The relation of surprisal and human processing difficulty

Information Theory Lecture

Vera Demberg and Matt Crocker

Information Theory Lecture, Universität des Saarlandes

April 19th, 2015

Information theory in der Psycholinguistics

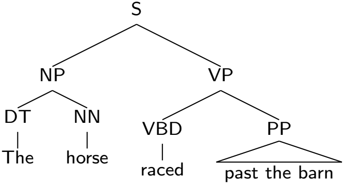
Surprisal allows us to estimate a measure of how much information is being conveyed by an utterance.

Psycholinguistic perspective:

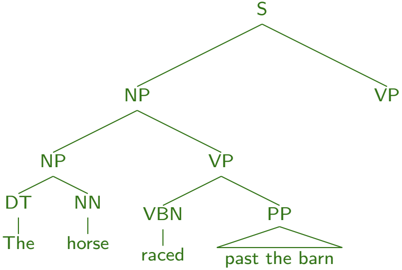
- Hypothesis: Processing difficulty is proportional to the amount of information conveyed.
- i.e., can we measure the difficulty of a sentence using information theoretic concepts?

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:

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

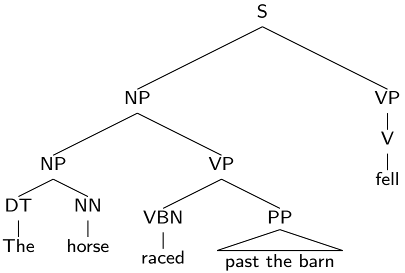
- Surprisal s of word w_n:
  $s_{w_n} = pp_{w_n} - pp_{w_{n-1}}$

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

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



How to calculate surprisal:

- Calculate prefix probabilities:

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

- Surprisal s of word w_n:

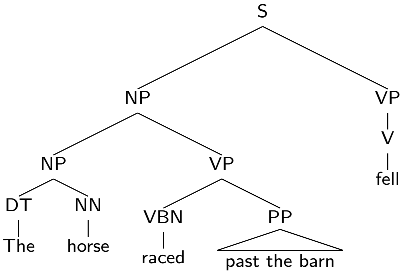
$$s_{w_n} = pp_{w_n} - pp_{w_{n-1}}$$

Example PCFG:

Rule	**Probability of rule**
S $\rightarrow$ 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:

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

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

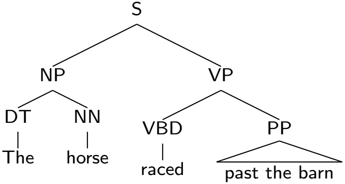
Example PCFG:

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	$p = 0.7$

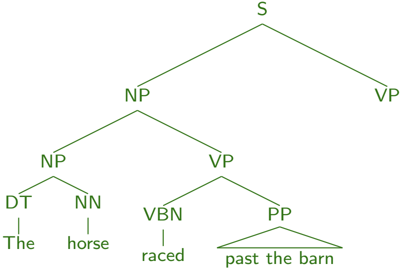
- Predictions also depend on parametrization of the grammar, training

Lexical vs. structural surprisal

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

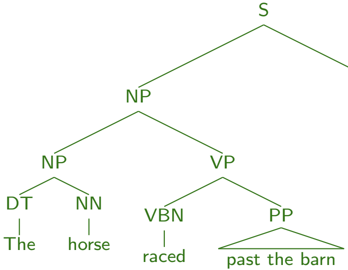


$-\log(1.06596 \times 10^{-15}) = 49.736$



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

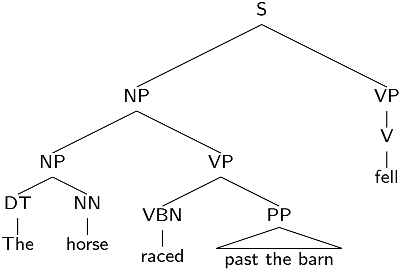
$pp_{w_{n+1}} = -\log(1.06596 \times 10^{-15} \times 0.003)$
$= 58.12$



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

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



$$S_{w_n} = -\log \sum_{T \in Trees} \frac{p(T|w_1 \ldots w_n)}{p(T|w_1 \ldots w_{n-1})}$$

$$structS_{w_n} = -\log \sum_{\mathrm{POS}_n \in POS} \sum_{T \in Trees} \frac{p(T|w_1 \ldots \mathrm{POS_n})}{p(T|w_1 \ldots w_{n-1})}$$

$$lexS_{w_n} = -\log \sum_{\mathrm{POS}_n \in POS} \sum_{T \in Trees} \frac{p(T|w_1 \ldots w_n)}{p(T|w_1 \ldots \mathrm{POS}_n)}$$

Table of Contents

Corpus-based results

Support from reading times in naturalistic texts

- on Dundee Corpus (Demberg and Keller, 2008; Frank, 2009; Fossum and Levy, 2012; Smith and Levy, 2013)
- on English stories with long dependencies (Roark et al., 2009)
- on Potsdam Sentence Corpus (German) (Boston et al., 2008)
- on Brown SPR Corpus (Smith and Levy 2013)

Reading times

linking theory: reading times reflect processing difficulty; if we find a correlation, then surprisal predicts behaviour.

Corpus-based results

Support from reading times in naturalistic texts

- on Dundee Corpus (Demberg and Keller, 2008; Frank, 2009; Fossum and Levy, 2012; Smith and Levy, 2013)
- on English stories with long dependencies (Roark et al., 2009)
- on Potsdam Sentence Corpus (German) (Boston et al., 2008)
- on Brown SPR Corpus (Smith and Levy 2013)

Reading times

linking theory: reading times reflect processing difficulty; if we find a correlation, then surprisal predicts behaviour.

Support from EEG:

- surprisal predictive of N400 amplitudes (Frank et al., 2013)

N400

N400 has been linked to predictability, difficulty in retrieving / integrating a word.

Corpus-based results

Support from reading times in naturalistic texts

- on Dundee Corpus (Demberg and Keller, 2008; Frank, 2009; Fossum and Levy, 2012; Smith and Levy, 2013)
- on English stories with long dependencies (Roark et al., 2009)
- on Potsdam Sentence Corpus (German) (Boston et al., 2008)
- on Brown SPR Corpus (Smith and Levy 2013)

Reading times

linking theory: reading times reflect processing difficulty; if we find a correlation, then surprisal predicts behaviour.

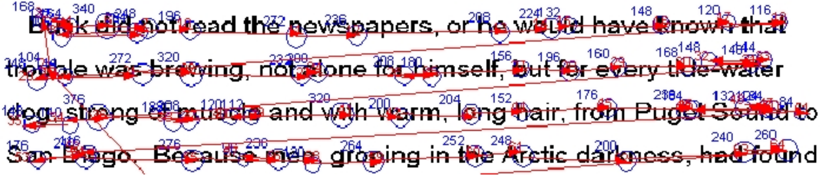
Support from EEG:

- surprisal predictive of N400 amplitudes (Frank et al., 2013)

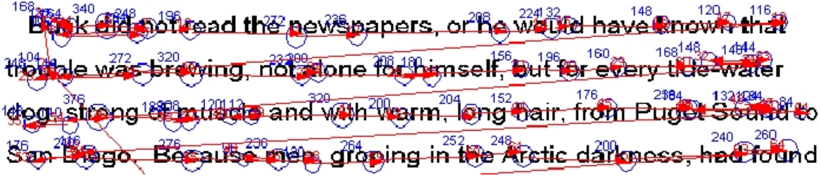
N400

N400 has been linked to predictability, difficulty in retrieving / integrating a word.

The Dundee Corpus (Kennedy and Pynte 2005)

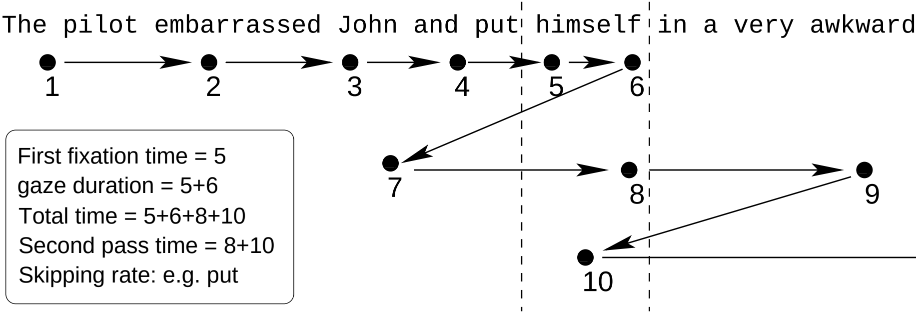


The Dundee Corpus (Kennedy and Pynte 2005)



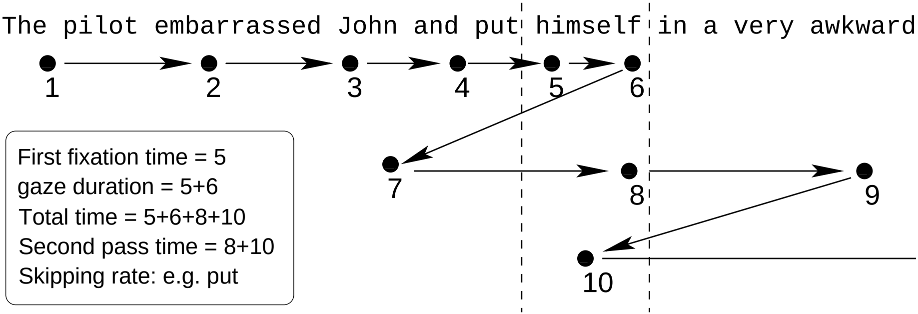
- 51,000 words of British newspaper articles (The Independent)
- 10 subjects read the whole text and answered comprehension questions
- Eye-movements recorded
- Data Cleaning:
  - exclude first and last word of a line
  - exclude words adjacent to punctuation
  - remove tracklosses
  - remove words including numbers

Reading Time Measures



First fixation time = 5
gaze duration = 5+6
Total time = 5+6+8+10
Second pass time = 8+10
Skipping rate: e.g. put

Reading Time Measures



First fixation time = 5
gaze duration = 5+6
Total time = 5+6+8+10
Second pass time = 8+10
Skipping rate: e.g. put

- What are the different measures at “John”?
- Why should we distinguish between different measures?

Evaluation on Corpus Data

Use **Eye-tracking Corpora** as complementary evidence to experimental data:

- Sentences are read **in context**
- "real" language, **naturally occurring** text
- Test on many different constructions
- Evaluate **many theories on same data** to obtain better comparability
- But: **less control** over materials

Evaluation on Corpus Data

Use **Eye-tracking Corpora** as complementary evidence to experimental data:

- Sentences are read **in context**
- "real" language, **naturally occurring** text
- Test on many different constructions
- Evaluate **many theories on same data** to obtain better comparability
- But: **less control** over materials

Method:

- Calculated **Surprisal** for each word in the corpus based on Roark parser [Roark, 2001, 2009]
- Calculated **DLT** Integration Costs (IC) for each word based on MINIPAR [Lin, 1998]

Broad-Coverage Evaluation on Dundee Corpus

Correlation between Theories:

	Integration Cost	Lexical Surprisal
Lexical Surprisal	0.19	
Structural Surprisal	-0.09	0.36

Linear Mixed Effect Models

- All variables and binary interactions entered into a hierarchical linear mixed effects model
- Full random effects structure
- Stepwise removal of variables that decrease model quality (using AIC)

Linear Mixed Effect Models

- All variables and binary interactions entered into a hierarchical linear mixed effects model
- Full random effects structure
- Stepwise removal of variables that decrease model quality (using AIC)

Random variable:	Covariates:	Independent variable:
subject ID	word length	integration cost
	log frequency	lexical surprisal
Dependent variables:	word position	structural surprisal
first fixation duration	previous fixation	
gaze duration	launch distance	
total reading time	fixation land position	

Broad-Coverage Evaluation on Dundee Corpus

	Total Time	
Predictor	Coef	Sig
(Intercept)	254.07	***
WordLength	7.36	***
WordFrequency	-15.80	***
PreviousWordFrequency	-6.35	***
PreviousWordFixated	-35.60	***
LaunchDistance	-0.86	
LandingPosition	-21.39	***
SentencePosition	-0.28	***
ForwardBigramSurprisal	2.77	***
BackwardBigramSurprisal	-1.36	**
WordLength:WordFrequency	-4.15	***
IntegrationCost	-2.82	***
LexicalSurprisal	-0.16	
StructuralSurprisal	1.21	***

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

Methodological interlude

What is...

- Random intercept?
- Random slope for predictor?
- Full random effects structure?
- "conservative"

Watch out for

- Collinearity
- Model selection

A more problematic example from the literature

	Estimate	Std. Error	t-value
Open-class			
(Intercept)	$2.40 \times 10^{+00}$	2.39×10^{-02}	100.4*
Lexical Surprisal	-1.99×10^{-04}	7.28×10^{-04}	-0.3
Word Length	8.97×10^{-04}	4.62×10^{-04}	1.9
Bigram	4.18×10^{-04}	5.27×10^{-04}	0.8
Unigram Freq	-2.43×10^{-03}	1.20×10^{-03}	-2.0*
Derivation Steps	-1.17×10^{-03}	9.02×10^{-04}	-1.3
Syntactic Entropy	2.55×10^{-03}	6.19×10^{-04}	4.1*
Lexical Entropy	3.96×10^{-04}	6.68×10^{-04}	0.6
Syntactic Surprisal	3.28×10^{-03}	9.71×10^{-04}	3.4*
Order in narrative	-1.43×10^{-05}	4.34×10^{-06}	-3.3*
POS Surprisal	-6.84×10^{-04}	8.11×10^{-04}	-0.8
POS Entropy	1.47×10^{-03}	6.05×10^{-04}	2.4*

Table: Mixed effects models Roark (2009)

A more problematic example from the literature

Predictor	SynH	LexH	SynS	LexS	Freq	Bgrm	PosS	PosH	Step	WLen
Syntactic Entropy (SynH)	1.00	-0.26	0.00	0.24	-0.24	0.20	0.02	0.55	-0.05	0.18
Lexical Entropy (LexH)	-0.26	1.00	0.01	-0.40	0.43	-0.38	-0.03	0.02	0.11	-0.29
Syntactic Surprisal (SynS)	0.00	0.01	1.00	-0.12	0.08	0.18	0.77	0.21	0.38	-0.03
Lexical Surprisal (LexS)	0.24	-0.40	-0.12	1.00	-0.81	0.87	-0.10	-0.20	-0.35	0.64
Unigram Frequency (Freq)	-0.24	0.43	0.08	-0.81	1.00	-0.69	0.02	0.18	0.31	-0.72
Bigram Probability (Bgrm)	0.20	-0.38	0.18	0.87	-0.69	1.00	0.11	-0.11	-0.16	0.56
POS Surprisal (PosS)	0.02	-0.03	0.77	-0.10	0.02	0.11	1.00	0.22	0.32	0.02
POS Entropy (PosH)	0.55	0.02	0.21	-0.20	0.18	-0.11	0.22	1.00	0.16	-0.11
Derivation steps (Step)	-0.05	0.11	0.38	-0.35	0.31	-0.16	0.32	0.16	1.00	-0.24
Word Length (WLen)	0.18	-0.29	-0.03	0.64	-0.72	0.56	0.02	-0.11	-0.24	1.00

Table: Correlations of predictors for models in Roark (2009)

Note: very high correlations for

- Frequency, Lexical Surprisal, Bigram Prob, (Word Length)
- Syntactic surprisal and POS Surprisal

Watch out for terminology in the literature

- lexicalized surprisal refers to surprisal calculated based on a syntactic parser, combination of both lexical and structural surprisal; used to contrast with "POS surprisal". (Not what you should use anymore nowadays.)
- Syntactic surprisal used ambiguously: sometimes refers to surprisal calculated via a syntactic parser, sometimes only to the structural portion of it.

Conclusion so far:

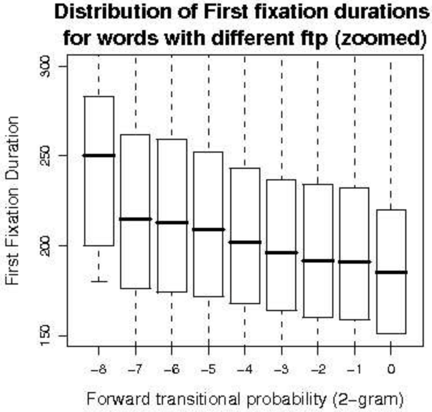
- Syntactic surprisal is predictive of reading times over and above simple word frequencies and bigram surprisal.
- Syntactic surprisal refers to the portion of surprisal that is caused by syntactic structure, ignoring lexical probability.
- Lexical surprisal is highly correlated with word frequency.

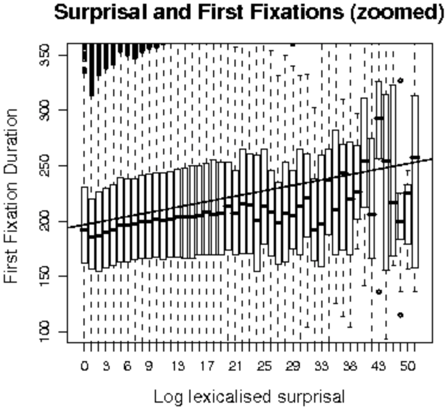
Conclusion so far:

- Syntactic surprisal is predictive of reading times over and above simple word frequencies and bigram surprisal.
- Syntactic surprisal refers to the portion of surprisal that is caused by syntactic structure, ignoring lexical probability.
- Lexical surprisal is highly correlated with word frequency.

Does this relationship between surprisal and reading times hold across the whole range of surprisal values? Or does it just flatten out at some point when the word is not among the very strongly predictable?

Effect of Surprisal on Reading times

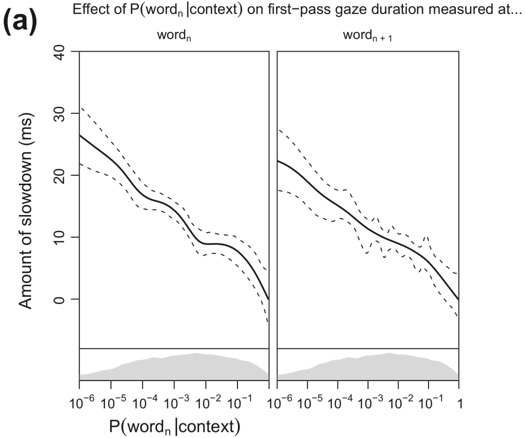




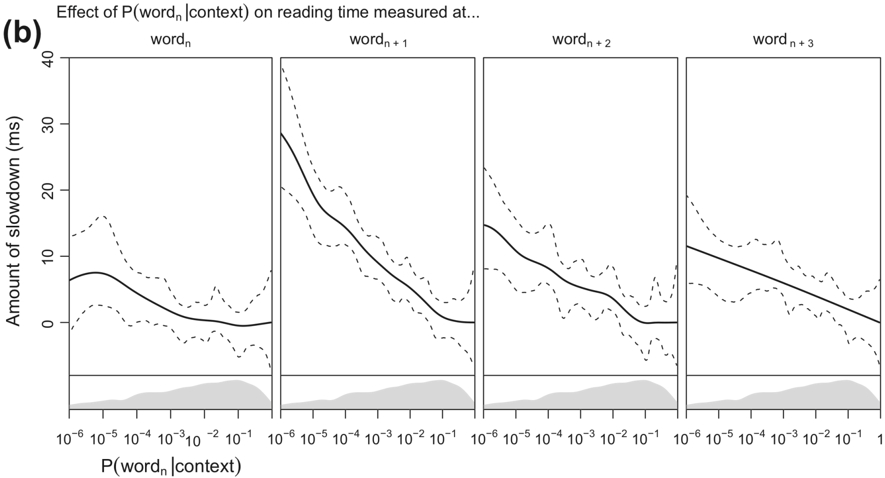
Demberg (2010), reading times on Dundee corpus.

Effect of Surprisal on Reading Times

Smith and Levy (2013) have a whole paper focussed on this question.



Effect of Surprisal on Reading Times



If you're using self-paced reading as a measure, make sure you analyse word n+1!

Surprisal and ERPs

Can we also correlate surprisal to the event related potentials we observe in EEG studies?

- N400 would be a good candidate, as it's long been known to respond to predictability
- Smith and Levy (2010) showed that cloze and corpus-estimated surprisal are at least somewhat similar ($\rho = 0.5$)

Surprisal and ERPs

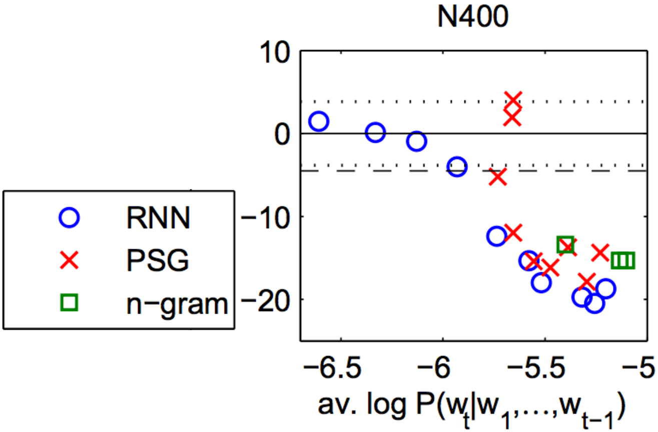
Can we also correlate surprisal to the event related potentials we observe in EEG studies?

- N400 would be a good candidate, as it's long been known to respond to predictability
- Smith and Levy (2010) showed that cloze and corpus-estimated surprisal are at least somewhat similar ($\rho = 0.5$)

Method: Linear mixed effects model with

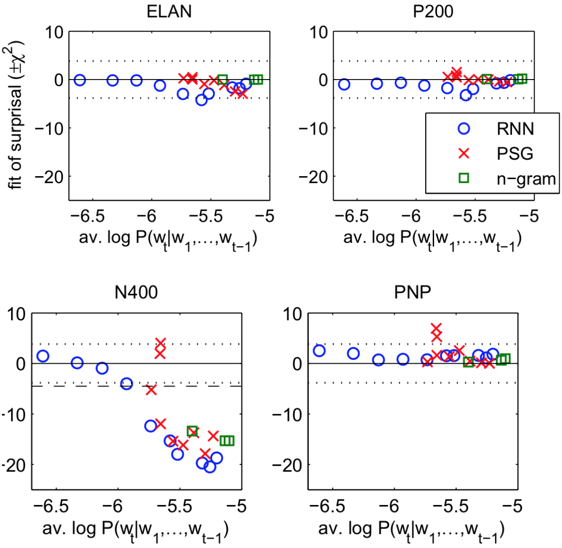
- baseline potential
- log-transformed word frequency
- word length (number of characters),
- word position in the sentence
- sentence position in the experiment

Modelling processing difficulty: Surprisal



This plotting method is very unusual:
it shows the χ^2 from comparing a model with vs. without surprisal as a predictor;
Positive / negative shows the direction of the regression coefficient.
This does not show whether the effect is linear.

No correlation with other ERP measures



Conclusions

- Surprisal is correlated with human reading times and the N400.
- i.e. there is evidence that this notion of the information to be processed has explanatory power for human language processing.
- How surprisal is estimated also matters!

Table of Contents

Information-theoretic measures

Different accounts of how predictability / uncertainty might affect sentence processing have also been suggested:

- Surprisal (aka *pointwise entropy*)
  How unexpected was the word?
- Entropy Reduction
  The amount by which a word reduces the uncertainty about the rest of the sentence.
- Entropy (one step vs. multi-step)
  The uncertainty about the next word / the rest of the sentence;
  related to competition models
- Commitment (higher difficulty for changing top-ranking analysis)
  Surprisal should have larger effect after highly-contraining contexts.

Entropy Reduction

Hale 2003, 2006:

- Hypothesis: a word is difficult to process if it greatly reduces the **uncertainty about the rest of the sentence**.
- Uncertainty is quantified as the entropy of the distribution over **complete parses of the sentence**; that is, if A_i is the set of all possible parses of the sentence after word w_i, then the uncertainty following wi is given by

$$H_{w_i} = -\sum_{a \in A^i} P(a) \log P(a)$$

- Processing load proportional to

$$ER(w_n) = \max\{H_{w_n} - H_{w_{n-1}}, 0\}$$

Entropy Reduction

Hale 2003, 2006:

- Hypothesis: a word is difficult to process if it greatly reduces the **uncertainty about the rest of the sentence**.
- Uncertainty is quantified as the entropy of the distribution over **complete parses of the sentence**; that is, if A_i is the set of all possible parses of the sentence after word w_i, then the uncertainty following wi is given by

$$H_{w_i} = -\sum_{a \in A^i} P(a) \log P(a)$$

- Processing load proportional to

$$ER(w_n) = \max\{H_{w_n} - H_{w_{n-1}}, 0\}$$

- Extremely hard to calculate for large grammars.

Entropy

But what about entropy itself as a measure?

- Hypothesis: word is difficult because there is lots of uncertainty about how the sentence will continue
- related to competition hypothesis (McRae et al., 1998; Tabor and Tanenhaus, 1999)
- Uncertainty about what? Complete rest of sentence or next word?
- Has been approximated by calculating the uncertainty about the next word (e.g., Roark, 2009).
- "One-step" vs. "multi-step" entropy

(Beware some sloppyness in use of terms in the literature, there sometimes seems to be some confusion regarding Entropy vs. ER hypotheses.)

Information-theoretic measures

Different accounts of how predictability / uncertainty might affect sentence processing have also been suggested:

- Surprisal (aka *pointwise entropy*)
  How unexpected was the word?
- Entropy Reduction
  The amount by which a word reduces the uncertainty about the rest of the sentence.
- Entropy (one step vs. multi-step)
  The uncertainty about the next word / the rest of the sentence;
  related to competition models
- Commitment (higher difficulty for changing top-ranking analysis)
  Surprisal should have larger effect after highly-contraining contexts.

Example (from Linzen and Jaeger, 2014)

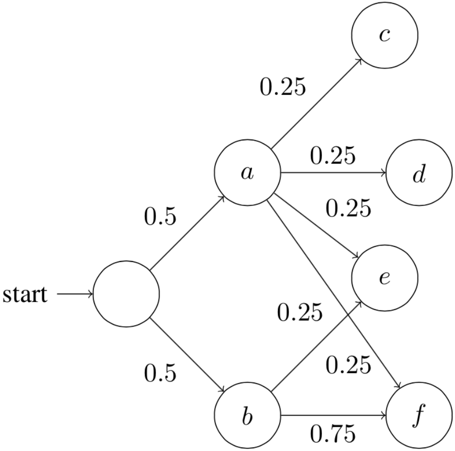


Figure 1: Example language. Output strings are indicated inside the nodes, and transition probabilities are indicated on the edges. For example, the probability of the sentence bf is 0.5×0.75.

Consider “sentences” **ae** vs. **be**:

- Surprisal?
- ER?
- Entropy?
- Commitment?

Example (from Linzen and Jaeger, 2014)

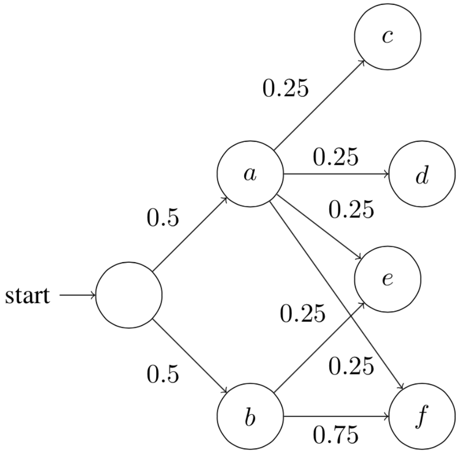


Figure 1: Example language. Output strings are indicated inside the nodes, and transition probabilities are indicated on the edges. For example, the probability of the sentence bf is 0.5×0.75.

Consider "sentences" **ae** vs. **be**:

- Surprisal? ae = be
- ER?
- Entropy?
- Commitment?

Example (from Linzen and Jaeger, 2014)

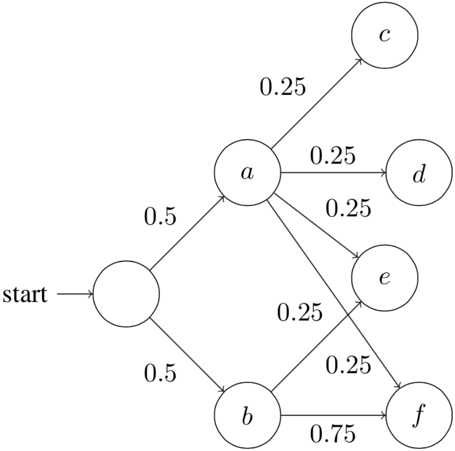


Figure 1: Example language. Output strings are indicated inside the nodes, and transition probabilities are indicated on the edges. For example, the probability of the sentence bf is 0.5×0.75.

Consider "sentences" **ae** vs. **be**:

- Surprisal? ae = be
- ER? b > a and a**e** > b**e**
- Entropy?
- Commitment?

Example (from Linzen and Jaeger, 2014)

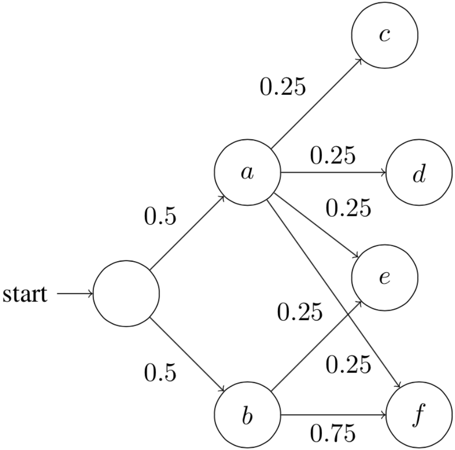


Figure 1: Example language. Output strings are indicated inside the nodes, and transition probabilities are indicated on the edges. For example, the probability of the sentence bf is 0.5×0.75.

Consider "sentences" **ae** vs. **be**:

- Surprisal? ae = be
- ER? b > a and a**e** > b**e**
- Entropy? a>b and e=e
- Commitment?

Example (from Linzen and Jaeger, 2014)

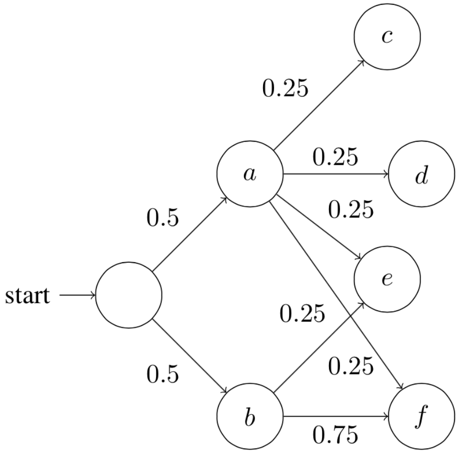


Figure 1: Example language. Output strings are indicated inside the nodes, and transition probabilities are indicated on the edges. For example, the probability of the sentence bf is 0.5×0.75.

Consider "sentences" **ae** vs. **be**:

- Surprisal? ae = be
- ER? b > a and a**e** > b**e**
- Entropy? a>b and e=e
- Commitment? **b**e > a**e**

Information-theoretic measures

Different accounts of how predictability / uncertainty might affect sentence processing have also been suggested:

- Surprisal (aka *pointwise entropy*)
  How unexpected was the word?
- Entropy Reduction
  The amount by which a word reduces the uncertainty about the rest of the sentence.
- Entropy (one step vs. multi-step)
  The uncertainty about the next word / the rest of the sentence;
  related to competition models
- Commitment (higher difficulty for changing top-ranking analysis)
  Surprisal should have larger effect after highly-contraining contexts.

High or low Surprisal / ER / Entropy / Commitment?

The horse raced past the barn fell.

Information-theoretic measures

Different accounts of how predictability / uncertainty might affect sentence processing have also been suggested:

- Surprisal (aka *pointwise entropy*)
  How unexpected was the word?
- Entropy Reduction
  The amount by which a word reduces the uncertainty about the rest of the sentence.
- Entropy (one step vs. multi-step)
  The uncertainty about the next word / the rest of the sentence;
  related to competition models
- Commitment (higher difficulty for changing top-ranking analysis)
  Surprisal should have larger effect after highly-contraining contexts.

High or low Surprisal / ER / Entropy / Commitment?

The children went inside to play.
The children went outside to play.

Information-theoretic measures

Different accounts of how predictability / uncertainty might affect sentence processing have also been suggested:

- Surprisal (aka *pointwise entropy*)
  How unexpected was the word?
- Entropy Reduction
  The amount by which a word reduces the uncertainty about the rest of the sentence.
- Entropy (one step vs. multi-step)
  The uncertainty about the next word / the rest of the sentence; related to competition models
- Commitment (higher difficulty for changing top-ranking analysis)
  Surprisal should have larger effect after highly-contraining contexts.

High or low Surprisal / ER / Entropy / Commitment?

The children went inside to play.
The children went outside to play.

Information-theoretic measures

Different accounts of how predictability / uncertainty might affect sentence processing have also been suggested:

- Surprisal (aka *pointwise entropy*)
  How unexpected was the word?
- Entropy Reduction
  The amount by which a word reduces the uncertainty about the rest of the sentence.
- Entropy (one step vs. multi-step)
  The uncertainty about the next word / the rest of the sentence; related to competition models
- Commitment (higher difficulty for changing top-ranking analysis)
  Surprisal should have larger effect after highly-contraining contexts.

High or low Surprisal / ER / Entropy / Commitment?

The children went inside to look.
The children went outside to look.

Comparison

Examples

	sent. comp. surpr.	subcat entropy
The men forgot the waterfall **had** dried up		
The men heard the waterfall **had** dried up		
The men claimed the waterfall **had** dried up		
The men sensed the waterfall **had** dried up		

Comparison

Surprisal at had?
Subcategorization frame entropy at the verb?

Examples

	sent. comp. surpr.	subcat entropy
The men forgot the waterfall **had** dried up		
The men heard the waterfall **had** dried up		
The men claimed the waterfall **had** dried up		
The men sensed the waterfall **had** dried up		

	NP	Inf	PP	SC
forget	0.55	0.14	0.2	0.09
hear	0.72	0	0.17	0.11
claim	0.36	0.12	0	0.45
sense	0.61	0	0.02	0.34

NP = noun phrase; Inf = infitive;
PP = PP completion; SC = sentence complement

Comparison

Surprisal at had?
Subcategorization frame entropy at the verb?

Examples

	sent. comp. surpr.	subcat entropy
The men forgot the waterfall **had** dried up	3.46	
The men heard the waterfall **had** dried up	3.22	
The men claimed the waterfall **had** dried up	1.15	
The men sensed the waterfall **had** dried up	1.55	

	NP	Inf	PP	SC
forget	0.55	0.14	0.2	0.09
hear	0.72	0	0.17	0.11
claim	0.36	0.12	0	0.45
sense	0.61	0	0.02	0.34

NP = noun phrase; Inf = infitive;
PP = PP completion; SC = sentence complement

Comparison

Surprisal at had?

Subcategorization frame entropy at the verb?

Examples

	sent. comp. surpr.	subcat entropy
The men forgot the waterfall **had** dried up	3.46	1.7
The men heard the waterfall **had** dried up	3.22	1.12
The men claimed the waterfall **had** dried up	1.15	1.71
The men sensed the waterfall **had** dried up	1.55	1.18

	NP	Inf	PP	SC
forget	0.55	0.14	0.2	0.09
hear	0.72	0	0.17	0.11
claim	0.36	0.12	0	0.45
sense	0.61	0	0.02	0.34

NP = noun phrase; Inf = infitive;
PP = PP completion; SC = sentence complement

Experiment

Experimental Items

locally ambiguous	SCS	SE
The men forgot the waterfall **had** dried up		
The men heard the waterfall **had** dried up		
The men claimed the waterfall **had** dried up		
The men sensed the waterfall **had** dried up		
unambiguous		
The men forgot **that** the waterfall had dried up		
The men heard **that** the waterfall had dried up		
The men claimed **that** the waterfall had dried up		
The men sensed **that** the waterfall had dried up		

Methods:

- Self-paced reading via Mechanical Turk
- 128 participants (4 excluded)
- 8 verbs in each condition (32 verbs total); 64 fillers

Experiment

Experimental Items

locally ambiguous	SCS	SE
The men forgot the waterfall **had** dried up		
The men heard the waterfall **had** dried up		
The men claimed the waterfall **had** dried up		
The men sensed the waterfall **had** dried up		
unambiguous		
The men forgot **that** the waterfall had dried up		
The men heard **that** the waterfall had dried up		
The men claimed **that** the waterfall had dried up		
The men sensed **that** the waterfall had dried up		

Methods:

- Self-paced reading via Mechanical Turk
- 128 participants (4 excluded)
- 8 verbs in each condition (32 verbs total); 64 fillers

Experiment

Experimental Items

locally ambiguous	SCS	SE
The men forgot the waterfall **had** dried up	3.46	
The men heard the waterfall **had** dried up	3.22	
The men claimed the waterfall **had** dried up	1.15	
The men sensed the waterfall **had** dried up	1.55	
unambiguous		
The men forgot **that** the waterfall had dried up		
The men heard **that** the waterfall had dried up		
The men claimed **that** the waterfall had dried up		
The men sensed **that** the waterfall had dried up		

Methods:

- Self-paced reading via Mechanical Turk
- 128 participants (4 excluded)
- 8 verbs in each condition (32 verbs total); 64 fillers

Experiment

Experimental Items

locally ambiguous	SCS	SE
The men forgot the waterfall **had** dried up	3.46	1.7
The men heard the waterfall **had** dried up	3.22	1.12
The men claimed the waterfall **had** dried up	1.15	1.71
The men sensed the waterfall **had** dried up	1.55	1.18

unambiguous

The men forgot **that** the waterfall had dried up

The men heard **that** the waterfall had dried up

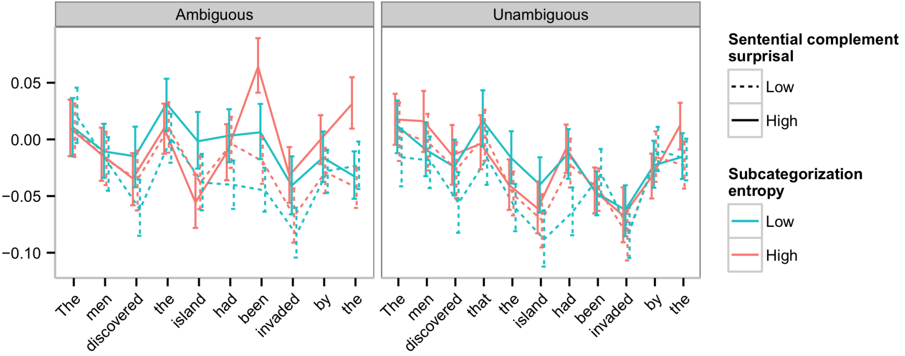
The men claimed **that** the waterfall had dried up

The men sensed **that** the waterfall had dried up

Methods:

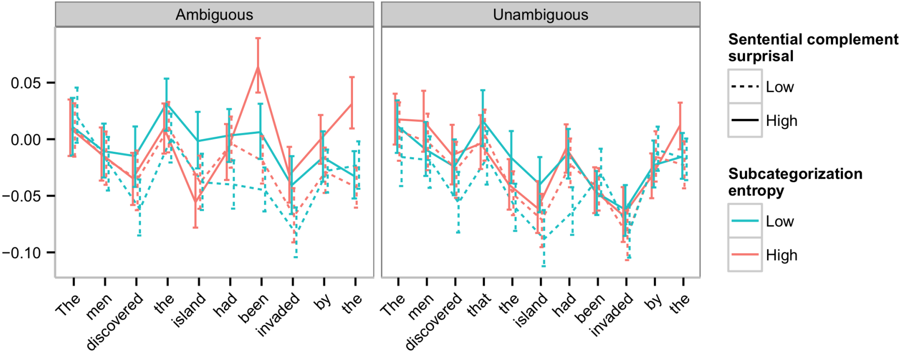
- Self-paced reading via Mechanical Turk
- 128 participants (4 excluded)
- 8 verbs in each condition (32 verbs total); 64 fillers

Results



- no differences on *the men* or *discovered*
- significantly higher RTs for high SC surprisal on *the island* only in unambiguous sentences (*that* present); (spill-over from *that*).
- significantly faster RTs for unambiguous conditions on *had been invaded*.
- RTs on same region significantly higher for high SC surprisal in ambiguous condition.
- no significant effects of subcat frame entropy (expected on verb)

Results



Discussion:

- Evidence for subcat frame surprisal: reading times higher when evidence for unexpected subcategorization frame is encountered.
- But did we calculate / conceptualize entropy in the right way when considering subcat frame entropy??

Discussion: single vs. multistep entropy

Subcat frame entropy is like “single step entropy”.

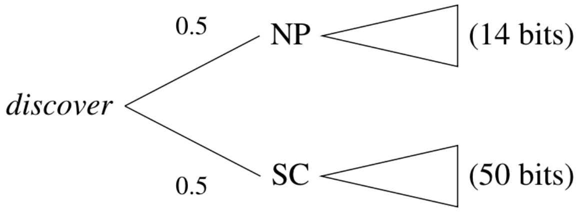
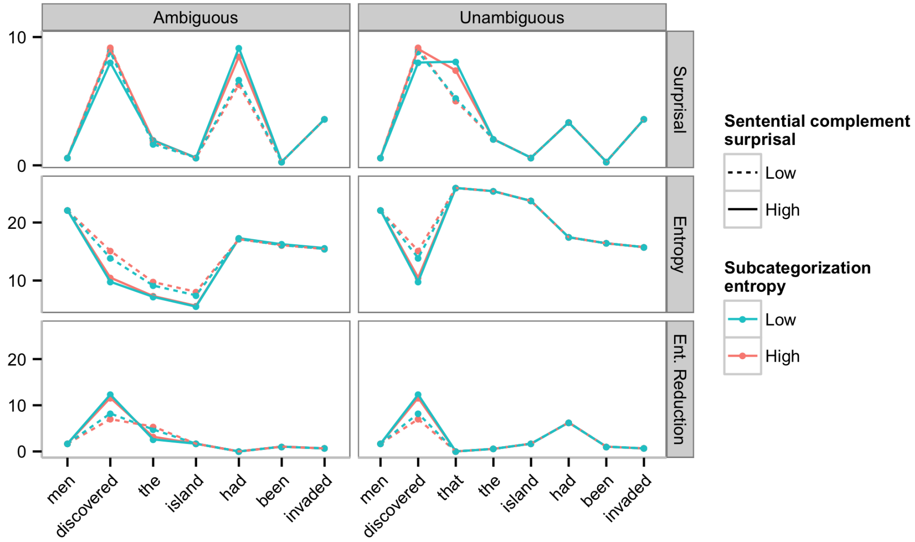
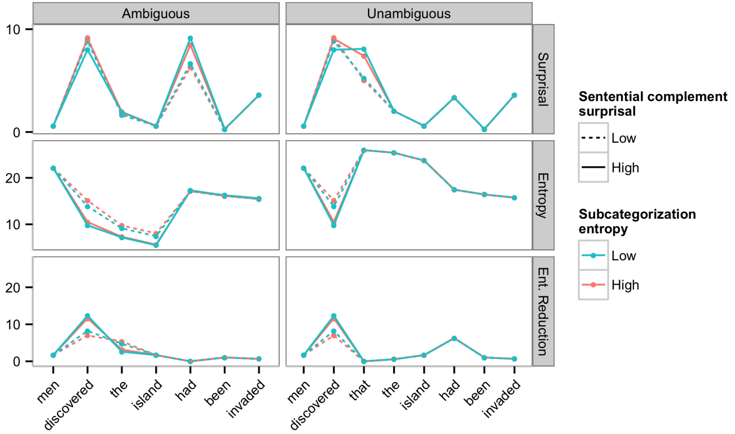


Figure 3: Entropy calculation example: the single step entropy after *discover* is 1 bit; the overall entropy is $1 + 0.5 \times 14 + 0.5 \times 50 = 33$ bits.

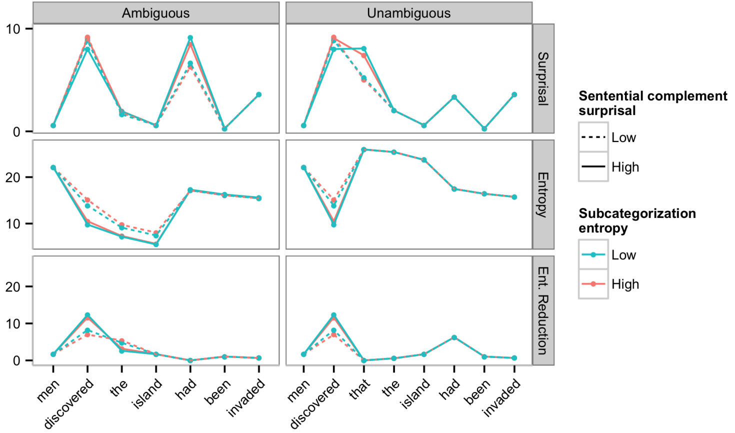
- Single step entropy can be very different from full entropy.
- Linzen and Jaeger calculate the full entropy using an adapted parser.

Model predictions





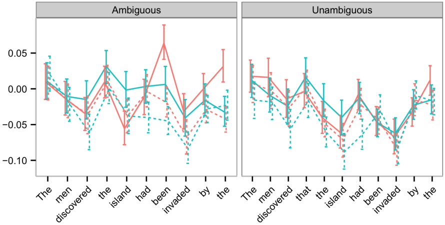
- Surprisal effects as seen in data
- Entropy / ER driven by SC surprisal, not much affected by Subcat entropy.
- Entropy and ER make opposite predictions on main verb:
  - Entropy: verbs that typically take complements are harder to process.
  - ER: verbs that typically take complements are easier to process.

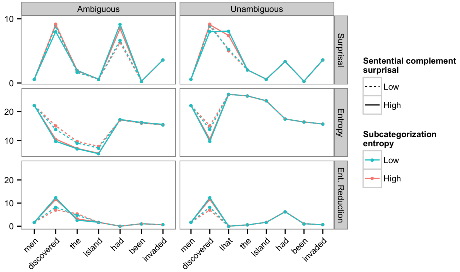


- Surprisal effects as seen in data
- Entropy / ER driven by SC surprisal, not much affected by Subcat entropy.
- Entropy and ER make opposite predictions on main verb:
  - Entropy: verbs that typically take complements are harder to process.
  - ER: verbs that typically take complements are easier to process.

Result: ER was a significant positive predictor of RTs on main verb in lme model.

Overall results





- overall predictions not comparable due to *partial* lexicalization of PCFG model.
- computational modelling was chosen to estimate full entropy compared to just single step entropy (subcat frame entropy).
- experimental design controlled for subcat frame entropy but not full entropy, therefore, needed mixed effects model to estimate effect of full entropy.
- no evidence in Linzen & Jaeger expt for single step entropy
- but RTs on verb were longer for verbs that don't take complement, i.e. when post-verb entropy is lower. This is consistent with entropy reduction.

Overall results / discussion

- Surprisal is not the only way in which information-theoretic concepts have been linked to processing difficulty.
- Surprisal doesn't exclusively explain all effects.
- Other studies have found additional effects of, e.g., level of constraint, entropy reduction (Frank, 2011).
- Distinguish: **Entropy** vs. **Entropy Reduction**
- Estimating entropy of **complete sentence** can be very different from **entropy of next step**