Information-theoretic approaches to syntactic processing

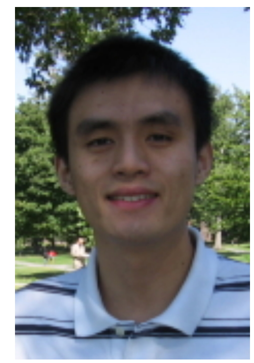
John Hale

Cornell University
Linguistics Department

Thank you:

Jiwon Yun

Zhong Chen



Tim Hunter

sentence comprehension

easy  hard

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

leading idea

Q. when is comprehension
more (vs less) difficult?

A. where more (vs less) information
is conveyed

choice of continuation informs the 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...

conditional entropy

avg uncertainty of this distribution

	probability (lets pretend)
the boy eats shy people for breakfast	1.0×10^{-25}
the boy eats using chopsticks on Tuesday	1.0×10^{-7}
the boy eats like a hippopotamous	1.0×10^{-6}
the boy eats the dog with a spoon	0.0001
the boy eats his sister's bicycle	0.0005
the boy eats at Denny's frequently	0.00001
the boy eats of the forbidden fruit	1.0×10^{-66}
the boy eats went for a walk	0.0

fluctuation

$$H(Derivation|Prefix = \text{"the boy eats"})$$

$$H(Derivation|Prefix = \text{"the boy eats his"})$$

any downward change quantifies
information gained from "his"

entropy reduction hypothesis

$$\boxed{\text{observed processing effort reflects decreases in } H_i}$$

where H_i abbreviates $H(Derivation|Prefix = w_{0 \ldots i})$

outline

Entropy reduction studies
relative clauses in English and Korean

How does it work?
computing $\downarrow H_i$

Why does it work?
reflections on information theory & linguistics

garden path sentences

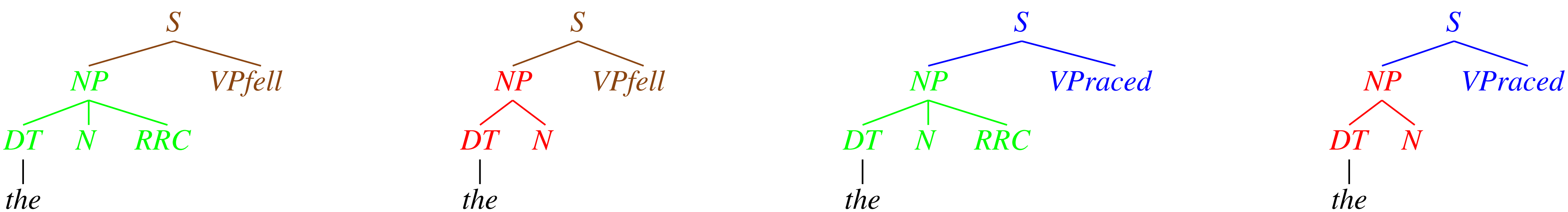
Bever 70

naive probabilistic grammar

1.00	S	$\rightarrow$	NP VP
0.88	NP	$\rightarrow$	DT NN
0.12	NP	$\rightarrow$	NP RRC
1.00	PP	$\rightarrow$	IN NP
1.00	RRC	$\rightarrow$	Vppart PP
0.50	VP	$\rightarrow$	Vpast
0.50	VP	$\rightarrow$	Vppart PP
1.00	DT	$\rightarrow$	the
0.50	NN	$\rightarrow$	horse
0.50	NN	$\rightarrow$	barn
0.50	Vppart	$\rightarrow$	groomed
0.50	Vppart	$\rightarrow$	raced
0.50	Vpast	$\rightarrow$	raced
0.50	Vpast	$\rightarrow$	fell
1.00	IN	$\rightarrow$	past

Hale JPR 03

the



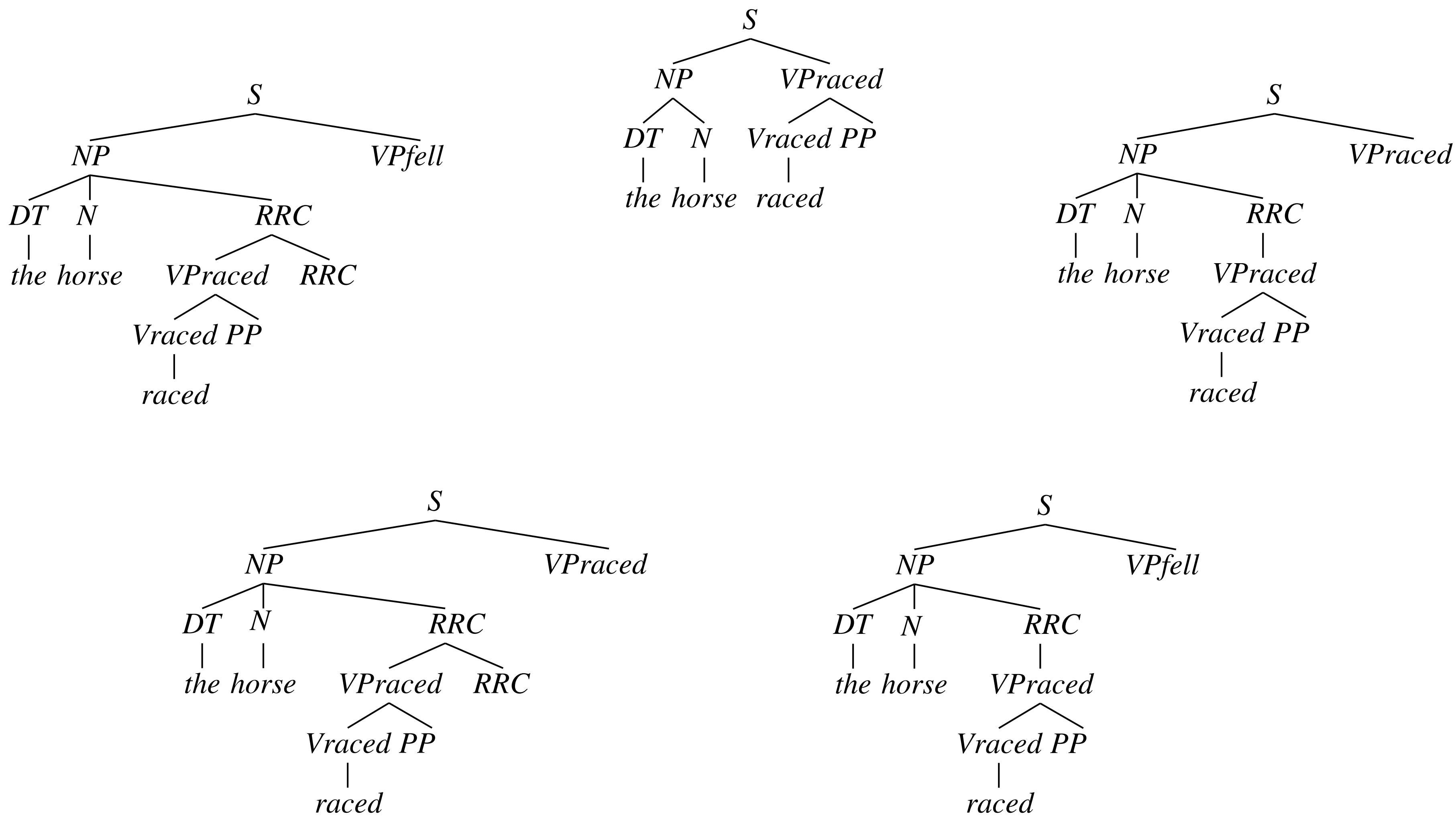
entropy: 4.65 bits

the horse

entropy: 3.65 bits

$\downarrow H = 1$

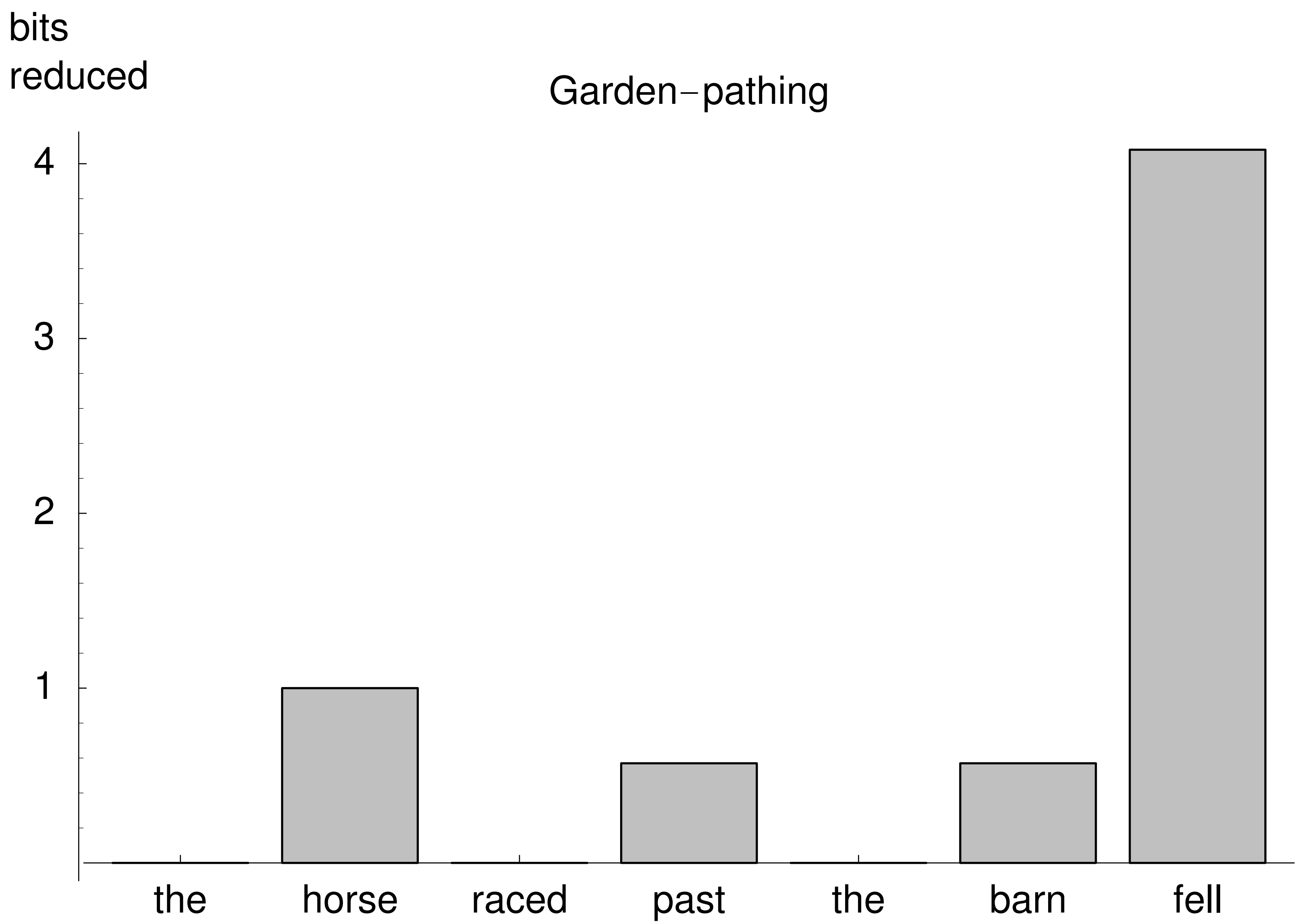
the horse raced



entropy: 5.2 bits

$\downarrow H$=*none*

last word gives 4 bits

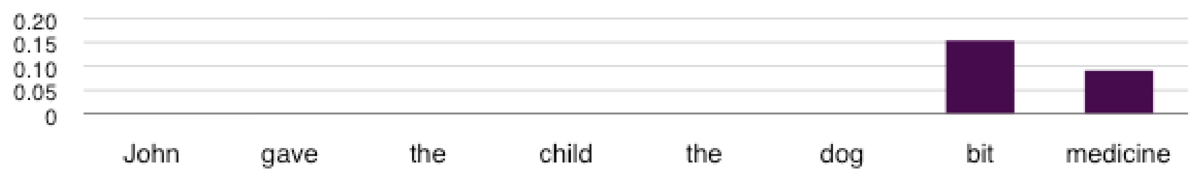


total: 6.2 bits

wide coverage dependency parser

Direct Object vs. Subject (late closure):

Double Object vs. Relative Clause:



Noun-phrase vs. Sentential Complement:

Compound Noun vs. Sentential Complement:

Hall & Hale AMLaP 07

GPSG-style fragment

0.20	NP	$\longrightarrow$	SPECNP NBAR
0.40	NP	$\longrightarrow$	I
0.40	NP	$\longrightarrow$	John
1.00	SPECNP	$\longrightarrow$	DT
0.50	NBAR	$\longrightarrow$	NBAR S[+R]
0.50	NBAR	$\longrightarrow$	N
1.00	S	$\longrightarrow$	NP VP
0.87	S[+R]	$\longrightarrow$	NP[+R] VP
0.13	S[+R]	$\longrightarrow$	NP[+R] S/NP
1.00	S/NP	$\longrightarrow$	NP VP/NP
0.50	VP/NP	$\longrightarrow$	V[SUBCAT2] NP/NP
0.50	VP/NP	$\longrightarrow$	V[SUBCAT3] NP/NP PP[to]
0.33	VP	$\longrightarrow$	V[SUBCAT2] NP
0.33	VP	$\longrightarrow$	V[SUBCAT3] NP PP[to]
0.33	VP	$\longrightarrow$	V[SUBCAT4] PP[for]
0.33	V[SUBCAT2]	$\longrightarrow$	met
0.33	V[SUBCAT2]	$\longrightarrow$	attacked
0.33	V[SUBCAT2]	$\longrightarrow$	disliked
1.00	V[SUBCAT3]	$\longrightarrow$	sent
1.00	V[SUBCAT4]	$\longrightarrow$	hoped

1.00	PP[to]	$\longrightarrow$	PBAR[to] NP
1.00	PBAR[to]	$\longrightarrow$	P[to]
1.00	P[to]	$\longrightarrow$	to
1.00	PP[for]	$\longrightarrow$	PBAR[for] NP
1.00	PBAR[for]	$\longrightarrow$	P[for]
1.00	P[for]	$\longrightarrow$	for
1.00	NP[+R]	$\longrightarrow$	who
0.50	DT	$\longrightarrow$	the
0.50	DT	$\longrightarrow$	a
0.17	N	$\longrightarrow$	editor
0.17	N	$\longrightarrow$	senator
0.17	N	$\longrightarrow$	reporter
0.17	N	$\longrightarrow$	photographer
0.17	N	$\longrightarrow$	story
0.17	N	$\longrightarrow$	ADJ N
1.00	ADJ	$\longrightarrow$	good
1.00	NP/NP	$\longrightarrow$	ϵ

center embedding

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 yet

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

subject vs object-extracted RC

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

selectively slower

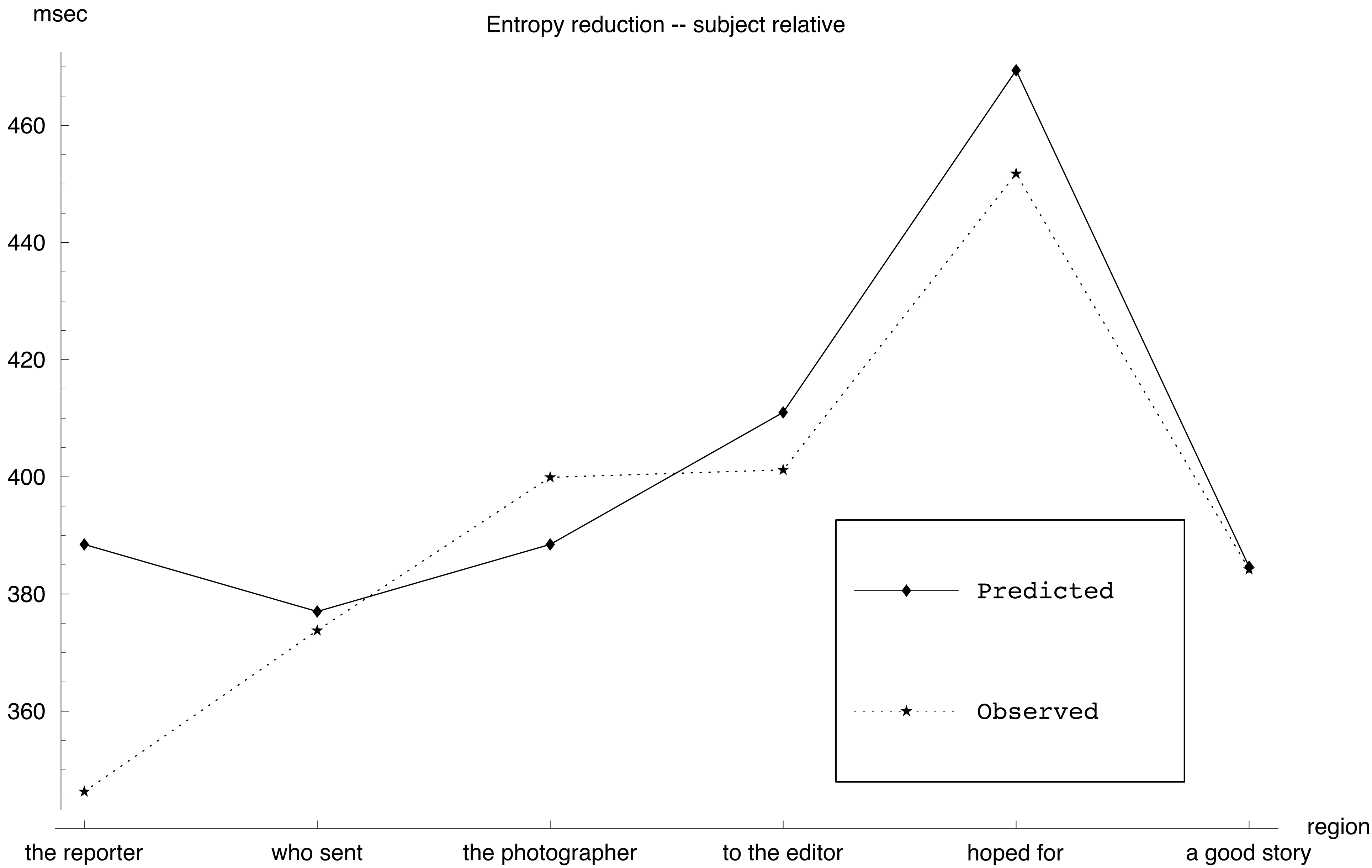
the reporter who the photographer $\boxed{\text{sent ∅ to the editor}}$ hoped for a good story

Grodner and Gibson CogSci 05
among others

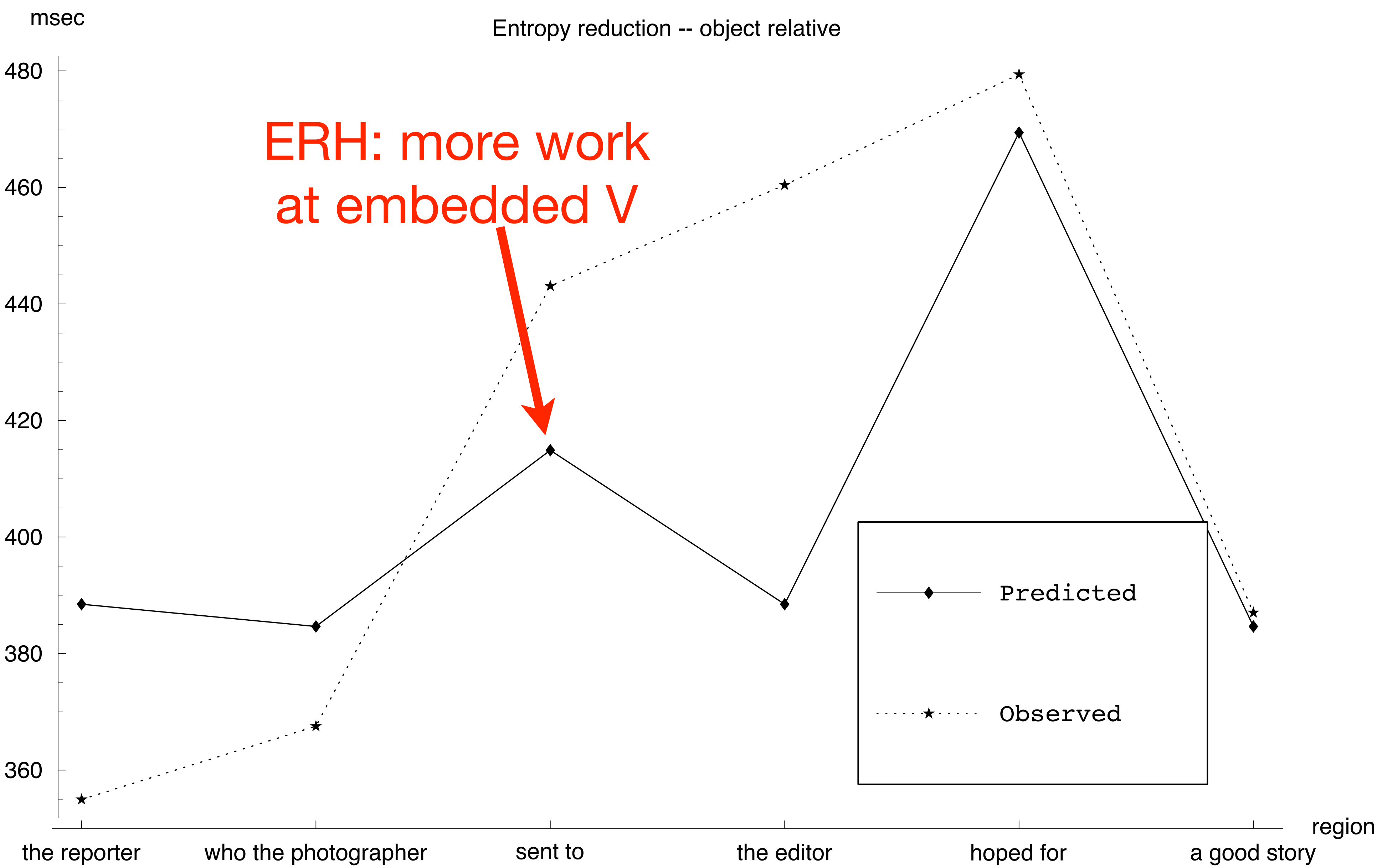
bits ↔ reading time

$$\mathsf{RT}(w_i) \;=\; \alpha\,(\downarrow H_i) + \beta$$

subject-extracted



object-extracted



bits ↔ reading time

$$\mathrm{RT}(w_i) \;=\; \alpha\,(\downarrow H_i) + \beta$$

$$\begin{aligned}\alpha &= 7.38\\ \beta &= 377\end{aligned}$$

$r^2 = 0.49, p < 0.01$

Many types of RCs

indirect object

the man who Stephen explained the accident to $\emptyset$ is kind

oblique

the girl who Sue wrote the story with $\emptyset$ is proud

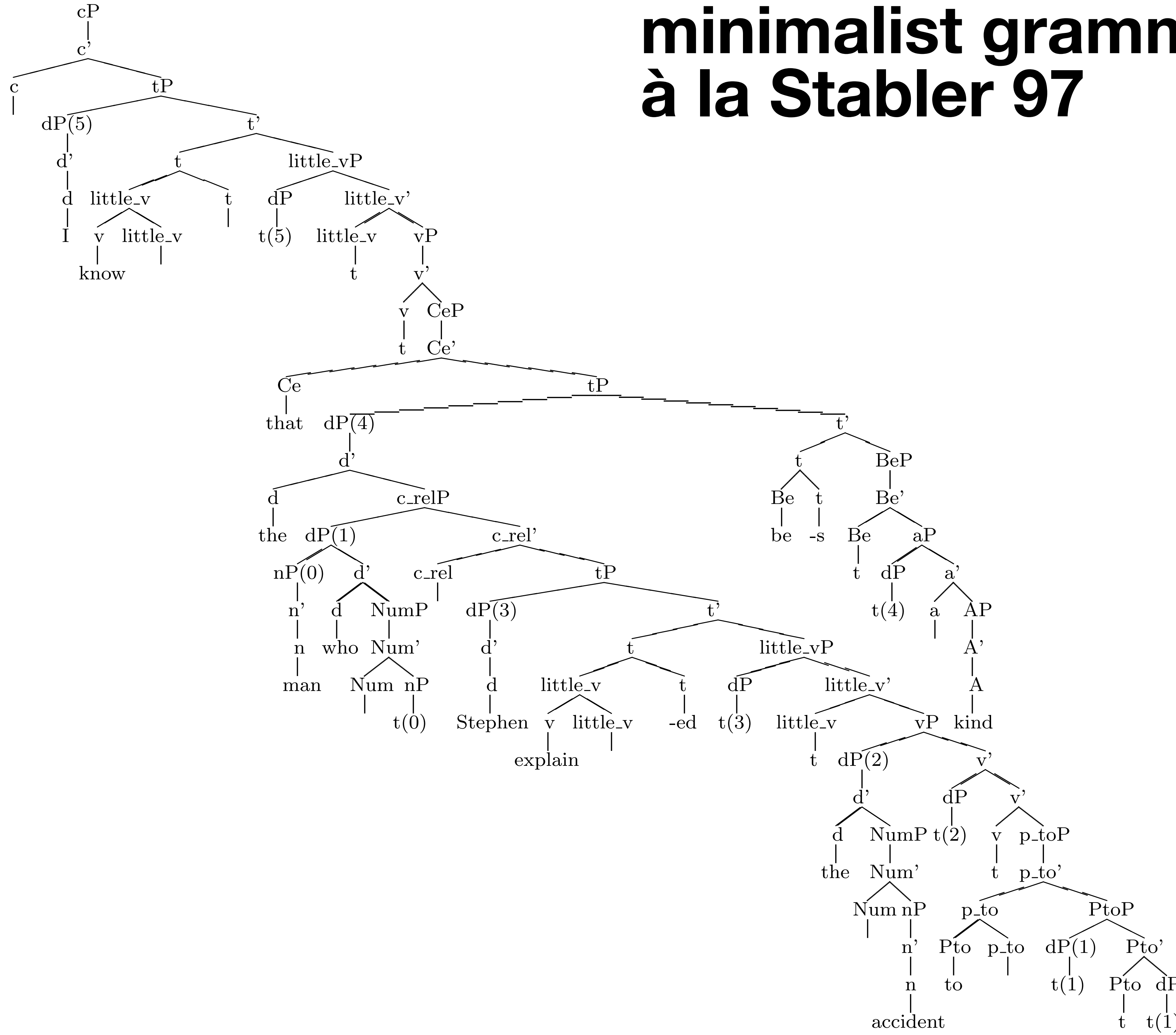
genitive subject

the boy whose brother $\emptyset$ tells lies is always honest

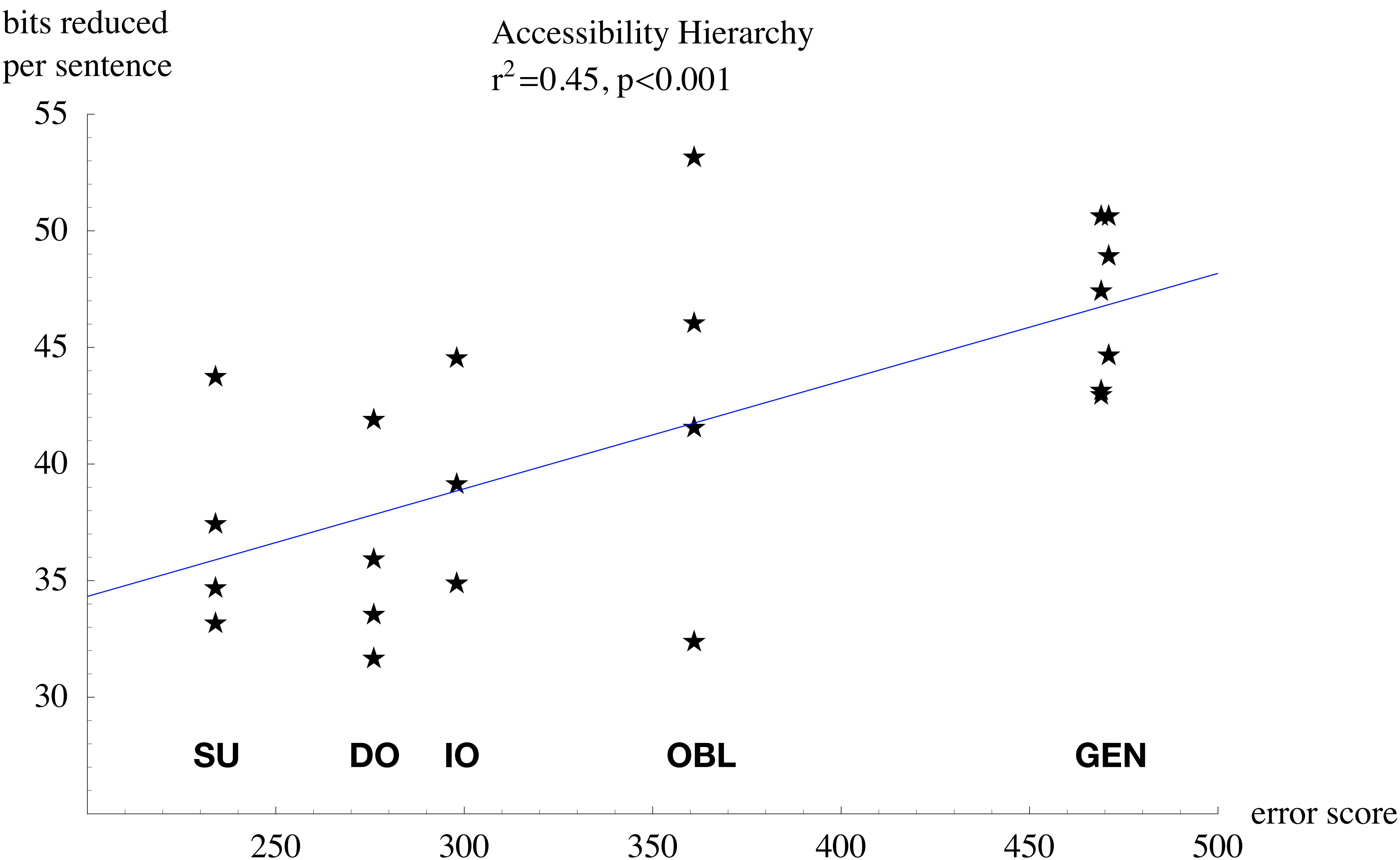
genitive object

the sailor whose ship Jim took $\emptyset$ had one leg

minimalist grammars à la Stabler 97

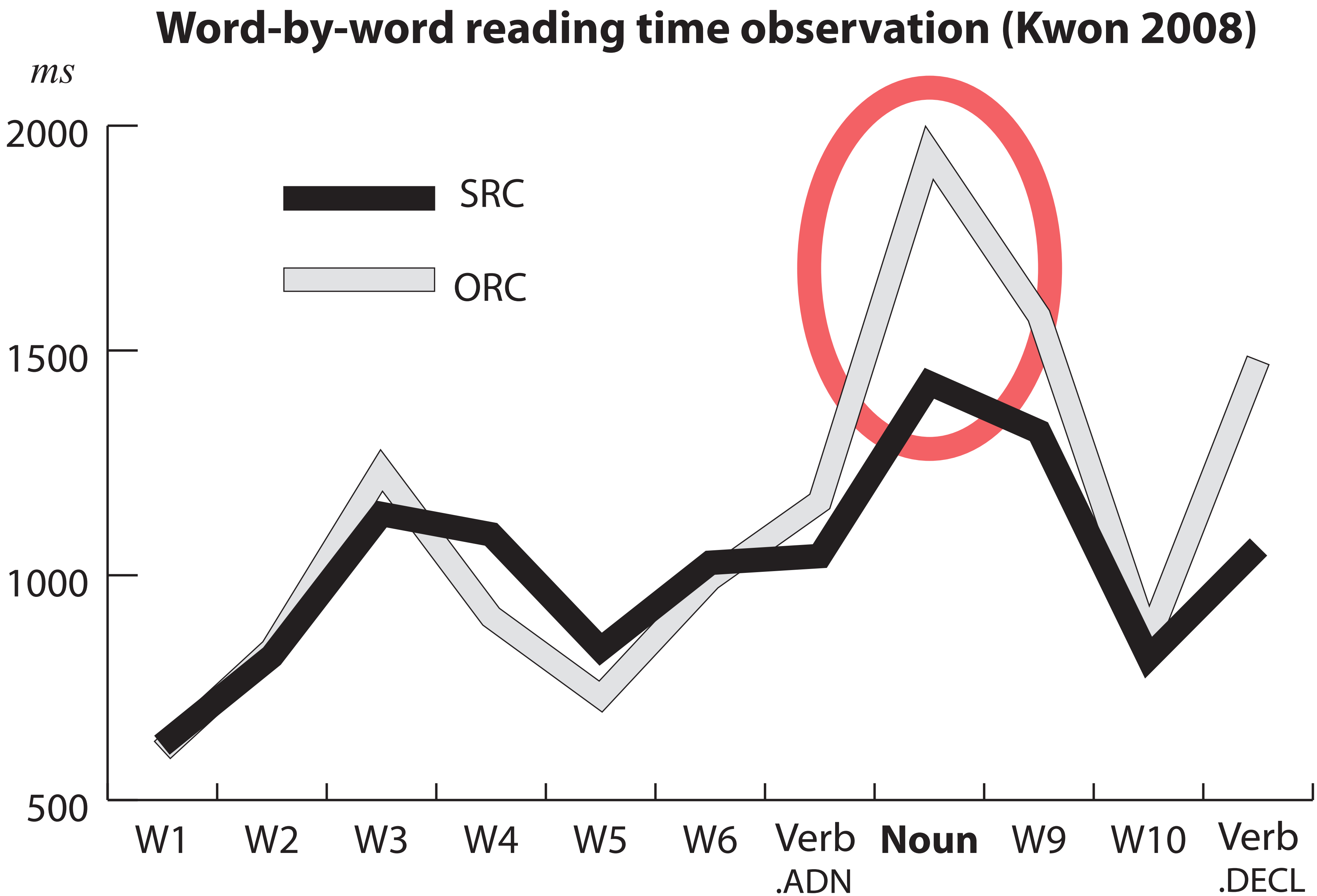


predicted work vs human accuracy



Keenan & S. Hawkins 87, Hale 06

Korean Subj-RC advantage

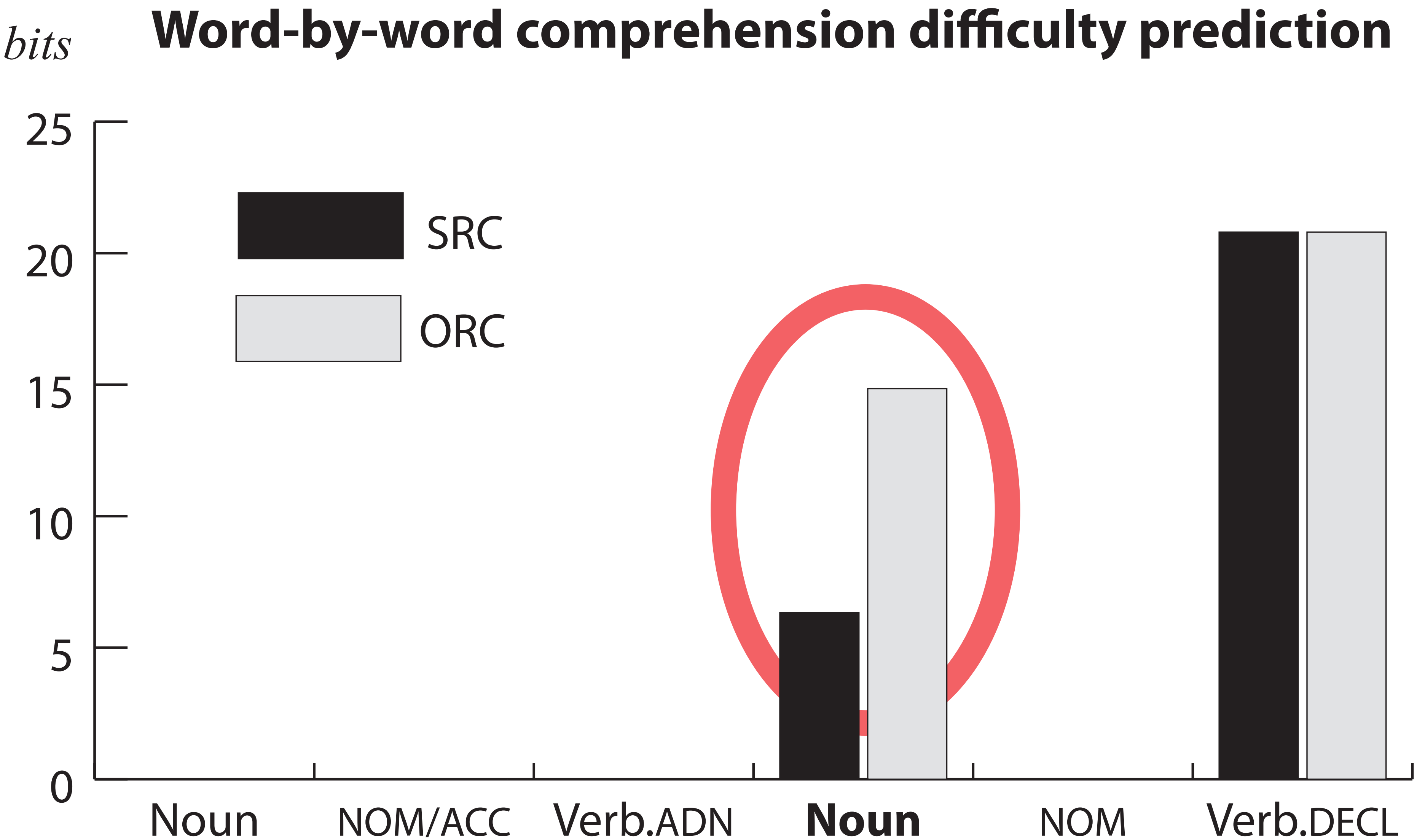


Dependency width doesn't derive it

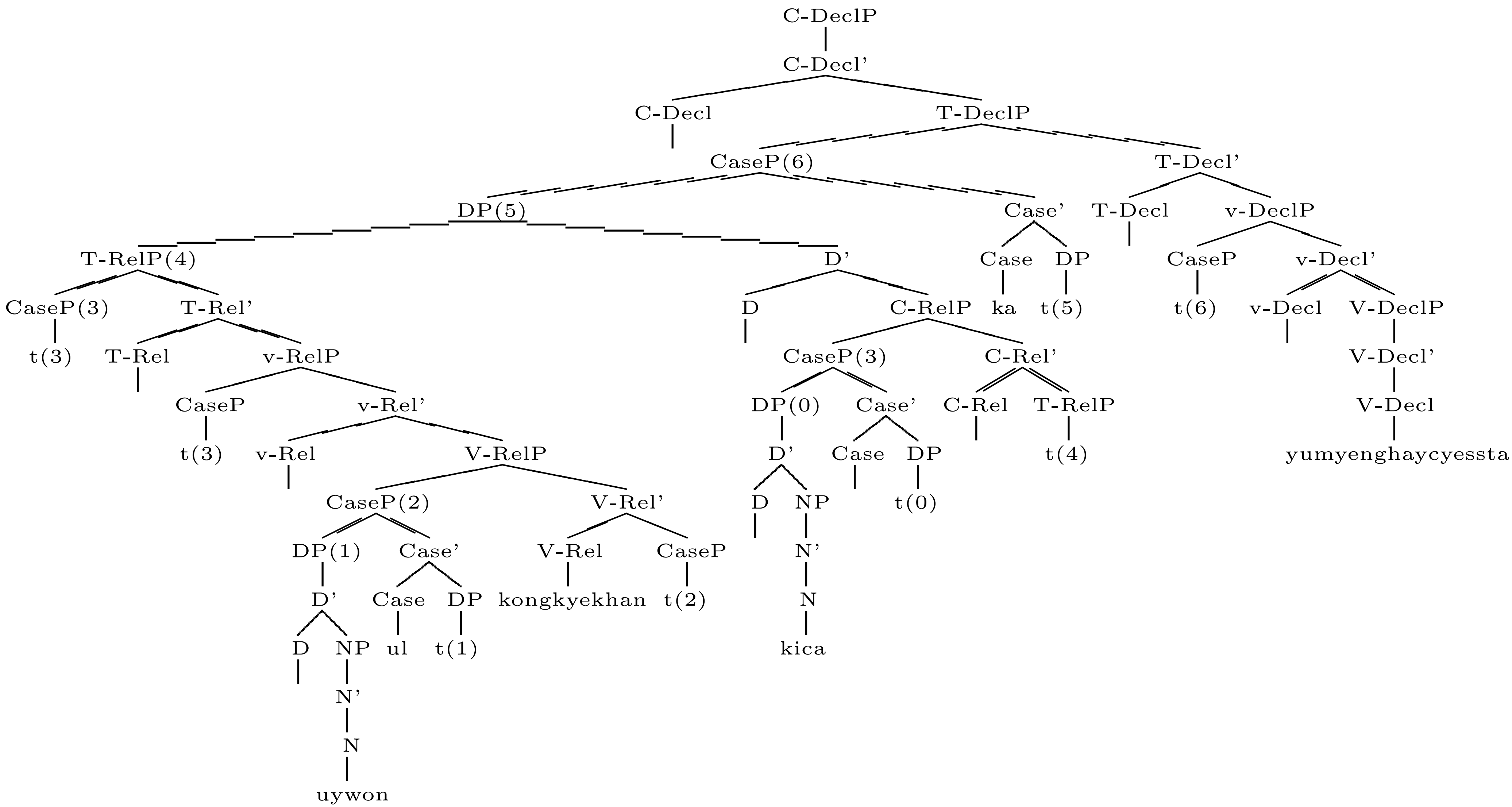
$$\textbf{SRC} \; \big[_{\mathrm{RC}} \; \emptyset \; \text{Object Verb} \; \big] \; \text{HeadNoun}$$

$$\textbf{ORC} \; \big[_{\mathrm{RC}} \; \text{Subject} \; \emptyset \; \text{Verb} \; \big] \; \text{HeadNoun}$$

ERH+MG *does* derive the SRC advantage

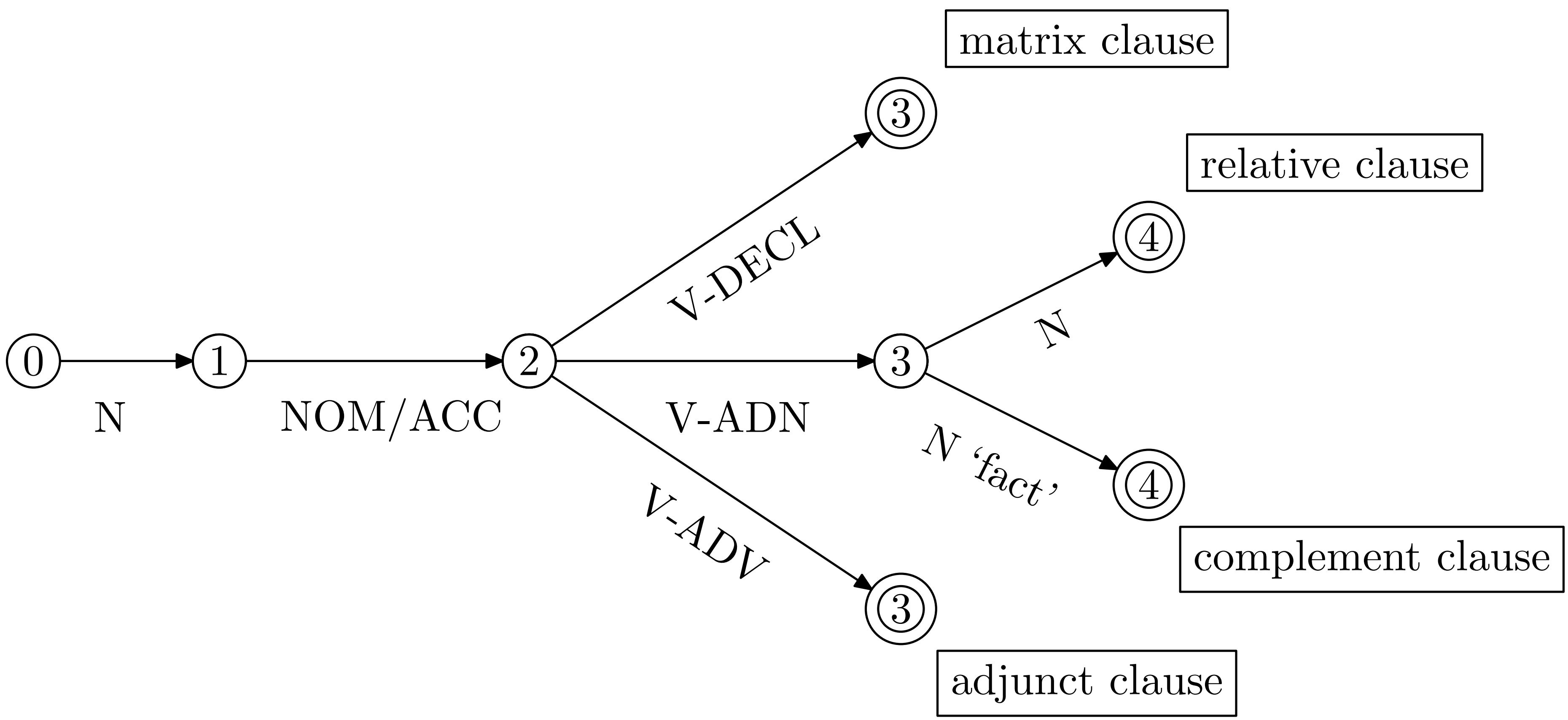


“The reporter who attacked the senator became famous”

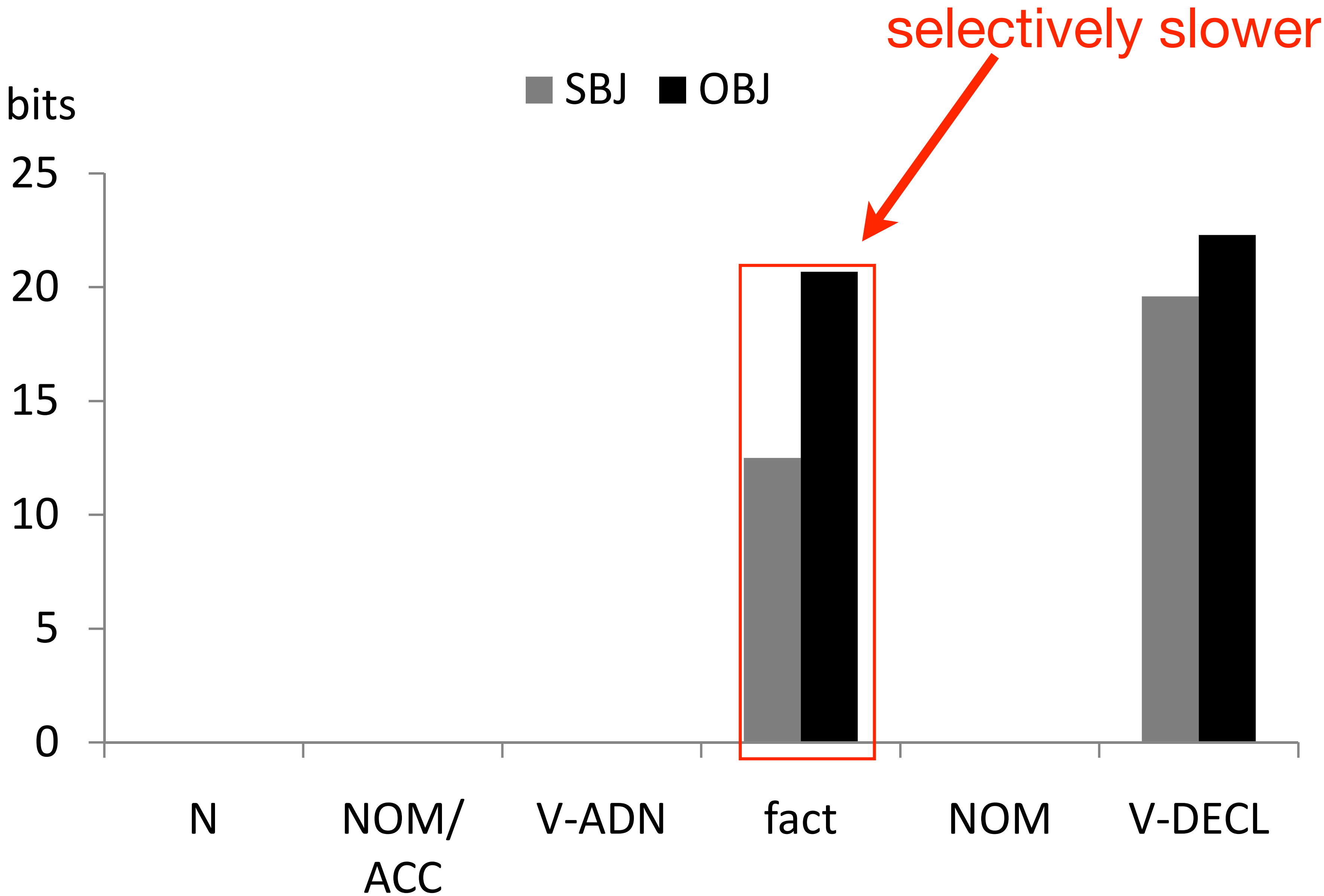


Yun, Whitman & Hale 10

One of four clause types

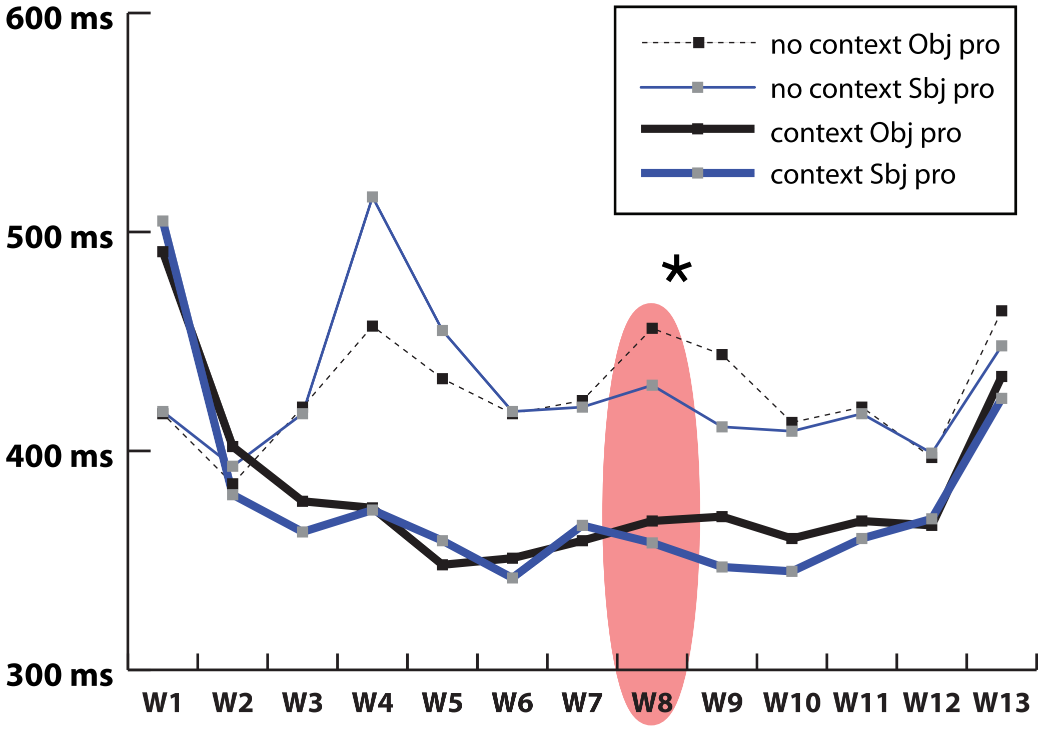


Novel prediction



(c) Complement Clause

Confirmed experimentally



object-“extracted” slower, $p < 0.007$

NCCNPs with subject *pro*

W1	W2		W3	W4	W5	W6	W7	**W8**	W9	**W10**	W11	W12	W13
지난	달	***pro***	편집장을	뇌물	수수	혐의로	협박한	**사실이**	밝혀지자	**총장은**	즉각	기자회견을	열었다.
Last	month	**he$_i$**	editor-ACC	bribe	taking	suspicion-with	threaten-ADN	**fact-NOM**	was.revealed-as	**chancellor$_i$-TOP**	immediately	press.conference-ACC	held

‘The chancellor$_i$ immediately held a press conference as the fact that he$_i$ threatened the editor for taking a bribe last month was revealed.’

NCCNPs with object *pro*

W1	W2	W3		W4	W5	W6	W7	**W8**	W9	**W10**	W11	W12	W13
지난	달	편집장이	***pro***	뇌물	수수	혐의로	협박한	**사실이**	밝혀지자	**총장은**	즉각	기자회견을	열었다.
Last	month	editor-NOM	**him$_i$**	bribe	taking	suspicion-with	threaten-ADN	**fact-NOM**	was.revealed-as	**chancellor$_i$-TOP**	immediately	press.conference-ACC	held

‘The chancellor$_i$ immediately held a press conference as the fact that the editor threatened him$_i$ for taking a bribe last month was revealed.’

Kwon, Yun et al CUNY2011

outline

Entropy reduction studies
relative clauses in English and Korean

How does it work?
computing $\downarrow H_i$

Why does it work?
reflections on information theory & linguistics

computing H_i

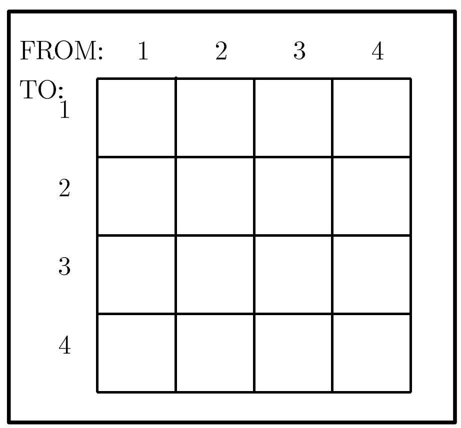
=c +nom agrD	ϵ	=i +rel c	ϵ
=>agrD droot	*the*	d	*I*
=d =d i	*met*	=n d -rel	*who*
n -nom	*boy*		

Minimalist Grammar *(or other formalism)*

```
S --> t157 [0,0;0,1;0,2] (* concatenation *)
t157 --> t0 t136 [0,0][0,1][0,2;1,0;1,1;1,2] (* r1' *)
t0 --> E t0_tmp2 [0,0][1,0][1,1]
t0_tmp2 --> t0_tmp1 E [0,0][1,0]
t0_tmp1 --> ""
t136 --> t135 [0,3;0,0][0,1][0,2] (* move' *)
t135 --> t42 t178 [0,0][1,1;0,1][0,2;1,0;1,2][1,3] (* r1left *)
t42 --> E t42_tmp2 [0,0][1,0][1,1]
t42_tmp2 --> t42_tmp1 E [0,0][1,0]
t42_tmp1 --> ""
t42_tmp1 --> "-ed"
t42_tmp1 --> "-s"
```

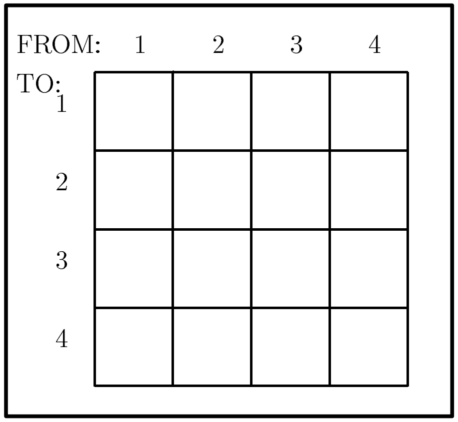
weighted MCFG G

input string w=$w_1 w_2 w_3$...w_i
prefix of a sentence in $L(G)$



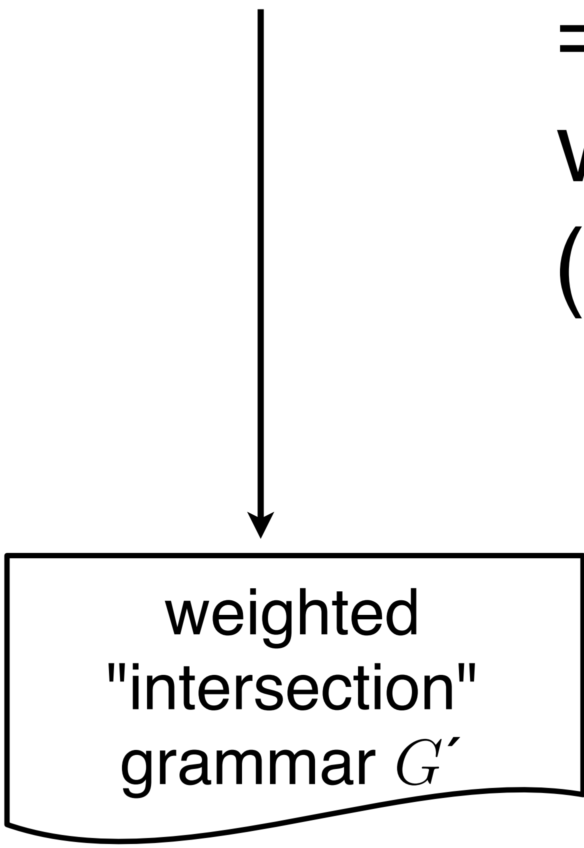
chart

computing H_i



chart's `items`
form a graph

= a system of equations,
whose solutions are
(sums of) probabilities



$$H(G') = H_i$$

entropy of probabilistic grammar

$\vec{h}$	vector of 1-step rewriting entropies
$\vec{H}$	vector of infinite step rewriting entropies
$\mathbf{M}$	"fertility matrix" giving expected number j symbols birthed by the i^{th} symbol

$$\vec{H} = \vec{h} + \mathbf{M}\vec{H}$$

Grenander 67

entropy of probabilistic grammar

$\vec{h}$	vector of 1-step rewriting entropies
$\vec{H}$	vector of infinite step rewriting entropies
$\mathbf{M}$	"fertility matrix" giving expected number j symbols birthed by the i^{th} symbol

$$\begin{aligned}\vec{H} &= \vec{h} + \mathbf{M}\vec{H} \\ \vec{h} &= \vec{H} - \mathbf{M}\vec{H} = (I - \mathbf{M})\vec{H}\end{aligned}$$

Grenander 67

entropy of probabilistic grammar

$\vec{h}$	vector of 1-step rewriting entropies
$\vec{H}$	vector of infinite step rewriting entropies
$\mathbf{M}$	"fertility matrix" giving expected number j symbols birthed by the i^{th} symbol
$\mathbf{I}$	the identity matrix with ones down the diagonal

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

Grenander 67

outline

Entropy reduction studies
relative clauses in English and Korean

How does it work?
computing $\downarrow H_i$

Why does it work?
reflections on information theory & linguistics

SCIENTIFIC AMERICAN

JULY 1949 VOL. 181, NO. 1

THE MATHEMATICS OF COMMUNICATION

An important new theory is based on the statistical character of language. In it the concept of entropy is closely linked with the concept of information

by Warren Weaver

Entropy reduction in 1953

(5) Non-linguistic considerations. We will assume that all of the 1152 sentences in this language are equiprobable, and that there are no dependencies *between* sentences in a string.

Application of entropy measurement

(1) Entropy of a single sentence: $H_S = \log_2 1152 = 10.17$.
(2) Entropy reduction of each phoneme, considered with respect to its position in the sentence is: $H_P = H_S - H_R$, where H_R is the entropy of the statements which are still possible after the transmission of phoneme P.

To illustrate: Consider the successive phonemes of the sentence abibibbabbi ‘see boy man not’ or ‘does not the boy see the man?’ (Free translation.)

Phoneme	Remarks	No. Possible Sentences Remaining	H	H_P
—	Any sentence possible before transmission begins	1152	10.17	—
a	Must be question with trans. vb., of which the no. of possibilities is	512	9.	1.17
b	Verb either *see* or *kill*	256	8.	1.0
a	Verb must be *see*	128	7.	1.0
b	N_s must be either *man, woman, boy* or *girl*	64	6.	1.0
i	Must be either *boy* or *girl*	32	5.	1.0
b	Must be *boy*	16	4.	1.0
b	N_o is *man, woman, boy,* or *girl*	8	3.	1.0
a	*Man* or *woman*	4	2.	1.0
b	*Man;* sentence either pos. or neg.	2	1.	1.0
b	Redundant; gives no information	2	1.	0.0
i	Sentence is negative	1	0.0	1.0
				10.17

(3) Entropy reduction of each morpheme, also considered in relation to its position in the sentence, is $H_M = H_S - H_R$ where H_S and H_R are as defined previously. The same sentence is used here as in the example above.

“Psycholinguistics” Sebeok and Osgood, eds.
§5.3 Applications of Entropy Measures to Problems of Sequence Structure

Review of General Psychology
2003, Vol. 7, No. 2, 183–188

1089-2680/03/$12.00 DOI: 10.1037/1089-2680.7.2.183

Whatever Happened to Information Theory in Psychology?

R. Duncan Luce
University of California, Irvine

"The elements of choice in information theory are absolutely neutral and lack any internal structure.
That is fine for a communication engineer ....[but]
by and large, however, the stimuli of psychological experiments are to some degree structured, and so, in a fundamental way, they are not in any sense interchangeable."

April, 2000

THE ROYAL SOCIETY

Formal grammar and information theory: together again?

By Fernando Pereira

"Probabilities can be assigned to complex linguistic events, even novel ones, by using the causal *structure of the underlying models* to propagate the uncertainty in the elementary decisions."

⇒ these models incorporate linguistic theories!

Conclusions

Information theory helps identify
which RCs are hard *where*

the account uses substantial syntactic claims

the difference since 1953 is the grammar

bonus slides

definitions

Let G be a (probabilistic) grammar, X a random variable whose outcomes x are derivations on G, and Y a related variable whose outcomes y are initial substring of sentences in $L(G)$.

mutual information of grammar and prefix string

$$I(X;Y) \quad = \quad H(X) - H(X|Y)$$

information conveyed by a particular prefix

$$I(X;y) \quad = \quad H(X) - H(X|y)$$

definitions

Let G be a (probabilistic) grammar, X a random variable whose outcomes x are derivations on G, and Y a related variable whose outcomes y are initial substring of sentences in $L(G)$.

mutual information of grammar and prefix string

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(X) - \mathbb{E}\left[H(X|y)\right] \end{aligned}$$

information conveyed by a particular prefix

$$I(X;y) = H(X) - H(X|y)$$

definitions

Let G be a (probabilistic) grammar, X a random variable whose outcomes x are derivations on G, and Y a related variable whose outcomes y are initial substring of sentences in $L(G)$.

info conveyed by increment

$$I(X; y_{\text{new}}) - I(X; y_{\text{old}})$$

definitions

Let G be a (probabilistic) grammar, X a random variable whose outcomes x are derivations on G, and Y a related variable whose outcomes y are initial substring of sentences in $L(G)$.

info conveyed by increment

$$
\begin{aligned}
& I(X; y_{\text{new}}) - I(X; y_{\text{old}}) \\
= \quad & \Big(H(X) - H(X|y_{\text{new}})\Big) - \Big(H(X) - H(X|y_{\text{old}})\Big)
\end{aligned}
$$

definitions

Let G be a (probabilistic) grammar, X a random variable whose outcomes x are derivations on G, and Y a related variable whose outcomes y are initial substring of sentences in $L(G)$.

info conveyed by increment

$$
\begin{aligned}
& I(X; y_{\text{new}}) - I(X; y_{\text{old}}) \\
&= \Big(H(X) - H(X|y_{\text{new}})\Big) - \Big(H(X) - H(X|y_{\text{old}})\Big) \\
&= H(X|y_{\text{old}}) - H(X|y_{\text{new}})
\end{aligned}
$$

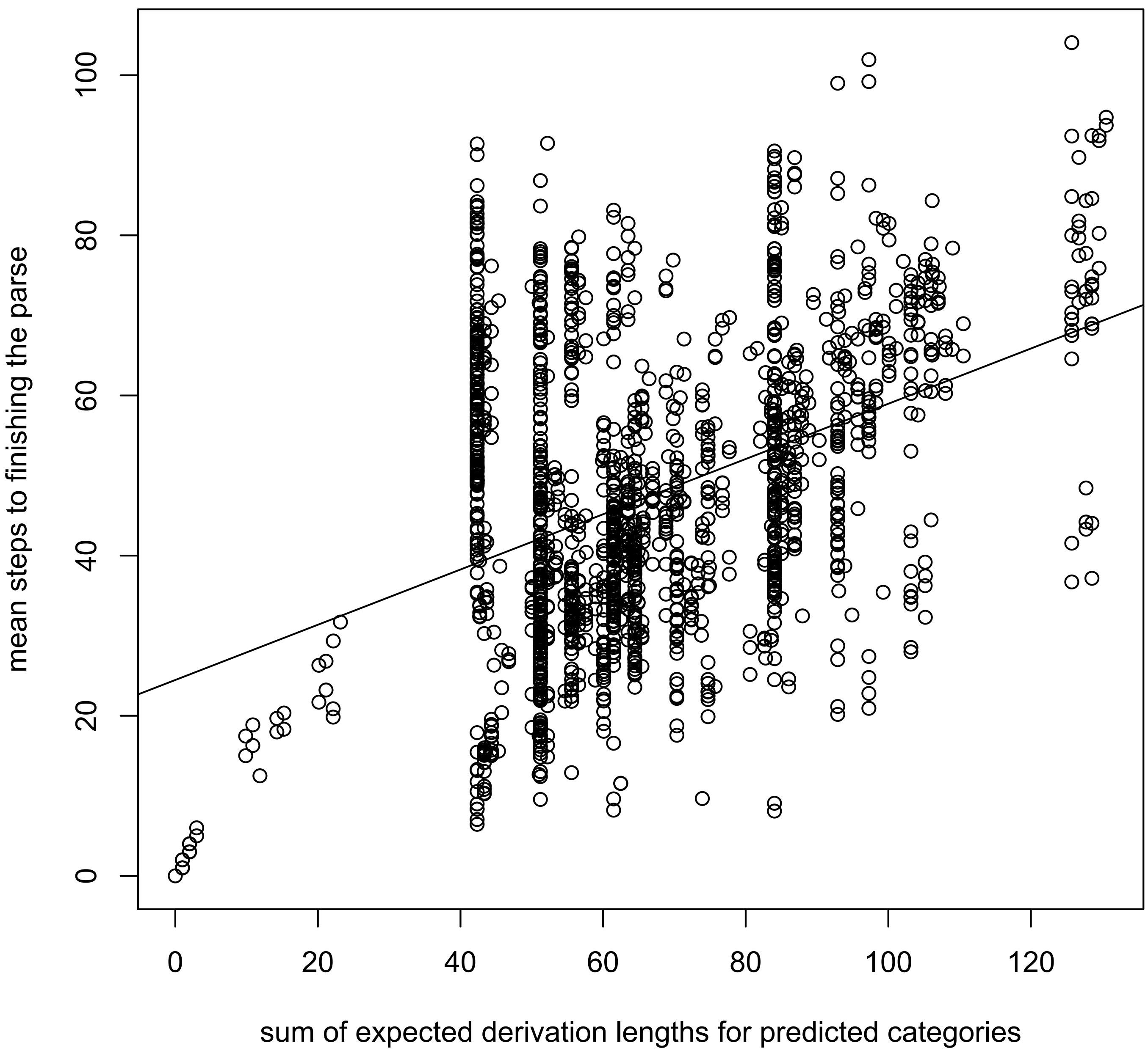
definitions

Let G be a (probabilistic) grammar, X a random variable whose outcomes x are derivations on G, and Y a related variable whose outcomes y are initial substring of sentences in $L(G)$.

info conveyed by increment

$$
\begin{aligned}
& I(X; y_{\text{new}}) - I(X; y_{\text{old}}) \\
&= \Big(H(X) - H(X|y_{\text{new}})\Big) - \Big(H(X) - H(X|y_{\text{old}})\Big) \\
&= H(X|y_{\text{old}}) - H(X|y_{\text{new}}) \\
&= \downarrow H_i \qquad \text{where } y_{\text{old}} = w_1 \cdots w_{i-1}, \\
& \qquad\qquad\qquad\quad\; y_{\text{new}} = w_1 \cdots w_i
\end{aligned}
$$

the best heuristic tracks uncertainty about the rest of the sentence



$r = 0.4,$
$p < 0.0001$

Hale 2011