Computational Psycholinguistics day 3: Surprisal-based sentence processing

Class notes

July 13, 2007

1 Background

Previous usage of predictability-based measures of sentence comprehension difficulty:

- Empirical observations that *Cloze* probability (Taylor, 1953) affects reading times (Ehrlich and Rayner, 1981) and event-related potentials (ERPs; Kutas and Hillyard, 1980, 1984)

My brother came inside to...	chat? eat? play? rest?
The children went outside to...	chat? eat? play? rest?

- Use of root mean-squared (RMS) word-prediction error to evaluate neural-net learning of natural language sentences (Elman, 1990, 1991; Christiansen and Chater, 1999; MacDonald and Christiansen, 2002; Rohde, 2003)
- Predictability is implicated in mathematical models of word reading, but usually on an absolute probability scale (Reichle et al., 1998; Engbert et al., 2005; see McDonald et al., 2005 for an exception)

2 Hale 2001

Proposals:

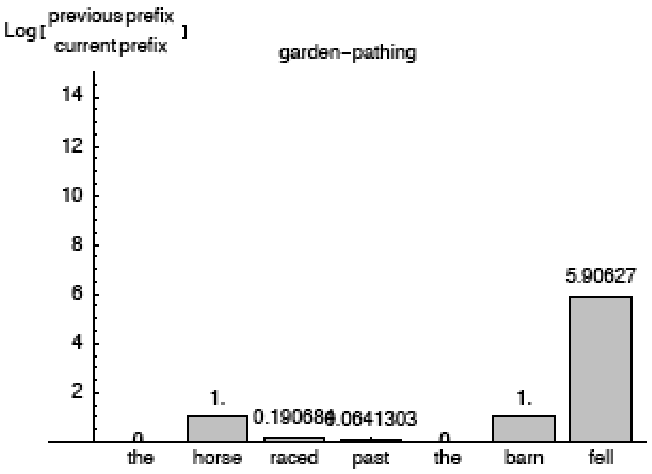
1. Probabilistic context-free grammars (PCFGs) are a good model of how human sentence comprehension works.
2. A probabilistic Earley parser is a good model of online *eager* sentence comprehension for PCFGs

3. The cognitive effort associated with a word in a sentence can be measured by the word's negative log conditional probability:

$$log\frac{1}{P(w_i|w_{1\cdots i-1})}$$

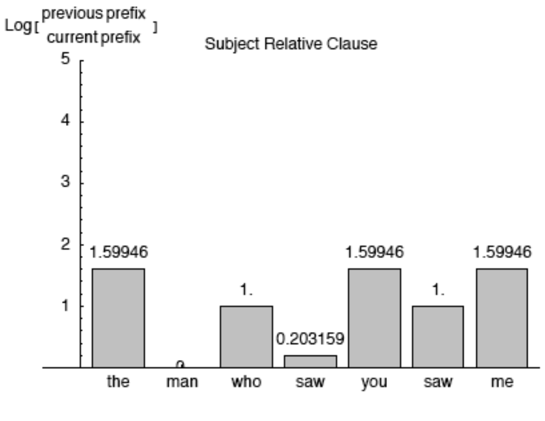
Results from this proposal:

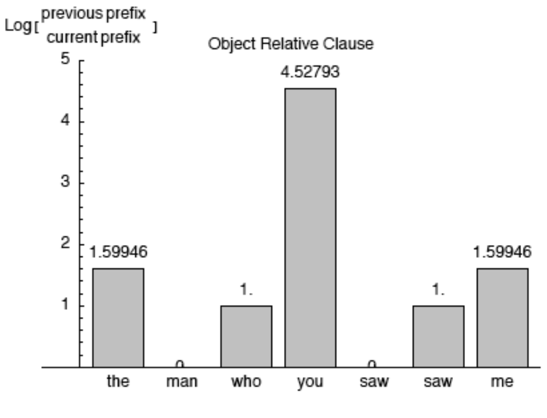
1. Garden path sentences: *the horse raced past the barn fell*



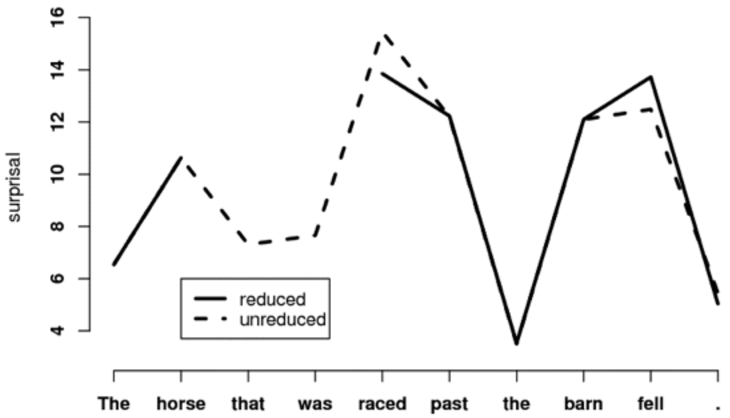
2. Subject/object RC ambiguities: one of the best-established processing asymmetries is the English subject/object RC asymmetry:

   The reporter who attacked the senator $<^{easier}$
   The reporter who the senator attacked





Caveat: these results are with a tiny, mostly hand-crafted grammar. Example using the entire Brown corpus:



3 Levy 2007

A different derivation of surprisal:

1. Let the probability distribution over *complete structures* T (e.g., context-free trees) given a string prefix $w_{1 \cdots i}$ be denoted as $P_i(T)$.
2. The *relative entropy*, or *Kullback-Leibler divergence* (Cover and Thomas, 1991), $D(q||p)$ between two probability distributions is
   - a natural (though asymmetric) measure of probabilistic distance;
   - can be thought of as the penalty incurred for using the distribution p to encode the finer-grained distribution q.
3. It turns out that the relative entropy between distributions before and after a word w is equivalent to the surprisal of w:

$$D(P_{k+1}||P_k) = \log \frac{1}{P_k(w_{k+1})}$$

4. If we think of probabilistic distance as the amount of work involved in reranking the candidate set T, then surprisal is also a measure of reranking work.

Results:

1. Constrained syntactic contexts.

   German verb-final clauses (Konieczny, 2000):

(1) a. Er hat den Abgeordneten begleitet, und ...
He has the delegate escorted, and ...
"He escorted the delegate, and ..."

b. Er hat den Abgeordneten ans Rednerpult begleitet, und ...
He has the delegate to_the lectern escorted, and ...
“He escorted the delegate to the lectern, and ...”

c. Er hat den Abgeordneten an das große Rednerpult begleitet, und ...
He has the delegate to the big lectern escorted, and ...
“He escorted the delegate to the large lectern, and ...”

	Average RT (ms)	Surprisal	DLT prediction
no PP	514	15.99	faster
short PP	477	15.41	slower
long PP	463	15.35	slower

2. Verb identity versus verb location (Jaeger et al., 2005):

(2) a. The player [that the coach met **at 8 o’clock**] bought the house...

b. The player [that the coach met *by the river* **at 8 o’clock**] bought the house...

c. The player [that the coach met NEAR THE GYM *by the river* **at 8 o’clock**] bought the house...

	Number of PPs intervening between embedded and matrix verb		
	1 PP	2 PPs	3 PPs
DLT prediction	Easier	Harder	Hardest
Surprisal	13.87	13.54	13.40
Mean Reading Time (ms)	510 ± 34	410 ± 21	394 ± 16

3. Facilitative ambiguity:

(3) (Traxler et al., 1998)

a. The daughter$_i$ of the colonel$_j$ who shot herself$_{i/*j}$ on the balcony had been very depressed.

b. The daughter$_i$ of the colonel$_j$ who shot himself$_{*i/j}$ on the balcony had been very depressed.

c. The son$_i$ of the colonel$_j$ who shot himself$_{i/j}$ on the balcony had been very depressed.

The ambiguous form can derive probability mass from both attachments; the unambiguous form can only derive mass from one attachment.

4 Other developments

- Other views of surprisal

  - Smith (2006) has shown that surprisal can be derived as a highly general optimization of a time/resource tradeoff, assuming only a *scale-free property* (that the cost of a unit U can be derived as the sum of the costs of the subunits $u_{1 \cdots n}$ that make it up)
  - This works because joint events are characterized by products of probabilities, and the log of a product is the sum of logs

- Surprisal and sentence production:
  - With some extra (empiricially testable) assumptions, surprisal can lead to the idea of *uniform information density* (UID)
  - Under UID, optimal communication involves smoothing out the surprisal profile of an utterance

References

Christiansen, M. H. and Chater, N. (1999). Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science*, 23(2):157–205.

Cover, T. and Thomas, J. (1991). *Elements of Information Theory.* John Wiley.

Ehrlich, S. F. and Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior*, 20:641–655.

Elman, J. (1990). Finding structure in time. *Cognitive Science*, 14:179–211.

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, 7(2-3):195–225.

Engbert, R., Nuthmann, A., Richter, E. M., and Kliegl, R. (2005). SWIFT: a dynamical model of saccade generation during reading. *Psychological Review*, 112(4):777–813.

Jaeger, F., Fedorenko, E., and Gibson, E. (2005). Dissociation between production and comprehension complexity. Poster Presentation at the 18th CUNY Sentence Processing Conference, University of Arizona.

Konieczny, L. (2000). Locality and parsing complexity. *Journal of Psycholinguistic Research*, 29(6):627–645.

Kutas, M. and Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, 207(4427):203–205.

Kutas, M. and Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, 307:161–163.

MacDonald, M. C. and Christiansen, M. H. (2002). Reassessing working memory: Comment on Just and Carpenter (1992) and Waters and Caplan (1996). *Psychological Review*, 109(1):35–54.

McDonald, S. A., Carpenter, R., and Shillcock, R. C. (2005). An anatomically constrained, stochastic model of eye movement control in reading. *Psychological Science*, 112(4):814–840.

Reichle, E. D., Pollatsek, A., Fisher, D. L., and Rayner, K. (1998). Toward a model of eye movement control in reading. *Psychological Review*, 105(1):125–157.

Rohde, D. (2003). The `tgrep2` manual.

Smith, N. (2006). Surprisal-based sentence processing as optimal behavior. M.S., UC San Diego.

Taylor, W. L. (1953). A new tool for measuring readability. *Journalism Quarterly*, 30:415.

Traxler, M. J., Pickering, M. J., and Clifton, C. (1998). Adjunct attachment is not a form of lexical ambiguity resolution. *Journal of Memory and Language*, 39:558–592.