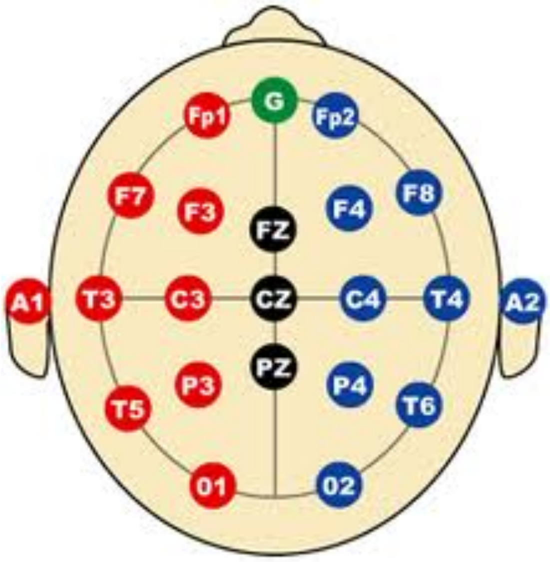
- Reading times are known to reflect processing difficulty due to lexical, syntactic and semantic factors … more on this later.
- Event-related potentials are a neurophysiological measure that indexes processes of lexical retrieval (N400) and integration (P600)
- The visual world paradigm.

ERP Components



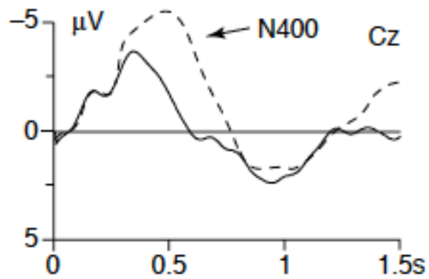
Topographical distribution

- **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



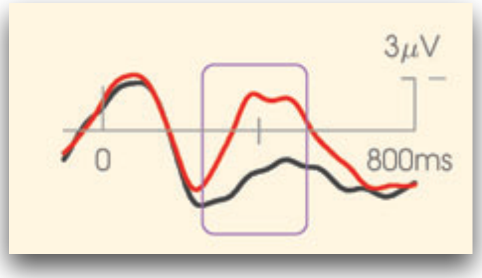
The N400

- Negative deflection peaking around 400ms after stimulus onset



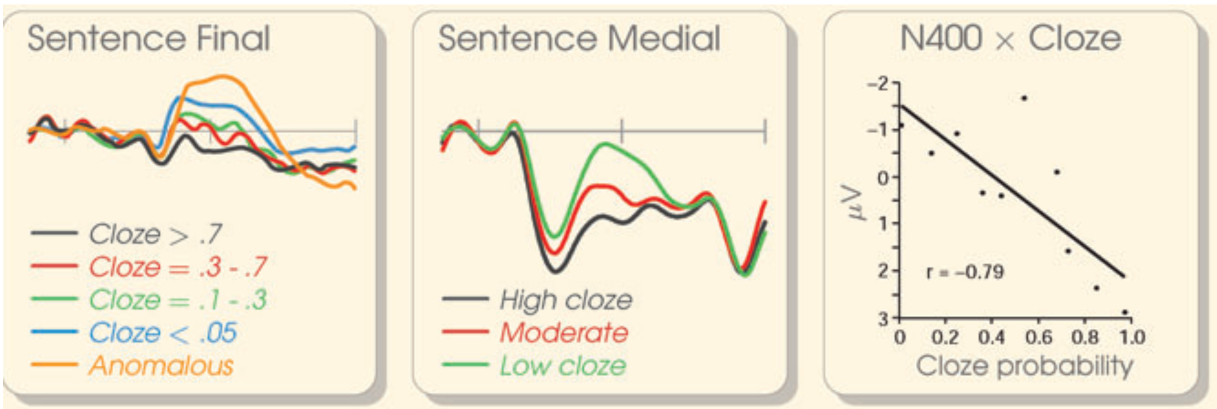
- Maximal over centro-posterior sites, bilateral
- Discovered by Kutas and Hillyard in the early 80s

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 confirmed predictions, or facilitated 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

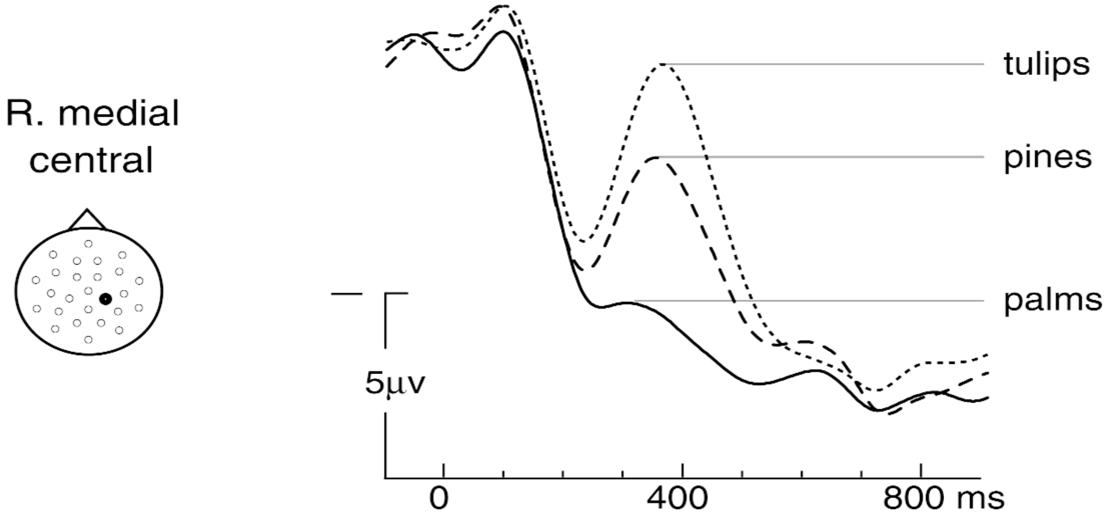
palms / pines / tulips
[tree] / [tree] / [flower]

Unexpected within-category violation
Unexpected between-category violation

Federmeier & Kutas (1999)

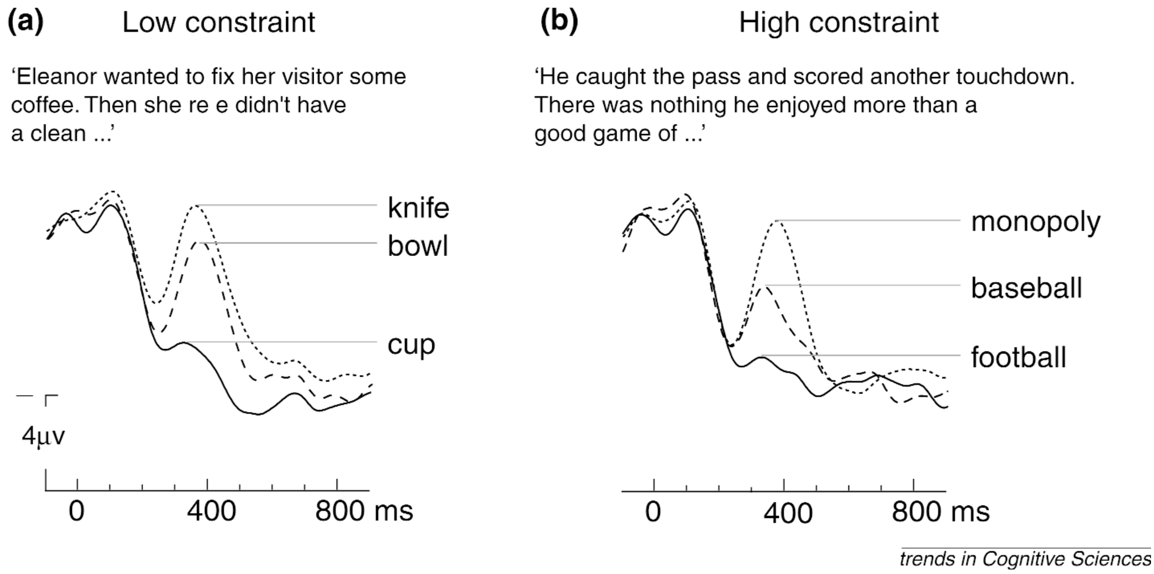
Results

'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 language processor pre-activates semantic features of the expected word
- 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”

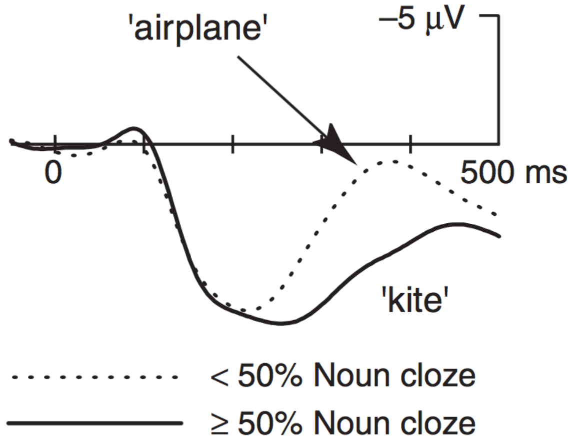
Delong, Urbach & Kutas, *Nature Neuroscience*, 2005

Lexical Prediction?

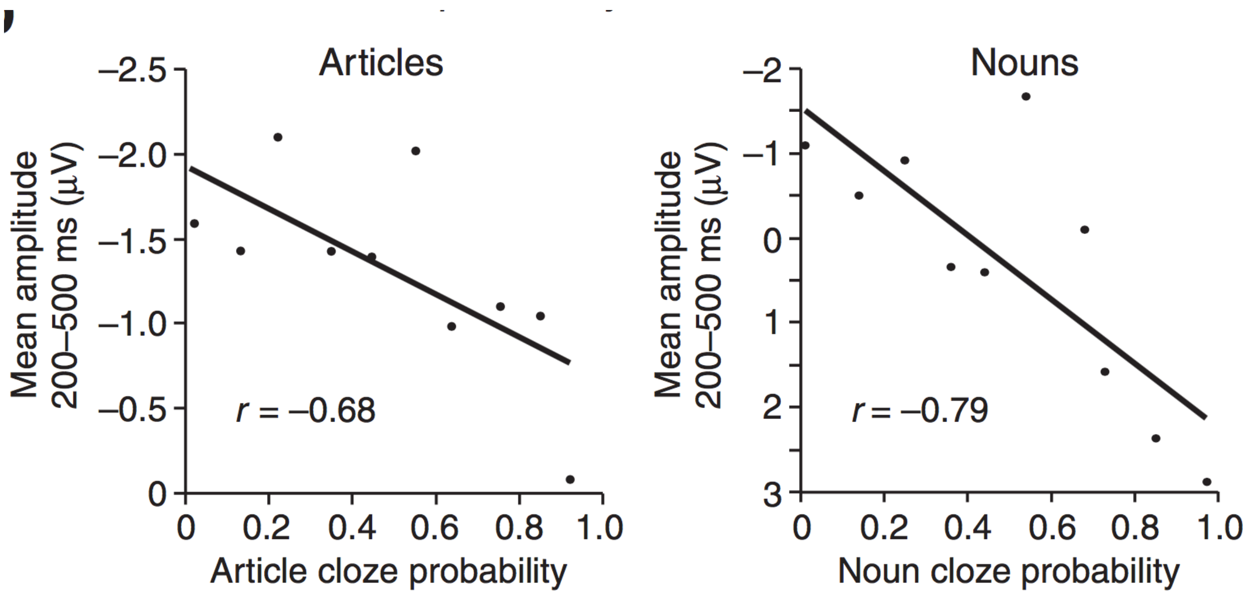
I

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

Nouns



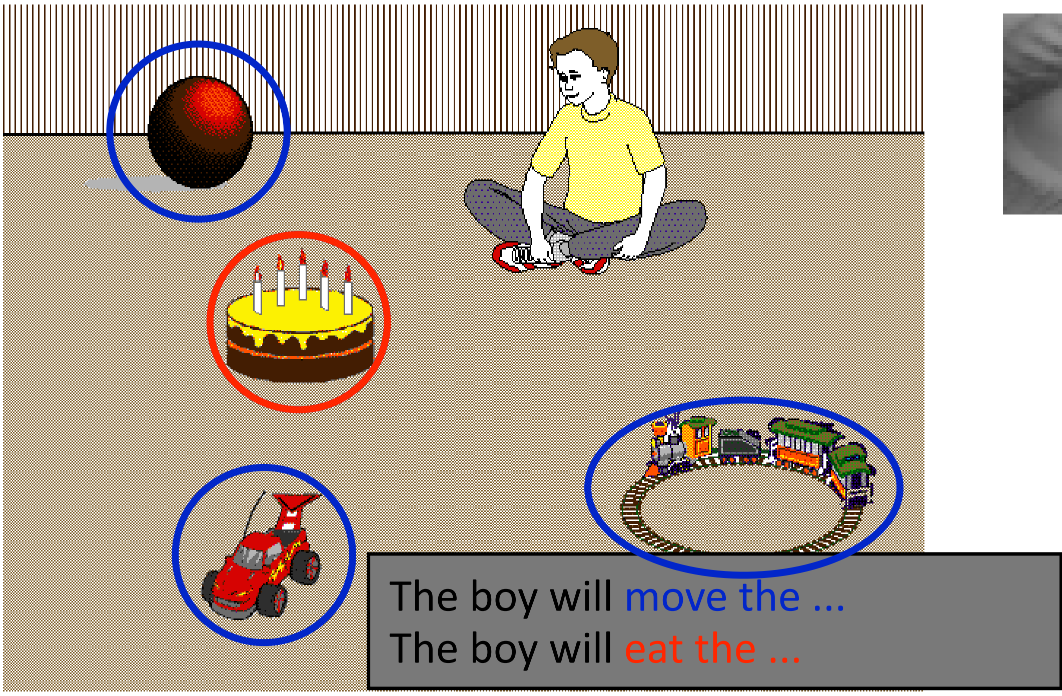
Lexical Prediction?

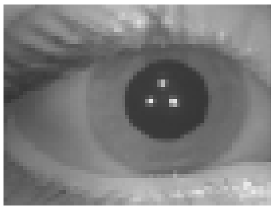


Evidence for On-line Prediction

- 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

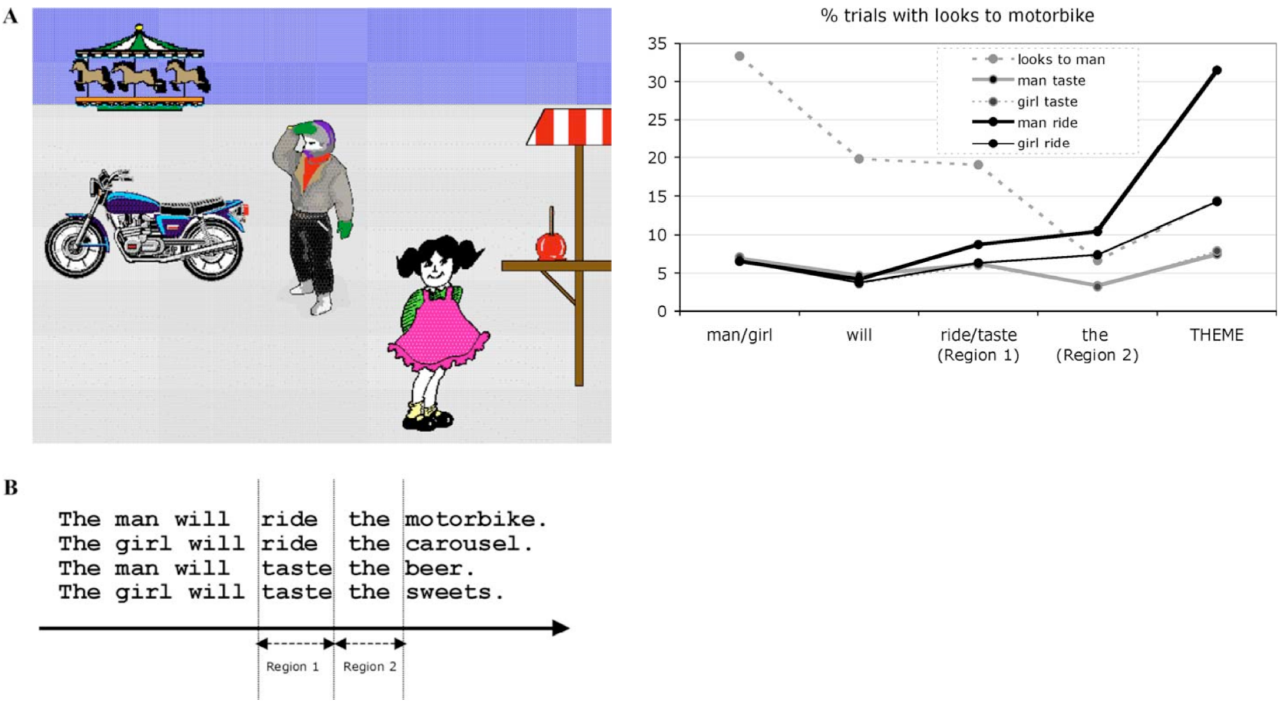




But hang on a second ..

- 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



Experimental Measures

- Reading times, N400, visual attention clearly index surprisal, but not perfectly.
  - They are influenced by other factors, and sometime to a greater extent
- These are also multi-dimensional measures, and surprisal effects can manifest themselves differently in different experiments.
- Cloze appears to offer a better estimate of “human” surprisal, than corpus based estimates

Interim Summary

- Surprisal theory unifies the notions of incremental parsing and expectations into a single account

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

- Broad empirical support for both aspects:
  - cost of syntactic disambiguation
  - ease of processing expected words