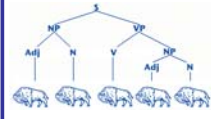


Relative Clauses Remodeled: The Problem with Mixed-Effects Models

Douglas Roland - University at Buffalo



Modeling eye-tracking corpus data

Demberg and Keller (2007) analyzed data from an eye-tracking corpus to show that object relative clauses are read more slowly than subject relative clauses in naturally occurring text.

This contradicts recent experimental work suggesting object relative clauses may not be more difficult than subject relative clauses in naturally occurring contexts (e.g., Mak et al., 2008; Roland et al., 2008; Sato et al., 2008)

We reanalyze the eye-tracking corpus data and find that the reported effect of relative clause type was due to a single extremely difficult sentence containing an object relative clause.

Why did a single sentence cause the Mixed-Effects model to report a significant effect?

Data

Dundee Eye-tracking Corpus (Kennedy & Pynte 2005)

- Data from 10 participants
- ~50,000 words of news paper text
- 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*

*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 39 where a prepositional phrase is extracted. Models with and without these 30 items produce similar results. Other minor differences are due to differences in the parsing/structure methods used.

Effect of relative clause type caused by single (difficult) sentence

Mixed-Effects model for predicting reading times for the embedded verb region of relative clause

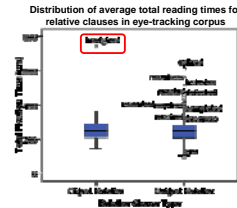
- Predictors (+ binary interactions)
 - Relative pronoun (who/which/that)
 - Word length
 - Log frequency (from BNC data)
 - Log forward transition probability (from BNC data)
 - Word landing position
 - Subject (random variable)
- Dependent variables (separate models for each)
 - Total fixation duration
 - First pass duration
 - First fixation duration
- p values calculated using Markov chain Monte Carlo (MCMC) sampling.

1. Replication of effect of relative clause type

(data from total reading time at embedded verb shown)

Predictor	Coefficient	p
(Intercept)	190.55	0.00
Relative Clause Type	-106.32	0.00
Length	20.12	0.00
Log BNC frequency	-4.22	0.47
Word landing position	-1.57	0.72
Log transition probability	-13.47	0.00
Relative Clause Type * log BNC frequency	12.18	0.00
Length * Word landing position	-1.90	0.00
Log BNC frequency * Log transition probability	1.30	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 inveighed against).



2. Effect of Relative Clause type goes away when 'outlier' sentence is removed

Predictor	Coefficient	p
(Intercept)	96.89	0.12
Relative Clause Type	20.69	0.56
Length	18.45	0.00
Log BNC frequency	6.17	0.29
Word landing position	-2.23	0.61
Log transition probability	-10.82	0.00
Relative Clause Type * log BNC frequency	-1.26	0.75
Length * Word landing position	-1.74	0.01
Log BNC frequency * Log transition probability	0.95	0.00

Results

- Analysis of the reading time data and corpus texts revealed an "outlier" object relative clause example in a difficult sentence
- Model 1 replicated the previous findings of an effect of relative clause type
- Model 2 showed no effect of relative clause type when the outlier example was excluded
- Model 3 showed no effect of relative clause type when an interaction between the random effect of item and the fixed effect of relative clause type was included in the model (see Baayen's comments in Forster 2008 advocating this procedure)

3. - or when interaction between item and relative clause type is included in model

Predictor	Coefficient	t value
(Intercept)	198.0413	2.095
Relative Clause Type	-57.51	-0.954
Length	15.8456	4.999
Log BNC frequency	-2.0599	-0.234
Word landing position	-2.8419	-0.579
Log transition probability	-11.1511	-2.409
Relative Clause Type * log BNC frequency	6.6734	0.967
Length * Word landing position	-1.525	-2.06
Log BNC frequency * Log transition probability	0.9246	2.216

Modeling typical experimental data

Data:

- 3 sets of simulated data
- Prepared using parameters (e.g., mean, SD) from a participant-paced reading time experiment
- Data sets differed only in which items reflected a difference between conditions

Analysis

- ANOVAs (F1, F2, and minF)
- Mixed-effects models
 - Experimental variable as a fixed factor
 - Items and subjects as random factors
 - p values calculated using Markov chain Monte Carlo (MCMC) sampling.

References

Baayen, R. H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics using R*. Cambridge University Press.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4), 390-412.

Demberg, V., & Keller, F. (2007). Eye-tracking Evidence for Integration Cost Effects in Corpus Data. *Proceedings of the 20th meeting of the Cognitive Science Society (CogSci-07)*.

Forster, K. I. (2008). What is F2 Good For? *Journal of Memory and Language*, 59(4), 380.

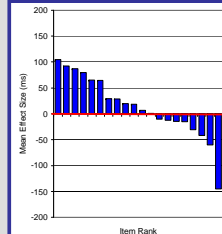
Kennedy, A., & Pynte, J. (2005). Parafoveal-on-foveal effects in normal reading. *Vision Research*, 45(2), 153-168.

Mai, W. M., Vank, W., & Schreuder, H. (2008). Discourse structure and relative clause processing. *Memory & Cognition*, 36(1), 170-181.

Roland, D., O'Mara, C., Yan, H., & Manner, G. (2008). *Discourse and object relative clauses: The effect of topic versus mention*. Paper presented at the CUNY Sentence Processing Conference, CUNY Heli.

Sato, A., Kahanman, B., & Sakai, H. (2008). Processing object relative clauses in context: Another support for the discourse structural account for the processing load asymmetry. *BEFC Technical Report 2208-08*, 108(134), 95-2006.

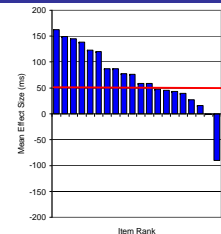
Null effect



ANOVA:
 $F_1(1,29) = 0.13, p = .76$
 $F_2(1,15) = 0.07, p = .79$
 $\min F(1,32) = 0.05, p = .83$

Mixed Effects Model:
 $P_{(MCMC)} = .7850$

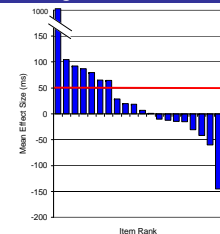
Small effect for all items



ANOVA:
 $F_1(1,29) = 22.91, p < .001$
 $F_2(1,15) = 13.46, p = .002$
 $\min F(1,32) = 8.48, p = .007$

Mixed Effects Model:
 $P_{(MCMC)} = .0001$

Large effect for one item



ANOVA:
 $F_1(1,29) = 17.51, p < .001$
 $F_2(1,15) = 1.05, p = .32$
 $\min F(1,17) = 0.99, p = .33$

Mixed Effects Model:
 $P_{(MCMC)} = .0008$

Including interactions between fixed and random effects

One way to reduce the risk of spurious effects is to include interactions between the fixed and random effects when fitting the model (see Baayen's comments in Forster 2008 advocating this procedure).

- Model with random intercepts for each item and subject (used above)


```
mymodel.lmer <- lmer(RT~RC+(1|Item)+(1|Subject))
```
- Model with random slopes and intercepts for each item and subject


```
mymodel.lmer <- lmer(RT~RC+(1+RC|Item)+(1+RC|Subject))
```

- Null effect and Small effect for all items datasets
 - Interactions not significant, removed during model fitting
- Large effect on one item dataset
 - Interaction with item is significant
 - When interaction with item is included, the effect of the experimental factor is non-significant

Predictor	Coefficient	t value
(Intercept)	500.51	30.450
RCExperimental	63.29	1.020

Conclusions

- Eye-tracking corpus data
 - Reanalysis shows no effect of relative clause type
 - Too few object relatives to examine the effects of animacy, information status, and pronominalization
- Models of corpus data
 - Statistical models of corpus data are particularly prone to misleading results due to (uncontrolled) factors which are not included in the models.
- ANOVAs
 - F_1 and F_2 provide a useful check against outliers when they can be calculated (but don't guarantee this - see Forster 2008)
- Mixed-Effects Models
 - Subject to misleading results when interactions between fixed and random effects are not included
 - But, when interactions are included
 - p values can not (yet?) be calculated
 - Model fitting is more time consuming
 - Models less likely to converge
- Recommendations
 - Provide separate F_1 and F_2 results along with Mixed-Effects results when possible
 - Include more detailed information about predictors removed during model fitting

Acknowledgements

- Jean-Pierre Koenig, Gail Mauner, Carolyn O'Meara, and Hongook Yun
- Roger Levy and the CUNY reviewers for helpful comments and discussion
- The members of the Psycholinguistics and Computational Linguistics Labs at the University at Buffalo