Computational Cognitive Science

Lecture 16: Surprisal 1

Frank Keller

12 November 2026

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

Overview

Reading: Hale (2001).

Background

Garden Paths and Odds

Traditional models of human parsing (e.g., Jurafsky 1996):

- all syntactic trees for a sentence are computed in parallel and assigned probabilities;
- the probabilities are used to rank the set of trees; low probability trees are pruned (no longer considered);
- the set of trees (and their probabilities) is updated incrementally (word by word) as the input comes in;
- if the tree that turns out to be ultimately correct has been pruned, then a garden path occurs.

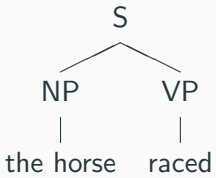
Garden Paths and Odds

Example garden path sentence:

(1) The horse raced past the barn fell.

First parse tree:

$P(\text{race}, \langle \text{agent} \rangle) = 0.92$



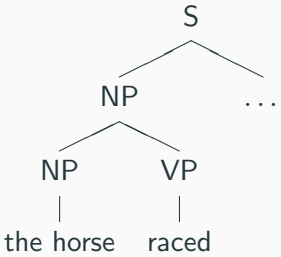
$P(t_1) = 0.92$ (preferred)

Garden Paths and Odds

Second parse tree:

$P(\text{race}, \langle \text{agent}, \ \text{theme} \rangle) = 0.08$

$\text{NP} \rightarrow \text{NP XP} \quad 0.14$



$P(t_2) = 0.0112$ (grossly dispreferred)

Garden Paths and Expectation

In the Jurafsky model, processing difficulty is caused by the ratio of the correct parse to the best parse. For example in (1):

$$\frac{P(t_1)}{P(t_2)} = \frac{0.92}{0.0112} = 82 : 1$$

Intuitively, a high ratio means that the parser has a strong expectation about the correct structure.

Maybe this expectation is based not only on two trees (most probable one and correct one) but on all trees of the sentence.

Syntactic surprisal: processing difficulty (including garden paths) occurs when the probability distribution over parse trees changes.

Expectations in Sentence Processing

There is evidence that expectation plays a role in sentence processing. For instance, when the parser sees an *either*, it expects an *or* (Staub and Clifton 2006):

(2) Peter read either a book or an essay in the school magazine.

(3) Peter read a book or an essay in the school magazine.

The region *an essay* is read faster in (2) than in (3).

The parser is surprised to see an *or* if it doesn't expect it, i.e., if there is no *either*. Surprisal leads to processing difficulty.

Expectations in Sentence Processing

Intuitively, this is also what is going on in garden paths:

(4) a. The horse raced past the barn fell.

b. The bird found in the room died.

In (4-a), the parser is surprised when it gets *fell*, as it expected the sentence to end at *barn*. In (4-b), the surprisal at *died* is lower.

(5) a. The complex houses married and single students.

b. The warehouse fires a dozen employees each year.

In (5-a), the parser is surprised when it gets *married*, as it expected a verb. In (5-b) it assumes *fires* is a verb, and is not surprised.

Beyond Ambiguity

Ambiguity resolution (and garden paths) is not the only thing we want to model in sentence processing.

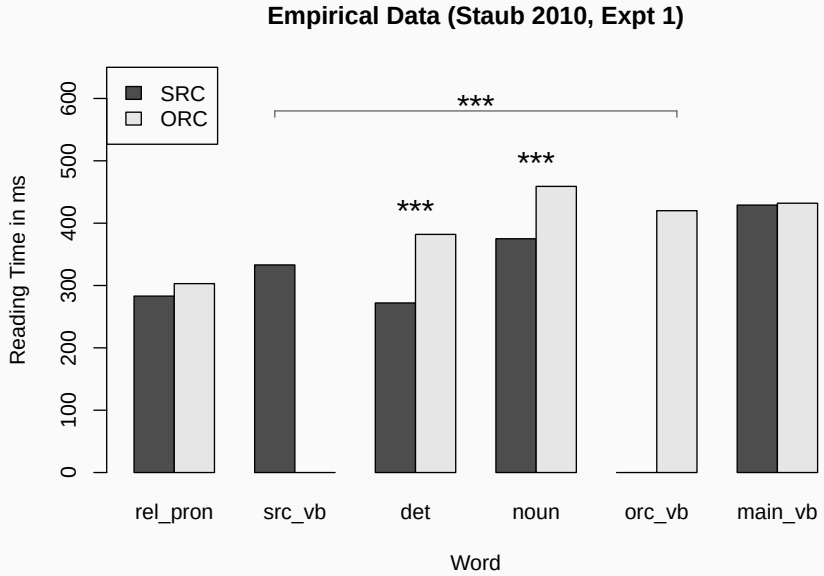
Some sentences cause difficulty even though not ambiguous:

(6) a. The reporter that attacked the senator admitted the error.
b. The reporter that the senator attacked admitted the error.

The object relative clause (ORC) in (6-b) is more difficult to process than the subject relative clause (SRC) in (6-a).

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

Beyond Ambiguity



The Surprisal Model

The Surprisal Model

The surprisal model (Hale 2001) assumes an incremental, parallel, probabilistic parser:

- all syntactic trees for a sentence prefix $w_1 \cdots w_k$ are computed at the same time, and assigned probabilities;
- the set of trees is updated as a new word w_{k+1} comes in; trees no longer compatible with the input are removed;
- surprisal measures the change in the probability distribution as trees are removed (disconfirmed) when w_{k+1} is processed;
- if w_{k+1} disconfirms trees with a large probability mass (high surprisal), then processing difficulty occurs.

The Surprisal Model

Surprisal is defined in terms of $P(T|w_1 \cdots w_k)$, the probability distribution over trees T given a sentence prefix $w_1 \cdots w_k$.

Surprisal measures the change in this distribution from w_k to w_{k+1}. This is formalized using the Kullback-Leibler divergence (relative entropy), see Levy (2008).

The KL divergence between two distributions P and Q is:

$$D(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \tag{1}$$

The Surprisal Model

The surprisal at word w_{k+1} is the KL divergence between $P(T|w_1 \cdots w_{k+1})$ and $P(T|w_1 \cdots w_k)$ is:

$$S_{k+1} = \sum_T P(T|w_1 \cdots w_{k+1}) \log \frac{P(T|w_1 \cdots w_{k+1})}{P(T|w_1 \cdots w_k)} \tag{2}$$

We can simply this because we know that:

$$P(T|w_1 \cdots w_k) = \frac{P(T, w_1 \cdots w_k)}{P(w_1 \cdots w_k)} = \frac{P(T)}{P(w_1 \cdots w_k)} \tag{3}$$

All the trees T contain the same words $w_1 \cdots w_k$, therefore $P(T, w_1 \cdots w_k) = P(T)$.

The Surprisal Model

We can now substitute Equation (3) into Equation (2): This can be simplified to:

$$S_{k+1} = \sum_T \frac{P(T)}{P(w_1 \cdots w_{k+1})} \cdot \log \frac{\frac{P(T)}{P(w_1 \cdots w_{k+1})}}{\frac{P(T)}{P(w_1 \cdots w_k)}} \tag{4}$$

$$= -\log \frac{P(w_1 \cdots w_{k+1})}{P(w_1 \cdots w_k)} = -\log P(w_{k+1} | w_1 \cdots w_k)$$

This simplification uses the fact that $\sum_T \frac{P(T)}{P(w_1 \cdots w_{k+1})} = 1$.

Therefore, surprisal is the negative log probability of a word given its preceding context. The syntactic tree T does not figure in the definition of surprisal. Surprisal is representationally agnostic.

Computing Surprisal

Surprisal doesn't depend on which representation we use; we just need to compute $P(w_{k+1}|w_1 \cdots w_k)$. We could use:

- an incremental parser which computes probabilities over trees;
- an n-gram model, which computes probabilities over sequences of words;
- intermediate cases, such as a model which computes probabilities over part of speech sequences (a tagger);
- a recurrent neural network.

However, when modeling human language processing, we are interested in the cognitive process that leads to surprisal (e.g., an n-gram model doesn't tell us much about that).

Computing Surprisal

The prefix probability $P(w_1 \cdots w_k)$ can be obtained from a parser by summing over all trees compatible with the prefix:

$$P(w_1 \cdots w_k) = \sum_T P(T, w_1 \cdots w_k) \tag{5}$$

We can now formulate Equation (4) in terms of prefix probabilities:

$$S_{k+1} = -\log \frac{P(w_1 \cdots w_{k+1})}{P(w_1 \cdots w_k)} = -\log \frac{\sum_T P(T, w_1 \cdots w_{k+1})}{\sum_T P(T, w_1 \cdots w_k)} \tag{6}$$

This is how surprisal is computed in practice.

Surprisal: Example

Assume we want to compute the prefix probability for:

(7) The reporter who ...

The prefix probability by definition is:

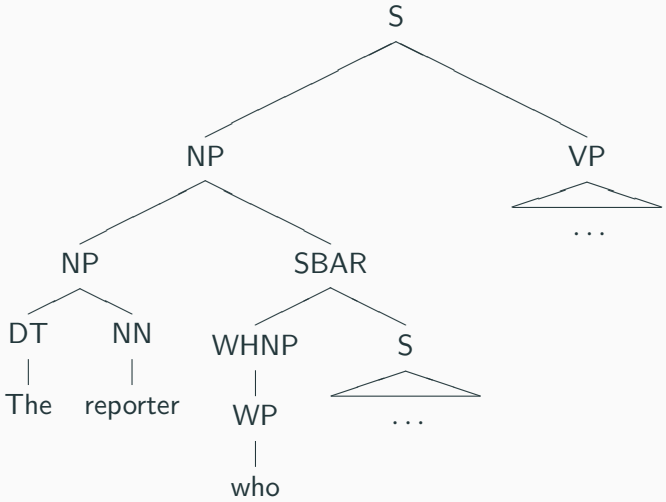
$$P(\text{the}, \text{reporter}, \text{who}) = \sum_{T} P(T, \text{the}, \text{reporter}, \text{who})$$

Assume that there is only one tree. We compute its probability using a PCFG, i.e., by multiplying the probs of the rules in T:

$$P(T, \text{the}, \text{reporter}, \text{who}) = \prod_{i} P(rule_i)$$

Surprisal: Example

Assume the following syntactic tree:



Surprisal: Example

An example for a PCFG that generates this tree:

Example	**Rule**	**Rule probability**
The reporter who ...	S $\rightarrow$ VP NP	$p = 0.6$
The reporter who ...	NP $\rightarrow$ NP SBAR	$p = 0.004$
The reporter	NP $\rightarrow$ DT NN	$p = 0.5$
The	DT $\rightarrow$ the	$p = 0.7$
reporter	NN $\rightarrow$ reporter	$p = 0.0002$
who ...	SBAR $\rightarrow$ WHNP S	$p = 0.12$
who	WHNP $\rightarrow$ WP	$p = 0.2$
who	WP $\rightarrow$ who	$p = 0.8$

Results

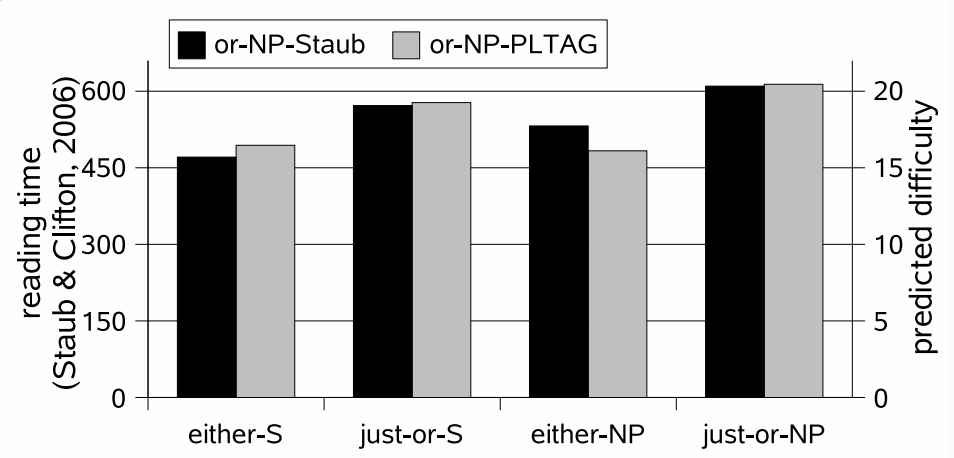
To evaluate surprisal, we need an probabilistic parser. The incremental top-down parser of Roark (2001) is often used.

Evaluation procedure:

- train the parser on a training corpus (e.g., Penn Treebank);
- take experimental materials from psycholinguistic experiments;
- parse them using the parser, compute the surprisal values for each sentence;
- compare these to the reading time results for the sentence (typically by-condition averages).

Results: either . . . or

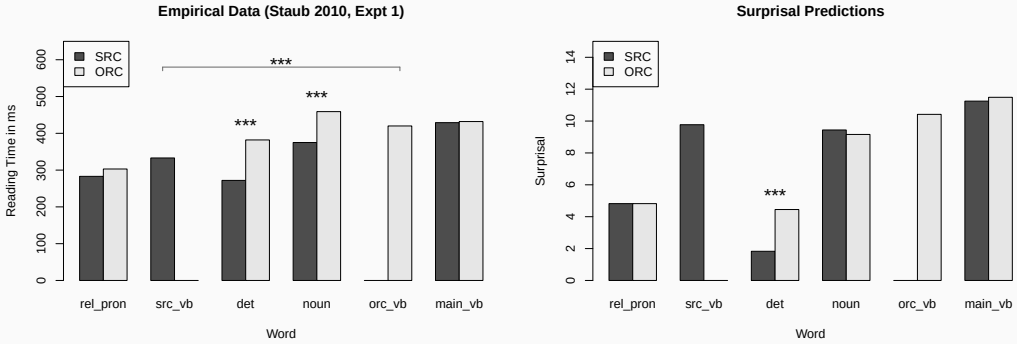
Compare reading times for *either . . . or* sentences against surprisal:



Surprisal successfully models the data.

Results: Relative Clauses

Compare reading times for relative clauses against surprisal values:

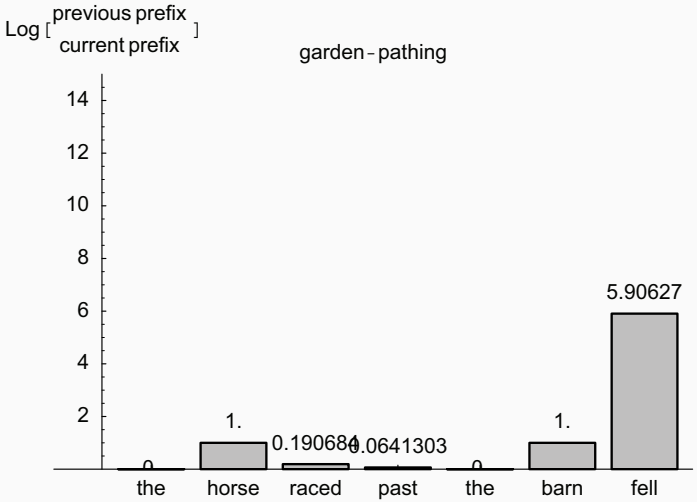


Surprisal successfully models only the difference at the NP.

To model the difference at the verb, we need to add a distance-based memory cost component (Demberg et al. 2013).

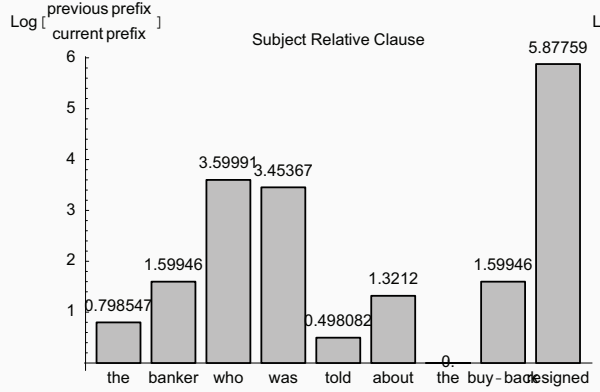
Results: Garden Paths

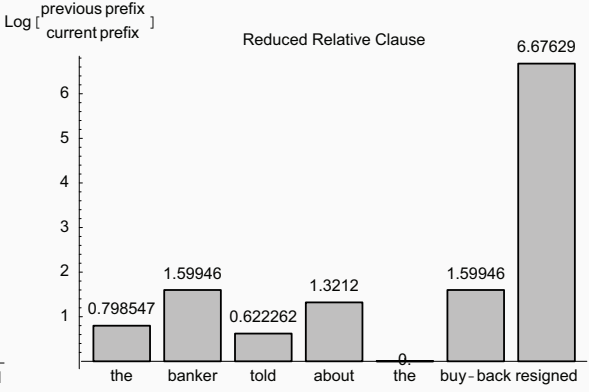
Garden paths still work (Hale 2001):



Results: Garden Paths

Compare reduced relative clause (garden path) against unreduced relative clause (not a garden path):





Summary

- The human sentence processor builds up expectations about the input;
- if these expectations are not met, surprisal ensues, which manifests itself as processing difficulty;
- mathematically, surprisal is the change in the probability distribution over possible trees from one word to the next;
- it can be computed based on the prefix probabilities returned by a probabilistic parser;
- the surprisal model accounts for garden path sentences, but also for processing difficulty not related to ambiguity.

References

Demberg, Vera, Frank Keller, and Alexander Koller. 2013. Incremental, predictive parsing with psycholinguistically motivated tree-adjoining grammar. *Computational Linguistics* 39(4):1025–1066.

Hale, John. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the 2nd Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Pittsburgh, PA, volume 2, pages 159–166.

Jurafsky, Daniel. 1996. A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science* 20(2):137–194.

Levy, Roger. 2008. Expectation-based syntactic comprehension. *Cognition* 106(3):1126–1177.

Roark, Brian. 2001. Probabilistic top-down parsing and language modeling. *Computational Linguistics* 27(2):249–276.

Staub, Adrian. 2010. Eye movements and processing difficulty in object relative clauses. *Cognition* 116:71–86.

Staub, Adrian and Charles Clifton. 2006. Syntactic prediction in language comprehension: Evidence from either . . . or. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 32:425–436.