Uncertainty and surprisal in sentence processing

Tal Linzen[1] and T. Florian Jaeger[2]

New York University, Department of Linguistics[1] and University of Rochester, Brain and Cognitive Sciences[2]

Surprisal theory and entropy

- In surprisal theory, processing cost is determined by the conditional probability of the element being processed (Hale 2001, Demberg and Keller 2008, Levy 2008):

$$RT \propto -\log P(w_i \mid w_1, \ldots, w_{i-1})$$

- Surprisal is independent of the distribution of alternatives: ambiguity and disambiguation do not entail a processing cost
- Ambiguity can be quantified using the entropy over parses:

$$H_i = -\sum_{j=1}^{n} p_j \log p_j$$

- Is there an effect of entropy when surprisal is kept constant?

High entropy context

$$\xrightarrow[w_1]{} \begin{cases} 0.25 & a_1 \\ 0.25 & a_2 \\ 0.25 & a_3 \\ 0.25 & a_4 \end{cases} \xrightarrow{w_2\,(ambiguity)} \begin{cases} 0.25 & a_1 \\ 0.25 & a_2 \\ 0.25 & a_3 \\ 0.25 & a_4 \end{cases} \xrightarrow{w_3\,(disambiguation)} \begin{cases} 1 & a_1 \\ 0 & a_2 \\ 0 & a_3 \\ 0 & a_4 \end{cases}$$

Low entropy context

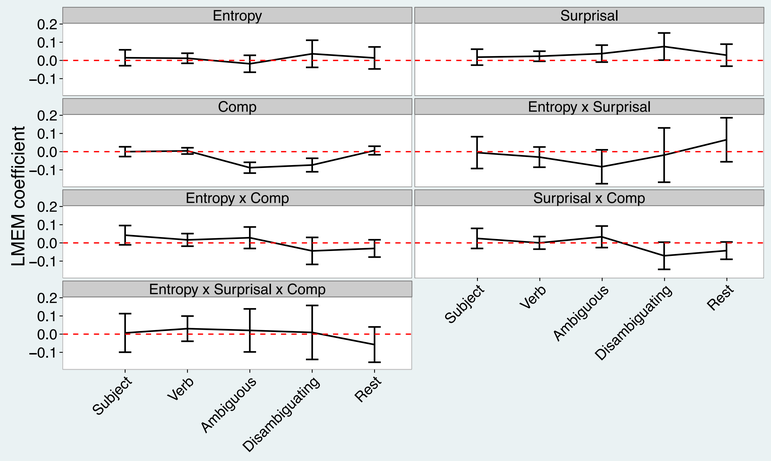
$$\xrightarrow[w_1]{} \begin{cases} 0.25 & a_1 \\ 0.75 & a_2 \end{cases} \xrightarrow{w_2\,(ambiguity)} \begin{cases} 0.25 & a_1 \\ 0.75 & a_2 \end{cases} \xrightarrow{w_3\,(disambiguation)} \begin{cases} 1 & a_1 \\ 0 & a_2 \end{cases}$$

Self-paced reading experiment

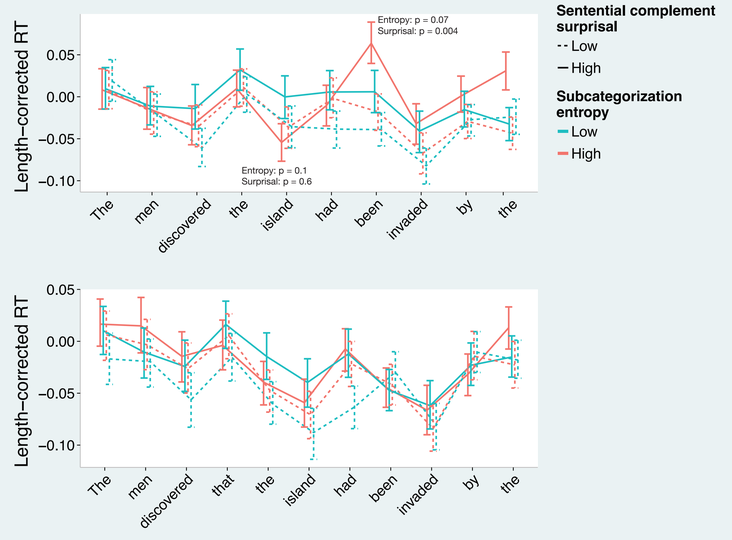
- Materials inspired by Garnsey et al (1997):

The men discovered (that) the island had been invaded by the enemy.
subject *verb* *ambiguous* *disambiguating* *rest*

- Orthogonally varied entropy of subcategorization frame distribution and surprisal of sentential complement (SC)
- Subcategorization frequencies from Gahl et al (2004)
- Presence of complementizer varied within item (across subjects)
- 32 items, 64 fillers
- 128 participants (recruited on Amazon Mechanical Turk)



	NP	Inf (to)	PP	SC (that)	SC surprisal	Entropy
forget	0.55	0.14	0.2	0.09	3.46	1.7
hear	0.72	0.0	0.17	0.11	3.22	1.12
claim	0.36	0.12	0	0.45	1.15	1.71
sense	0.61	0.0	0.02	0.34	1.55	1.18



Entropy in alternative theories

- In its strong form (Levy 2008), surprisal theory predicts no effect of entropy
- The **Entropy Reduction Hypothesis** (**ERH**, Hale 2006) states that disambiguation entails a processing cost:

$$RT \propto \max(0, H_{i-1} - H_i)$$

- Calculating the entropy over all possible parses in a realistic (lexicalized) grammar is difficult; Roark et al. (2009) propose to approximate it using the entropy over the next single step in the derivation ("single-step entropy": h_i)
- **Ambiguity Cost Hypothesis:** Roark et al. (2009) find that reading times on a word correlate positively with single-step entropy after the word: syntactic ambiguity is costly to create (or maintain)
- **Entropy-mediated Surprisal Hypothesis:** Higher ambiguity could lead to delayed commitment to a parse, which may reduce reading times at the ambiguous region and attenuate the effect of surprisal at the disambiguating region

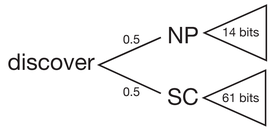
Evaluating the ERH

- How good is the single-step approximation to complete entropy?
- We estimated the complete entropy for each nonterminal category using a PCFG acquired from the WSJ section of the Penn TreeBank (Grenander 1967, Hale 2006)
- At least for subcategorization decisions, single-step entropy is dwarfed by complete entropy:

$$H(a_i) = h(a_i) + \sum_{r \in \Pi_i} p_r \sum_{j=1}^{k_r} H(a_{rj})$$

$$h(\text{discover}) = -2 \log_2 0.5 = 1$$

$$H(\text{discover}) = 1 + 0.5 \times 14 + 0.5 \times 61 = 38.5$$



- In our materials, single-step and complete entropy are correlated (r = 0.46), but not as much as SC surprisal and complete entropy (r = -0.62)
- Assuming that pre-verb entropy is equal to the entropy of a VP, most verbs in our materials (27 out of 32) caused a reduction in entropy
- However, entropy reduction was not a significant predictor of reading times either on the verb or on the following word

Conclusions

- Surprisal effect at disambiguating region replicated (Garnsey et al 1997)
- No support for the Ambiguity Cost Hypothesis (trend in wrong direction)
- Mixed support for the Entropy-mediated Surprisal Hypothesis
- Single-step entropy is a poor approximation of complete entropy
- ERH predicts that SC-biased verbs should be much harder to process than NP-biased verbs; no clear support for this prediction

References

Demberg, V. and Keller, F. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*.

Gahl, S., Jurafsky, D., and Roland, D. 2004. Verb subcategorization frequencies: American English corpus data, methodological studies, and cross-corpus comparisons. *BRM*.

Garnsey, S., Pearlmutter, N., Myers, E., and Lotocky, M. 1997 The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *JML*.

Grenander, U. 1967. Syntax-controlled probabilities. Tech. Rep., Brown University.

Hale, J. 2001. A probabilistic Earley parser as a psycholinguistic model. *NAACL*.

Hale, J. 2006. Uncertainty about the rest of the sentence. *Cognitive Science*.

Levy, R. 2008. Expectation-based syntactic comprehension. *Cognition*.

Roark, B., Bachrach, A., Cardenas, C., and Pallier, C. 2009. Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. *EMNLP*.

This work was supported by an Alfred P. Sloan Fellowship to T. Florian Jaeger.