Computational Psycholinguistics

Lecture 9: **Information Theoretic Approaches**

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

Summary of Informativity

- Optimal function incorporates aspects of earlier models:
  - Basic cognitive limitations: serial interpretation + reanalysis
  - Maximising success of reaching correct interpretation

$$P(\text{global success}) = \prod_{i=1}^{n} P(\text{success at } L_i)$$

$$I(H_i) = P(H_i) \cdot S(H_i)$$

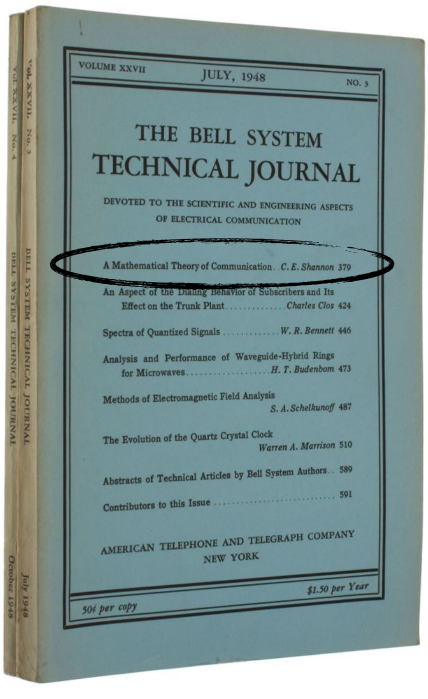
$$S(H_i) = \frac{1}{P(\text{Confirm } H_i)}$$

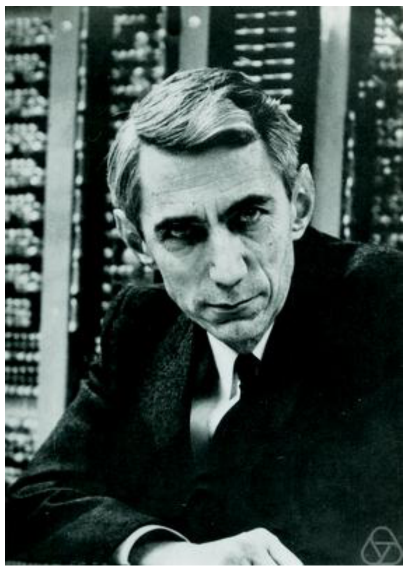
- Explains why people don't always follow likelihood alone
  - Prefer to form testable (interpretable) dependencies
  - These can be evaluated as plausible, or trigger reanalysis *quickly*
- Informativity is an idealisation of what the HSPM should approximate

Rational Models and Linking Hypotheses

- Rational Hypothesis 1: $\arg\max_i P(s_i)$ for all $s_i \in S$
- Rational Hypothesis 2: $\arg\max_i P(s_i) \cdot S(s_i)$ for all $s_i \in S$
- Implementing and evaluating more plausible “optimal functions”:
    - More linguistically informed probabilistic models (lexical, semantic ...)
    - Integration with non-probabilistic factors (recency, memory load)
- Richer linking functions between parser and human processing measures
    - Relate the parsing mechanisms to observed processing difficulty, i.e. reading measures, event-related potentials, fMRI

1948





Claude Shannon

Information Theoretic Approaches

- We can think of language as a communication system, in which information is transmitted from speaker to hearer
- Rationality suggests that language, and language use, will be optimized to transmit information as efficiently as possible (speaker) while taking into account cognitive limitations of the hearer.
- The average amount of information conveyed by a linguistic unit
    - Uncertainty of a random variable is measured by its *entropy*
- Information Theory (Shannon)
    - Finding the best "code" for sending messages of a language

A Mathematical Theory of Communication

By C. E. SHANNON

INTRODUCTION

THE recent development of various methods of modulation such as PCM and PPM which exchange bandwidth for signal-to-noise ratio has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important papers of Nyquist[1] and Hartley[2] on this subject. In the present paper we will extend the theory to include a number of new factors, in particular the effect of noise in the channel, and the savings possible due to the statistical structure of the original message and due to the nature of the final destination of the information.

The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have *meaning*; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. 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.

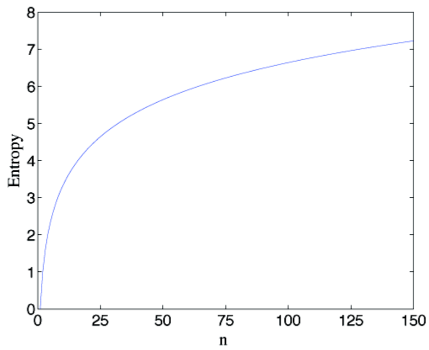
H is Entropy = …

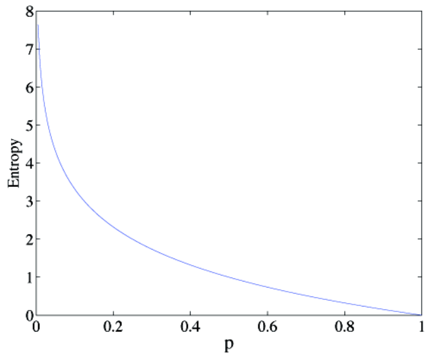
- How much information is conveyed by a particular message, event, outcome?
- The number of yes-no questions (or *bits*) required to specify the state of the system
- If n is the number of **equally** likely states of the system:

$$H = \log_2[n]$$

$$H = -\log_2\left[\frac{1}{n}\right]$$

$$H = -\log_2[p]$$





Entropy for Non-Uniform Events

- Information: for a given language
  - The number of bits needed to send a message, on average
- Optimal code for an event having probability $p(x)$ is: $\left\lceil \log_2 \frac{1}{p(x)} \right\rceil$
- The average number of bits needed to transmit a message in a language X is:
  - Entropy: $H(X) = \sum_{x \in X} p(x) \log_2 \frac{1}{p(x)}$

Example 1: 8-sided die

- Let x represent the result of rolling a (fair) 8-sided die.
- Entropy: $H(X) = \sum_{x \in X} p(x) \log_2 \frac{1}{p(x)}$

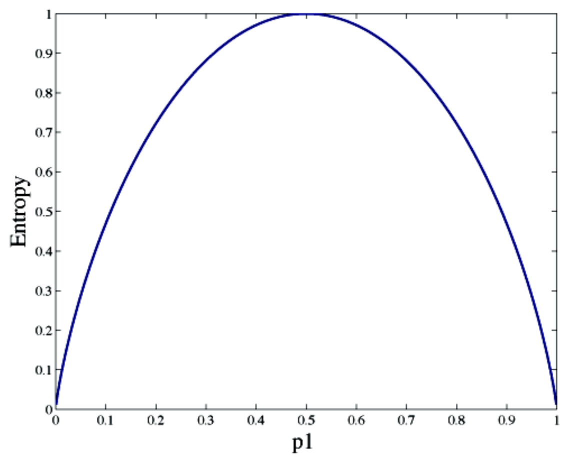
$$H(X) = \sum_{x \in X} \frac{1}{8} \log_2 \frac{1}{\frac{1}{8}} = \log_2 8 = 3$$

- The average length of the message required to transmit one of 8 equiprobable outcomes is 3 bits.

“1”	“2”	“3”	“4”	“5”	“6”	“7”	“8”
001	010	011	100	101	110	111	000

Entropy of a Weighted Coin

$$H(X) = \sum_{x \in X} p(x) \log_2 \frac{1}{p(x)}$$



- The more uncertain the result, the higher the entropy.
  - Fair coin: H(X) = 1.0
- The more certain the result, the lower the entropy.
  - Completely biased coin: H(X) = 0.0

Example: Simplified Polynesian

P	T	K	A	I	U
0,125	0,25	0,125	0,25	0,125	0,125

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x)$$

$$= -[4 \times \frac{1}{8} \log \frac{1}{8} + 2 \times \frac{1}{4} \log \frac{1}{4}]$$

$$= 2\frac{1}{2} bits$$

Example 2: Simplified Polynesian

P	T	K	A	I	U
0,125	0,25	0,125	0,25	0,125	0,125

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x)$$

$$= -[4 \times \frac{1}{8} \log \frac{1}{8} + 2 \times \frac{1}{4} \log \frac{1}{4}]$$

$$= 2\frac{1}{2} bits$$

Recall: H = log2(6)
= 2.585 bits

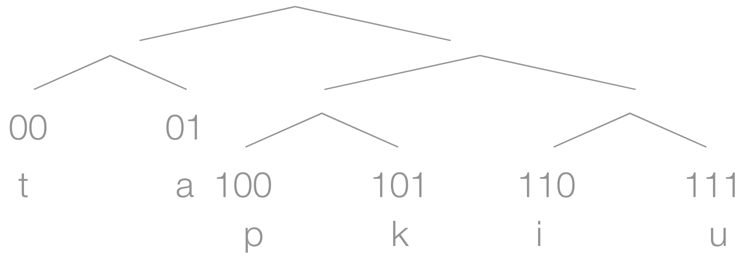
Example 2: Simplified Polynesian

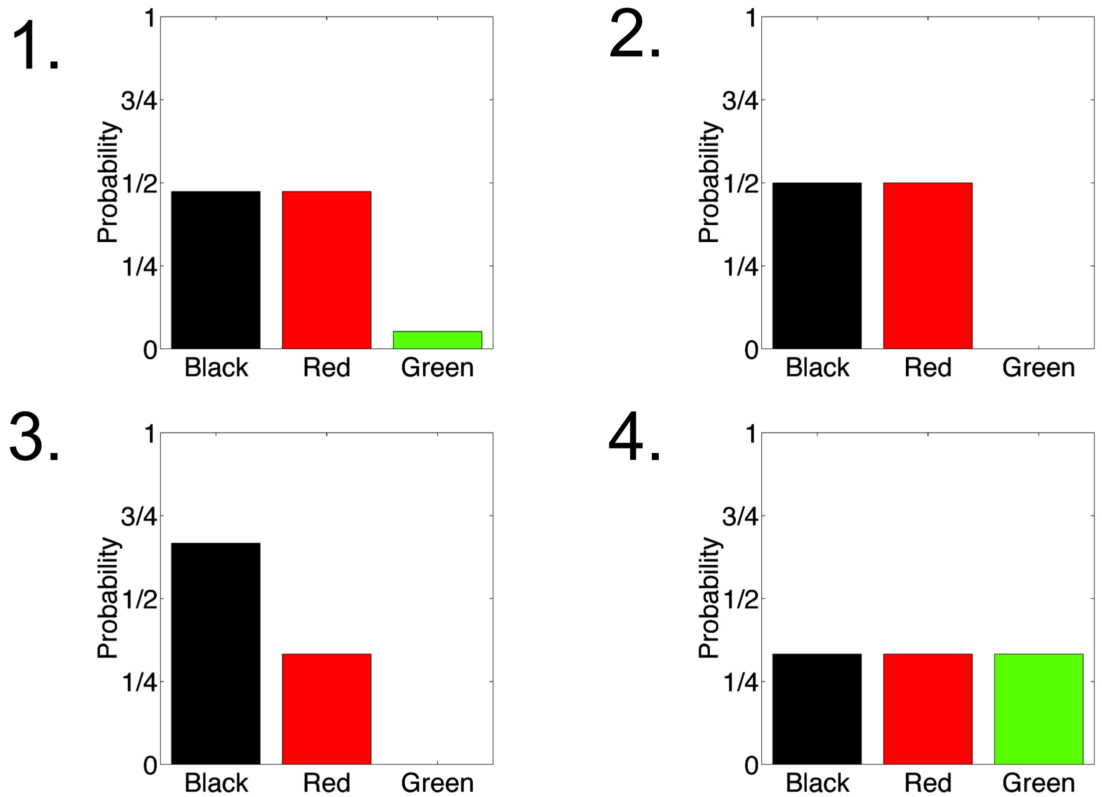
- Simplified Polynesian:

P	T	K	A	I	U
0,125	0,25	0,125	0,25	0,125	0,125

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x)$$
$$= -[4 \times \frac{1}{8} \log \frac{1}{8} + 2 \times \frac{1}{4} \log \frac{1}{4}]$$
$$= 2\frac{1}{2} bits$$

- Coding Tree:





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).

Predictability & Integration

- Surprisal theory: expected words will be easier to process:
  - their predictability reflects amount of information conveyed
- This has broad empirical support from psycholinguistics, where Cloze probability (Taylor, 1953) correlate with reading times and N400 ERPs:
  - *My brother came inside to ...* *chat? eat? play? rest?*
  - *The children went outside to ...* *chat? eat? **play**? rest?*
- Evidence of anticipatory processing is also found in visual world experiments, where people look at the visual referents of words likely to be mentioned next:
  - *The boy will eat the ...* [more looks to **cake**, than other objects]

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:

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

Parse Surprisal

- We can also show how define surprisal in terms of the probabilities recovered by a probabilistic grammar/parser:

$$\begin{aligned}
\text{Surprisal}_{k+1} &= -\log_2 P(w_{k+1} \mid w_1 \ldots w_k) \\
&= -\log_2 \frac{P(w_1 \ldots w_{k+1})}{P(w_1 \ldots w_k)} \\
&= \log_2 P(w_1 \ldots w_k) - \log_2 P(w_1 \ldots w_{k+1}) \\
&= \log_2 \sum_T P(T, w_1 \ldots w_k) - \log_2 \sum_T P(T, w_1 \ldots w_{k+1}) \\
&= prefprob_{w_k} - prefprob_{w_{k+1}}
\end{aligned}$$

Hale 2001

- Hale proposed that surprisal measures determined by an incremental probabilistic Earley parser offer a psychologically plausible index of effort.

$$prefprob_{w_n} = -\log_2 \sum_T p(T \mid w_1 \ldots w_n)$$

$$\text{Surprisal}_{w_n} = prefprob_{w_{n-1}} - prefprob_{w_n}$$

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*

Hale 2001

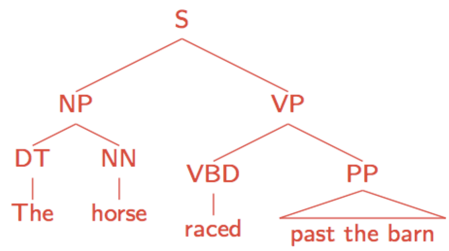
- Hale proposed that surprisal measures determined by an incremental probabilistic Earley parser offer a psychologically plausible index of effort.

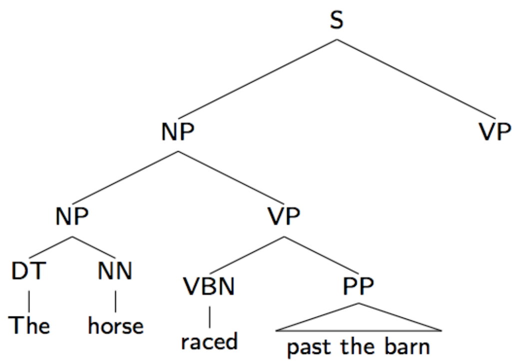
  - $$prefprob_{w_n} = \log_2 \sum_T p(T \mid w_1 \ldots w_n)$$

    $$\text{Surprisal}_{w_n} = prefprob_{w_n-1} - prefprob_{w_n}$$

- When *fell* is encountered, the higher probability parse is eliminated.

- This results in a large drop in the prefix probability as we process word *n*



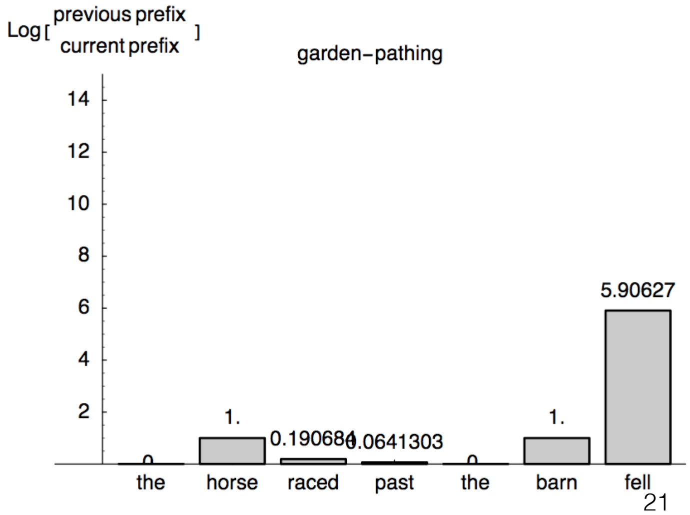


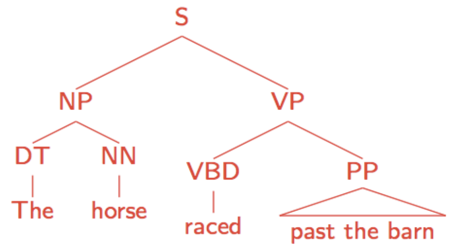
Hale 2001: Results (toy)

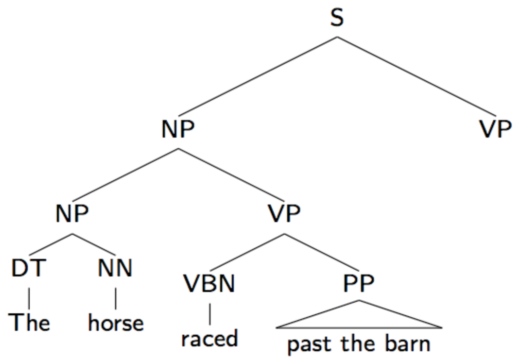
- Hale proposed that surprisal measures determined by an incremental probabilistic Earley parser offer a psychologically plausible index of effort.

$$prefprob_{w_n} = \log_2 \sum_T p(T \mid w_1 \ldots w_n)$$

$$\text{Surprisal}_{w_n} = prefprob_{w_n - 1} - prefprob_{w_n}$$







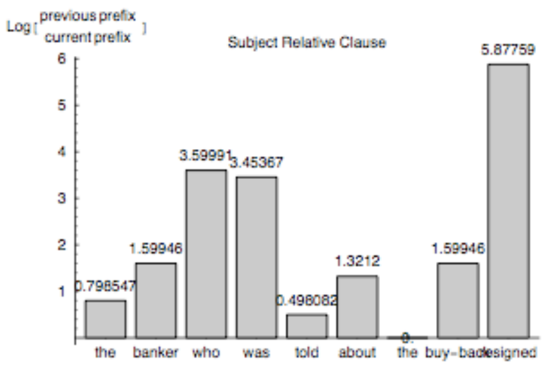


Figure 4: Mean 10.5

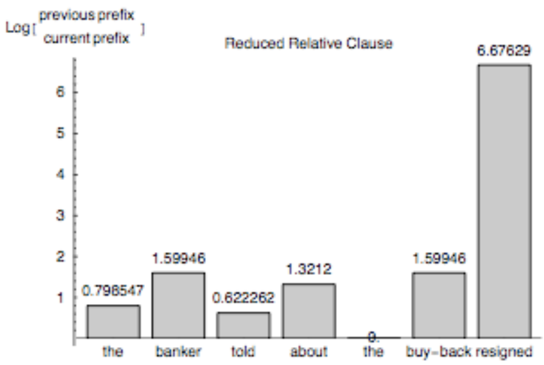


Figure 5: Mean: 16.44

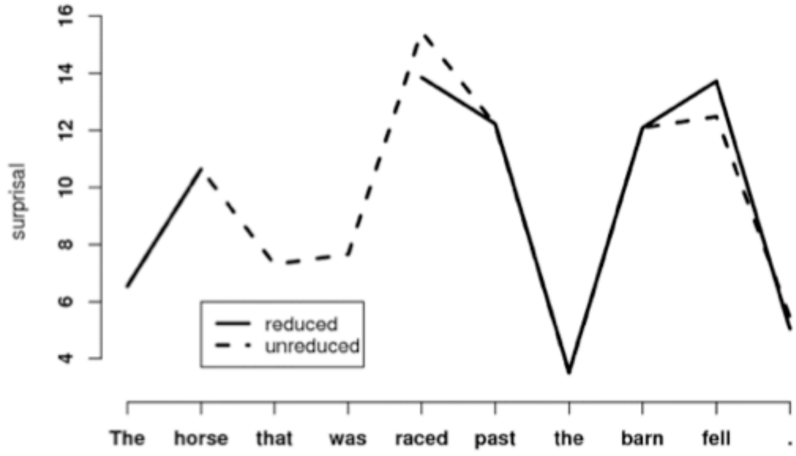
0.574927953937	S	→	NP VP
0.425072046063	S	→	VP
1.0	SBAR	→	WHNP S
0.80412371161	NP	→	DT NN
0.082474226966	NP	→	NP SBAR
0.113402061424	NP	→	NP VP
0.11043	VP	→	VBD PP
0.141104	VP	→	VBD NP PP
0.214724	VP	→	AUX VP
0.484663	VP	→	VBN PP
0.0490798	VP	→	VBD
1.0	PP	→	IN NP
1.0	WHNP	→	*who*
1.0	DT	→	*the*
0.33	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*

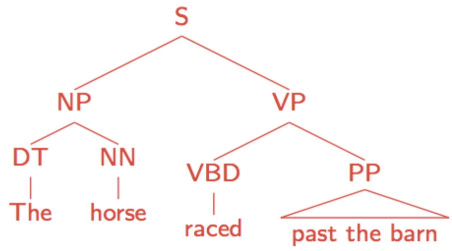
Hale 2001: Results (Brown)

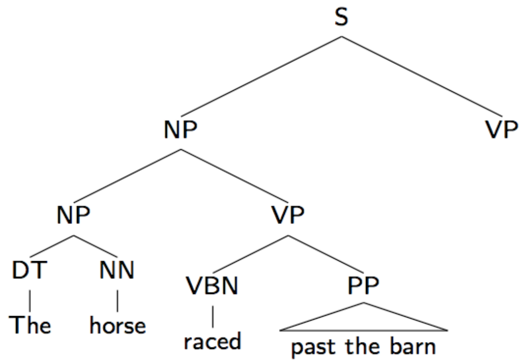
- Hale proposed that surprisal measures determined by an incremental probabilistic Earley parser offer a psychologically plausible index of effort.

$$prefprob_{w_n} = \log_2 \sum_T p(T \mid w_1 \ldots w_n)$$

$$\text{Surprisal}_{w_n} = prefprob_{w_n - 1} - prefprob_{w_n}$$



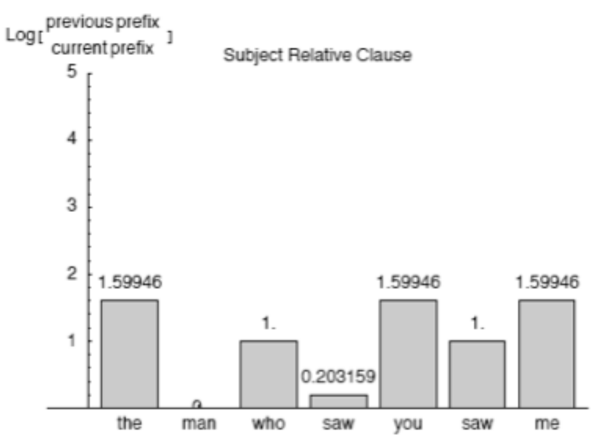


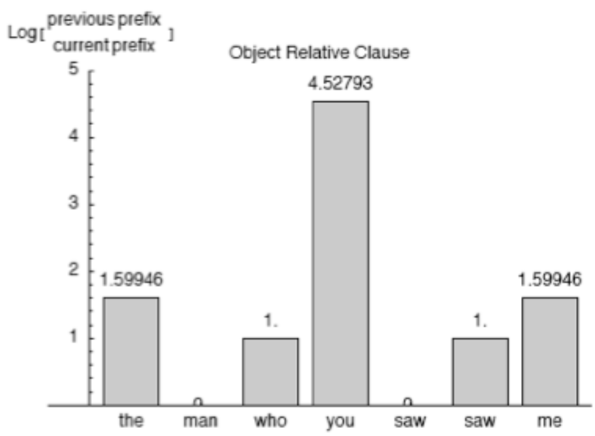


Unambiguous example

- For example, it is well known that subject relative clauses are processed more easily than object relatives:

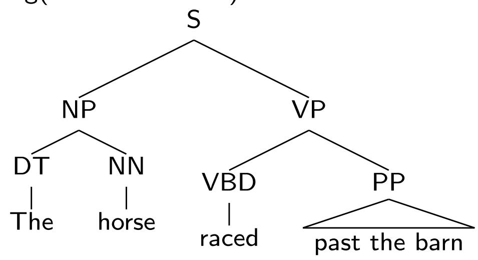
The reporter who attacked the senator $<^{easier}$
The reporter who the senator attacked



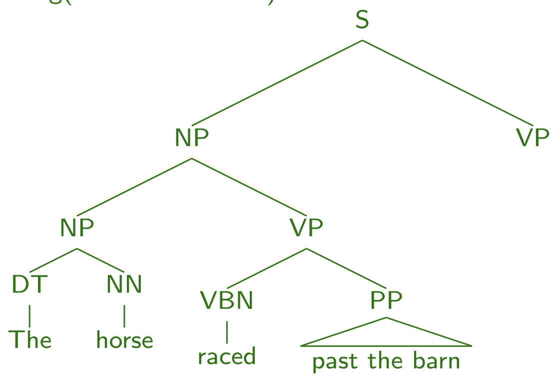


Syntactic Surprisal

$-\log(1.7766 \times 10^{-11}) = 35.712$



$-\log(1.06596 \times 10^{-15}) = 49.736$



sum of both: $pp_{w_n} = 35.712$

How to calculate surprisal:

- Calculate prefix probabilities:

$$pp_{w_n} = -\log \sum_{T \in Trees} p(T|w_1 \dots w_n)$$

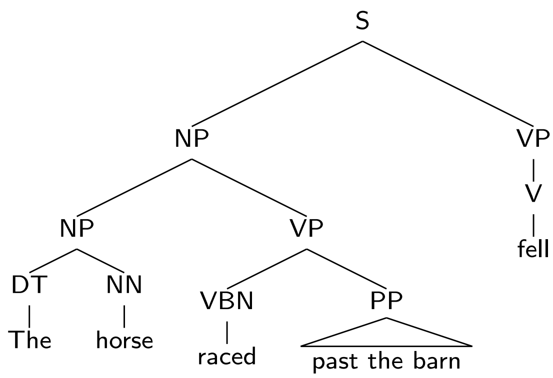
- Surprisal s of word w_n:
  $s_{w_n} = pp_{w_n} - pp_{w_{n-1}}$

Example PCFG:

Rule	**Probability of rule**
S $\rightarrow$ NP VP	$p = 0.6$
VBD $\rightarrow$ raced	$p = 0.0005$
VBN $\rightarrow$ raced	$p = 0.000001$
DT $\rightarrow$ the	$p = 0.7$

Syntactic Surprisal

$pp_{w_{n+1}} = -\log(1.06596 \times 10^{-15} \times 0.003)$
$= 58.12$



How to calculate surprisal:

- Calculate prefix probabilities:

$$pp_{w_n} = -\log \sum_{T \in Trees} p(T|w_1 \dots w_n)$$

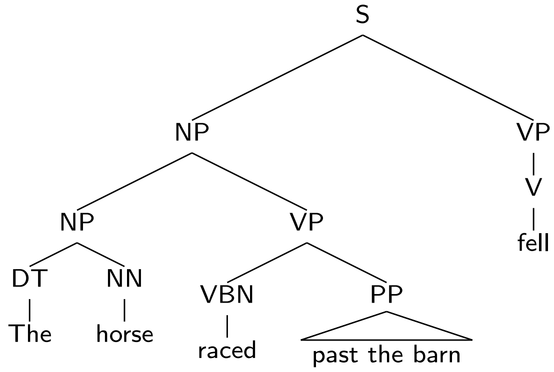
- Surprisal s of word w_n:
  $s_{w_n} = pp_{w_n} - pp_{w_{n-1}}$

Example PCFG:

Rule	**Probability of rule**
S $\rightarrow$ NP VP	$p = 0.6$
VBD $\rightarrow$ raced	$p = 0.0005$
VBN $\rightarrow$ raced	$p = 0.000001$
DT $\rightarrow$ the	$p = 0.7$

Syntactic Surprisal

$$pp_{w_{n+1}} = -\log(1.06596 \times 10^{-15} \times 0.003) = 58.12$$



$pp_{w_{n-1}} = 35.712$
$pp_{w_n} = 58.12$
$surprisal(w_n) = 22.41$

How to calculate surprisal:

- Calculate prefix probabilities:
$$pp_{w_n} = -\log \sum_{T \in Trees} p(T|w_1 \ldots w_n)$$
- Surprisal s of word w_n:
$s_{w_n} = pp_{w_n} - pp_{w_{n-1}}$

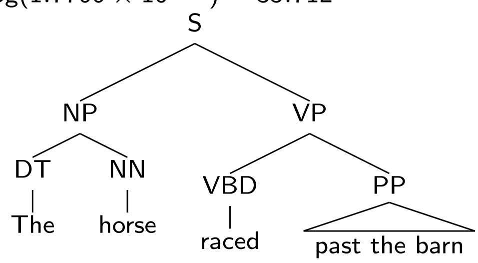
Example PCFG:

Rule	**Probability of rule**
S → NP VP	$p = 0.6$
VBD → raced	$p = 0.0005$
VBN → raced	$p = 0.000001$
DT → the	$p = 0.7$

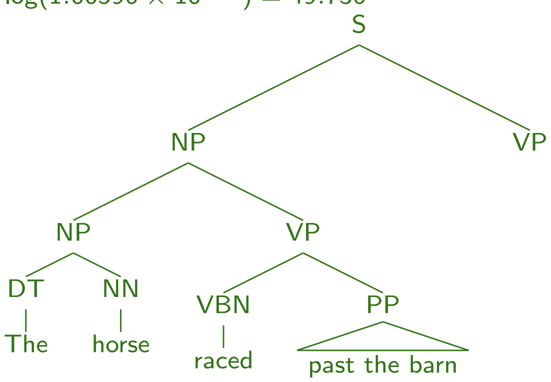
- Predictions also depend on parametrization of the grammar, training

Lexical vs. structural surprisal

$$-\log(1.7766 \times 10^{-11}) = 35.712$$

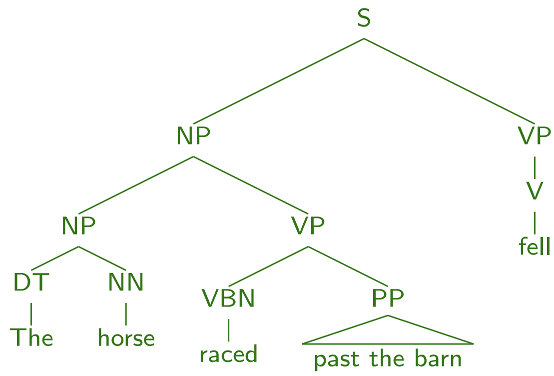


$$-\log(1.06596 \times 10^{-15}) = 49.736$$



sum of both: $pp_{w_n} = 35.712$

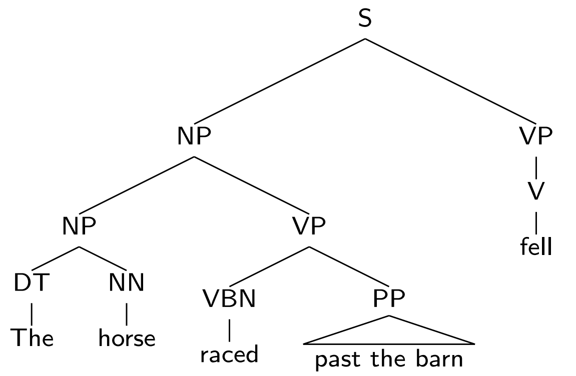
$$pp_{w_{n+1}} = -\log(1.06596 \times 10^{-15} \times 0.003) = 58.12$$



$pp_{w_{n-1}} = 35.712$
$pp_{w_n} = 58.12$
$surprisal(w_n) = 22.41$

Some of the surprisal is due to the lexical identity of *fell*, and some of it is due to the syntactic structural information conveyed by that word.

Lexical vs. structural surprisal



$$
\begin{aligned}
S_{w_n} &= -\log \sum_{T \in Trees} \frac{p(T|w_1 \dots w_n)}{p(T|w_1 \dots w_{n-1})} \\
structS_{w_n} &= -\log \sum_{\mathrm{POS}_n \in POS} \sum_{T \in Trees} \frac{p(T|w_1 \dots \mathrm{POS_n})}{p(T|w_1 \dots w_{n-1})} \\
lexS_{w_n} &= -\log \sum_{\mathrm{POS}_n \in POS} \sum_{T \in Trees} \frac{p(T|w_1 \dots w_n)}{p(T|w_1 \dots \mathrm{POS}_n)}
\end{aligned}
$$