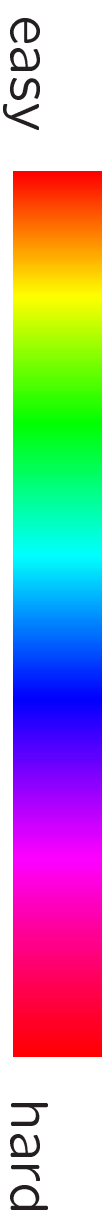


sentence processing



- reading times
- error scores
- eye fixations
- evoked potentials

processing difficulty

Q. Why is sentence processing **more** or **less** difficult?

A. Because more or less information is transacted.

choice of continuation informs hearer

the boy eats...

the boy eats shy...

the boy eats **using**...

the boy eats **like**...

the boy eats **the**...

the boy eats **his**...

the boy eats **at**...

the boy eats **of**...

the boy eats went...

terms

missing information

entropy degree of uncertainty

expected surprisal of unknown outcome

information conveyed when entropy is reduced

reducing entropy



sides binary names entropy

4

00
01
10
11

2 bits



8

000
001
010
011
100
101
110
111

3 bits



16

0000
0001
0010
⋮

4 bits

plan of talk

1. gallery of graded processing difficulty
 - garden-path sentences
 - center-embedded sentences
 - subject and object relativization
 - the Accessibility Hierarchy entire
2. deriving the predictions
 - formalizing information transacted as entropy reduction
 - calculating conditional entropy reduction

garden-path sentences

the horse raced past the barn

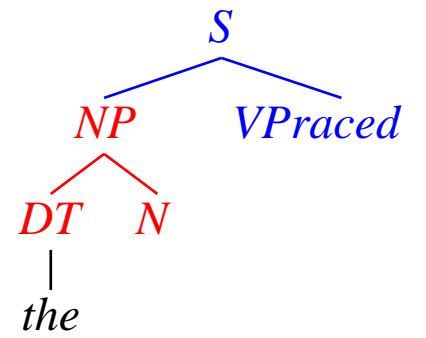
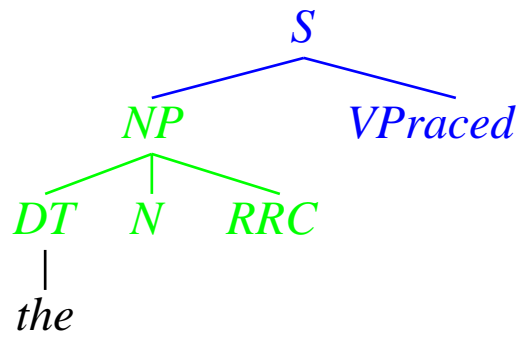
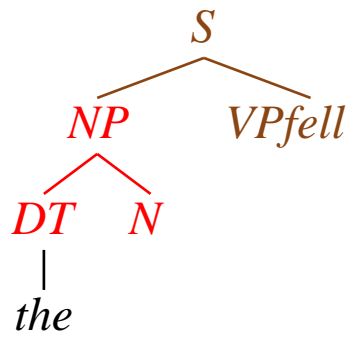
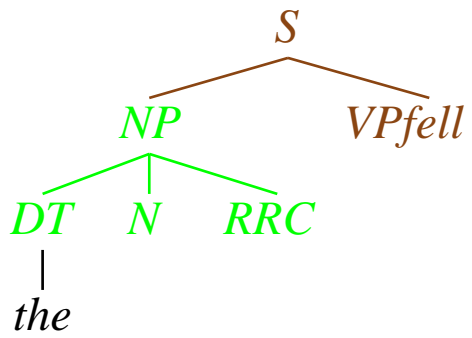
the horse raced past the barn fell

Bever 70

raced $\stackrel{?}{=}$ main verb

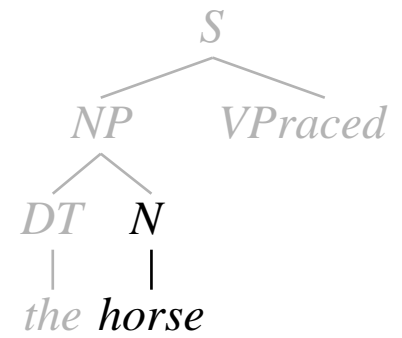
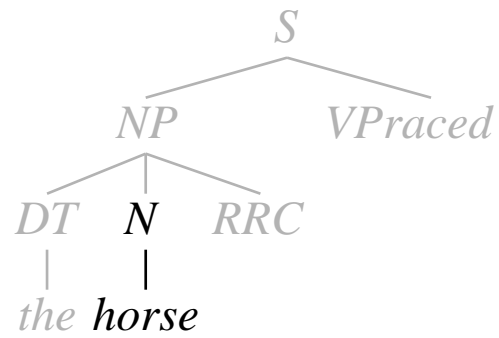
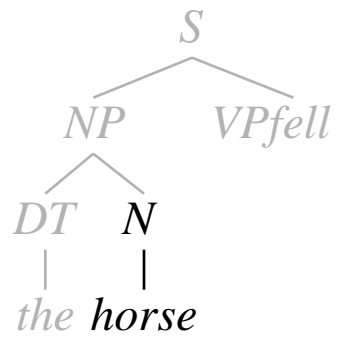
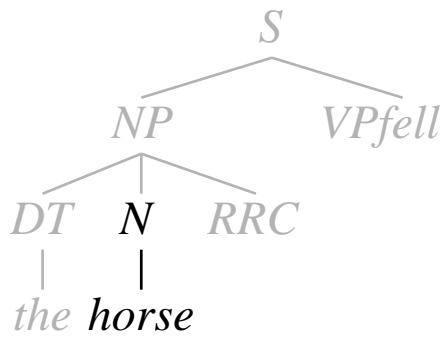
raced $\stackrel{?}{=}$ head of reduced relative clause

main verb/reduced relative 1/4



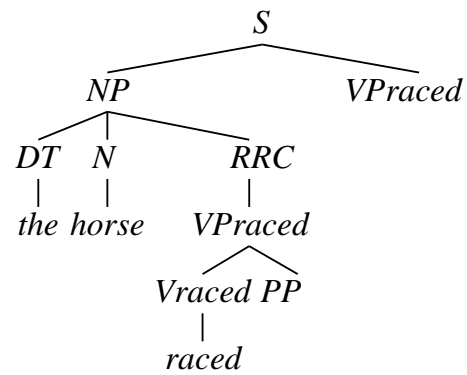
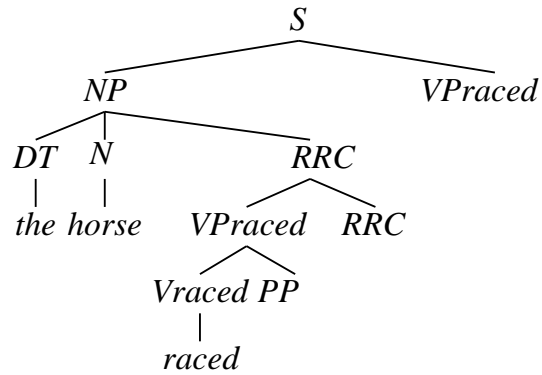
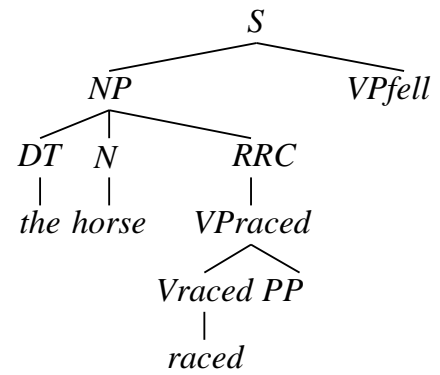
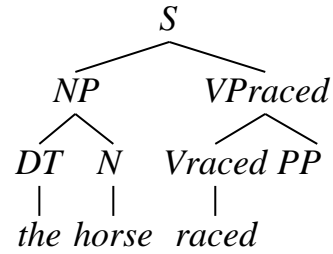
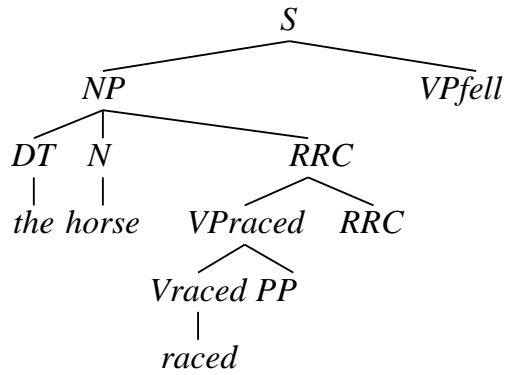
entropy: 4.65 bits

main verb/reduced relative 2/4



entropy: 3.65 bits

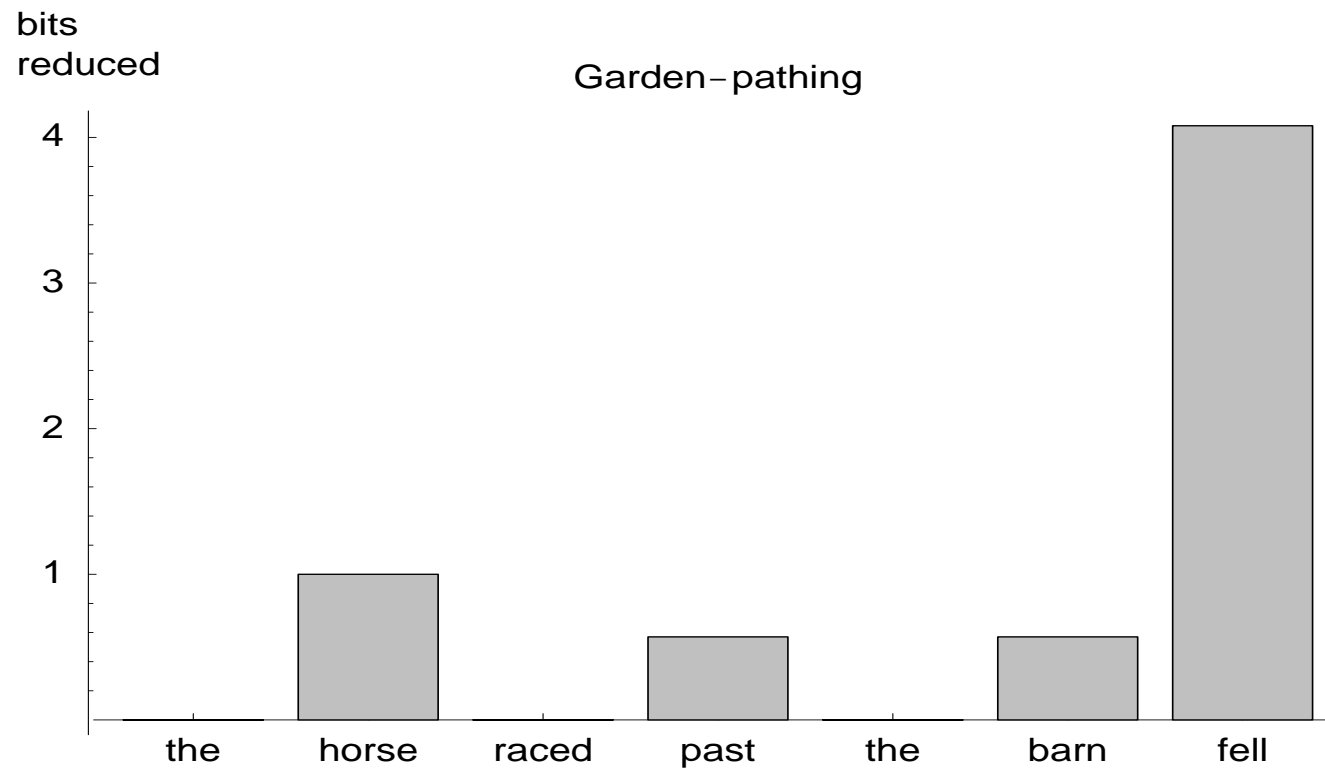
main verb/reduced relative 3/4



entropy: 5.2 bits

main-verb/reduced relative 4/4

At “fell” the hearer gets 4 bits



total: 6.2 bits

center-embedding

the reporter disliked the editor

? the reporter [who the senator attacked] disliked the editor

*? the reporter [who the senator [who John met] attacked]
disliked the editor

Chomsky 57, Yngve 60, Chomsky & Miller 63,.....

summed entropy reductions

21 bits the reporter disliked the editor

39 bits the reporter [who the senator attacked] disliked the editor

48 bits the reporter [who the senator [who John met] attacked]
disliked the editor

but

24 bits John met the senator [who attacked the reporter
[who disliked the editor]]

subject/object processing asymmetry

the reporter who *t* sent the photographer to the editor
hoped for a good story

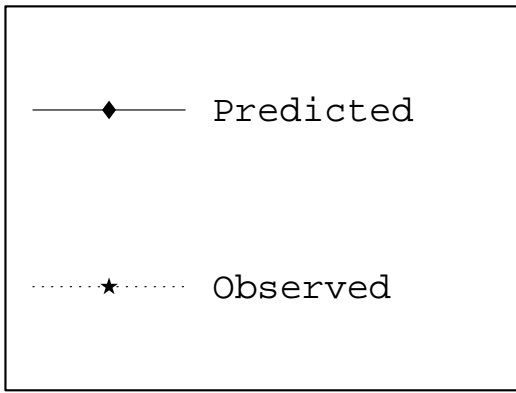
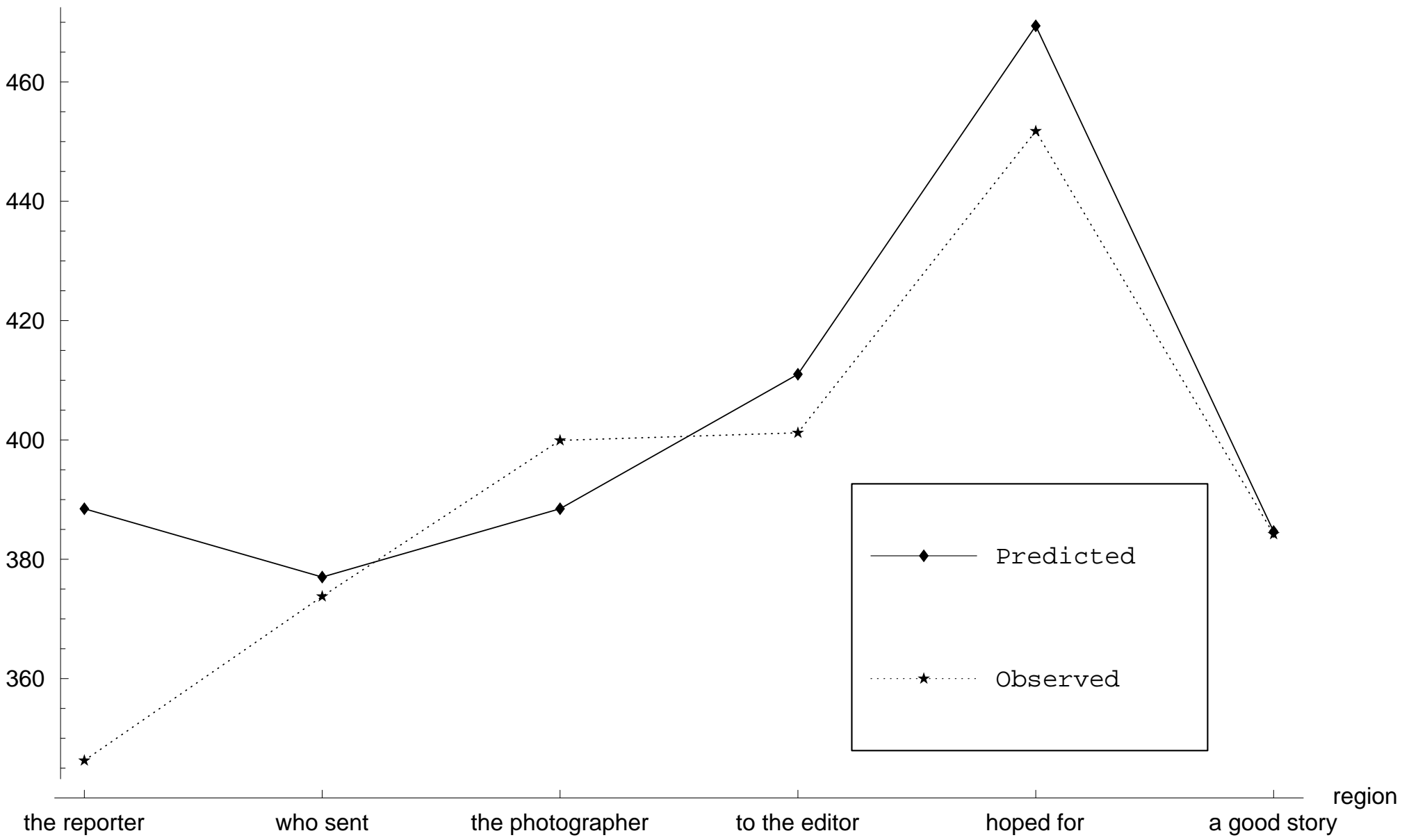
< *easier*

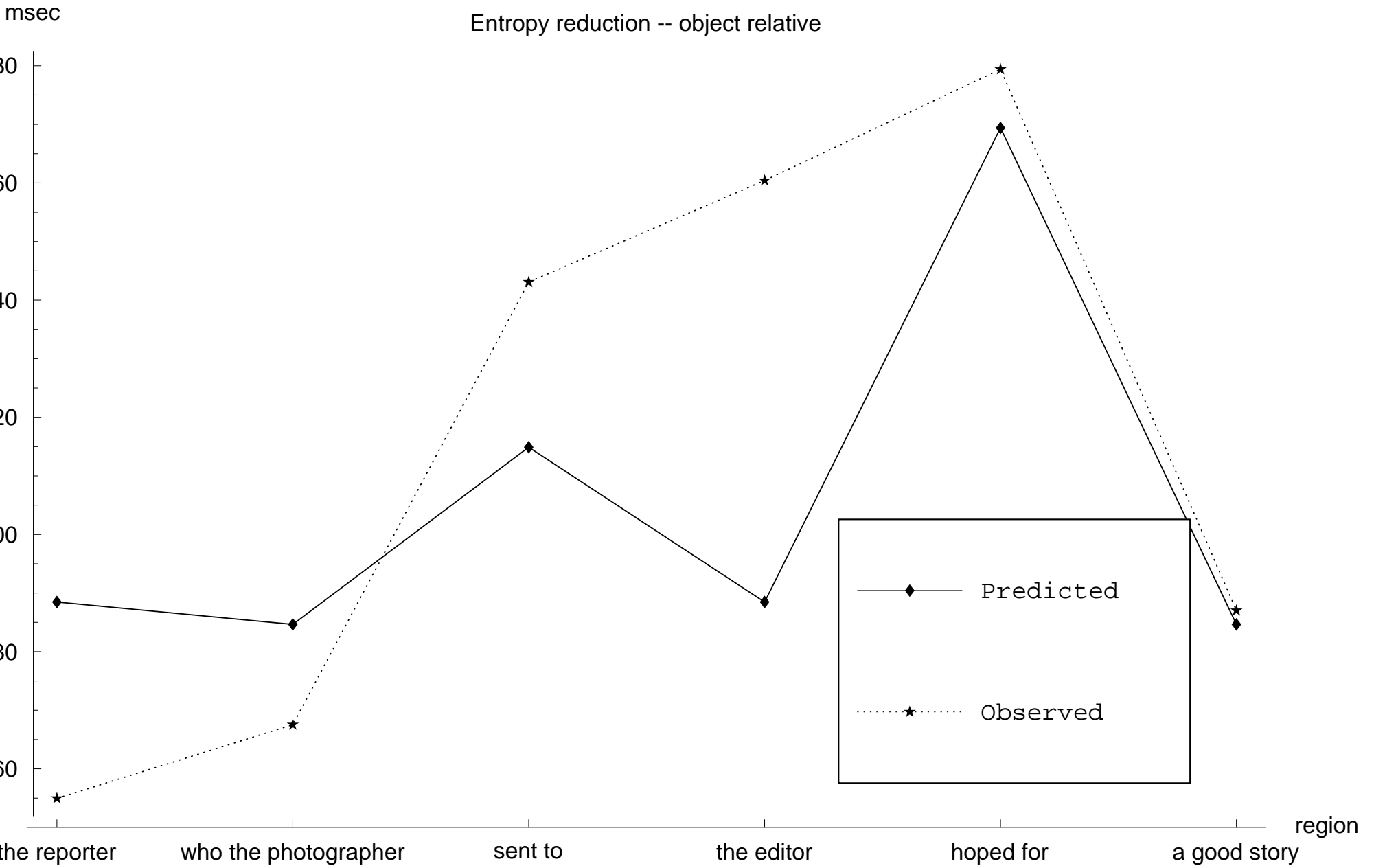
the reporter who the photographer *sent t* to the editor
hoped for a good story

see references in Gibson 98
Grodner, Watson and Gibson 00

msec

Entropy reduction -- subject relative





beyond subject and object relativization

indirect object the boy who Paul sold the book to *t* hates reading

oblique the girl who Sue wrote the story with *t* is proud

genitive subject the boy whose brother *t* tells lies is always honest

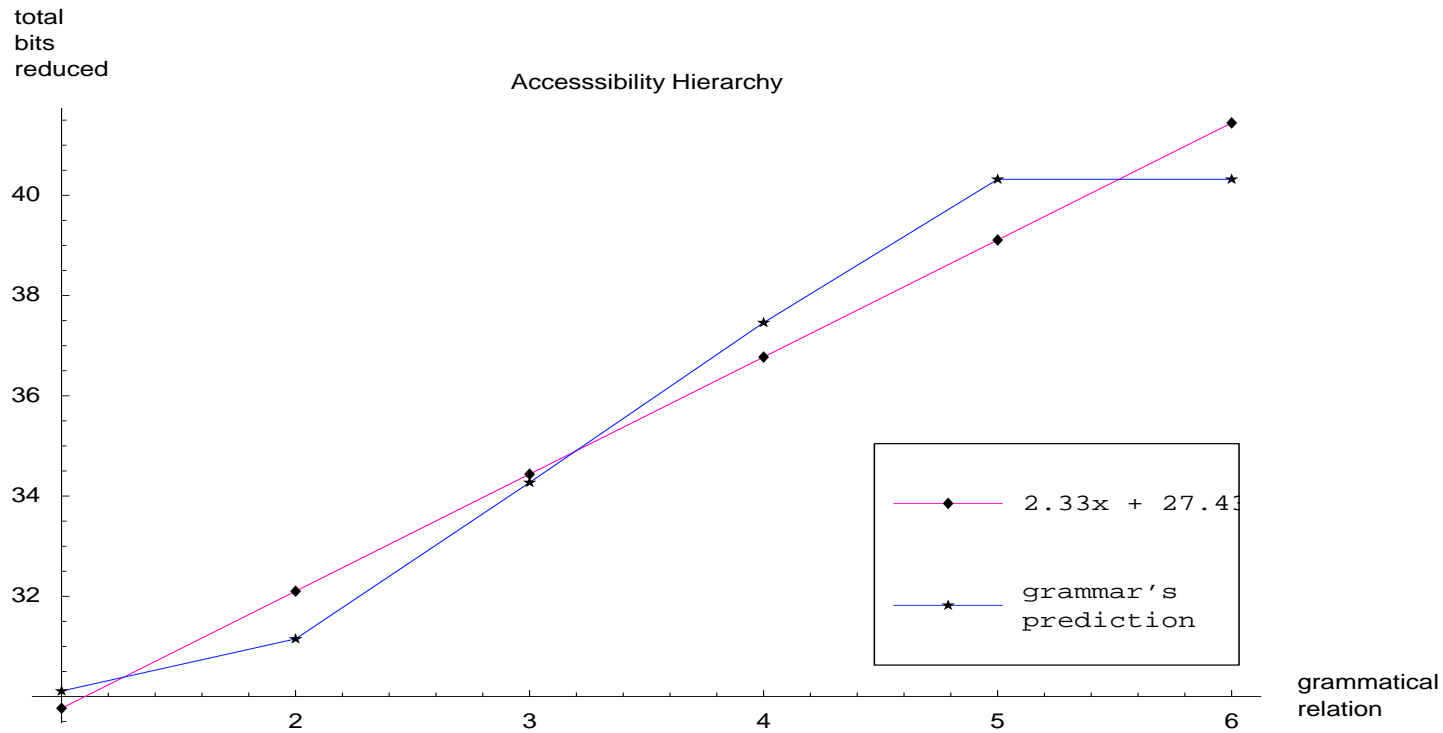
genitive object the sailor whose ship Jim took *t* had one leg

the accessibility hierarchy

1	2	3	4	5/6
Subject	Direct Object	Indirect Object	Oblique	Genitive
(406)	(364)	(342)	(279)	(169)

repetition accuracy results
S. Hawkins and Keenan 87

entropy reduction increases with position on AH



$$r^2 = 0.96, p < 0.001$$

formalization

defines

grammar
probabilities
entropy
conditional entropy

possible structures
structural expectations
difficulty of guessing structure
uncertainty of hearer

kinds of probabilistic grammar

grammar

influence order of rule-application (Salomaa 69, Smith 73)
probabilistic tree automata (Ellis 69)
attach probabilities to generative rules (Grenander 67, Suppes 70)
conceive of parse as Bayesian net (Jurafsky 96)
optimal sequencing of reanalyses (Chater, Crocker & Pickering 98)
implicitly represented in hidden units (Elman, Tabor, others)

see also Jurafsky survey in Bod, Hay & Jannedy 02

rule choice in a probabilistic grammar

probabilities

0.87 NP \rightarrow the boy

0.13 NP \rightarrow the tall boy

- Particular rule choices are *alternative outcomes*.
- Nonterminal symbols are *random variables*

entropy of rule choice

entropy

$$H(X) = - \sum_{x \in X} p(x) \log_2 p(x)$$

0.87 NP → the boy

0.13 NP → the tall boy

$$\begin{aligned} H(\text{NP}) &= -[(0.87 \times \log_2 0.87) + (0.12 \times \log_2 0.12)] \\ &\approx 0.55 \text{bits} \end{aligned}$$

entropy of derivations

conditional entropy

$H_G(\mathbf{s})$ difficulty of guessing derivation in G

$H_i \doteq H_G(\mathbf{s} | w_0 \dots w_i)$ difficulty of guessing derivation
that generates words $w_0 w_1 \dots w_i$

formalized claim

word-by-word reading time is linearly related to entropy reduction

$$\text{RT}(w_i) = \alpha [\text{reduction}(H_{i-1}, H_i)] + \beta$$

$$\alpha = 7.38$$

$$\beta = 377$$

$$r^2 = 0.49, p < 0.01$$

deriving predictions

the entropy of a nonterminal XP is the sum of
the **rule choice entropy**
and the **expected entropy of XP's children**

$$h(\text{XP}) = - \sum_{\text{XP} \rightarrow \text{X}' \text{ ZP} \in \text{rules}(\text{XP})} p_{\text{XP} \rightarrow \text{X}' \text{ ZP}} \log_2 p_{\text{XP} \rightarrow \text{X}' \text{ ZP}}$$
$$H(\text{XP}) = h(\text{XP}) + \sum_{\text{XP} \rightarrow \text{X}' \text{ ZP} \in \text{rules}(\text{XP})} p_{\text{XP} \rightarrow \text{X}' \text{ ZP}} [H(\text{X}') + H(\text{ZP})]$$

(Grenander 67)

solution to recursion relation

Assuming a well-defined probability model
and letting M be the expectation matrix with components

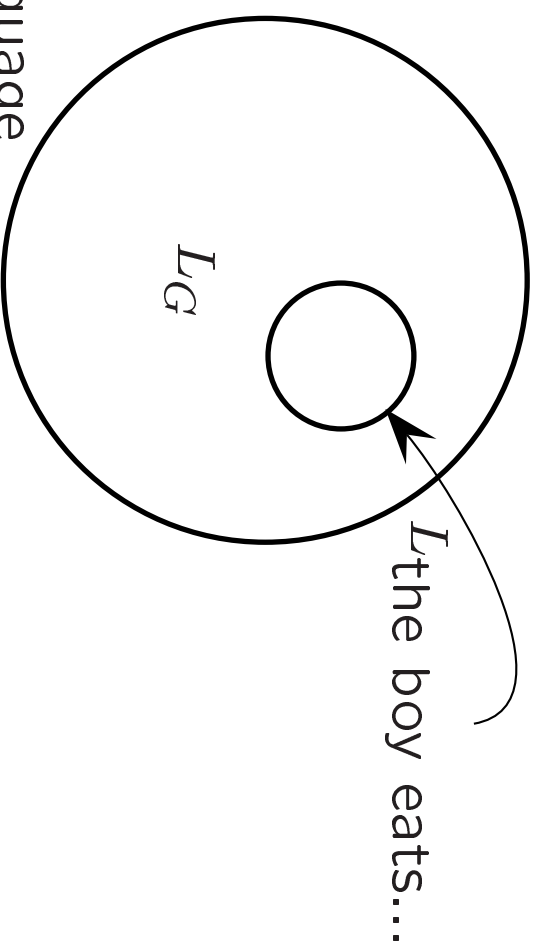
$$m_{XP, YP} = \sum_{XP \rightarrow \alpha \in \text{rules}(XP)} p_{XP \rightarrow \alpha} \#_{YP}(\alpha)$$

then

$$\begin{aligned} H &= h + MH \\ h &= H - MH = (I - M)H \\ H &= (I - M)^{-1}h \end{aligned}$$

(Grenander 67)

entropy of parser state



grammar \equiv derivations in the language

parser state \equiv derivations consistent with words seen so far

a parser state *is* a grammar

(Lang 74, 88, Billot and Lang 89)

0.20	NP	→	SPECNP NBAR
0.40	NP	→	I
0.40	NP	→	John
1.00	SPECNP	→	DT
0.50	NBAR	→	NBAR S[+R]
0.50	NBAR	→	N
1.00	S	→	NP VP
0.87	S[+R]	→	NP[+R] VP
0.13	S[+R]	→	NP[+R] S/NP
1.00	S/NP	→	NP VP/NP
0.50	VP/NP	→	V[SUBCAT2] NP/NP
0.50	VP/NP	→	V[SUBCAT3] NP/NP PP[to]
0.33	VP	→	V[SUBCAT2] NP
0.33	VP	→	V[SUBCAT3] NP PP[to]
0.33	VP	→	V[SUBCAT4] PP[for]
0.33	V[SUBCAT2]	→	met
0.33	V[SUBCAT2]	→	attacked
0.33	V[SUBCAT2]	→	disliked
1.00	V[SUBCAT3]	→	sent
1.00	V[SUBCAT4]	→	hoped
1.00	PP[to]	→	PBAR[to] NP
1.00	PBAR[to]	→	P[to]
1.00	P[to]	→	to
1.00	PP[for]	→	PBAR[for] NP
1.00	PBAR[for]	→	P[for]
1.00	P[for]	→	for
1.00	NP[+R]	→	who
0.50	DT	→	the
0.50	DT	→	a
0.17	N	→	editor
0.17	N	→	senator
0.17	N	→	reporter
0.17	N	→	photographer
0.17	N	→	story
0.17	N	→	ADJ N
1.00	ADJ	→	good
1.00	NP/NP	→	ε

grammatical issues

“It is necessary only to make explicit the relational character of these notions by defining ‘Subject-of,’ for English, as the relation holding between the NP of a sentence of the form $NP \frown Aux \frown VP$ and the whole sentence, ‘Object-of’ as the relation between the NP of a VP of the form $V \frown NP$ and the whole VP, etc” Chomsky 65, page 69

“grammatical relations are independent of phrase structure configuration”
Perlmutter and Postal 74

promotion analysis of relative clauses

[*DP* the [*Agr_D* [*CP* I met [*DP* who [*NP* boy]]]]]]

[*DP* the [*Agr_D* [*CP* [*DP* who [*NP* boy]]]_{*i*} [*IP* I met *t_i*]]]]

A curved arrow points from the trace *t_i* in the IP complement to the relative pronoun *who* in the CP specifier, indicating movement.

[*DP* the [*Agr_P* boy [*Agr_D* [*CP* [*DP* who [*NP*]]_{*i*} [*IP* I met *t_i*]]]]]

A curved arrow points from the trace *t_i* in the IP complement to the noun *boy* in the Agr_P specifier, indicating movement.

[*DP* Agr_D+the [*Agr_P* boy [*t_{Agr}* [*CP* [*DP* who [*NP*]]_{*i*} [*IP* I met *t_i*]]]]]]

A curved arrow points from the trace *t_{Agr}* in the Agr_P specifier to the Agr_D head, indicating movement.

Minimalist Grammars

MGs formalize certain aspects of Chomsky's Minimalist Program

- words are bundles of features
- syntactic structure is created bottom-up by rules named Merge and Move
- these rules respect a simple version of Chomsky's "Shortest Move Condition"
- there are provisions for both phrasal movement and head movement
- movement (as well as merger) always 'checks' features, meaning that checked features are removed from the computation
- ubiquitous empty categories handled analogously to overt ones

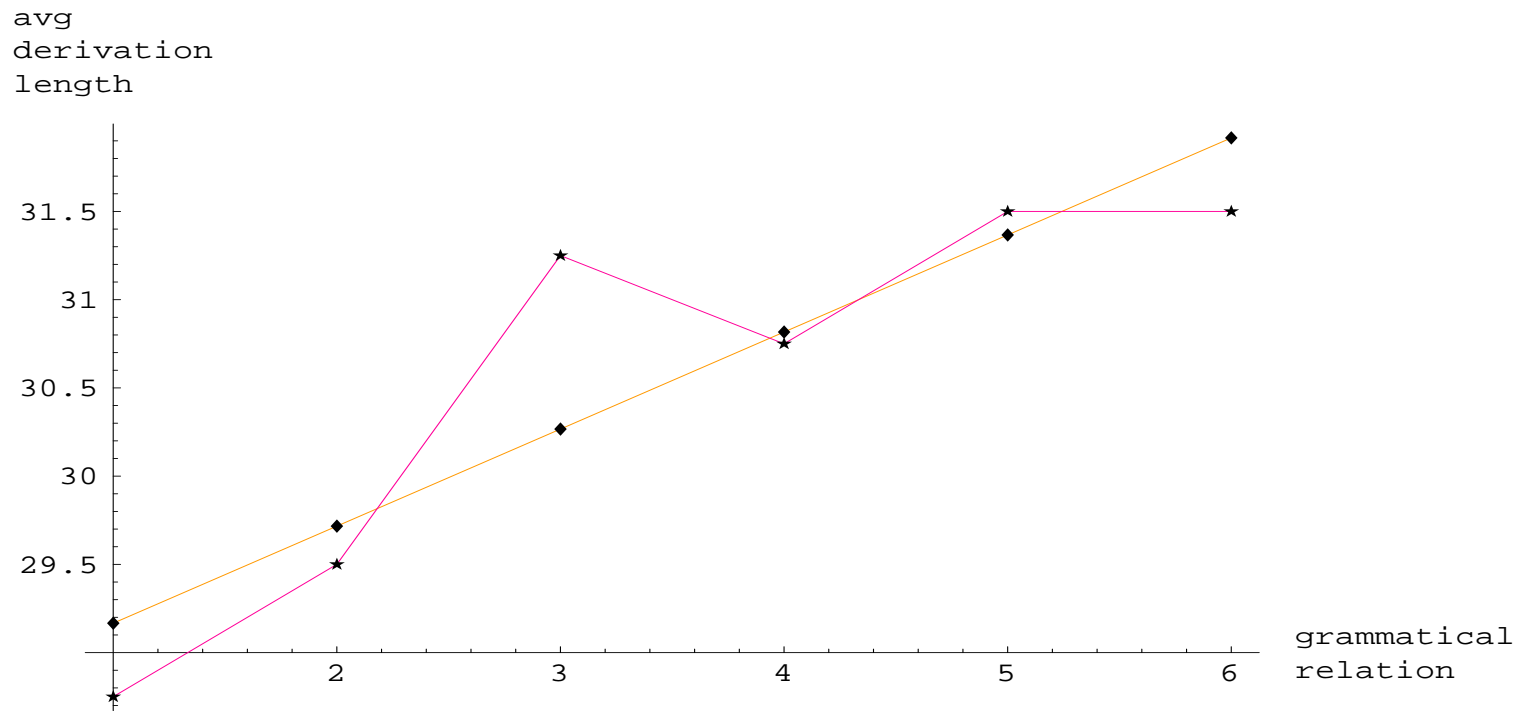
MGs are *mildly context-sensitive*: Michaelis, Harkema

Larsonian grammar

- VP shells for double objects following Larson 88
- promotion analysis of relative clauses after Kayne 94
- dative shift as in Baker 97
- affix-hopping analysis of English auxiliaries from Chomsky 57
- verb, tense, preposition case assigners
- number agreement
- adjoined temporal and superlative modifiers
- lacking gender agreement, *do*-support, negation, semantics,...

derivations longer according to hierarchy

J. Hawkins 99: Filler-Gap Domain is larger further along the hierarchy.



$$r^2 = 0.79, p < 0.02$$

computing predictions

1. use chart parser to find all derivations consistent with prefix (Lang)
2. represent result as a probabilistic grammar
3. invert a matrix to find conditional entropy of start symbol (Grenander)
4. subtract values for adjacent prefixes if entropy goes down
5. fit linking theory parameters α, β

conclusion

Entropy reduction predicts graded processing

- important role for grammatical knowledge
- important role for statistical knowledge