Do theories of syntactic processing difficulty scale up to naturally occurring text?

Vera Demberg and Frank Keller

School of Informatics
University of Edinburgh

AMLaP
August 25, 2007

Eye-tracking Corpora as Evidence

Experimental approach:

- careful design to eliminate potential confounds
- full control over task and materials
- **but:** materials constructed by experimenter, presented out of context, one construction repeated many times during the experiment

Eye-tracking corpora as complementary evidence:

- naturally occurring text, sentences read in context
- more data points (for frequent constructions)
- wide range of constructions
- data collected under lab conditions
- **but:** less control over materials

Evaluating Theories of Syntactic Processing Complexity

Dependency Locality Theory (DLT, Gibson 1998, 2000):

- structural approach
- “backward-looking processing”
- difficulty triggered by integration of previous material; increases with distance
- explains locality effects

Surprisal (Hale 2001, Levy 2007):

- statistical approach
- “forward-looking processing”
- difficulty triggered when surprising (not predicted) structures are encountered; increases with log probability
- explains (amongst others) anti-locality effects

Research Questions

- Can DLT and surprisal explain processing complexity in naturally occurring, contextualized text?
- How do DLT and surprisal relate to each other? Do they explain complementary aspects of processing complexity?
- Do aspects of the theories need to be modified in the light of corpus data?

Overview

1. The Dundee Corpus
2. Linear Mixed Effect Models
3. Dependency Locality Theory
4. Surprisal

The Dundee Corpus (Kennedy and Pynte 2005)

- 51,000 words of British newspaper articles (The Independent)
- 10 subjects read the whole text and answered comprehension questions
- eye-movements recorded with Dr. Bouis eye-tracker
- corpus not part of a treebank, therefore parsed with:
  - Minipar (Lin, 1998), dependency structures for computing integration cost
  - Roark parser (Roark, 2001), incremental processing for calculating surprisal
- exclude first and last word of a line, and words adjacent to punctuation; remove tracklosses

Linear Mixed Effect Models

- All variables and binary interactions entered into a hierarchical linear mixed effects model
- Stepwise removal of variables that decrease model quality (using AIC)

Random variable:
subject ID

Dependent variables:
first fixation duration
gaze duration
total reading time

Covariates:
word length
log frequency
word position
previous fixation
launch distance
fixation land position

Independent variables:
integration cost
forward transitional prob.
backward transitional prob.
surprisal

Linear Mixed Effect Models

- All variables and binary interactions entered into a hierarchical linear mixed effects model
- Stepwise removal of variables that decrease model quality (using AIC)

Random variable:
subject ID

Dependent variables:
first fixation duration
gaze duration
total reading time

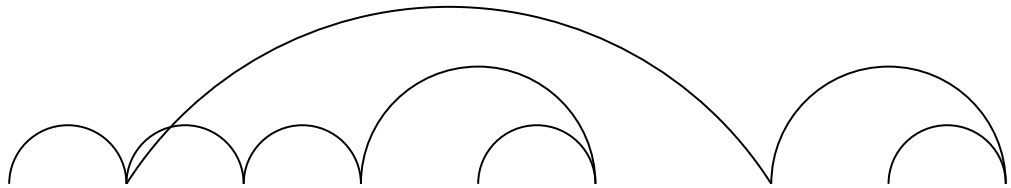
Covariates:
word length
log frequency
word position
previous fixation
launch distance
fixation land position

Independent variables:
integration cost
forward transitional prob.
backward transitional prob.
surprisal

DLT Integration Cost (Gibson 2000)

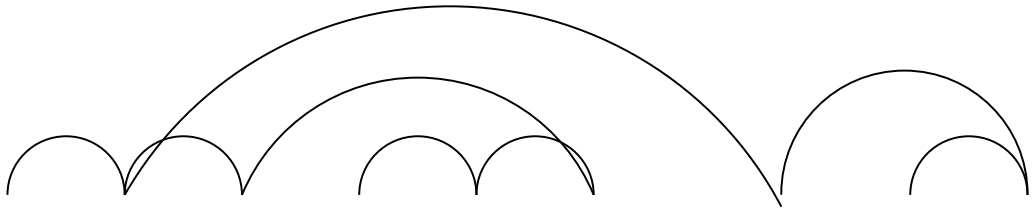
- **Theory:** integration cost occurs at heads of phrases when dependents are integrated; number of discourse referents crossed determines cost
- **Implementation:** use Minipar to compute dependency structures for the Dundee corpus, determine DRs, compute integration for every word

SRC:



	The	reporter	who	attacked	the	senator	admitted	the	error.
Disc.ref.		x		x		x	x		x
Integ. Cost	I(0)	I(1)	I(0)	I(1)	I(0)	I(1)	I(1)+I(3)	I(0)	I(1)

ORC:



	The	reporter	who	the	senator	attacked	admitted	the	error.
Disc.ref.		x			x	x	x		x
Integ. Cost	I(0)	I(1)	I(0)	I(0)	I(1)	I(1)+I(2)	I(1)+I(3)	I(0)	I(1)

Integration Cost: Results

Table: First pass durations

Predictor	Coefficient	Significance
(Intercept)	199.72	***
WORDLEN	13.05	***
LOGFREQ	-1.16	***
PREVFIX	-21.14	***
LAUNDIST	0.10	***
LANDPOS	-9.62	***
WORDPOS	-0.22	***
INTEGCOST	-1.94	***
FORWTRANS	-6.99	***
BACKTRANS	0.78	***

$*p < 0.05$, $**p < 0.01$, $***p < 0.001$

Contrary to prediction: words with higher integration cost are read faster

Integration Cost: Results

Table: First pass durations

Predictor	Coefficient	Significance
(Intercept)	199.72	***
WORDLEN	13.05	***
LOGFREQ	-1.16	***
PREVFIX	-21.14	***
LAUNDIST	0.10	***
LANDPOS	-9.62	***
WORDPOS	-0.22	***
INTEGCOST	-1.94	***
FORWTRANS	-6.99	***
BACKTRANS	0.78	***

$*p < 0.05$, $**p < 0.01$, $***p < 0.001$

Contrary to prediction: words with higher integration cost are read faster

Integration Cost: Results

Table: First pass durations

Predictor	Coefficient	Significance
(Intercept)	199.72	***
WORDLEN	13.05	***
LOGFREQ	-1.16	***
PREVFIX	-21.14	***
LAUNDIST	0.10	***
LANDPOS	-9.62	***
WORDPOS	-0.22	***
INTEGCOST	-1.94	***
FORWTRANS	-6.99	***
BACKTRANS	0.78	***

$*p < 0.05$, $**p < 0.01$, $***p < 0.001$

Contrary to prediction: words with higher integration cost are read faster

Integration Cost: Results

Table: First pass durations

Predictor	Coefficient	Significance
(Intercept)	199.72	***
WORDLEN	13.05	***
LOGFREQ	-1.16	***
PREVFIX	-21.14	***
LAUNDIST	0.10	***
LANDPOS	-9.62	***
WORDPOS	-0.22	***
INTEGCOST	-1.94	***
FORWTRANS	-6.99	***
BACKTRANS	0.78	***

$*p < 0.05$, $**p < 0.01$, $***p < 0.001$

Contrary to prediction: words with higher integration cost are read faster

Problems with Integration Cost

Contrary to prediction: words with higher integration cost are read faster

But: Demberg & Keller (2007) show that DLT makes correct predictions for relative clauses in Dundee corpus

Possible explanations:

- DLT makes incomplete predictions: integration costs assigned only to nouns and verbs
- many words in corpus with IC = 0; could explain negative effect

Open questions:

- Does DLT make correct predictions for nouns and verbs?
- Can DLT be extended to cover words with IC = 0?

Problems with Integration Cost

Contrary to prediction: words with higher integration cost are read faster

But: Demberg & Keller (2007) show that DLT makes correct predictions for relative clauses in Dundee corpus

Possible explanations:

- DLT makes incomplete predictions: integration costs assigned only to nouns and verbs
- many words in corpus with IC = 0; could explain negative effect

Open questions:

- Does DLT make correct predictions for nouns and verbs?
- Can DLT be extended to cover words with IC = 0?

Problems with Integration Cost

Contrary to prediction: words with higher integration cost are read faster

But: Demberg & Keller (2007) show that DLT makes correct predictions for relative clauses in Dundee corpus

Possible explanations:

- DLT makes incomplete predictions: integration costs assigned only to nouns and verbs
- many words in corpus with IC = 0; could explain negative effect

Open questions:

- Does DLT make correct predictions for nouns and verbs?
- Can DLT be extended to cover words with IC = 0?

Integration Cost at Nouns

As predicted: nouns with lower integration cost are read faster

Table: First pass durations

Predictor	Coefficient	Significance
⋮	⋮	⋮
log(INTEGCOST+1)	17.21	***
⋮	⋮	⋮

$*p < 0.05$, $**p < 0.01$, $***p < 0.001$

Integration Cost at Verbs

- contrary to prediction: verbs with higher integration cost read faster
- **new analysis:** integration cost by type of dependent being integrated
- auxiliaries facilitate integration at verbs
- compatible with argument attraction in linguistic theories, e.g., HPSG

Dependent	Coef	Sign
-PRP-AUXRB-	-34.47	***
-PRPAUX-PRPAUX-	-31.31	***
-NNS-MD-AUX-	-27.06	*
-PRP-AUX-	-21.55	***
-AUXG-	-21.36	*
-RB-NN-	-20.90	*
-NN-MD-AUX-	-20.81	*
-NNP-AUX-	-16.40	*
-TO-PRP-	-12.81	*
-NNP-	13.14	**
-WP-	23.18	*
-NNS-	23.77	*
-NN-RB-	25.24	*

$*p < 0.05$, $**p < 0.01$, $***p < 0.001$

Integration Cost at Verbs

- contrary to prediction: verbs with higher integration cost read faster
- **new analysis:** integration cost by type of dependent being integrated
- auxiliaries facilitate integration at verbs
- compatible with argument attraction in linguistic theories, e.g., HPSG

Dependent	Coef	Sign
-PRP-AUXRB-	-34.47	***
-PRPAUX-PRPAUX-	-31.31	***
-NNS-MD-AUX-	-27.06	*
-PRP-AUX-	-21.55	***
-AUXG-	-21.36	*
-RB-NN-	-20.90	*
-NN-MD-AUX-	-20.81	*
-NNP-AUX-	-16.40	*
-TO-PRP-	-12.81	*
-NNP-	13.14	**
-WP-	23.18	*
-NNS-	23.77	*
-NN-RB-	25.24	*

$*p < 0.05$, $**p < 0.01$, $***p < 0.001$

Summary DLT

Findings:

- DLT predicts zero integration for most words in the corpus
- makes correct predictions for noun reading times
- but not a good predictor for verb reading times
- auxiliaries facilitate integration at main verbs

Future work:

- investigate if underlying dependency structures have an influence
- extend DLT to other word categories (prepositions, adverbs, adjectives)

Surprisal (Hale 2001, Levy 2007)

Idea: processing difficulty $\propto$ the word's surprisal

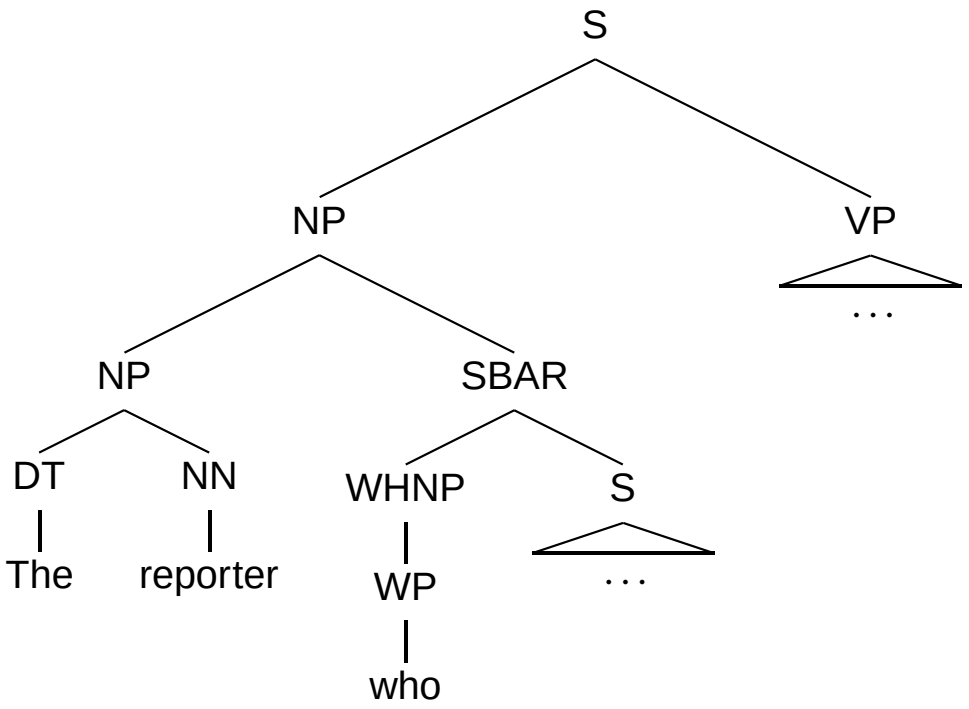
Example	Rule	Probability
The reporter who ...	S $\longrightarrow$ NP VP	$p = 0.6$
The reporter who ...	NP $\longrightarrow$ NP SBAR	$p = 0.004$
The reporter	NP $\longrightarrow$ DT NN	$p = 0.5$
The	DT $\longrightarrow$ the	$p = 0.7$
...		

Example:
surprisal for *who* =
prefix prob. of *who* $-$
prefix prob. of *reporter*

Two versions:
lexicalized
unlexicalized

Surprisal (Hale 2001, Levy 2007)

Idea: processing difficulty $\propto$ the word's surprisal



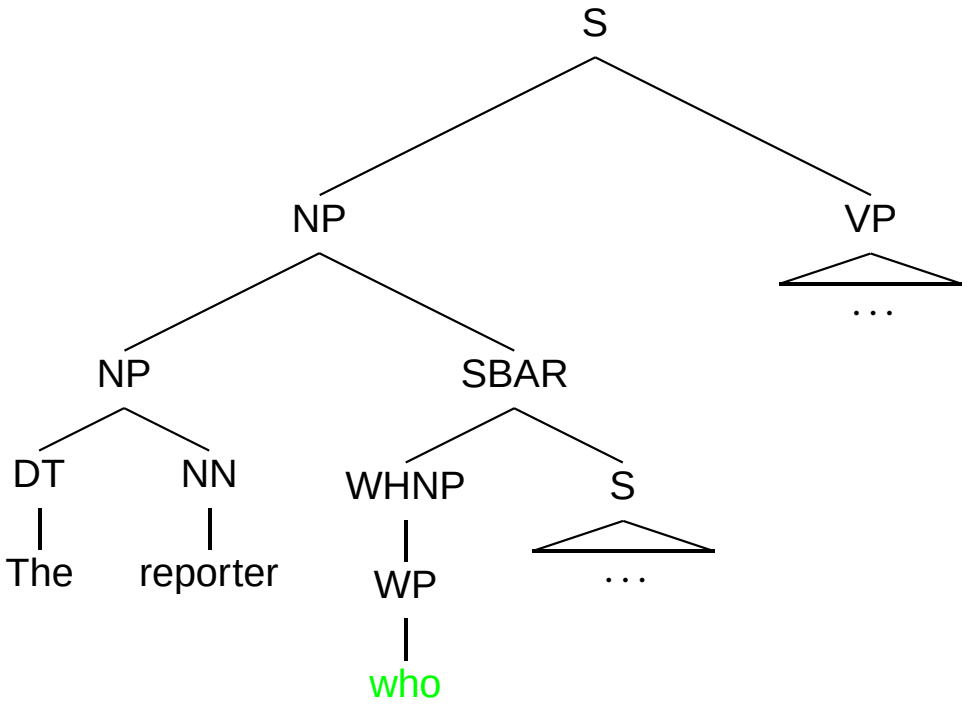
Example	Rule	Probability
The reporter who ...	S $\rightarrow$ NP VP	$p = 0.6$
The reporter who ...	NP $\rightarrow$ NP SBAR	$p = 0.004$
The reporter	NP $\rightarrow$ DT NN	$p = 0.5$
The	DT $\rightarrow$ the	$p = 0.7$
...		

Example:
surprisal for *who* =
prefix prob. of *who* $-$
prefix prob. of *reporter*

Two versions:
lexicalized
unlexicalized

Surprisal (Hale 2001, Levy 2007)

Idea: processing difficulty $\propto$ the word's surprisal



Example	Rule	Probability
The reporter who ...	S $\rightarrow$ NP VP	$p = 0.6$
The reporter who ...	NP $\rightarrow$ NP SBAR	$p = 0.004$
The reporter	NP $\rightarrow$ DT NN	$p = 0.5$
The	DT $\rightarrow$ the	$p = 0.7$
...		

Example:

surprisal for *who* =

prefix prob. of *who* —

prefix prob. of *reporter*

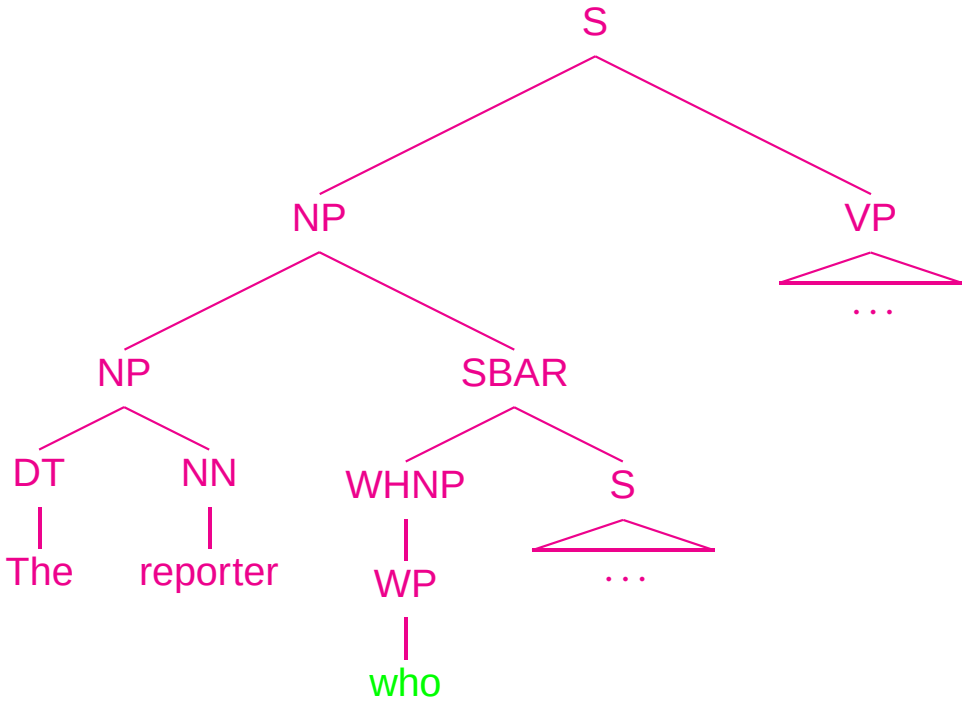
Two versions:

lexicalized

unlexicalized

Surprisal (Hale 2001, Levy 2007)

Idea: processing difficulty $\propto$ the word's surprisal



Example	Rule	Probability
The reporter who ...	S $\longrightarrow$ NP VP	$p = 0.6$
The reporter who ...	NP $\longrightarrow$ NP SBAR	$p = 0.004$
The reporter	NP $\longrightarrow$ DT NN	$p = 0.5$
The	DT $\longrightarrow$ the	$p = 0.7$
...		

Example:

surprisal for *who* =

prefix prob. of *who* $-$

prefix prob. of *reporter*

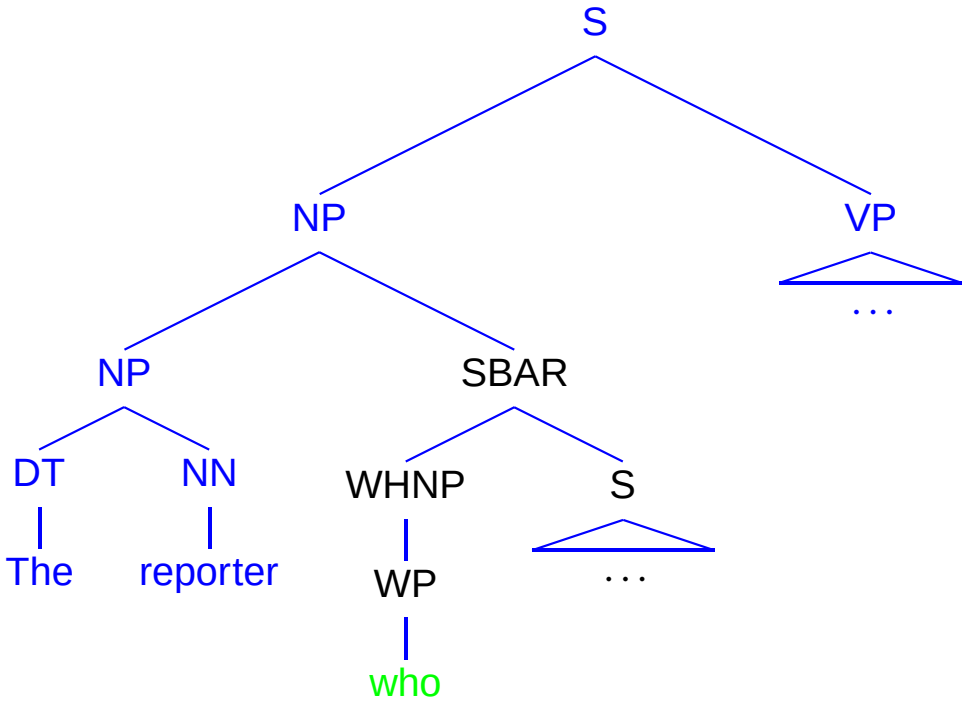
Two versions:

lexicalized

unlexicalized

Surprisal (Hale 2001, Levy 2007)

Idea: processing difficulty $\propto$ the word's surprisal



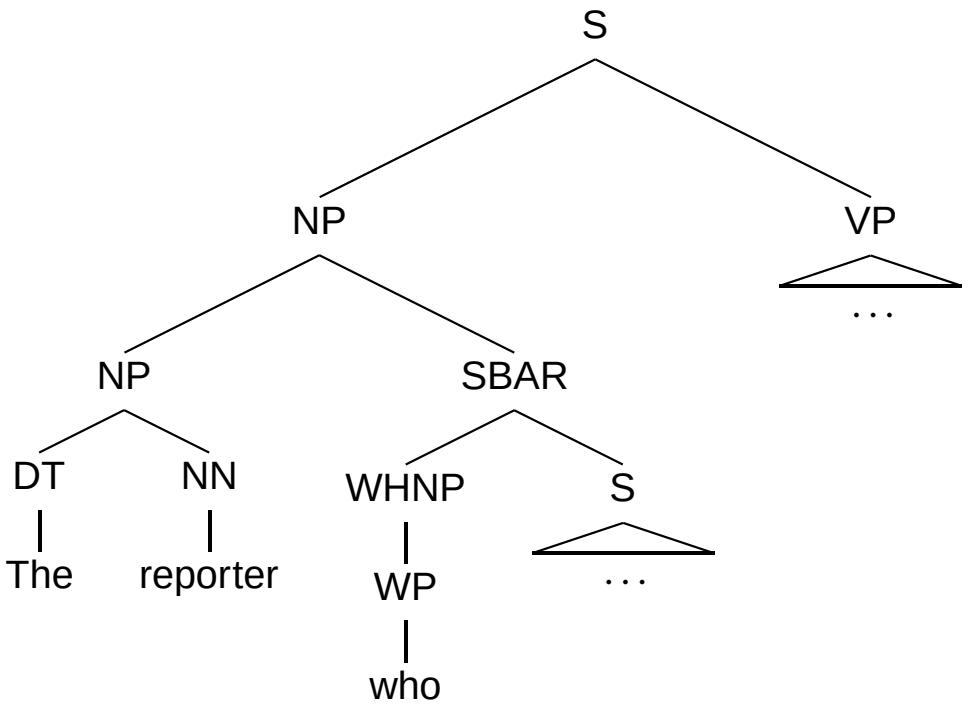
Example	Rule	Probability
The reporter who ...	S $\longrightarrow$ NP VP	$p = 0.6$
The reporter who ...	NP $\longrightarrow$ NP SBAR	$p = 0.004$
The reporter	NP $\longrightarrow$ DT NN	$p = 0.5$
The	DT $\longrightarrow$ the	$p = 0.7$
...		

Example:
surprisal for *who* =
prefix prob. of *who* $-$
prefix prob. of *reporter*

Two versions:
lexicalized
unlexicalized

Surprisal (Hale 2001, Levy 2007)

Idea: processing difficulty $\propto$ the word's surprisal



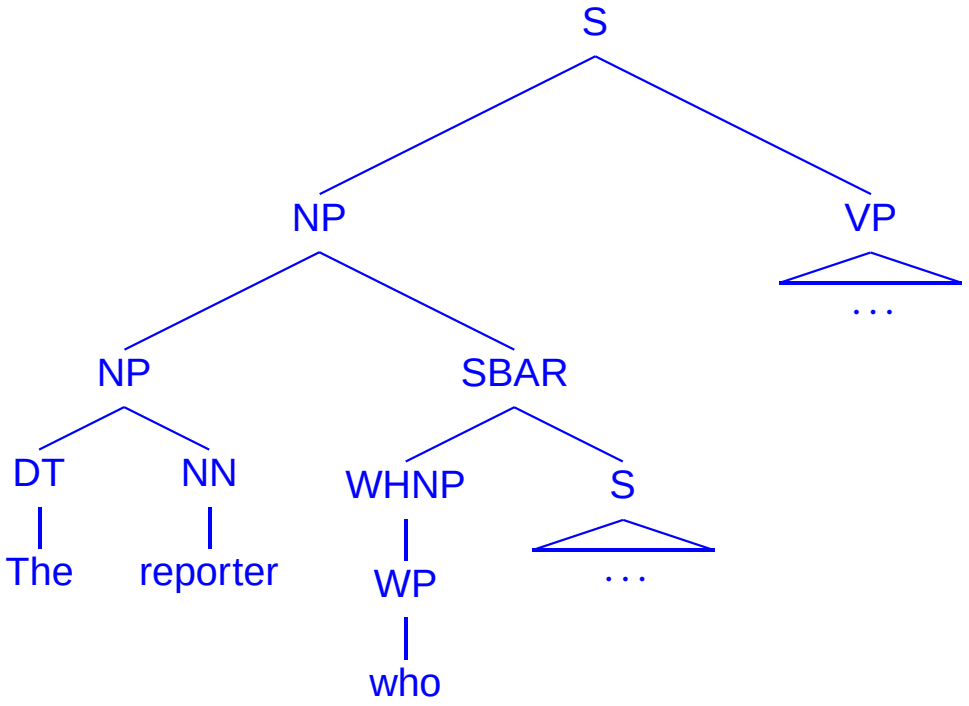
Example	Rule	Probability
The reporter who ...	S $\rightarrow$ NP VP	$p = 0.6$
The reporter who ...	NP $\rightarrow$ NP SBAR	$p = 0.004$
The reporter	NP $\rightarrow$ DT NN	$p = 0.5$
The	DT $\rightarrow$ the	$p = 0.7$
...		

Example:
surprisal for *who* =
prefix prob. of *who* $-$
prefix prob. of *reporter*

Two versions:
lexicalized
unlexicalized

Surprisal (Hale 2001, Levy 2007)

Idea: processing difficulty $\propto$ the word's surprisal



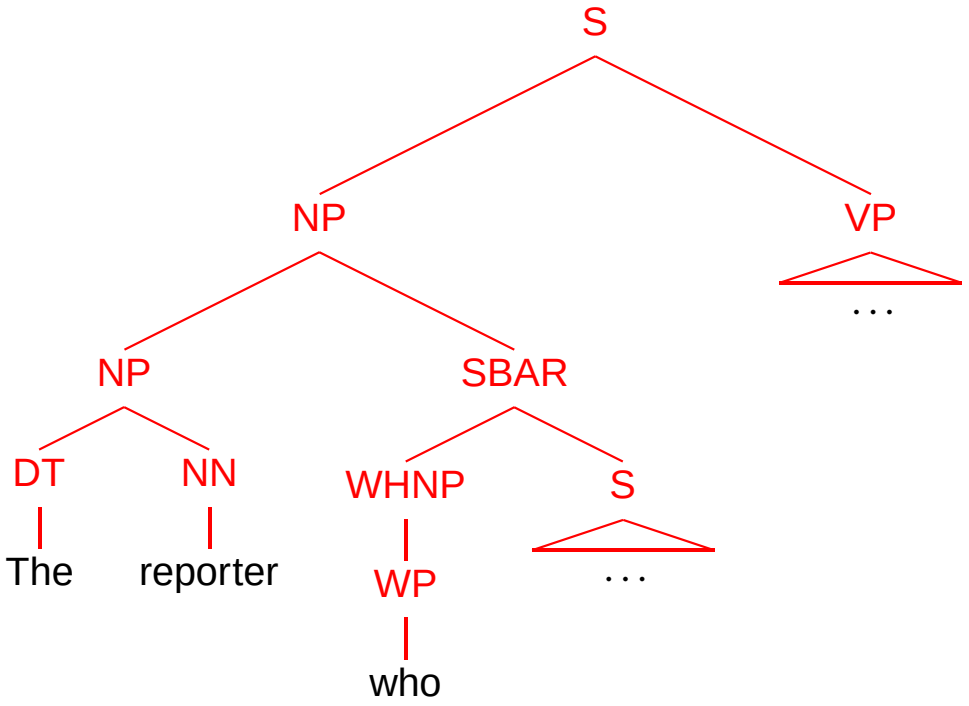
Example	Rule	Probability
The reporter who ...	S $\longrightarrow$ NP VP	$p = 0.6$
The reporter who ...	NP $\longrightarrow$ NP SBAR	$p = 0.004$
The reporter	NP $\longrightarrow$ DT NN	$p = 0.5$
The	DT $\longrightarrow$ the	$p = 0.7$
...		

Example:
surprisal for *who* =
prefix prob. of *who* $-$
prefix prob. of *reporter*

Two versions:
lexicalized
unlexicalized

Surprisal (Hale 2001, Levy 2007)

Idea: processing difficulty $\propto$ the word's surprisal



Example	Rule	Probability
The reporter who ...	S $\longrightarrow$ NP VP	$p = 0.6$
The reporter who ...	NP $\longrightarrow$ NP SBAR	$p = 0.004$
The reporter	NP $\longrightarrow$ DT NN	$p = 0.5$
The	DT $\longrightarrow$ the	$p = 0.7$
...		

Example:
surprisal for *who* =
prefix prob. of *who* $-$
prefix prob. of *reporter*

Two versions:
lexicalized
unlexicalized

Surprisal: Results

Both lexicalized and unlexicalized surprisal are significant predictors of reading times.

Table: First pass durations

Predictor	Coefficient	Significance
⋮	⋮	⋮
SURPRIS	0.75	***
ULXSPRS	2.46	***
⋮	⋮	⋮

$*p < 0.05$, $**p < 0.01$, $***p < 0.001$

Correlations Between Predictors

Correlation	ULXSPRS	SURPRIS	INTEGCOST
SURPRIS	0.22		
INTEGCOST	0.04	0.20	
LOGFREQ	0.003	-0.60	-0.25

- lexicalized surprisal correlated with log frequency
- integration cost orthogonal to surprisal: no correlation, does not compete for variance in regression

Correlations Between Predictors

Correlation	ULXSPRS	SURPRIS	INTEGCOST
SURPRIS	0.22		
INTEGCOST	0.04	0.20	
LOGFREQ	0.003	-0.60	-0.25

- lexicalized surprisal correlated with log frequency
- integration cost orthogonal to surprisal: no correlation, does not compete for variance in regression

Correlations Between Predictors

Correlation	ULXSPRS	SURPRIS	INTEGCOST
SURPRIS	0.22		
INTEGCOST	0.04	0.20	
LOGFREQ	0.003	-0.60	-0.25

- lexicalized surprisal correlated with log frequency
- integration cost orthogonal to surprisal: no correlation, does not compete for variance in regression

Conclusions

- Dependency locality theory:
  - theory incomplete: predictions only for verbs and nouns
  - correct predictions for nouns, but not for verbs
  - auxiliaries facilitate processing at main verb
- Surprisal:
  - broad coverage model: correct predictions for full corpus
  - surprisal goes beyond transitional probability and frequency
- surprisal orthogonal to integration cost: full theory of processing complexity needs to combine integration cost and surprisal
- methodological contribution: evaluated theories of processing difficulty on naturally occurring, contextualized text

Integration Cost on other types of words?

Part of speech	Mean residual RT	Corpus freq.
Sentence adjectives (but, because, although)	9.98	3320
Expletives (it, there)	7.26	1070
Adjectives, adverbs	3.42	56308
Prepositions	2.14	48827
Nouns (only those with IC = 0)	−1.89	6551
Complementizers (if, that, whether)	−1.92	4720
Postdeterminers (last, next, first, same)	−2.81	1166
Determiners	−2.82	43693
Auxiliaries	−3.09	24303
Predeterminers (all, such, even)	−4.93	793

Table: First pass durations in the Dundee Corpus: residual reading times partitioned with respect to parts of speech, for non-verbal and non-nominal parts of speech.

Integration Cost: Results for all words

Predictor	Coefficients	Significance
(Intercept)	102.43	***
WORDLEN	31.56	***
LOGFREQ	6.38	***
FORWTRANS	-8.16	***
BACKTRANS	-0.71	***
WORDPOS	-0.23	***
PREVFIX	-21.54	***
LAUNDIST	0.10	***
LANDPOS	2.50	***
INTEGCOST	-1.55	***
FORWTRANS:BACKTRANS	-0.43	***
WORDLEN:LANDPOS	-1.66	***
WORDLEN:LOGFREQ	-1.73	***

$*p < 0.05$, $**p < 0.01$, $***p < 0.001$

Contrary to prediction: words with higher integration cost are read faster

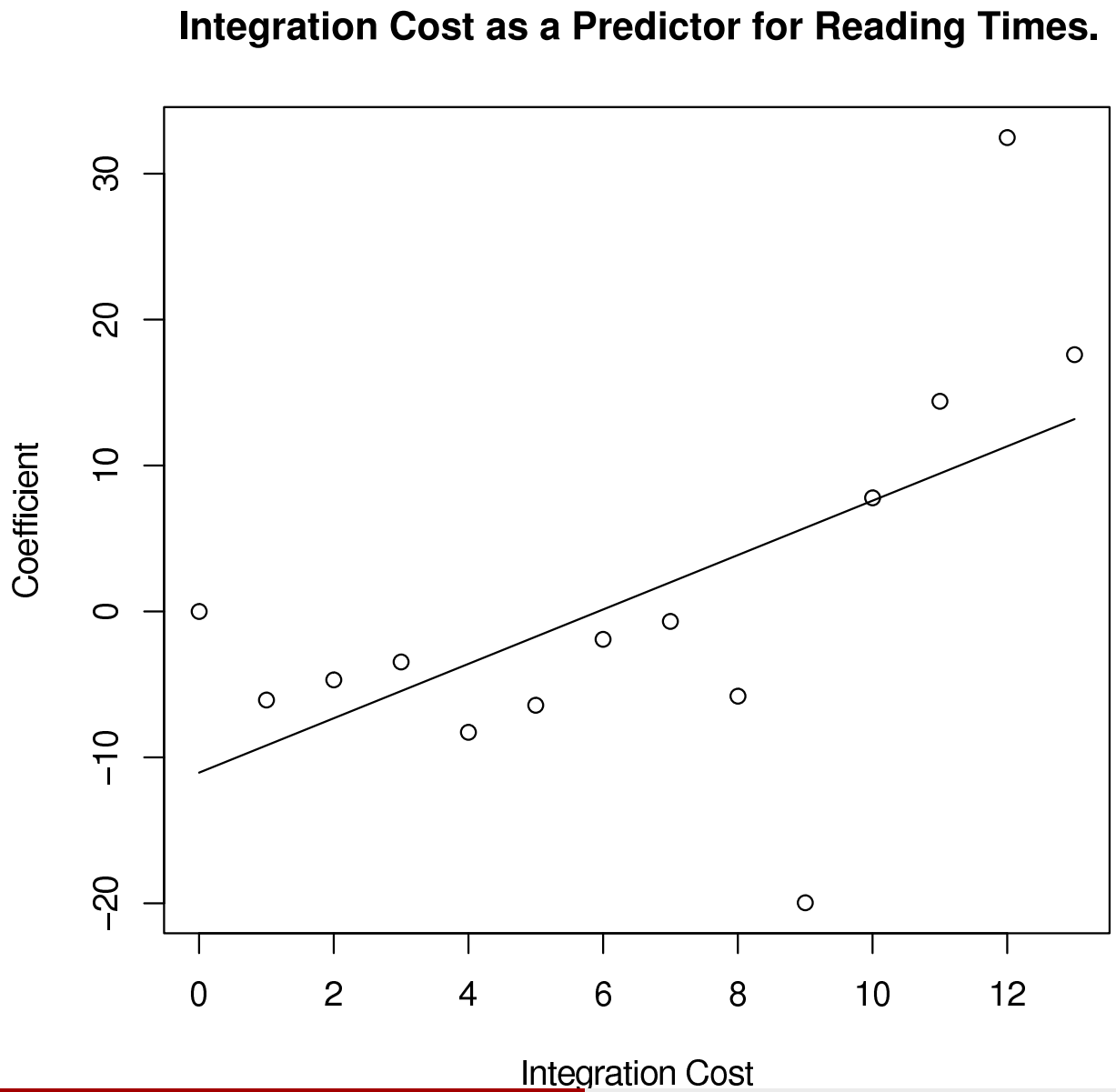
Integration Cost: Results for all words

Predictor	Coefficients	Significance
(Intercept)	102.43	***
WORDLEN	31.56	***
LOGFREQ	6.38	***
FORWTRANS	-8.16	***
BACKTRANS	-0.71	***
WORDPOS	-0.23	***
PREVFIX	-21.54	***
LAUNDIST	0.10	***
LANDPOS	2.50	***
INTEGCOST	-1.55	***
FORWTRANS:BACKTRANS	-0.43	***
WORDLEN:LANDPOS	-1.66	***
WORDLEN:LOGFREQ	-1.73	***

$*p < 0.05$, $**p < 0.01$, $***p < 0.001$

Contrary to prediction: words with higher integration cost are read faster

Problems with Integration Cost



New analysis: estimated separate coefficient for each integration cost value:

- high number of words with IC = 0
- negative coefficients for words with IC < 10
- but: overall tendency correct: words with higher IC have larger coefficients

Integration Cost at Nouns

Predictor	Coefficient	Significance
(Intercept)	74.40	***
WORDLEN	34.13	***
LOGFREQ	8.89	***
FORWTRANS	-7.30	***
BACKTRANS	0.15	
WORDPOS	-0.21	***
PREVFIX	-19.64	***
LANDPOS	-3.13	***
log(INTEGCOST+1)	17.21	***
FORWTRANS:BACKTRANS	-0.25	*
WORDLEN:LANDPOS	-0.99	***
WORDLEN:LOGFREQ	-2.13	***

$*p < 0.05$, $**p < 0.01$, $***p < 0.001$

As predicted: nouns with lower integration cost are read faster

Integration Cost at Nouns

Predictor	Coefficient	Significance
(Intercept)	74.40	***
WORDLEN	34.13	***
LOGFREQ	8.89	***
FORWTRANS	-7.30	***
BACKTRANS	0.15	
WORDPOS	-0.21	***
PREVFIX	-19.64	***
LANDPOS	-3.13	***
log(INTEGCOST+1)	17.21	***
FORWTRANS:BACKTRANS	-0.25	*
WORDLEN:LANDPOS	-0.99	***
WORDLEN:LOGFREQ	-2.13	***

$*p < 0.05$, $**p < 0.01$, $***p < 0.001$

As predicted: nouns with lower integration cost are read faster

Integration Cost at Nouns

Predictor	Coefficient	Significance
(Intercept)	74.40	***
WORDLEN	34.13	***
LOGFREQ	8.89	***
FORWTRANS	-7.30	***
BACKTRANS	0.15	
WORDPOS	-0.21	***
PREVFIX	-19.64	***
LANDPOS	-3.13	***
log(INTEGCOST+1)	17.21	***
FORWTRANS:BACKTRANS	-0.25	*
WORDLEN:LANDPOS	-0.99	***
WORDLEN:LOGFREQ	-2.13	***

$*p < 0.05$, $**p < 0.01$, $***p < 0.001$

As predicted: nouns with lower integration cost are read faster

Integration Cost at Nouns

Predictor	Coefficient	Significance
(Intercept)	74.40	***
WORDLEN	34.13	***
LOGFREQ	8.89	***
FORWTRANS	-7.30	***
BACKTRANS	0.15	
WORDPOS	-0.21	***
PREVFIX	-19.64	***
LANDPOS	-3.13	***
log(INTEGCOST+1)	17.21	***
FORWTRANS:BACKTRANS	-0.25	*
WORDLEN:LANDPOS	-0.99	***
WORDLEN:LOGFREQ	-2.13	***

$*p < 0.05$, $**p < 0.01$, $***p < 0.001$

As predicted: nouns with lower integration cost are read faster

Surprisal: Results

Predictor	Coef	Sign
(Intercept)	132.93	***
WORDLEN	32.25	***
LOGFREQ	4.33	***
WORDPOS	-0.25	***
PREVFIX	-19.23	***
LANDPOS	2.68	***
SURPRIS	0.75	***
ULXSPRS	2.46	***
SURPRIS:ULXSPRS	-0.11	**
LEN:LANDPOS	-1.71	***
LEN:LOGFREQ	-1.78	***

$*p < 0.05$, $**p < 0.01$, $***p < 0.001$

Predictor	Coef	Sign
(Intercept)	99.16	***
WORDLEN	31.68	***
LOGFREQ	6.32	***
WORDPOS	-0.23	***
PREVFIX	-19.78	***
LANDPOS	2.41	***
FORWTRANS	-7.50	***
BACKTRANS	-0.42	
ULXSPRS	1.05	***
FORWTR:BACKTR	-0.36	***
LEN:LANDPOS	-1.67	***
LEN:LOGFREQ	-1.73	***

$*p < 0.05$, $**p < 0.01$, $***p < 0.001$

- lexicalized surprisal is a significant predictor only in the absence of transitional probabilities
- unlexicalized surprisal is independent of transitional probabilities