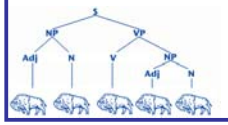
Relative Clauses and Surprisal



Douglas Roland
University at Buffalo
droland@buffalo.edu

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 is at odds with recent experimental work in Dutch, English, and Japanese showing that object relative clauses are not more difficult when appropriate discourse contexts are provided (Mak et al. 2008, Roland et al. 2007,2008, Sato et al. 2008).

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 from consideration, 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 $\propto$ - log P(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):
  - Predicts 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 reading times for a variety of structures of psycholinguistic interest (e.g., Hale 2001, Levy 2008)
  - Predicts reading times of words in running text (e.g., Boston et al., 2008; Demberg & Keller, 2008)
  - Surprisal (when based on Treebank structural probabilities) can not account for effects of thematic fit, animacy, pronominalization, and discourse structure on object relative clause processing (e.g., Gennari & MacDonald, 2008; Mak et al., 2008; Reali & Christiansen, 2007; Traxler et al., 2005)
  - Lexicalized Treebank probabilities are not better than structure only probabilities (e.g., Demberg and Keller 2008), probably because of training data scarcity

Discourse Context Affects Processing

Observation

- Embedded NP usually old information in object relative clauses

Theory

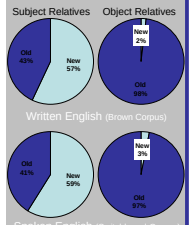
- Object relative clauses are more "unnatural" in isolation
  - The lady that visited the banker enjoyed the dinner very much.
  - The lady that the banker visited enjoyed the dinner very much.

New information OK here

New information *unexpected* here

See Fox and Thompson (1990) more about why object relatives are used this way

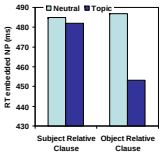
Relative Clauses differ in information status



Providing discourse topic reduces difficulty

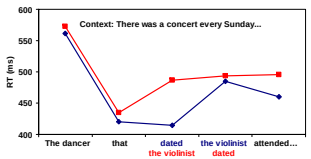
- Participant-paced moving window paradigm
- Neutral vs. Topic preceding context
  - There was a concert every Sunday afternoon
  - The violinist was very famous
- Subject vs. Object relative clause
  - The dancer that dated the violinist attended every one of his performances.
  - The dancer that the violinist dated attended every one of his performances.

Repeated NP *benefit* in object relative

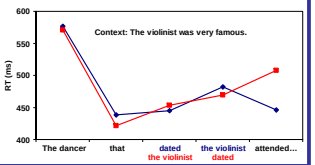


While a pronoun would be preferred to the full NP *the violinist*, a reference to *violinist* is clearly the expected in the object relative clause.

Classic object relative clause difficulty after neutral context



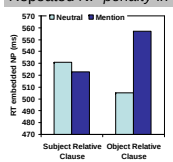
Topic discourse context eliminates object relative difficulty during relative clause



Reference to topic expected in object relative

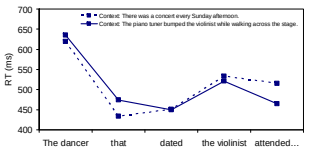
- Participant-paced moving window paradigm
- Neutral vs. Mention preceding context
  - There was a concert every Sunday afternoon.
  - The piano tuner bumped the violinist while walking across the stage.
- Subject vs. Object relative clause
  - The dancer that dated the violinist attended every one of his performances.
  - The dancer that the violinist dated attended every one of his performances.

Repeated NP *penalty* in object relative

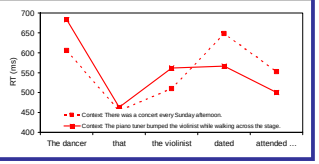


The violinist is read more slowly in the object relative/mention condition, because *piano tuner* (the topic of the context sentence) is expected. Subject relatives appear to have no such expectations.

Subject relatives (again) show no effect of context



After "wrong NP" penalty, object relatives are read more quickly after Mention context



Object 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)
  - Eye-tracking data for ~50,000 words of news paper text
  - Data from 10 participants
  - Includes ~ 400 relative clauses in 'natural' contexts

Relative Pronoun	Subject Relative	Object Relative
that	158	18
which	77	9*
who	130	1+4 whom
Total	365	32*

Regression modeling for prediction reading times for the embedded verb region

- Predictors (+ binary interactions)
  - Relative pronoun (who|which|that)
  - 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 from consideration, there is no difference between the reading times of subject and object relative clauses.

(data from total reading time at embedded verb shown)

Predictor	Coefficient	p
(Intercept)	190.55	0.00
RC.typeSR	-106.32	0.00
length	20.12	0.00
log.bnc.freq	-4.22	0.47
wordlandingposition	-1.57	0.72
log.transition.probability	-13.47	0.00
RC.typeSR:log.bnc.freq	12.18	0.00
length:wordlandingposition	-1.90	0.00
log.bnc.freq:log.transition.probability	1.30	0.00

Predictor	Coefficient	p
(Intercept)	96.89	0.12
RC.typeSR	20.69	0.56
length	18.45	0.00
log.bnc.freq	6.17	0.29
wordlandingposition	-2.23	0.61
log.transition.probability	-10.82	0.00
RC.typeSR:log.bnc.freq	-1.26	0.75
length:wordlandingposition	-1.74	0.01
log.bnc.freq:log.transition.probability	0.95	0.00

He is, it is said, much more sympathetic to the argument that to wait until then would sap Britain's credibility in Europe; would reduce, perhaps fatally, Britain's chances of filling the leadership vacuum in the EU, and would make it more difficult to reshape policy (not least, to take one topical example, on the over-rigid stability pact, which Gordon Brown has rightly investigated against).

Acknowledgements

We would like to thank:
- The University at Buffalo Research Foundation
- Carolyn O'Meara, Hongoak Yun, Gail Mauner, Jean-Pierre Koenig, and the other members of the Psycholinguistics Lab at the University at Buffalo
- The Computational Psycholinguistics Lab at the University of Buffalo

References

Boston, M. F., Hale, J., Kliegl, R., Patil, U., & Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Research*.

Demberg, V., & Keller, F. (2007). Eye-tracking Evidence for Integration Cost Effects in Corpus Data. In *Proceedings of the 29th meeting of the Cognitive Science Society* (CogSci-07), Nashville, Tennessee, US, August 2007.

Demberg, V., & Keller, F. (2008). Data from Eye-tracking Corpora as Evidence for Theories of Syntactic Processing Complexity. *To appear in Cognition*.

Fox, B. A., & Thompson, S. A. (1990). A discourse explanation of the grammar of relative clauses in English conversation. *Language*, 66(2), 297-316.

Gennari, S. P., & MacDonald, M. C. (2008). Semantic indeterminacy in object relative clauses. *Journal of Memory and Language*, 58, 161-187.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1), 1-76.

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In Y. Miyashita, A. Marantz & W. O'Neil (Eds.), *Image, Language, Brain* (pp. 95-126). Cambridge, MA: MIT Press.

Hale, J. (2001). A Probabilistic Earley Parser as a Psycholinguistic Model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Kennedy, A. & Pynte, J. (2005). Parafoveal-on-Foveal effects in normal reading. *Vision Research*, v.45, p153 - 168

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106, 1126-1177.

Mak, W. M., Vonk, W., & Schriefers, H. (2008). Discourse structure and relative clause processing. *Memory & Cognition*, 36(1), 170-181.

Reali, F., & Christiansen, M. H. (2007). Processing of relative clauses is made easier by frequency of occurrence. *Journal of Memory and Language*, 57(1), 1-23.

Sato, A., Kahraman, B., & Sakai, H. (2008). Processing object relatives clauses in context: Another support for the discourse functional account for the processing load asymmetry. *IEICE Tech. Rep.*, vol. 108, no. 184, TL2008-20, pp. 95-100, Aug. 2008.

Traxler, M. J., Williams, R. S., Blozis, S. A., & Morris, R. K. (2005). Working memory, animacy, and verb class in the processing of relative clauses. *Journal of Memory and Language*, 53(2), 204-224.

* Demberg and Keller report the presence of 61 object relative clauses in the Dundee Corpus, including 39 with the relative pronoun which. Hand checking of all instances of the word which in the corpus reveal that there are 9 instances which are part of relative clauses involving object extraction, and 30 where a prepositional phrase is extracted. Models with and without these 30 extra items produce similar results. Other minor differences are due to differences in the parsing/extraction methods used.