Lexical Effects in Structural Forgetting
Evidence for Experience-Based Accounts and a Neural Network Model

Michael Hahn (Stanford), Richard Futrell (UC Irvine), Edward Gibson (MIT)

CUNY Conference 2020
The patient who the nurse who the clinic had hired met Jack.
The patient who the nurse who the clinic had hired met Jack.

The patient who the nurse who the clinic had hired admitted met Jack.
The patient who the nurse who the clinic had hired met Jack.

The patient who the nurse who the clinic had hired admitted met Jack.
The patient who the nurse who the clinic had hired met Jack.

The patient who the nurse who the clinic had hired admitted met Jack.
* The **patient** who **the nurse** who **the clinic had hired** met Jack.

The **patient** who **the nurse** who **the clinic had hired** admitted met Jack.
* The patient who the nurse who the clinic had hired met Jack.

The patient who the nurse who the clinic had hired admitted met Jack.

**Structural Forgetting:** Dropping middle verb “admitted” makes sentence easier! (Frazier 1985, Gibson & Thomas 1999)
Experience-Based Account

Idea (Christiansen and Chater 2001, Frank et al., 2016, Futrell & Levy 2017)

1. **Nested verb-final embedding** is rare in English.
2. Comprehenders `regularize’ towards more common structures.
3. Explains why effect is **absent** (or weaker) in **German, Dutch**
   (Engelmann and Vasishth, 2009; Frank et al., 2016, Futrell & Levy 2017).

* The patient who the nurse who the clinic had hired met Jack.

The patient who the nurse who the clinic had hired admitted met Jack.
Experience-Based Account

Idea (Christiansen and Chater 2001, Frank et al., 2016, Futrell & Levy 2017)

1. Nested verb-final embedding is rare in English.
2. Comprehenders `regularize’ towards more common structures.
3. Explains why effect is absent (or weaker) in German, Dutch
   (Engelmann and Vasishth, 2009; Frank et al., 2016, Futrell & Levy 2017).

**New Prediction:** Lexical differences in expectations should modulate strength of forgetting effect!

Words that make embedding more likely should decrease effect.
Lexical Differences

The {fact / report} that the student who the professor hated dropped the class made the professor happy.
Lexical Differences

The {fact / report} that the student who the professor hated dropped the class made the professor happy.
Lexical Differences

The \{\text{fact} / \text{report}\} that the \text{student} who the \text{professor} hated dropped the class made the professor happy.

- Sentential complements are about 10x more likely after `fact` than after `report`.

- \text{Prediction}: Forgetting effect should be stronger for with `report` than `fact`
Experimental Predictions: Comprehension

**Difficulty** for ungrammatical version

**Difficulty** for grammatical version

“fact”

log P(that|the NOUN)

“disclosure”

“report”
Experimental Predictions: Production

"disclosure"
"report"

Rate of dropping a verb in production

log P(that|the NOUN)
Experiment 1: Production
Production Experiment

- Via Mechanical Turk
- 144 participants
- Asked to complete a prefix to a complete sentences
- 12 critical items, 64 fillers

Please complete the following sentence in the text box below:

The declaration that the surgeon who the attorney...

You can complete it in any way you like, as long as you make a complete English sentence. You don't need to type the part of the sentence that is already given to you -- just your completion.
The belief that the violinist who the corporate sponsors...

...hired for their event, was overpaid is ridiculous.

Grammatical

...were eying for their products was false.

Dropped a verb
Experimental Predictions: Production

Rate of dropping a verb in production

log P(that|the NOUN)
Rate of Responses with Missing Verb

p < 0.01 (trial-by-trial analysis)

counts from English Wikipedia
Experiment 2: Ratings
Rating study

The disclosure that the marketing whiz who the artist had hired was a fraud shocked everyone.

How hard is this sentence to understand?

Easy to understand 0 1 2 3 4 5 Hard to understand
The disclosure that the marketing whiz who the artist had hired was a fraud shocked everyone.

How hard is this sentence to understand?

Easy to understand ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Hard to understand
Rating study

The disclosure that the marketing whiz who the artist had hired was a fraud shocked everyone.

How hard is this sentence to understand?

Easy to understand □ 1 □ 2 □ 3 □ 4 □ 5 Hard to understand
Rating study

Based on Gibson & Thomas 1999

Noun

The disclosure that the marketing whiz who the artist had hired was a fraud shocked everyone.

Continuation

How hard is this sentence to understand?

Easy to understand

1 2 3 4 5 Hard to understand

Pool of 35 nouns

Pool of 20 continuations
Rating study

The disclosure that the marketing whiz who the artist had hired was a fraud shocked everyone.

How hard is this sentence to understand?

Easy to understand  ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Hard to understand
The disclosure that the marketing whiz who the artist had hired was a fraud shocked everyone.

How hard is this sentence to understand?

Easy to understand  ○ 1  ○ 2  ○ 3  ○ 4  ○ 5 Hard to understand
Rating Experiment

- 120 participants recruited on Mechanical Turk
- 20 critical trials per participant
- Nouns randomly matched with continuations
- 10 trials in grammatical, ungrammatical condition each
Experimental Predictions:
Comprehension

"fact"

"disclosure"
"report"

Difficulty for ungrammatical version

Difficulty for grammatical version

log P(that|the NOUN)
Difficulty for ungrammatical version

“disclosure”

Difficulty for grammatical version

“fact”
“disclosure”

Difficulty for ungrammatical version

“fact”

Difficulty for grammatical version
Production Predictions

Rate of dropping a verb in production

log P(that|the NOUN)
Comprehension Predictions

log P(that|the NOUN)

Rating Study

Mean Rating (Difficulty of Understanding)

grammatical
- FALSE
- TRUE

log P(that|the NOUN)
Computational Model
Noisy Context Surprisal (Futrell and Levy, 2017)

Computational-level formalization of the experience-based account

Noisy-context surprisal as a human sentence processing cost model

Richard Futrell and Roger Levy
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology
{futrell, rplevy}@mit.edu

Abstract
We use the noisy-channel theory of human sentence comprehension to develop an incremental processing cost model, as in Lewis and Vasishth, 2005; a major prediction of these models is locality effects, where processing a word is difficult when it is far from other words with which it must be syntactically integrated. Expectation-based models do not intrin...
Surprisal Theory: processing cost is given by surprisal (Hale 2001, Levy 2007)

\[
\text{Cost} = -\log P(X_i | X_1, ..., X_{i-1})
\]
Surprisal Theory: processing cost is given by surprisal (Hale 2001, Levy 2007)

\[ \text{Cost} = -\log P(X_i|X_{1\ldots i-1}) \]

Noisy-Context Surprisal: processing cost of word is surprisal given a noisy representation of preceding context:

\[ \text{Cost} = -\log P(X_i|\text{Memory}) \]
Surprisal Theory: processing cost is given by surprisal (Hale 2001, Levy 2007)

\[
\text{Cost} = -\log P(X_i | X_1 \ldots i-1)
\]

**Noisy-Context Surprisal**: processing cost of word is surprisal given a noisy representation of preceding context:

\[
\text{Cost} = -\log P(X_i | \text{Memory})
\]
Surprisal Theory: processing cost is given by surprisal (Hale 2001, Levy 2007)

\[ \text{Cost} = -\log P(X_i|X_1^{i-1}) \]

**Noisy-Context Surprisal:** processing cost of word is surprisal given a noisy representation of preceding context:

\[ \text{Cost} = -\log P(X_i|\text{Memory}) \]
Surprisal Theory: processing cost is given by surprisal (Hale 2001, Levy 2007)

\[ \text{Cost} = -\log P(X_i|X_1\ldots i-1) \]

**Noisy-Context Surprisal:** processing cost of word is surprisal given a noisy representation of preceding context:

\[ \text{Cost} = -\log P(X_i|\text{Memory}) \]

**Consequences:**

- More **frequent contexts** give rise to more accurate predictions
- **Infrequent contexts** may be “misremembered” as similar but more frequent contexts.
the fact that the lawyer who the doctor knew
the fact that the lawyer who the doctor knew
the fact that the lawyer who the doctor knew

Preliminary modeling assumption (Futrell & Levy, 2017):
Words are lost uniformly at random, at some fixed rate.
the fact that the lawyer who the doctor knew

Preliminary modeling assumption (Futrell & Levy, 2017):
Words are lost uniformly at random, at some fixed rate.

the fact ??? the lawyer who the ??? knew
the fact that ??? lawyer ??? the doctor ???
??? fact that the ??? ??? the doctor knew
the fact ??? the lawyer ??? the doctor knew

...
Preliminary modeling assumption (Futrell & Levy, 2017):
Words are lost uniformly at random, at some fixed rate.

For now, we focus on this example
the fact that the lawyer who the doctor knew

the fact ??? the lawyer who the ??? knew
the fact that the lawyer who the doctor knew
the fact that the lawyer who the doctor knew

the fact ??? the lawyer who the ??? knew

the fact that the lawyer who the witness knew
Actual Context: the fact that the lawyer who the doctor knew

Imperfect Memory: the fact ??? the lawyer who the ??? knew

Reconstructed Distribution: the fact that the lawyer who the witness knew
the fact that the lawyer who the parent knew
the fact that the lawyer who the doctor knew

the fact ??? the lawyer who the ??? knew

the fact that the lawyer who the witness knew
the fact that the lawyer who the parent knew
the fact about the lawyer who the professor knew
Actual Context

the fact that the lawyer who the doctor knew

Imperfect Memory

the fact ??? the lawyer who the ??? knew

Reconstructed Distribution

the fact that the lawyer who the witness knew
the fact that the lawyer who the parent knew
the fact about the lawyer who the professor knew
the fact regarding the lawyer who the crocodile knew
...
the fact that the lawyer who the doctor knew

the fact ??? the lawyer who the ??? knew

the fact that the lawyer who the witness knew
the fact that the lawyer who the parent knew
the fact about the lawyer who the professor knew
the fact regarding the lawyer who the crocodile knew

P(that|the fact) ~ 0.7
Actual Context

the fact that the lawyer who the doctor knew

Imperfect Memory

the fact ??? the lawyer who the ??? knew

Reconstructed Distribution

the fact that the lawyer who the witness knew
the fact that the lawyer who the parent knew
the fact about the lawyer who the professor knew
the fact regarding the lawyer who the crocodile knew

P(that|the fact) ~ 0.7
the fact that the lawyer who the doctor knew

the fact ??? the lawyer who the ??? knew

the fact that the lawyer who the witness knew

the fact that the lawyer who the parent knew

the fact about the lawyer who the professor knew

the fact regarding the lawyer who the crocodile knew

...
the fact that the lawyer who the doctor knew

the fact ??? the lawyer who the ??? knew

Expect to see two verbs!
the disclosure that the lawyer who the doctor knew
the disclosure that the lawyer who the doctor knew

the disclosure ??? the lawyer who the ??? knew
Actual Context
the disclosure that the lawyer who the doctor knew

Imperfect Memory
the disclosure ??? the lawyer who the ??? knew

the disclosure that the lawyer who the witness knew

the disclosure that the lawyer who the parent knew

the disclosure about the lawyer who the professor knew

the disclosure regarding the lawyer who the crocodile knew

...
the disclosure that the lawyer who the doctor knew

the disclosure ??? the lawyer who the ??? knew

the disclosure that the lawyer who the witness knew

the disclosure that the lawyer who the parent knew

the disclosure about the lawyer who the professor knew

the disclosure regarding the lawyer who the crocodile knew

\[ P(\text{that}|\text{the disclosure}) \sim 0.03 \]
the disclosure that the lawyer who the doctor knew

the disclosure ??? the lawyer who the ??? knew

the disclosure that the lawyer who the witness knew

the disclosure that the lawyer who the parent knew

the disclosure about the lawyer who the professor knew

the disclosure regarding the lawyer who the crocodile knew

P(that|the disclosure) ~ 0.03
the disclosure that the lawyer who the doctor knew

the disclosure ??? the lawyer who the ??? knew

- the disclosure that the lawyer who the witness knew
- the disclosure that the lawyer who the parent knew
- the disclosure about the lawyer who the professor knew
- the disclosure regarding the lawyer who the crocodile knew
...
the disclosure that the lawyer who the doctor knew

the disclosure ??? the lawyer who the ??? knew

Ungrammatical continuation has relatively high probability!
the fact that the lawyer who the doctor knew

Verb Verb
Verb

the fact ??? the lawyer who the ??? knew

Verb Verb
Verb

P(that|that fact) ≈ 0.7

the disclosure that the lawyer who the doctor knew

Verb Verb

the disclosure ??? the lawyer who the ??? knew

Verb Verb

P(that|the disclosure) ≈ 0.03
Higher-probability contexts give rise to more precise predictions.
Surprisal of ungrammatical continuation

Surprisal of grammatical continuation

"disclosure"

"fact"

log \( P(\text{that}|\text{the NOUN}) \)
Production Predictions

Rate of dropping a verb in production

log P(\text{that} \mid \text{the NOUN})

Production Study

Rate of responses with missing verb

log P(\text{that} \mid \text{the NOUN})
Comprehension Predictions

log P(that|the NOUN)

Rating Study

Mean Rating (Difficulty of Understanding)

grammatical
FALSE
TRUE
Neural Network Model
Actual Context

the fact that the lawyer who the doctor knew

Imperfect Memory

the fact ??? the lawyer who the ??? knew

- the fact that the lawyer who the witness knew
- the fact that the lawyer who the parent knew
- the fact about the lawyer who the professor knew
- the fact regarding the lawyer who the crocodile knew

...
the fact that the lawyer who the doctor knew

Neural Network 1 ("Noise Model")

the fact ??? the lawyer who the ??? knew

- the fact that the lawyer who the witness knew
- the fact that the lawyer who the parent knew
- the fact about the lawyer who the professor knew
- the fact regarding the lawyer who the crocodile knew
the fact that the lawyer who the doctor knew

Neural Network 1 ("Noise Model")

the fact ??? the lawyer who the ??? knew

Neural Network 2 ("Denoiser")

the fact that the lawyer who the witness knew

the fact that the lawyer who the parent knew

the fact about the lawyer who the professor knew

the fact regarding the lawyer who the crocodile knew

...
the fact that the lawyer who the doctor knew

the fact ??? the lawyer who the ??? knew

Neural Network 1 ("Noise Model")

the fact that the lawyer who the witness knew

the fact that the lawyer who the parent knew

the fact about the lawyer who the professor knew

the fact regarding the lawyer who the crocodile knew

Neural Network 2 ("Denoiser")

Neural Network 3 ("Prediction Model")
Rational Loss Model

Idea: Learn representations that trade off reconstruction accuracy with economy of memory.
Rational Loss Model

Idea: Learn representations that trade off reconstruction accuracy with economy of memory.

Formalize:
- **Economy of memory** ~ number of remembered words
- **Informativity**:
  - How accurately can lost words be recovered?
  - How accurately can next word be predicted?
Rational Loss Model

\[
\log p(\text{actual sequence} \mid \text{noised sample})
\]

maximize

Reconstruct the original sequence as well as possible
Rational Loss Model

\[
\log p(\text{actual sequence} \mid \text{noised sample}) + \log p(\text{next word} \mid \text{noised sample})
\]

maximize

Make sure the next word is easy to predict

Reconstruct the original sequence as well as possible
Rational Loss Model

\[
\log p(\text{actual sequence} | \text{noised sample}) \\
+ \log p(\text{next word} | \text{noised sample}) \\
+ \lambda \#\{\text{erased words}\}
\]

maximize

Reconstruct the original sequence as well as possible

Make sure the next word is easy to predict

Forget as many words as possible
Rational Loss Model

\[
\log p(\text{actual sequence} \mid \text{noised sample}) + \lambda \#\{\text{erased words}\} + \log p(\text{next word} \mid \text{noised sample})
\]

- **maximize**
- **Forget as many words as possible**
- **Make sure the next word is easy to predict**
- **Reconstruct the original sequence as well as possible**
the fact shocked everyone

Sequences from Corpus (English Wikipedia, 2 Billion words)

Erasure choices made by a feedforward network for each word individually.

Sequences from Corpus (English Wikipedia, 2 Billion words)
Sequences from Corpus (English Wikipedia, 2 Billion words)
Sequences from Corpus (English Wikipedia, 2 Billion words)

Trained to maximize:

\[
\log p(\text{actual sequence} \mid \text{noised sample}) + \log p(\text{next word} \mid \text{noised sample}) + \lambda \#\{\text{erased words}\}
\]
Examples

$\lambda < 4$

Low
Economy
Pressure

Samples from
the Denoising
Network

the report that the senator who the diplomat opposed

??? report that ??? senator who the diplomat opposed

the report that the senator who the diplomat opposed
the report that the senator who the diplomat opposed
the report that the senator who the diplomat opposed
the report that the senator who the diplomat opposed
the report that the senator who the diplomat opposed
the report that the senator who the diplomat opposed
the report that the senator who the diplomat opposed
the report that the senator who the diplomat opposed
the report that the senator who the diplomat opposed
Examples

\( \lambda = 4.5 \)
Medium
Economy
Pressure

Samples from the Denoising Network

- the report that the senator who the diplomat opposed
- ??? report ??? ??? senator who the diplomat opposed
- the report by the senator who the diplomat opposed
- the report included a senator who the diplomat opposed
- a report by the senator who the diplomat opposed
- the report of a senator who the diplomat opposed
- "report" to senator who the diplomat opposed
- a report by a senator who the diplomat opposed
- the report of the senator who the diplomat opposed
- the report of the senator who the diplomat opposed
- a report by the senator who the diplomat opposed
- a report to the senator who the diplomat opposed
Examples

\( \lambda = 4.5 \)
Medium
Economy
Pressure

Samples from the Denoising Network

the report that the senator who the diplomat opposed

??? report ??? ??? senator who the diplomat opposed

the report by the senator who the diplomat opposed

the report included a senator who the diplomat opposed

a report by the senator who the diplomat opposed

the report of a senator who the diplomat opposed

" report " to senator who the diplomat opposed

a report by a senator who the diplomat opposed

the report of the senator who the diplomat opposed

the report of the senator who the diplomat opposed

a report by the senator who the diplomat opposed

a report to the senator who the diplomat opposed
Fraction of reconstruction samples containing “the NOUN that”

$\lambda > 4$

Log-likelihood $P(\text{that} | \text{the NOUN})$
Surprisal of ungrammatical continuation

Surprisal of grammatical continuation

log $P(\text{that}|\text{the NOUN})$
$\lambda = 4.5$

Surprisal of ungrammatical continuation

Surprisal of grammatical continuation
Model (λ=4.5)

- Model replicates the crucial interaction
- But, also predicts a main effect of log P(that|the NOUN).
Comparison with Uniform Loss Model

Uniform Loss Model

Rational Loss Model
Comparison with Uniform Loss Model

- Interaction is reproduced by rational loss model, but not by uniform loss model.
Conclusion

1. Structural forgetting is attenuated by frequency effects
Conclusion

1. Structural forgetting is attenuated by frequency effects
2. This can be accounted for by **lossy-context surprisal**
Conclusion

1. Structural forgetting is attenuated by frequency effects
2. This can be accounted for by lossy-context surprisal
3. Implemented this using neural networks trained on large-scale corpus data
Conclusion

1. Structural forgetting is attenuated by frequency effects
2. This can be accounted for by lossy-context surprisal
3. Implemented this using neural networks trained on large-scale corpus data
4. Rationally optimized loss model reproduces observed lexical effect.
Thank you!
Rational Loss Model

\[
\log p(\text{actual sequence} | \text{noised sample}) + \lambda \text{#erased words}
\]

maximize

Reconstruct the original sequence as well as possible

Do we need both predictability and reconstructuctibility terms?

Forget as many words as possible

Make sure the next word is easy to predict

Reconstruct the original sequence as well as possible
Rational Loss Model

\[ \log p(\text{actual sequence} | \text{noised sample}) + \lambda \# \{\text{erased words}\} \]

maximize

Reconstruct the original sequence as well as possible

Forget as many words as possible
Examples

\( \lambda = 4.5 \)

Samples from the Denoising Network

- a report by the senator, the diplomat opposed
- the report that the senator and a diplomat opposed
- a report that the senator and the diplomat opposed
- an report that the senator and the diplomat opposed
- a report of the senator, the diplomat opposed
- a report by the senator and the diplomat opposed
- he report of the senator was the diplomat opposed
- a report by the senator and the diplomat opposed
- his report of a senator as a diplomat opposed
- her report that the senator as a diplomat opposed

Dropping Predictability Term

\( \lambda = 4.5 \)

??? report ??? ??? senator ??? ??? diplomat opposed

- a report by the senator, the diplomat opposed
- the report that the senator and a diplomat opposed
- a report that the senator and the diplomat opposed
- an report that the senator and the diplomat opposed
- a report of the senator, the diplomat opposed
- a report by the senator and the diplomat opposed
- he report of the senator was the diplomat opposed
- a report by the senator and the diplomat opposed
- his report of a senator as a diplomat opposed
- her report that the senator as a diplomat opposed
Examples

\( \lambda = 4.5 \)

Samples from the Denoising Network

Dropping Predictability Term

Overpredicts forgetting of embedding!

- a report by the senator, the diplomat opposed
  - a report that a senator and a diplomat opposed
  - an report that the senator and the diplomat opposed
  - a report of the senator, the diplomat opposed
  - a report by the senator and the diplomat opposed
  - he report of the senator was the diplomat opposed
  - a report by the senator and the diplomat opposed
  - his report of a senator as a diplomat opposed
  - her report that the senator as a diplomat opposed
Rational Loss Model

Maximize

\[ \log p( \text{next word} | \text{noised sample}) + \lambda \#\{\text{erased words}\} \]

- Forget as many words as possible
- Make sure the next word is easy to predict
- Reconstruct the original sequence as well as possible

Dropping Reconstrucitbility Term
Examples

$\lambda = 4.5$

the report that the senator who the diplomat opposed

??? ??? ??? ??? ??? who ??? diplomat opposed

Samples from the Denoising Network

Dropping Reconstructibility Term

Overpredicts forgetting of embedding!
Per-Noun Coefficient of Grammaticality in:

Rating ~ Grammatical + (1+Grammatical|Noun) + (1+Grammatical|Participant) + (1+Grammatical|Suffix)

**Weak** Effect of Grammaticality

Strong Effect of Grammaticality

$log P(\text{that} | \text{the noun})$
Weak Effect of Grammaticality

Strong Effect of Grammaticality

$R = 0.18$, $p = 0.29$
Weak Effect of Grammaticality

Strong Effect of Grammaticality

$R = -0.18, p = 0.31$

log $P(\text{the noun that})$
Loss Rate: 0.2

Fraction of reconstruction samples containing “the NOUN that”

log P(that|the NOUN)