Computational Cognitive Science

Lecture 17: Sentence Processing 2

Frank Keller

16 November 2026

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

Overview

From Surprisal to Large Language Models

Broad Evidence for Surprisal

The Inverse-Scaling Puzzle

Limits of Surprisal

Assessment

Reading: Oh and Schuler (2023).

From Surprisal to Large Language Models

From Lecture 16 to Lecture 17

Last time, we introduced word surprisal:

$$S_k = -\log P(w_k | w_1, \ldots, w_{k-1})$$

The surprisal hypothesis predicts:

higher surprisal $\Longrightarrow$ greater processing difficulty

We also separated two questions:

1. What probabilities does the comprehender assign to upcoming words?
2. How are those probabilities related to processing difficulty?

Today: what happens when the probabilities come from modern language models?

Language Models as Surprisal Estimators

To compute surprisal, we need an estimate of

$$P(w_k | w_1, \ldots, w_{k-1})$$

Historically, these probabilities have come from:

Prob. Parsers → n-grams → RNNs → Transformers → LLMs

Modern autoregressive LLMs are especially attractive because they:

- make token-by-token predictions in context;
- exploit much longer contexts;
- achieve much lower perplexity than older models.

Natural hypothesis: better next-word prediction ⇒ better model of human expectations

From Token Surprisal to Word Surprisal

LLMs usually predict subword tokens, not words.

Suppose a word w_k is split into tokens $t_1, \ldots, t_m$. Then its word surprisal is the sum of the token surprisals:

$$S_k = -\sum_{j=1}^{m} \log P(t_j | t_{<j}, \text{previous context}).$$

For a one-token word this reduces to the usual definition.

This gives us a way to compare LLM predictions with word-level behavioral measures:

- self-paced reading times;
- gaze duration;
- total fixation time.

The language model supplies the probabilities; surprisal supplies the linking hypothesis.

Broad Evidence for Surprisal

Does Surprisal Generalize Across Languages?

Most early evidence for surprisal came from English.

Wilcox et al. (2023) ask whether the theory generalizes across languages.

They test three predictions:

1. Surprisal: less predictable words take longer to process.
2. Contextual entropy: uncertainty about upcoming words also affects processing.
3. Linearity: each additional bit of surprisal produces approximately the same increase in reading time.

This is a much stronger test of a supposedly general theory of language processing.

Wilcox et al.: Data and Models

The study uses the Multilingual Eye Movement Corpus (MECO):

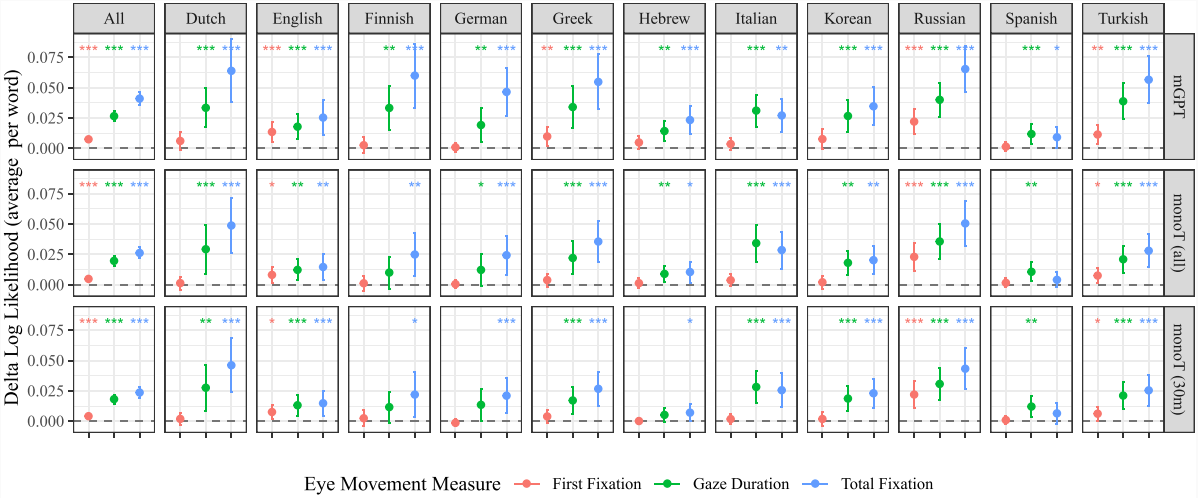
- 11 languages in the analyses;
- five language families;
- the same Wikipedia-style texts translated across languages;
- eye-tracking measures from native speakers.

Surprisal is estimated using:

- a multilingual GPT-style model;
- monolingual transformer models;
- monolingual models trained on a smaller amount of data.

Regression models also control for basic lexical factors such as word length and word frequency.

Result: Surprisal Predicts Reading Times Across Languages



Result: Surprisal Predicts Reading Times Across Languages

Wilcox et al.'s (2023) results show:

- Surprisal improves prediction for gaze duration and total fixation time in essentially every language.
- The effect is less consistent for first fixation, an earlier measure.

The surprisal effect is strikingly robust cross-linguistically.

Beyond the Current Word: Contextual Entropy

Surprisal asks: How unexpected was the word that actually occurred?

But before the word arrives, the comprehender may also be sensitive to how uncertain the context is:

$$H(W_k|w_{<k}) = -\sum_w P(w|w_{<k}) \log P(w|w_{<k}).$$

Wilcox et al. (2023) find that adding contextual entropy often improves reading-time predictions, although its effect is weaker than surprisal.

Processing may be partly anticipatory: uncertainty before the next word can matter.

Is the Surprisal Effect Linear?

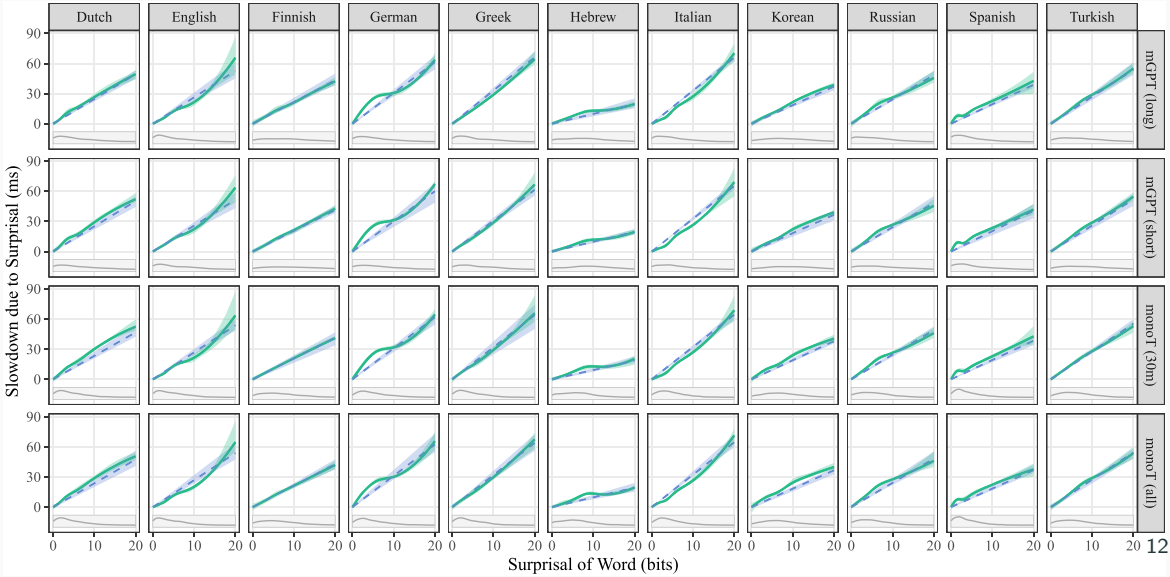
Surprisal theory is often taken to predict a linear linking function:

$$\text{reading time} = a + b\, S_k$$

This means that an extra bit of surprisal should incur roughly the same additional cost at low and high surprisal.

The flexible models recover an approximately linear relationship across languages.

Is the Surprisal Effect Linear?



Wilcox et al.: Take-Home Message

Across languages with very different grammatical properties:

- surprisal consistently predicts reading time;
- the effect is graded rather than all-or-none;
- contextual entropy provides weaker but additional predictive power;
- the surprisal–reading-time relationship is approximately linear.

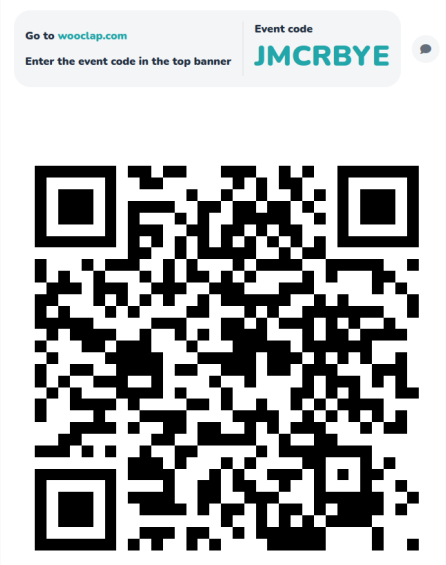
This provides broad support for the basic linking hypothesis from Lecture 16:

$$\boxed{\text{higher surprisal}} \implies \boxed{\text{greater processing difficulty}}$$

But it leaves open an important question:

Which language model gives the most human-like surprisal estimates?

Time for a short quiz on WooClap!



The Inverse-Scaling Puzzle

Should Better Language Models Be Better Cognitive Models?

Suppose we compare several language models.

A better LLM has lower perplexity:

$$\text{perplexity} \downarrow \iff \text{next-word prediction improves}$$

A natural prediction is therefore:

$$\begin{aligned} \text{perplexity} \downarrow &\Rightarrow \text{surprisal becomes more human-like} \\ &\Rightarrow \text{fit to human reading times} \uparrow \end{aligned}$$

Oh and Schuler (2023) test this prediction using increasingly large transformer language models.

Oh and Schuler: Models and Reading-Time Data

They compare variants from three autoregressive transformer families:

- GPT-2: approximately 124M–1.6B parameters;
- GPT-Neo family: approximately 125M–20B parameters;
- OPT: approximately 125M–66B parameters.

Two naturalistic reading datasets:

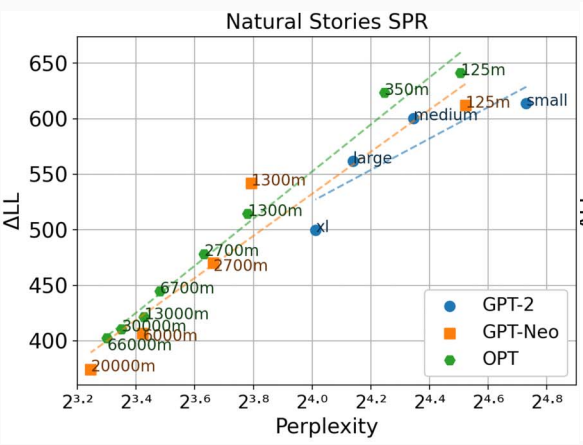
- Natural Stories: self-paced reading times;
- Dundee Corpus: eye-tracking, using go-past durations.

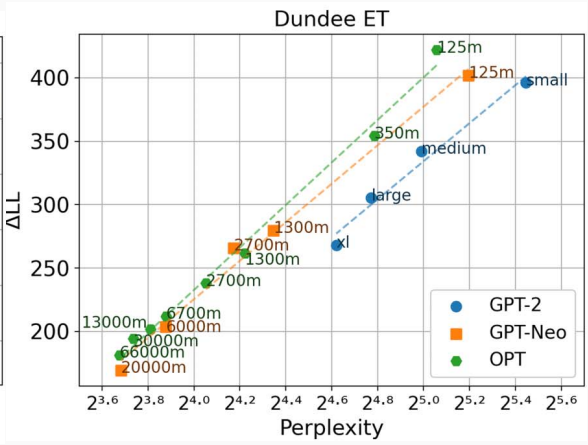
For each LLM:

1. calculate word surprisal;
2. add it to a baseline regression model;
3. measure improvement in model fit: $\Delta LL = LL(\text{baseline} + \text{surprisal}) - LL(\text{baseline})$

Larger ΔLL means LLM surprisal estimates are better predictors of human reading times.

Result: Better LLMs Fit Human Reading Times Worse





Within each model family: larger model/lower perplexity $\Longrightarrow$ poorer fit to reading times

This is the inverse-scaling effect.

Why Is Inverse Scaling Surprising?

The result separates two notions of "better":

NLP performance	**Cognitive-model performance**
predict the next word accurately	predict human processing difficulty accurately
low perplexity	surprisal explains human reading times

The largest LLMs are better at the first task but worse at the second.

Good text prediction does not guarantee cognitive plausibility.

Where Do Larger Models Go Wrong?

Oh et al. (2024) evaluate LLMs on naturalistic reading-time corpora. They find two main explanations for inverse scaling:

1. Word frequency is a major source of the mismatch:

- The inverse-scaling effect is strongest for the least frequent words.
- As the LLM gets larger, it predicts even rare words accurately.
- This means the predicted S_k for rare words becomes very small.
- But humans still show substantial difficulty on those words.

2. Larger models benefit more from context:

- Larger models effectively exploit longer context.
- They exploit local associations even when context is restricted.

Cognitive models may need to reproduce human limitations, not eliminate them.

Limits of Surprisal

A Different Question: Is Surprisal Itself Enough?

The inverse-scaling result concerns the probability model:

> *Are language-model probabilities human-like?*

But there is a second possible problem:

> *Even with good probabilities, is surprisal a complete theory of processing cost?*

Recall the garden path from Lecture 16:

(1) The horse raced past the barn fell.

At *fell*, readers must not only encounter an unexpected word; they may also need to revise an existing syntactic analysis.

Huang et al. (2024) test whether surprisal can explain this extra difficulty.

The Syntactic Ambiguity Processing Benchmark

Huang et al. construct the SAP Benchmark:

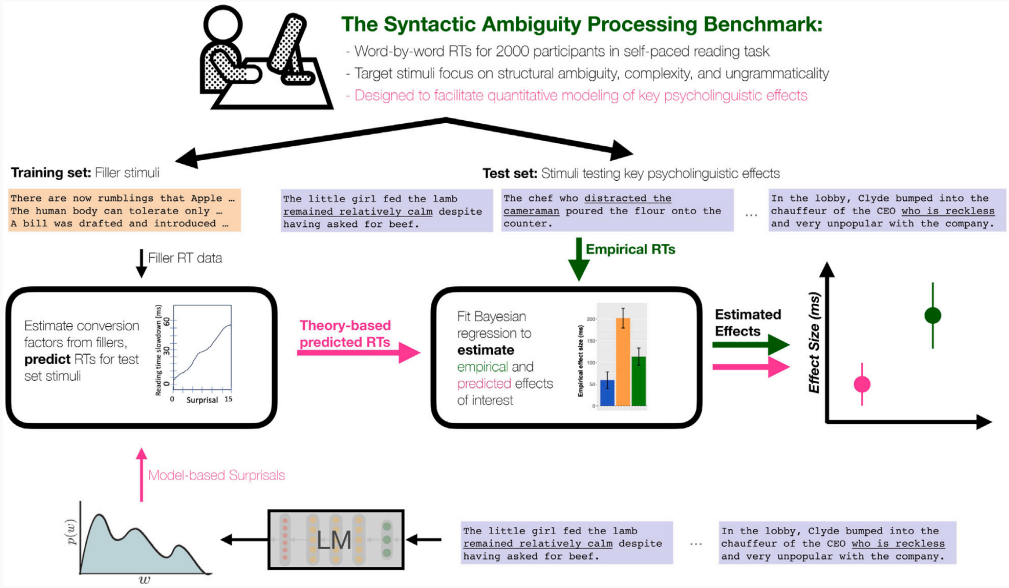
- large-scale self-paced reading experiment;
- approximately 2,000 participants;
- seven classic syntactic effects;
- carefully controlled experimental items plus filler sentences.

The constructions include:

- three types of garden path;
- subject versus object relative clauses;
- high versus low attachment;
- subject–verb agreement violations.

The goal is to test precise quantitative predictions, not just broad correlations.

How Do Huang et al. Test Surprisal?



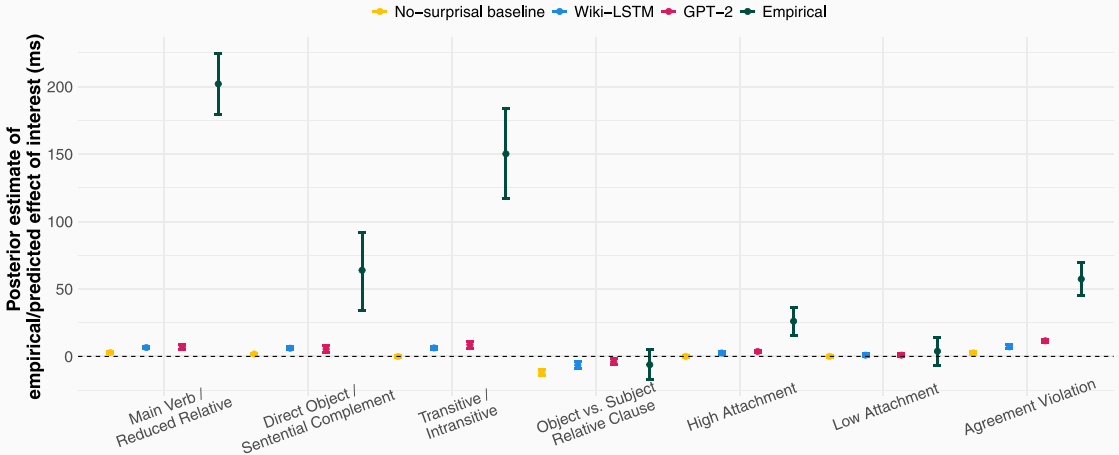
How Do Huang et al. Test Surprisal?

The crucial procedure is:

1. learn the mapping from language model surprisal to reading time using filler sentences;
2. use that mapping to predict reading times in the controlled syntactic items;
3. compare the predicted effect with the actual human effect.

This prevents the linking function from being tuned specifically to garden-path effects.

Result: Surprisal Underestimates Syntactic Difficulty



Surprisal predicts the direction and location of an effect, but badly underpredicts its magnitude.

For the main-verb/reduced-relative garden path, the human effect is about 31 times larger than the surprisal prediction.

Surprisal Also Misses Relative Difficulty

The problem is not simply a constant scaling factor.

Humans show a clear ranking of garden-path difficulty:

Main Verb/Reduced Relative > Transitive/Intransitive > Direct Object/Sentential Complement

The language-model surprisal predictions do not reproduce this ranking reliably.

At the item level, surprisal also explains only a limited proportion of the reliable variation for the critical syntactic constructions.

Unexpectedness is important, but syntactic reanalysis may create additional costs.

Assessment

Two Places a Surprisal Model Can Fail

A modern surprisal model has several distinct components:

$$\boxed{\text{language model}} \rightarrow \boxed{P(w_k|w_1,\ldots,w_{k-1})} \rightarrow \boxed{\text{surprisal}} \rightarrow \boxed{\text{reading time}}$$

Two important mismatches emerge:

1. **Probability-model mismatch:**
   LLMs may exploit more data, longer contexts, and stronger associations than humans (Oh and Schuler 2023; Oh et al. 2024).
2. **Linking-theory mismatch:**
   surprisal may omit costs from memory, retrieval, or syntactic reanalysis (Huang et al. 2024).

Summary

- Modern autoregressive LLMs provide powerful estimates of contextual word probabilities for surprisal models.
- Surprisal robustly predicts reading times on naturalistic text for many diverse languages.
- Recent work finds an inverse-scaling effect: within model families, increasing model size and lowering perplexity can worsen the fit to human reading times.
- This mismatch is largely driven by rare words, which larger models predict "superhumanly" well.
- Moreover, surprisal severely underestimates the magnitude of syntactic processing effects such as garden paths.
- A useful predictor is not necessarily a complete cognitive theory.

References

Huang, Kuan-Jung, Suhas Arehalli, Mari Kugemoto, Christian Muxica, Grusha Prasad, Brian Dillon, and Tal Linzen. 2024. Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty. *Journal of Memory and Language* 137:104510.

Oh, Byung-Doh and William Schuler. 2023. Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics* 11:336–350.

Oh, Byung-Doh, Shisen Yue, and William Schuler. 2024. Frequency explains the inverse correlation of large language models' size, training data amount, and surprisal's fit to reading times. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, St. Julian's, Malta, pages 2644–2663.

Wilcox, Ethan G., Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P. Levy. 2023. Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics* 11:1451–1470.