Processing Difficulties: Surprisal or Complexity?

Situations where surprisal/expectation/frequency-based accounts make the wrong predictions are frequently used to argue for complexity-based accounts of processing difficulty. Recent modeling of eye-tracking corpus data shows that object relative clauses are read more slowly even when frequency factors are taken into account, providing support for complexity accounts (Demberg and Keller 2007).

This poster presents data suggesting:

- The difficulties observed in processing object relative clauses in the eye-tracking corpus are the result of a single (difficult) sentence, and that when that sentence is removed, there is no difference between the reading times of subject and object relative clauses.
- The processing of object relative clauses is highly influenced by discourse expectations.
- Testing structures in isolation may provide misleading results.
- Models of surprisal/expectation/frequency need to take complex factors such as discourse into account.
- Phenomena presently attributed to memory limitations may be explainable in terms of more complete models of surprisal/expectation/frequency.

What is Surprisal?

Surprisal is a formalization of the notion that unexpected (or less frequent) things are more difficult to process.

Reading Time \sim - \log P(\text{word/context})

- While the idea unexpected words take longer to read is simple, actually determining how likely a word is in a given context can be very hard.
- Incorrect difficulty predictions by surprisal models may be due to inadequacies in calculating P(word/context), rather than other factors such as structural complexity.
- Treebank structural probabilities are the most common source of information for P(word/context): - Priors due to general bias against object relatives in both head final and head initial languages
- Complexity theories make the wrong prediction for head final languages (e.g., Dependency Locality Theory - Gibson 1998, 2000)
- Predicts for English (e.g., Catford's Complexity Theory - Catford 1961, 1967)
- Predicts for word order in running text (e.g., Boston et al., 2001; Demberg & Keller, 2008)
- Surprisal (when based on Treebank structural probabilities) cannot account for effects of thematic fit, animacy, pronounization, and discourse structure on object relative clause processing (e.g., Gernsbacher & Musolino, 2008; Hale et al., 2008; Hauk & Christiansen, 2007; Traxler et al., 2000)
- Localized Treebank probabilities are not better than structure only probabilities (e.g., Demberg and Keller 2008), probably because of training data scarcity.

Objects Relative difficulty in eye-tracking corpus caused by single (difficult) sentence

Recent modeling shows that object relative clauses are read more slowly even when frequency factors are taken into account, providing support for complexity accounts (Demberg and Keller 2007).

- Data – Dundee Corpus (Kennedy & Pynte 2005)
  - Data from 60 participants
  - Includes ~ 400 relative clauses in “natural” contexts

Regression modeling for prediction reading times for the embedded verb region

- Predictors (+ binary interactions)
  - Relative pronoun (who/whom/wh)
  - Word length
  - Log frequency (from BNC data)
  - Log forward/transition probability (from BNC data)
  - Word landing position
- Subject (random variable)
- Dependent variables (separate models for each)
  - Total fixation duration
  - First pass duration
  - First fixation duration

Results: The difficulties observed in processing object relative clauses in the eye-tracking corpus are the result of a single (difficult) sentence, and that when that sentence is removed, there is no difference between the reading times of subject and object relative clauses.