Eye-tracking Evidence for Frequency and Integration Cost Effects in Corpus Data

Vera Demberg[1], Frank Keller[1] and Roger Levy[2]

[1]School of Informatics
University of Edinburgh
[2]Department of Linguistics
University of California, San Diego

CUNY 2007, San Diego, CA
March 31, 2007

Introduction – Experimental approach

Advantages of experimental approach:

- controlled conditions
- established reliability and validity

Drawbacks of experimental approach:

- sentences presented out of context
- constructed manually by the experimenter
- bias: do subjects develop special strategies when presented with the same construction many times? (even when there are fillers)
- only few items from any experiment

Main objectives of this work

Use an eye-tracking corpus as complementary evidence to experimental data

- reading in context; sentences occur in natural context
- “real” language, naturally occurring text
- more data points (for frequent constructions)
- test on many different constructions
- but: less controlled conditions

Test predictions for reading times on relative clauses from

- SPLT (Syntactic Prediction Locality Theory, (Gibson, 1998))
- Transitional probabilities (McDonald & Shillcock, 2003)

Question: Can we find well-established complexity effects in corpus data?

Main objectives of this work

Use an eye-tracking corpus as complementary evidence to experimental data

- reading in context; sentences occur in natural context
- “real” language, naturally occurring text
- more data points (for frequent constructions)
- test on many different constructions
- but: less controlled conditions

Test predictions for reading times on relative clauses from

- SPLT (Syntactic Prediction Locality Theory, (Gibson, 1998))
- Transitional probabilities (McDonald & Shillcock, 2003)

Question: Can we find well-established complexity effects in corpus data?

Main objectives of this work

Use an eye-tracking corpus as complementary evidence to experimental data

- reading in context; sentences occur in natural context
- “real” language, naturally occurring text
- more data points (for frequent constructions)
- test on many different constructions
- but: less controlled conditions

Test predictions for reading times on relative clauses from

- SPLT (Syntactic Prediction Locality Theory, (Gibson, 1998))
- Transitional probabilities (McDonald & Shillcock, 2003)

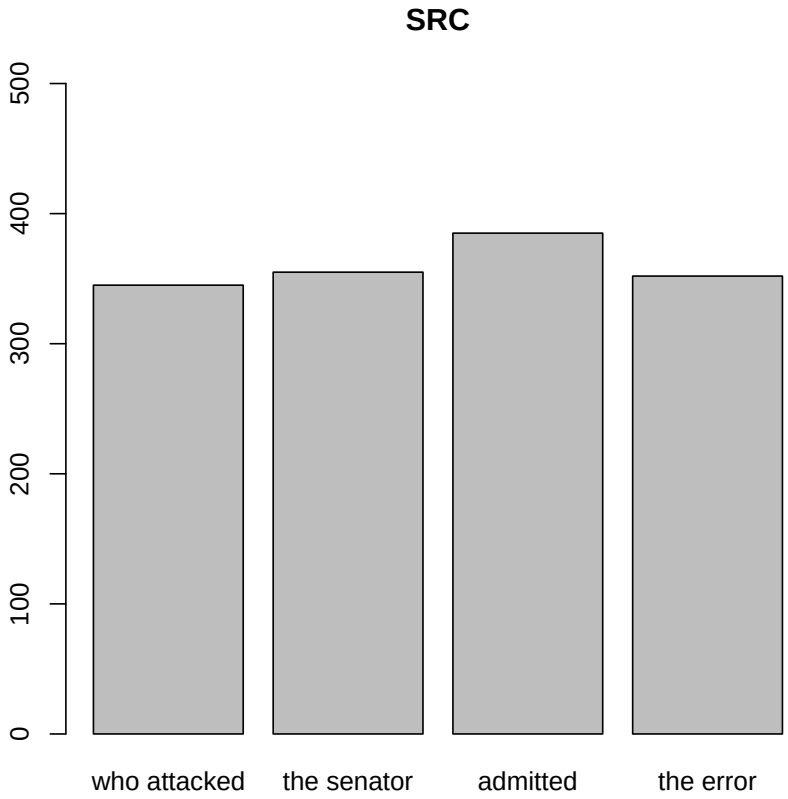
Question: Can we find well-established complexity effects in corpus data?

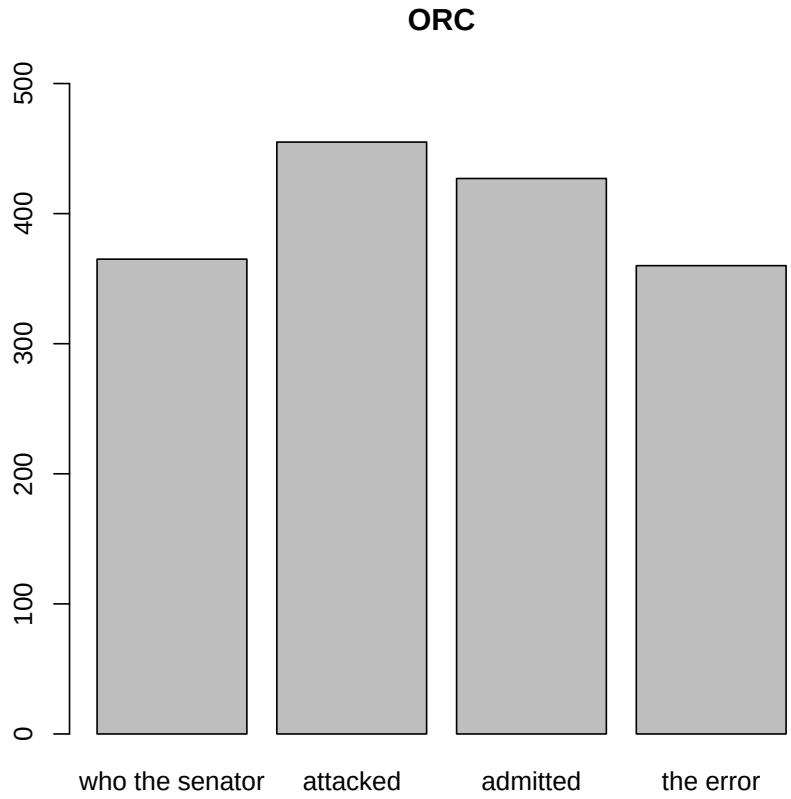
Overview

1. Subject vs. Object Relative Clauses
2. Background: Theories predicting RC reading times
3. The Dundee Corpus
4. Methods: Multiple Hierarchical Linear Regression
5. Results
6. Conclusions

Processing Difficulty and Relative Clauses

Reading times longer on object relative clauses (ORCs) than on subject relative clauses (SRCs), e.g. (King & Just, 1991; Gibson, 1998).

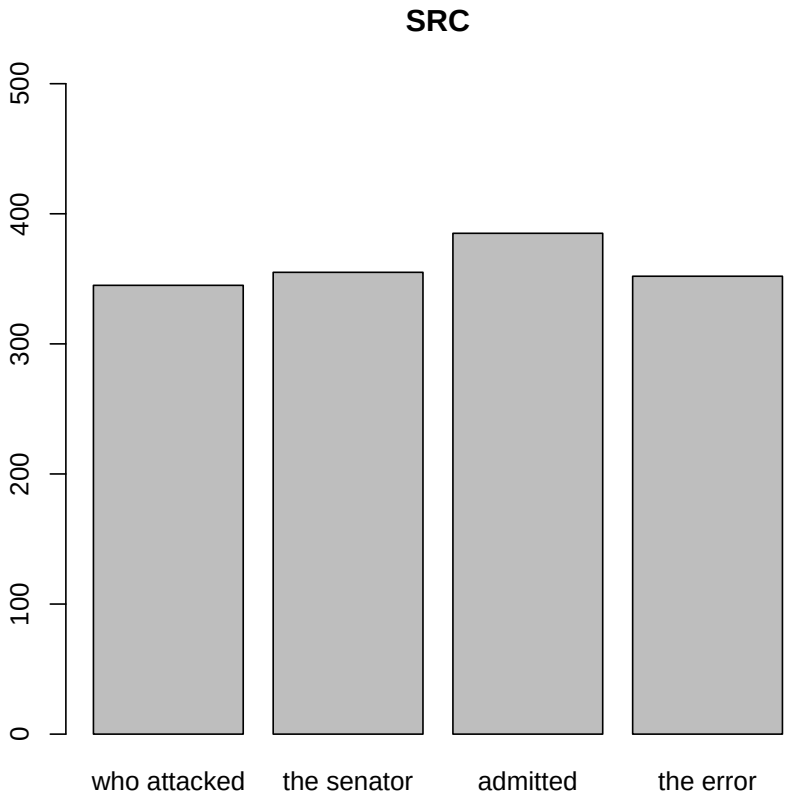


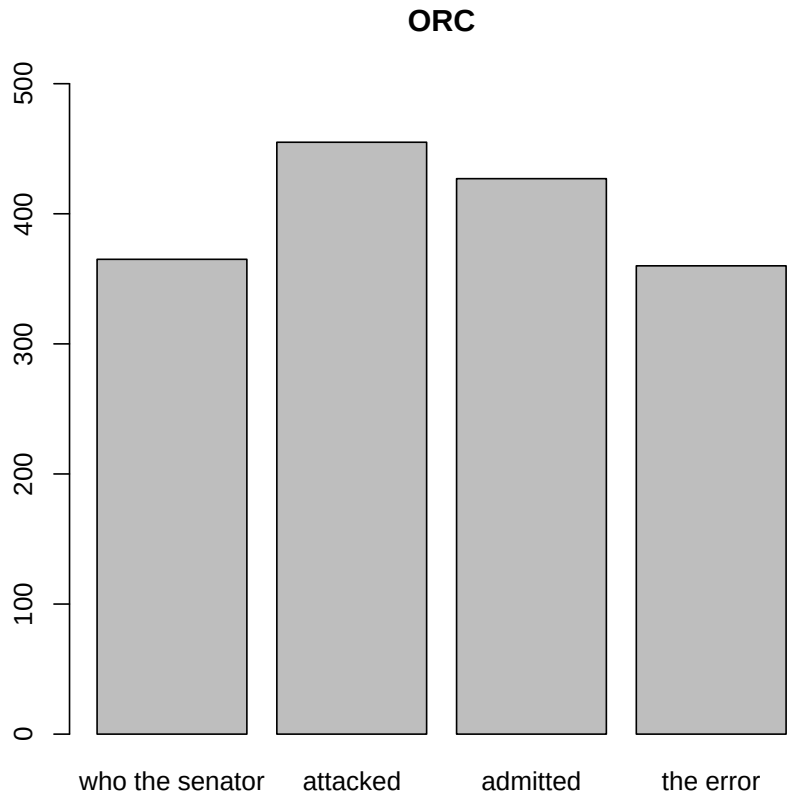


- SRC: The reporter *who attacked the senator* admitted the error.
- ORC: The reporter *who the senator attacked* admitted the error.

Processing Difficulty and Relative Clauses

We compare reading times on the main verb within the relative clause.

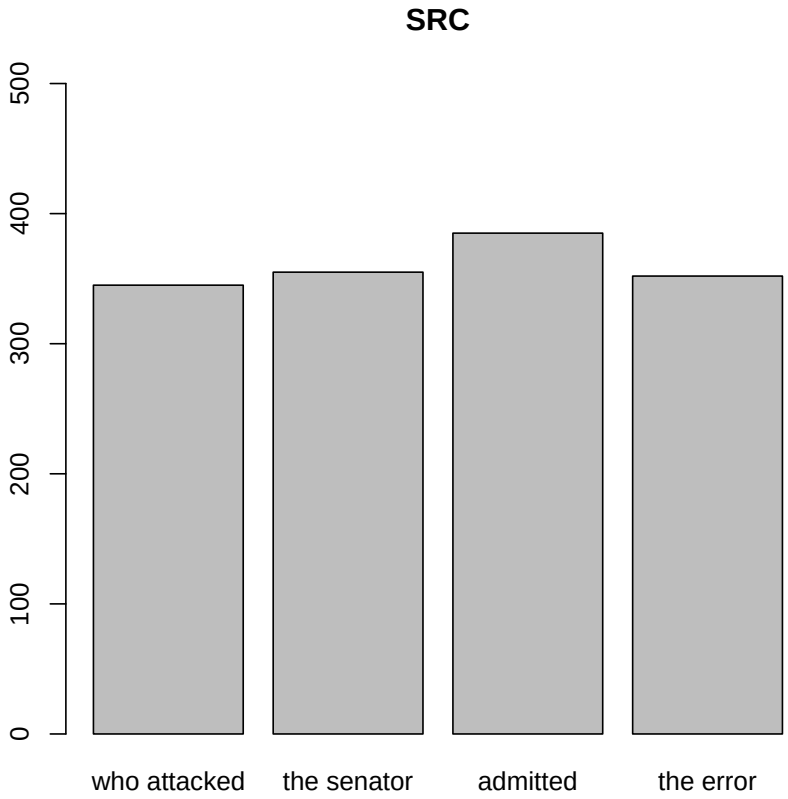


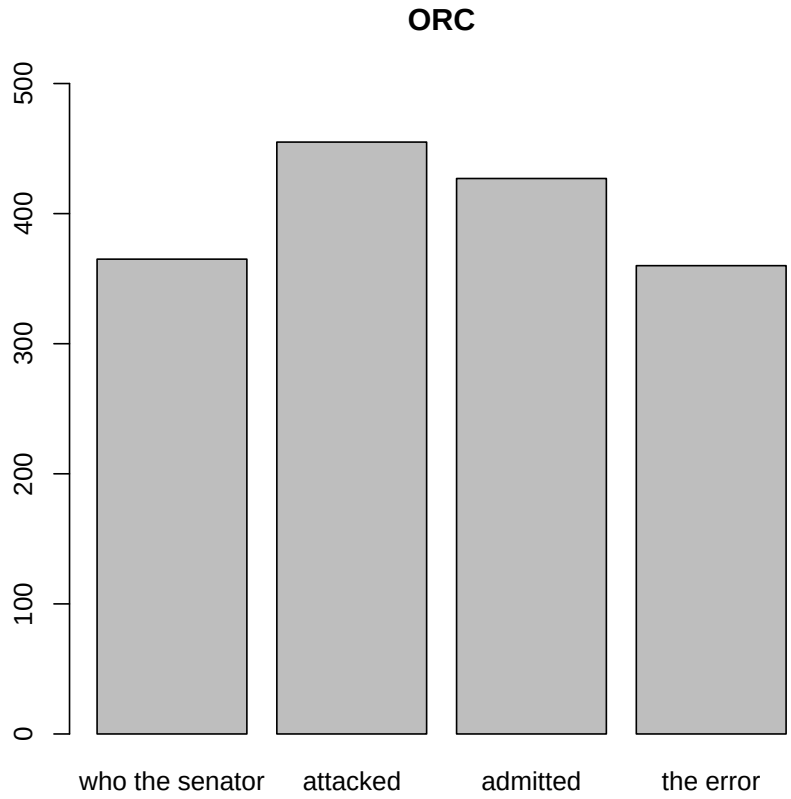


- SRC: The reporter *who attacked the senator* admitted the error.
- ORC: The reporter *who the senator attacked* admitted the error.

Processing Difficulty and Relative Clauses

We compare reading times in the disambiguating region, i.e. on the first word of the RC where the ambiguity between SRC vs. ORC is resolved.





- SRC: The reporter *who attacked the senator* admitted the error.
- ORC: The reporter *who the senator attacked* admitted the error.

Theories for Reading Times in RCs

A number of theories have been developed that account for RC reading times:

- Gibson (1998); Lewis et al. (2006): Locality
- King & Just (1991): Storage and Role changes
- McDonald & Shillcock (2003): Transitional Probabilities
- Hale (2001); Levy (2007): Surprisal

We pick out just two theories as an example here: Integration cost from SPLT and forward transitional probabilities.

Syntactic Prediction Locality Theory

(Gibson, 1998, 20f) makes the following integration cost predictions for the relative clause regions:

SRC:	The	reporter	who	attacked	the	senator	admitted	the	error.
	–	I(0)		I(0)	I(0)+I(1)	I(0)	I(0)+I(1)	I(3)	I(0) I(0)+I(1)
ORC:	The	reporter	who	the	senator	attacked	admitted	the	error.
	–	I(0)		I(0)	I(0)	I(0)	I(1)+I(2)	I(3)	I(0) I(0)+I(1)

Integration costs occur at the heads of phrases.

Syntactic Prediction Locality Theory

(Gibson, 1998, 20f) makes the following integration cost predictions for the relative clause regions:

SRC:	The reporter	who	attacked	the	senator	admitted	the	error.
	–	I(0)	I(0)	I(0)+I(1)	I(0)	I(0)+I(1)	I(3)	I(0) I(0)+I(1)
ORC:	The reporter	who	the	senator	attacked	admitted	the	error.
	–	I(0)	I(0)	I(0)	I(0)	I(1)+I(2)	I(3)	I(0) I(0)+I(1)

The main verb in the SRC should be read faster than in the ORC.

Syntactic Prediction Locality Theory

(Gibson, 1998, 20f) makes the following integration cost predictions for the relative clause regions:

SRC:	The reporter	who	attacked	the	senator	admitted	the	error.
	–	I(0)	I(0)	I(0)+I(1)	I(0)	I(0)+I(1)	I(3)	I(0)	I(0)+I(1)
ORC:	The reporter	who	the	senator	attacked	admitted	the	error.
	–	I(0)	I(0)	I(0)	I(0)	I(1)+I(2)	I(3)	I(0)	I(0)+I(1)

The verb (in SRCs) is more expensive to integrate than the determiner or noun (in ORCs).

Transitional Probability

Alternative account:
Shorter reading times are due to higher transitional probabilities (McDonald & Shillcock, 2003).

Claim:
$P(w_n|w_{n-1})$ is predictive of reading times.

Example:

verb region:	P(attacked | who) $>$ P(attacked | senator)
disambig. region:	P(the | who) $>$ P(attacked | who)

These probabilities can be estimated from large corpora; we used the British National Corpus (BNC, 100-million-word collection).

The Dundee Corpus

Dundee eye-tracking corpus (Kennedy et al., 2003)

- ca. 51.000 words of British newspaper articles (The Independent)
- 10 subjects
- parsed automatically with Charniak parser (Charniak, 2000)
  recall: 96%, precision: 92% for detecting RCs on WSJ

Frequency of relative clause types in Dundee eye-tracking corpus:

pronoun	SRC	ORC	proportion of ORC
that	150	18	10.7%
which	86	39	31.7%
who	137	4	2.8%
total	373	61	14%

The Dundee Corpus

Dundee eye-tracking corpus (Kennedy et al., 2003)

- ca. 51.000 words of British newspaper articles (The Independent)
- 10 subjects
- parsed automatically with Charniak parser (Charniak, 2000)
  recall: 96%, precision: 92% for detecting RCs on WSJ

Frequency of relative clause types in Dundee eye-tracking corpus:

pronoun	SRC	ORC	proportion of ORC
that	150	18	10.7%
which	86	39	31.7%
who	137	4	2.8%
total	373	61	14%

Some Example RCs from the Corpus

SRCs:

- ...titles that seem to stretch the definition a little...
- ...bag searches that make you wonder whether you've come to an underground military center...
- ...the bodies that deal with the human detritus...

ORCs:

- ...services that people need or want from computers...
- ...this no-holds-barren approach to sex and its consequences that many people still associate with the original Cosmo...
- ...answer – that few of us remained with one employer for our working lives...

Some Example RCs from the Corpus

SRCs:

- ...titles that seem to stretch the definition a little...
- ...bag searches that make you wonder whether you've come to an underground military center...
- ...the bodies that deal with the human detritus...

ORCs:

- ...services that people need or want from computers...
- ...this no-holds-barren approach to sex and its consequences that many people still associate with the original Cosmo...
- ...answer – that few of us remained with one employer for our working lives... (parsing error)

Data Selection

434 RCs $\times$ 10 subjects $=$ 4340 data points

We excluded all data points

- where the critical region was the first or last word of a line
- where the critical region was preceded or followed by a punctuation mark
- within a region of 4 adjacent words that had not been fixated (tracking error)
- that contained contractions (e.g. that'll, who'd)

This left us with approximately 3000 data points.

Analyses were only conducted on the fixated data points:

- approx. 1900 for first fixation times
- approx. 2200 for total durations

Multiple hierarchical linear regression

Since we don't closely control the context, we need to regress out possibly confounding factors.

- Independent variables:
  - target factors:
    - RC type
    - log transitional prob.
  - confounding factors:
    - relative pronoun
    - word length
    - log word freq.
    - word's POS tag
    - fixation landing position
- Dependent variables:
  - first fixation duration
  - gaze duration
  - total reading time
- Random variable:
  - subject ID

We entered all variables and their interactions first and stepwise removed those that decreased model quality (according to AIC).

Multiple hierarchical linear regression

Since we don't closely control the context, we need to regress out possibly confounding factors.

- Independent variables:
  - target factors:
    - RC type
    - log transitional prob.
  - confounding factors:
    - relative pronoun
    - word length
    - log word freq.
    - word's POS tag
    - fixation landing position
- Dependent variables:
  - first fixation duration
  - gaze duration
  - total reading time
- Random variable:
  - subject ID

We entered all variables and their interactions first and stepwise removed those that decreased model quality (according to AIC).

Multiple hierarchical linear regression

Since we don't closely control the context, we need to regress out possibly confounding factors.

- Independent variables:
  - target factors:
    - RC type
    - log transitional prob.
  - confounding factors:
    - relative pronoun
    - word length
    - log word freq.
    - word's POS tag
    - fixation landing position
- Dependent variables:
  - first fixation duration
  - gaze duration
  - total reading time
- Random variable:
  - subject ID

We entered all variables and their interactions first and stepwise removed those that decreased model quality (according to AIC).

Multiple hierarchical linear regression

Since we don't closely control the context, we need to regress out possibly confounding factors.

- Independent variables:
  - target factors:
    - RC type
    - log transitional prob.
  - confounding factors:
    - relative pronoun
    - word length
    - log word freq.
    - word's POS tag
    - fixation landing position
- Dependent variables:
  - first fixation duration
  - gaze duration
  - total reading time
- Random variable:
  - subject ID

We entered all variables and their interactions first and stepwise removed those that decreased model quality (according to AIC).

Multiple hierarchical linear regression

Since we don't closely control the context, we need to regress out possibly confounding factors.

- Independent variables:
  - target factors:
    - RC type
    - log transitional prob.
  - confounding factors:
    - relative pronoun
    - word length
    - log word freq.
    - word's POS tag
    - fixation landing position
- Dependent variables:
  - first fixation duration
  - gaze duration
  - total reading time
- Random variable:
  - subject ID

We entered all variables and their interactions first and stepwise removed those that decreased model quality (according to AIC).

Methods for Linear Regression

- all data points are entered directly
- averaging over items or subjects not necessary due to use of a more powerful regression method
- standard approach (Lorch & Myers, 1990):
  - separate regression for each subject
  - t-test over coefficients
- we used hierarchical linear regression (Richter, 2006):
  - account for variance that is due to subjects on a first “level”
  - the coefficients for the other independent variables are estimated in the second level
  - aka linear mixed effect models

Results – Main RC Verb

SRC: The reporter *who attacked the senator* admitted the error.
ORC: The reporter *who the senator attacked* admitted the error.

Total reading times:

Predictor	Coeff.	Sign.
(Intercept)	263.42	***
RC type(SRC)	-177.04	***
Log transitional prob	-24.73	***
Length	21.47	***
Log frequency	-11.66	**
Word landing position	6.39	
Length:landing position	-2.94	***
Log. freq:length	2.65	***
RC type(SRC):log. freq	18.65	***

$^{**}p < 0.01$, $^{***}p < 0.001$; R^2 = 15.6%

- Verbs read faster in SRC condition (as predicted by SPLT).
- Significant effect of transitional probability **in addition** to RC type effect.

Results – Main RC Verb

SRC: The reporter *who attacked the senator* admitted the error.
ORC: The reporter *who the senator attacked* admitted the error.

Total reading times:

Predictor	Coeff.	Sign.
(Intercept)	263.42	***
RC type(SRC)	-177.04	***
Log transitional prob	-24.73	***
Length	21.47	***
Log frequency	-11.66	**
Word landing position	6.39	
Length:landing position	-2.94	***
Log. freq:length	2.65	***
RC type(SRC):log. freq	18.65	***

$^{**}p < 0.01$, $^{***}p < 0.001$; R^2 = 15.6%

- Verbs read faster in SRC condition (as predicted by SPLT).
- Significant effect of transitional probability **in addition** to RC type effect.

Results – Main RC Verb

SRC: The reporter *who attacked the senator* admitted the error.
ORC: The reporter *who the senator attacked* admitted the error.

Total reading times:

Predictor	Coeff.	Sign.
(Intercept)	263.42	***
RC type(SRC)	-177.04	***
Log transitional prob	-24.73	***
Length	21.47	***
Log frequency	-11.66	**
Word landing position	6.39	
Length:landing position	-2.94	***
Log. freq:length	2.65	***
RC type(SRC):log. freq	18.65	***

$^{**}p < 0.01$, $^{***}p < 0.001$; R^2 = 15.6%

- Verbs read faster in SRC condition (as predicted by SPLT).
- Significant effect of transitional probability **in addition** to RC type effect.

Results – Main RC Verb

SRC: The reporter *who attacked the senator* admitted the error.
ORC: The reporter *who the senator attacked* admitted the error.

Total reading times:

Predictor	Coeff.	Sign.
(Intercept)	263.42	***
RC type(SRC)	-177.04	***
Log transitional prob	-24.73	***
Length	21.47	***
Log frequency	-11.66	**
Word landing position	6.39	
Length:landing position	-2.94	***
Log. freq:length	2.65	***
RC type(SRC):log. freq	18.65	***

$^{**}p < 0.01$, $^{***}p < 0.001$; R^2 = 15.6%

- Verbs read faster in SRC condition (as predicted by SPLT).
- Significant effect of transitional probability **in addition** to RC type effect.

Results – Main RC Verb

SRC: The reporter *who attacked the senator* admitted the error.
ORC: The reporter *who the senator attacked* admitted the error.

Total reading times:

Predictor	Coeff.	Sign.
(Intercept)	263.42	***
RC type(SRC)	-177.04	***
Log transitional prob	-24.73	***
Length	21.47	***
Log frequency	-11.66	**
Word landing position	6.39	
Length:landing position	-2.94	***
Log. freq:length	2.65	***
RC type(SRC):log. freq	18.65	***

$^{**}p < 0.01$, $^{***}p < 0.001$; R^2 = 15.6%

- Verbs read faster in SRC condition (as predicted by SPLT).
- Significant effect of transitional probability **in addition** to RC type effect.

Results – Main RC Verb

SRC: The reporter *who attacked the senator* admitted the error.
ORC: The reporter *who the senator attacked* admitted the error.

Total reading times:

Predictor	Coeff.	Sign.
(Intercept)	263.42	***
RC type(SRC)	-177.04	***
Log transitional prob	-24.73	***
Length	21.47	***
Log frequency	-11.66	**
Word landing position	6.39	
Length:landing position	-2.94	***
Log. freq:length	2.65	***
RC type(SRC):log. freq	18.65	***

$^{**}p < 0.01$, $^{***}p < 0.001$; R^2 = 15.6%

- Verbs read faster in SRC condition (as predicted by SPLT).
- Significant effect of transitional probability **in addition** to RC type effect.

Results – Main RC Verb

SRC: The reporter *who attacked the senator* admitted the error.
ORC: The reporter *who the senator attacked* admitted the error.

First pass times:

Predictor	Coeff.	Sign.
(Intercept)	216.1205141	***
RC type(SRC)	-42.8087717	*
Length	7.6596253	**
Log frequency	-2.7113107	
Log freq:length	-0.8476891	**
RC type(SRC):log freq	5.3769450	**

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$; R^2 = 9.9.%

- RC type effect essentially identical to total reading times
- no effect of transitional probability
- got equivalent results for first fixations

Results – Main RC Verb

SRC: The reporter *who attacked the senator* admitted the error.
ORC: The reporter *who the senator attacked* admitted the error.

First pass times:

Predictor	Coeff.	Sign.
(Intercept)	216.1205141	***
RC type(SRC)	-42.8087717	*
Length	7.6596253	**
Log frequency	-2.7113107	
Log freq:length	-0.8476891	**
RC type(SRC):log freq	5.3769450	**

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$; R^2 = 9.9.%

- RC type effect essentially identical to total reading times
- no effect of transitional probability
- got equivalent results for first fixations

Results – Main RC Verb

SRC: The reporter *who attacked the senator* admitted the error.
ORC: The reporter *who the senator attacked* admitted the error.

First pass times:

Predictor	Coeff.	Sign.
(Intercept)	216.1205141	***
RC type(SRC)	-42.8087717	*
Length	7.6596253	**
Log frequency	-2.7113107	
Log freq:length	-0.8476891	**
RC type(SRC):log freq	5.3769450	**

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$; R^2 = 9.9.%

- RC type effect essentially identical to total reading times
- no effect of transitional probability
- got equivalent results for first fixations

Results – Disambiguating Region

SRC: The reporter *who attacked the senator* admitted the error.
ORC: The reporter *who the senator attacked* admitted the error.

Total reading times:

Predictor	Coeff.	Sign.
(Intercept)	-205.8891	
RC type(SRC)	393.1053	**
Transitional prob	-44.7011	***
Landing pos	9.8672	*
Logarithmic frequency	22.0477	**
Length	28.4211	***
simplePOS-VP	-31.6457	*
type(SRC):Trans.prob	43.4744	**
type(SRC):Log.freq	-20.2642	*
Log.freq:Length	-1.3892	*
Landing pos:Length	-3.1838	***

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$; R^2 = 10.1%

- diambiguating region read faster in ORCs (consist. with SPLT)
- transitional probability also facilitates reading
- strong correlation between RC type and transitional prob (r = 0.91)

Results – Disambiguating Region

SRC: The reporter *who attacked the senator* admitted the error.
ORC: The reporter *who the senator attacked* admitted the error.

Total reading times:

Predictor	Coeff.	Sign.
(Intercept)	-205.8891	
RC type(SRC)	393.1053	**
Transitional prob	-44.7011	***
Landing pos	9.8672	*
Logarithmic frequency	22.0477	**
Length	28.4211	***
simplePOS-VP	-31.6457	*
type(SRC):Trans.prob	43.4744	**
type(SRC):Log.freq	-20.2642	*
Log.freq:Length	-1.3892	*
Landing pos:Length	-3.1838	***

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$; R^2 = 10.1%

- diambiguating region read faster in ORCs (consist. with SPLT)
- transitional probability also facilitates reading
- strong correlation between RC type and transitional prob (r = 0.91)

Results – Disambiguating Region

SRC: The reporter *who attacked the senator* admitted the error.
ORC: The reporter *who the senator attacked* admitted the error.

Total reading times:

Predictor	Coeff.	Sign.
(Intercept)	-205.8891	
RC type(SRC)	393.1053	**
Transitional prob	-44.7011	***
Landing pos	9.8672	*
Logarithmic frequency	22.0477	**
Length	28.4211	***
simplePOS-VP	-31.6457	*
type(SRC):Trans.prob	43.4744	**
type(SRC):Log.freq	-20.2642	*
Log.freq:Length	-1.3892	*
Landing pos:Length	-3.1838	***

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$; R^2 = 10.1%

- diambiguating region read faster in ORCs (consist. with SPLT)
- transitional probability also facilitates reading
- strong correlation between RC type and transitional prob (r = 0.91)

Results – Disambiguating Region

SRC: The reporter *who attacked the senator* admitted the error.
ORC: The reporter *who the senator attacked* admitted the error.

Total reading times:

Predictor	Coeff.	Sign.
(Intercept)	-205.8891	
RC type(SRC)	393.1053	**
Transitional prob	-44.7011	***
Landing pos	9.8672	*
Logarithmic frequency	22.0477	**
Length	28.4211	***
simplePOS-VP	-31.6457	*
type(SRC):Trans.prob	43.4744	**
type(SRC):Log.freq	-20.2642	*
Log.freq:Length	-1.3892	*
Landing pos:Length	-3.1838	***

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$; R^2 = 10.1%

- diambiguating region read faster in ORCs (consist. with SPLT)
- transitional probability also facilitates reading
- strong correlation between RC type and transitional prob (r = 0.91)

Results – Disambiguating Region

SRC: The reporter *who attacked the senator* admitted the error.
ORC: The reporter *who the senator attacked* admitted the error.

First fixation durations:

Predictor	Coeff.	Sign.
(Intercept)	195.541736	***
RC type(SRC)	18.902473	***
Log frequency	-1.486510	**

$^{**}p < 0.01$, $^{***}p < 0.001$; $R^2 = 8.1\%$

- Only RC type and frequency were found to be significant predictors for first fixation times.
- No significant effect for transitional probabilities here.
- The first word of the SRC (first word of VP) is read more slowly than the first word of the ORC (first word of NP).

Results – Disambiguating Region

SRC: The reporter *who attacked the senator* admitted the error.
ORC: The reporter *who the senator attacked* admitted the error.

First fixation durations:

Predictor	Coeff.	Sign.
(Intercept)	195.541736	***
RC type(SRC)	18.902473	***
Log frequency	-1.486510	**

$^{**}p < 0.01$, $^{***}p < 0.001$; $R^2 = 8.1\%$

- Only RC type and frequency were found to be significant predictors for first fixation times.
- No significant effect for transitional probabilities here.
- The first word of the SRC (first word of VP) is read more slowly than the first word of the ORC (first word of NP).

Results – Disambiguating Region

SRC: The reporter *who attacked the senator* admitted the error.
ORC: The reporter *who the senator attacked* admitted the error.

First fixation durations:

Predictor	Coeff.	Sign.
(Intercept)	195.541736	***
RC type(SRC)	18.902473	***
Log frequency	-1.486510	**

$^{**}p < 0.01$, $^{***}p < 0.001$; $R^2 = 8.1\%$

- Only RC type and frequency were found to be significant predictors for first fixation times.
- No significant effect for transitional probabilities here.
- The first word of the SRC (first word of VP) is read more slowly than the first word of the ORC (first word of NP).

Conclusions

- New type of evidence for locality-based theories (like SPLT).
- Transitional probability also predicts reading times, but independent of RC type effect.
- The RC type effect occurs in both the late measures and the early measures, while transitional probabilities were only predictive of the late measures.
- Regression method allows regions to be compared when they are different words, because potentially confounding variables are regressed out.
- Corpus-based methodology can easily be applied for evaluating other theories and testing them on different constructions.
- Corpus studies as complementary evidence to traditional experimental methods.

References

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

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Asssociation for Computational Linguistics*.

Kennedy, A., Hill, R., & Pynte, J. (2003). The Dundee corpus. Poster at the 12th European Conference on Eye Movements, Dundee.

King, J., & Just, M. A. (1991). Individual differences in syntactic processing: The role of working memory. *Journal of Memory and Language*, *30*, 580–602.

Levy, R. (2007). Expectation-based syntactic comprehension. *Cognition*. accepted.

Lorch, R. F., & Myers, J. L. (1990). Regression analyses of repeated measures data in cognitive research. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 149–157.

McDonald, S. A., & Shillcock, R. C. (2003). Low-level predictive inference in reading: The influence of transitional probabilities on eye movements. *Vision Research*, *43*, 1735–1751.

Richter, T. (2006). What is wrong with ANOVA and multiple regression? analyzing sentence reading times with hierarchical linear models. *Discourse Processes*, *41*, 221–250.