Probabilistic Models of Human Sentence Processing

Frank Keller

School of Informatics
University of Edinburgh
keller@inf.ed.ac.uk

IPAM Workshop on Probabilistic Models of Cognition
January 24–28, 2005

Overview of the Talk

1. Models of Sentence Processing
   - Properties of the Human Parser
   - Probabilistic Grammars
   - A Probabilistic Model of Human Parsing
2. Models of Text Processing
   - From Sentence to Text
   - Entropy and Sentence Length
   - Entropy in Context
   - Entropy out of Context

Part I

Models of Sentence Processing

The Garden Path View of Human Parsing

Parsing: extracting syntactic structure from a string; prerequisite for assigning a meaning to the string.

Structures are built *incrementally* (word by word) as the input comes in (Tanenhaus et al. 1995), which leads to *local ambiguity.*

Example:

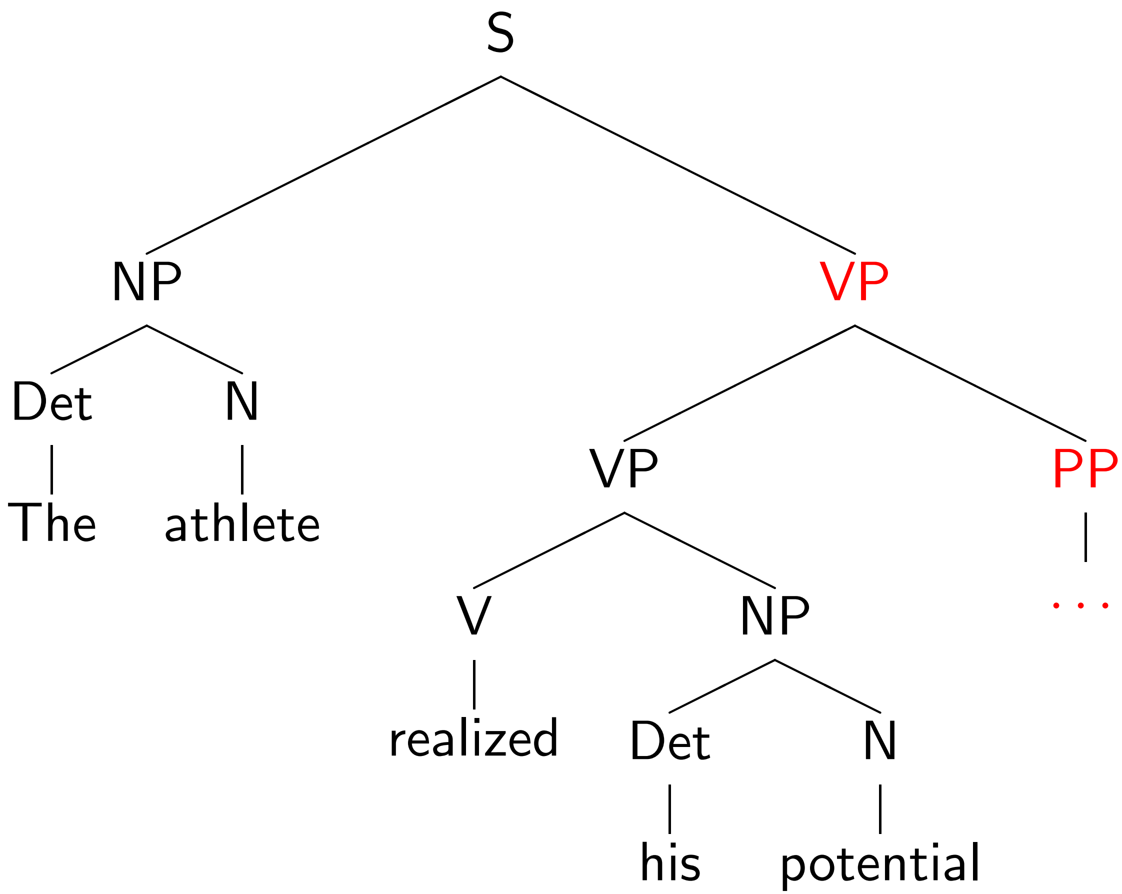
(1) The athlete realized his potential . . .

a. . . . at the competition.

b. . . . would make him a world-class sprinter.

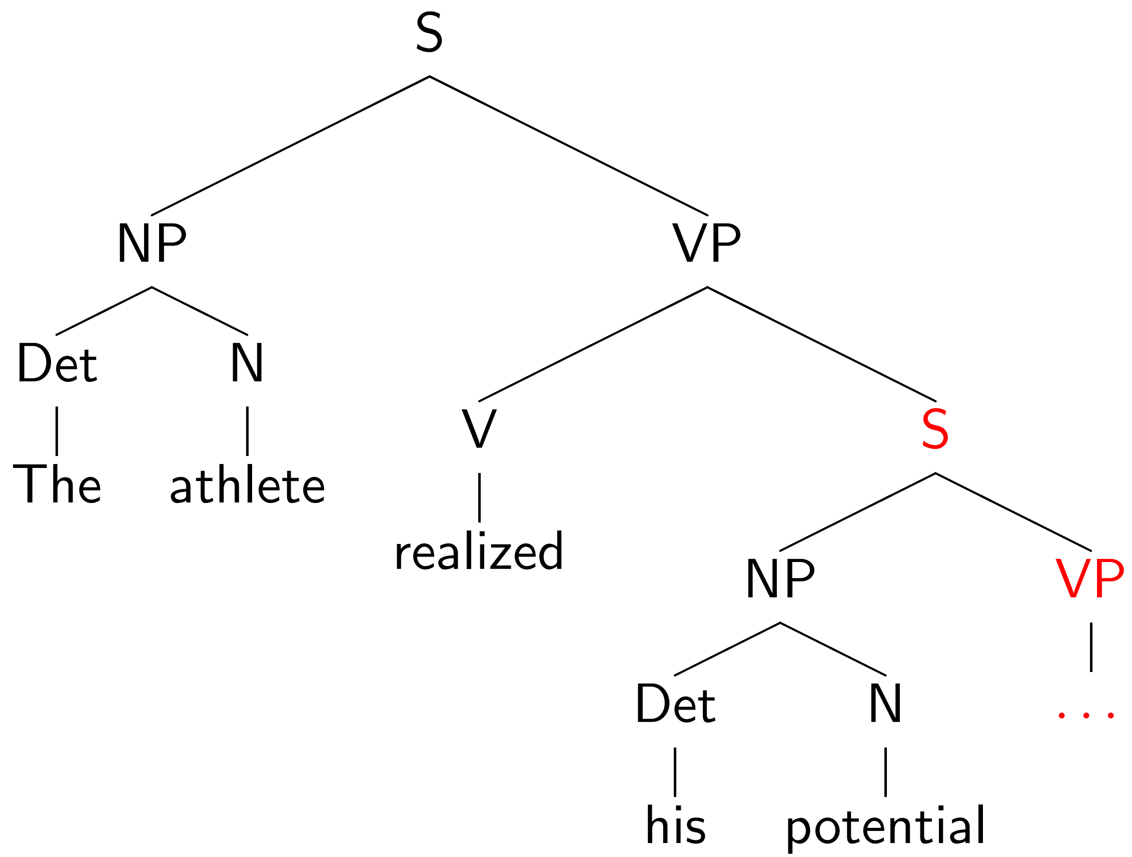
The Garden Path View of Human Parsing

Structure 1 (NP reading):



The Garden Path View of Human Parsing

Structure 2 (S reading):



The Garden Path View of Human Parsing

- *Early commitment:* when it reaches *potential*, the processor has to decide which structure to build.
- If the parser makes the wrong choice (e.g., NP reading for sentence (1-b)) it needs to backtrack and revise the structure.
- A *garden path* occurs, which typically results in longer reading times (and reverse eye-movements).
- Garden paths traditionally the main object of study in psycholinguistics:
  - determine experimentally under which conditions they occur;
  - draw conclusions about the architecture of the human parser;
  - build models that explain garden pathing.

Garden Path vs. Garden Variety

Limitations of current models of human parsing:

- Most models deals with *processing breakdown* (garden paths as main object of study).
- However, processing breakdown is exceedingly rare in naturally occurring speech and text.
- Under normal conditions, human parsing is *extremely robust:*
  - *accurate:* recovers the correct interpretation;
  - *coverage:* deals with most types of sentences, including ungrammatical and noisy input;
  - *efficient:* processes utterances in real-time, incrementally.

Challenge: model *garden variety* parsing, i.e., parsing that occurs naturally and doesn't lead to processing breakdown.

Modeling Robustness in Human Parsing

Goal: our models must account for, and explain:

- processing difficulty in specific circumstances;
- robustness in general.

Probabilistic approach: the processor computes $\hat{t}$, the most probable parse for a sentence S:

$$\hat{t} = \arg\max_t P(t, S)$$

Estimate $P(t, S)$ using *probabilistic grammars.*

Probabilistic Context-Free Grammars

- A PCFG is a standard CFG where each grammar rule $N \rightarrow \zeta$ is annotated with a probability $P(N \rightarrow \zeta)$;
- the probabilities of all rules with the same lefthand side sum to one:

$$\sum_j P(N \rightarrow \zeta^j) = 1$$

- the probability of a parse t is the product of the probabilities of all rules applied in that parse:

$$P(t) = \prod_{(N \rightarrow \zeta) \in t} P(N \rightarrow \zeta)$$

Probabilistic Context-Free Grammars

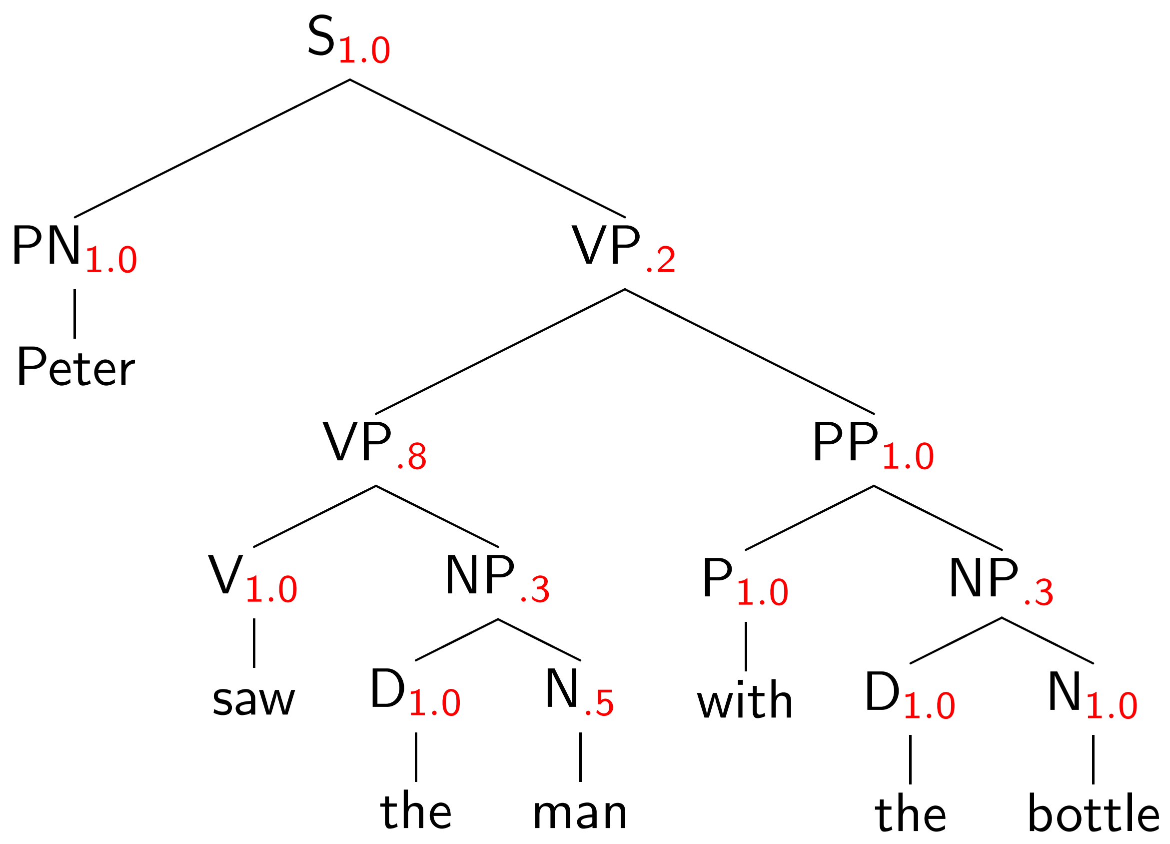
Example for an ambiguous sentences:

(2) Peter saw the man with the bottle.

Grammar that generates two readings for this sentence:

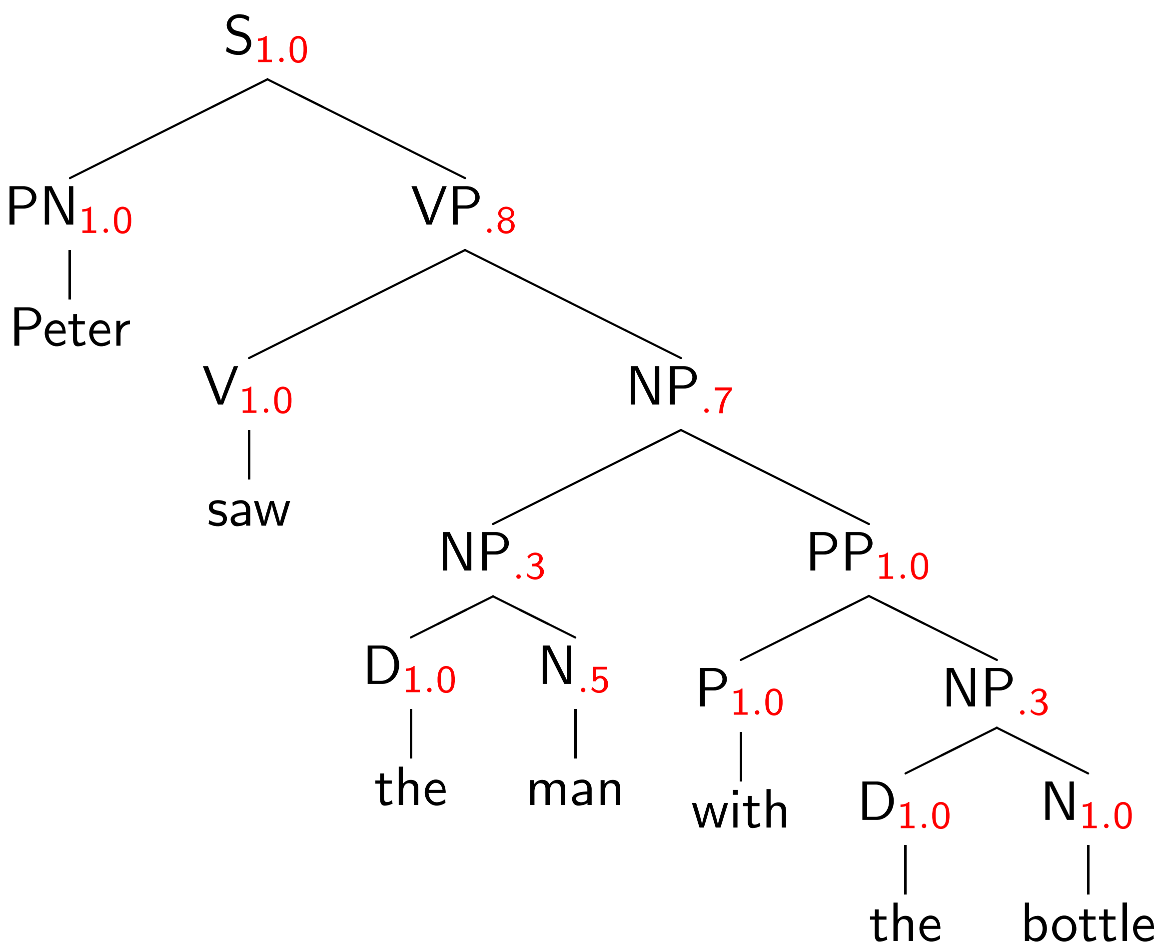
S	$\rightarrow$	PN VP	1.0	PN	$\rightarrow$	Peter	1.0
VP	$\rightarrow$	V NP	.8	V	$\rightarrow$	saw	1.0
VP	$\rightarrow$	VP PP	.2	D	$\rightarrow$	the	1.0
NP	$\rightarrow$	NP PP	.7	N	$\rightarrow$	man	.5
NP	$\rightarrow$	D N	.3	N	$\rightarrow$	bottle	.5
PP	$\rightarrow$	P NP	1.0	P	$\rightarrow$	with	1.0

Probabilistic Context-Free Grammars



$$P(t) = 1.0 \cdot 1.0 \cdot .2 \cdot .8 \cdot 1.0 \cdot 1.0 \cdot .3 \cdot 1.0 \cdot .3 \cdot 1.0 \cdot .5 \cdot 1.0 \cdot 1.0 = .0072$$

Probabilistic Context-Free Grammars



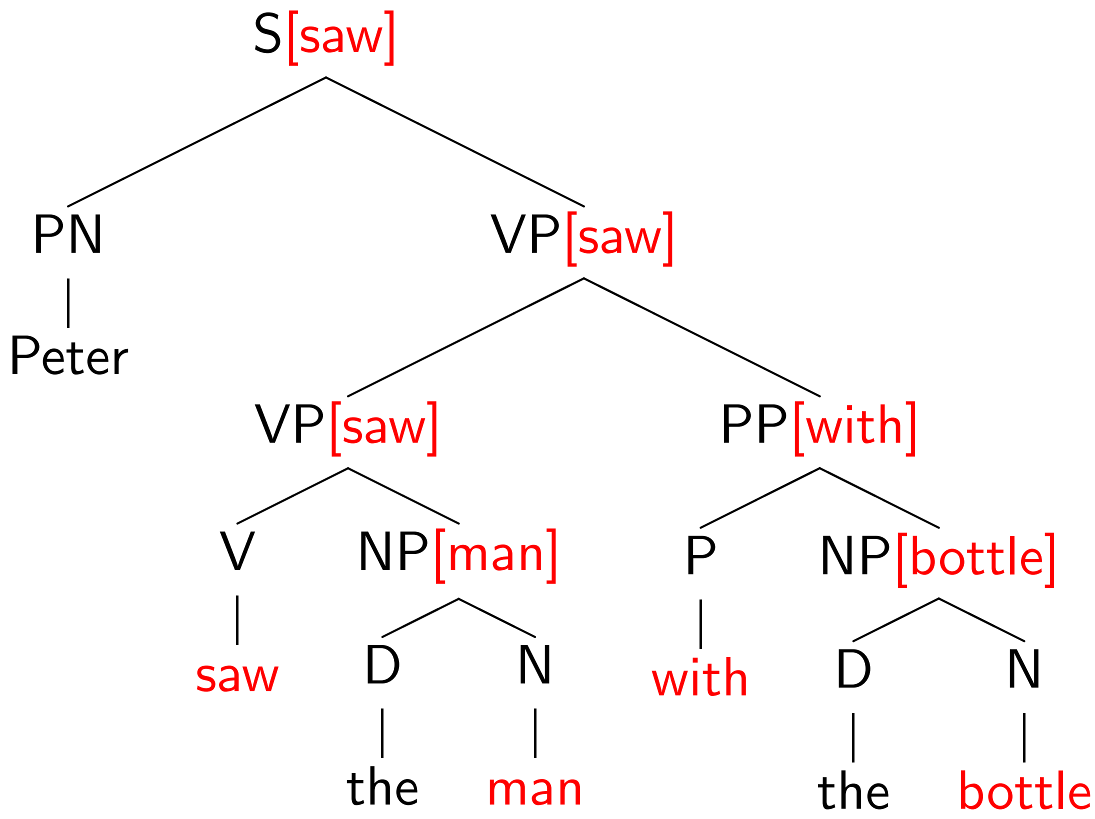
$$P(t) = 1.0 \cdot 1.0 \cdot .8 \cdot 1.0 \cdot .7 \cdot .3 \cdot 1.0 \cdot 1.0 \cdot .5 \cdot 1.0 \cdot .3 \cdot 1.0 \cdot 1.0 = .0252$$

Lexicalized PCFGs

Incorporate lexical information into the grammar:

- project lexical items along the head projection in the tree;
- augment grammar rules accordingly;
- approximates fine-grained syntactic (e.g., agreement) and semantic information (e.g., selectional restrictions);
- lexicalization can dramatically improve parsing accuracy (Charniak 1997, 2000; Collins 1997);
- convincing evidence that the human parser makes use of lexical information (Trueswell 1996; MacDonald et al. 1994).

Lexicalized PCFGs



New rule format: S[saw] $\rightarrow$ PN VP[saw]

Towards Robust Models of Human Parsing

Aim: build and evaluate models of human parsing that perform well on garden variety text:

- train probabilistic grammars on a syntactically annotated corpus;
- compare models that differ in the amount of information they extract from the corpus (unlexicalized vs. lexicalized);
- evaluate model fit against eye-movement data.

How do we evaluate model fit?

Evaluating the Robustness of a Model

Current practice in evaluation:

- take a small set of hand-picked examples; test if model makes the right predictions for these sentences; (e.g., Jurafsky 1996; Crocker and Brants 2000).
- take a small set of experimental conditions; test if the model produces the same pattern of results (e.g., Narayanan and Jurafsky 2002).

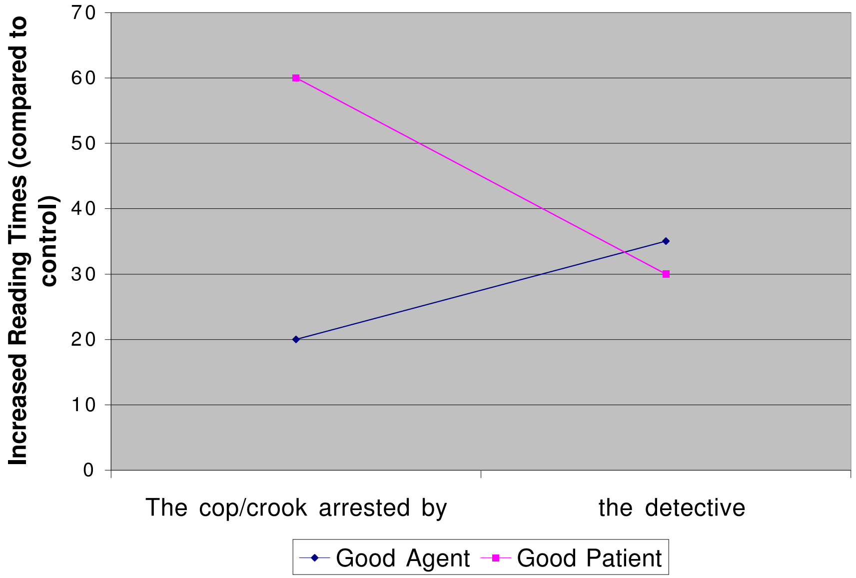
Instead, we should test if the model accounts for the *robustness* of the human parser:

Test *quantitative* predictions on a *random sample* of realistic sentence material.

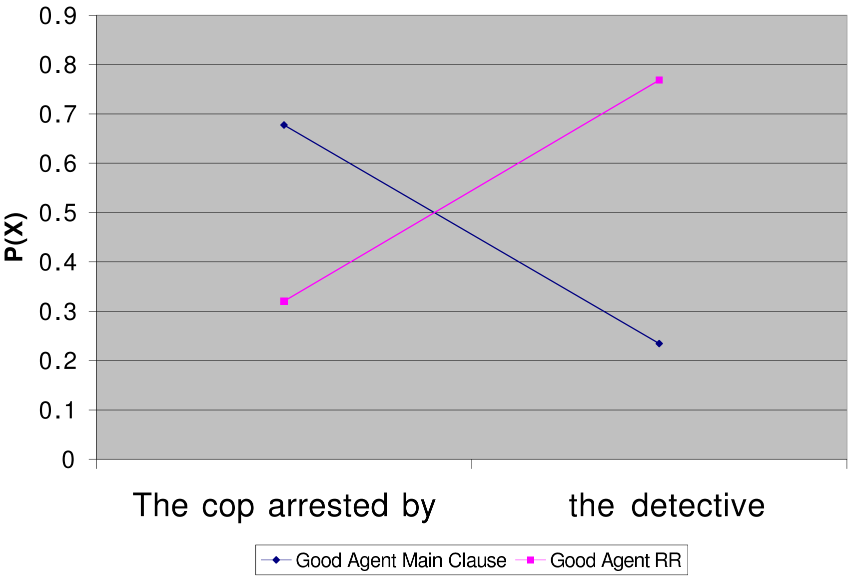
Evaluating the Robustness of a Model

Typical example (Narayanan and Jurafsky 2002):

Subjects:



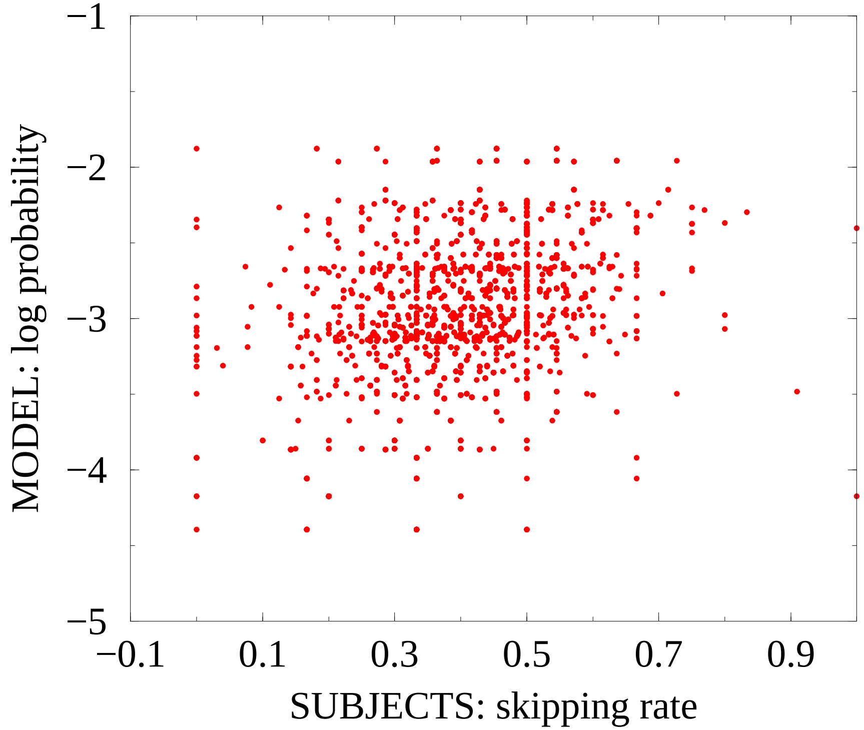
Model:



Qualitative evaluation on four experimental conditions; model produces roughly the same pattern as the experiment.

Evaluating the Robustness of a Model

Alternative: evaluation on large sample of sentences:



Regression analysis provides quantitative comparison of model predictions and experimental measures.

Training Data

Parsing models trained on the same training set: Penn Treebank (Marcus et al. 1993):

- approximately 1 million words of newspaper text (from the Wall Street Journal);
- manually part of speech tagged and annotated with phrase structure trees;
- standard division into training set (ca. 50,000 sentences), and development and test set (ca. 2,000 sentences each).
- train an unlexicalized and a lexicalized PCFG (Charniak 2000); rule probabilities obtained using maximum likelihood estimation.

Test Data

Embra Corpus (McDonald and Shillcock 2003):

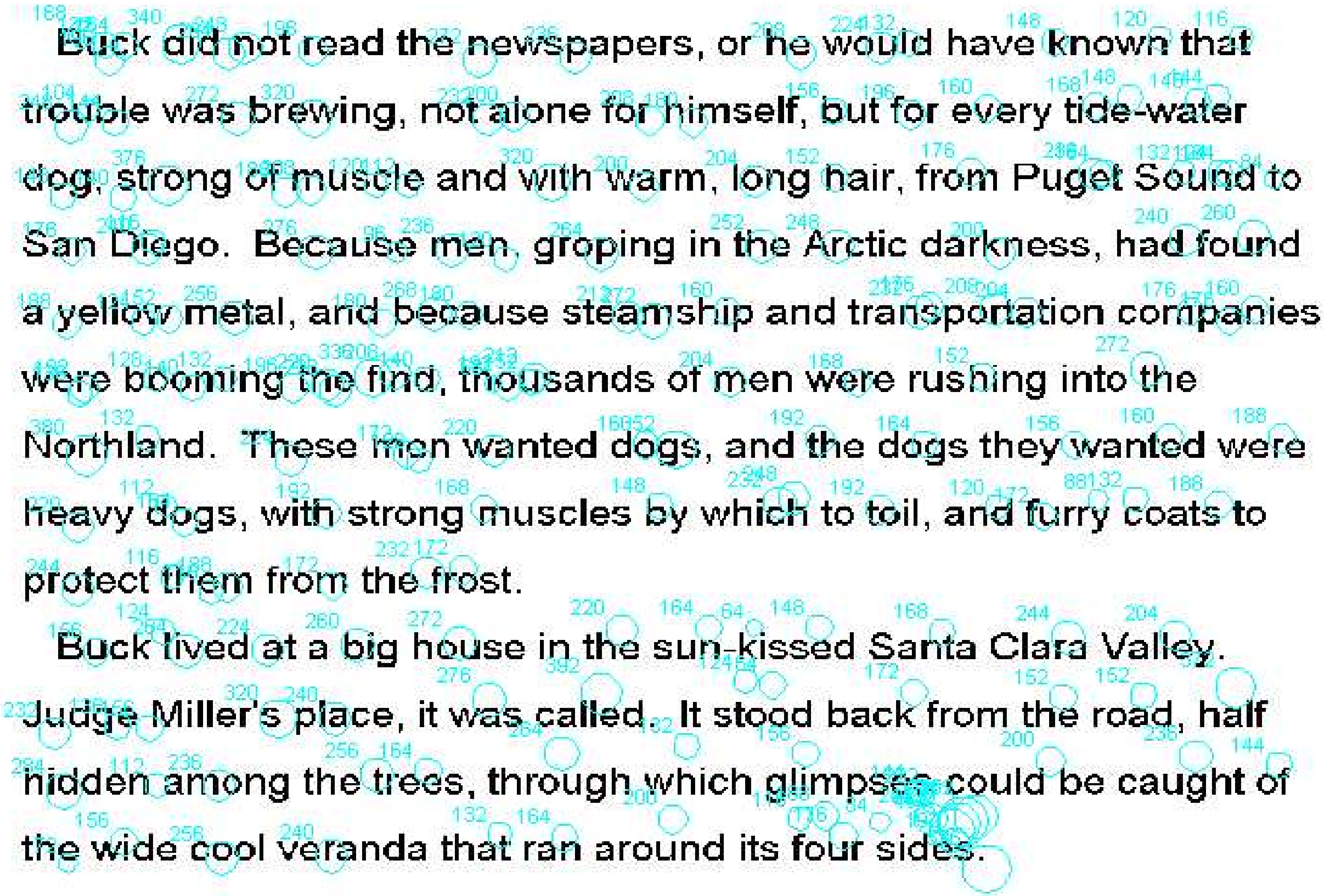
- material: 10 articles from UK broadsheet newspapers, wide range of topics;
- 97–405 words per article, 2,262 words in total;
- subjects: 23 native speakers of British English;
- texts presented on computer screen, 65 characters per line, 10 lines, 23 pages in total;
- eye-movements recorded using Dual Purkinje Image eye-tracker;
- comprehension question presented after each article.

Interlude: Eye-tracking Experiments

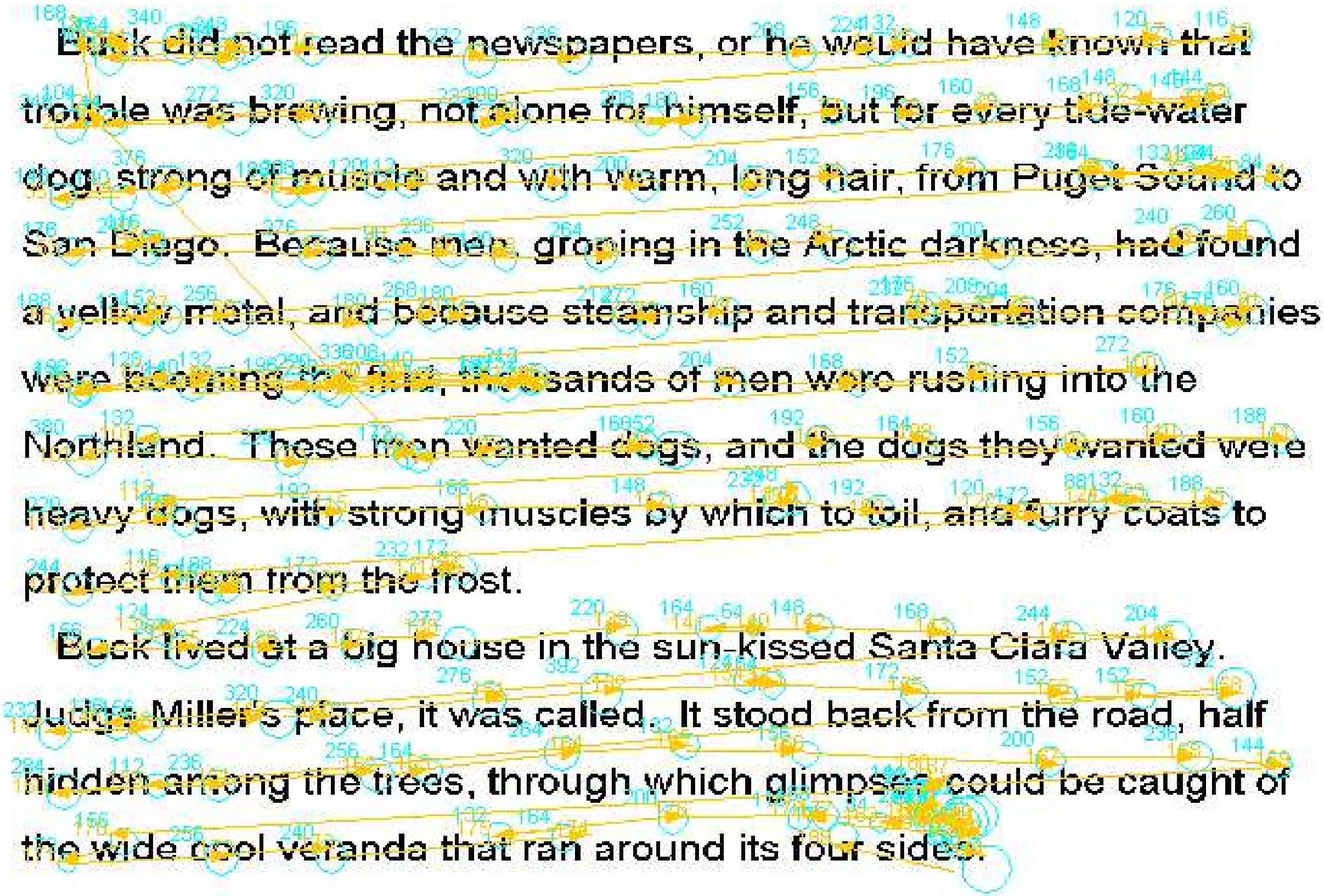
Buck did not read the newspapers, or he would have known that trouble was brewing, not alone for himself, but for every tide-water dog, strong of muscle and with warm, long hair, from Puget Sound to San Diego. Because men, groping in the Arctic darkness, had found a yellow metal, and because steamship and transportation companies were booming the find, thousands of men were rushing into the Northland. These men wanted dogs, and the dogs they wanted were heavy dogs, with strong muscles by which to toil, and furry coats to protect them from the frost.

Buck lived at a big house in the sun-kissed Santa Clara Valley. Judge Miller's place, it was called. It stood back from the road, half hidden among the trees, through which glimpses could be caught of the wide cool veranda that ran around its four sides.

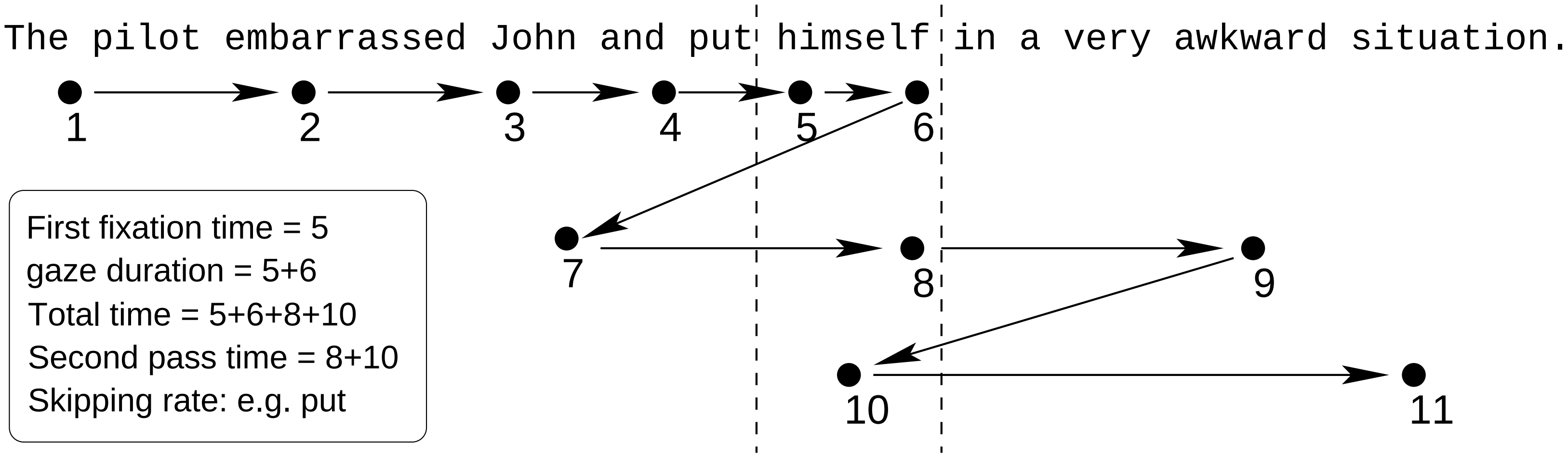
Interlude: Eye-tracking Experiments



Interlude: Eye-tracking Experiments



Interlude: Eye-tracking Experiments



- *Early measures* (first fixation time, gaze duration, skipping rate) are informative about early processes;
- *Later measures* (total time, second pass time) tell you more about processes that occur after some delay.

Modeling Methodology

Test models against experimental data (unseen test set):

- parse the experimental stimulus;
- compute the probability of best parse;
- correlate probability with reading time.

Note: all reading time measures are *by-sentence means* (e.g., mean first fixation time for all words in a sentence).

Sentence probabilities normalized in the same way (i.e., per-word means are computed).

Modeling Methodology

All analyses computed using regression analyses including the following *predictors:*

- mean word length;
- mean word frequency (logs);
- mean sentence probability (logs).

Word length and word frequency are known to correlate with reading times.

By including them in the regression we can assess if sentence probability is an *independent* predictor of reading measures.

All analyses conducted based on Lorch and Myers's (1990) recommendations for *repeated measures regression.*

Modeling Results

Baseline model: word frequency

Correlations (r)

	first	skip	gaze	total	length
length	$.365^{***}$	$-.436^{***}$	$.335^{***}$	$.279^{***}$	
freq	$-.227^{*}$	$.303^{**}$	$-.197$	$-.169$	$-.556$

Significant independent predictors in the regression equation:
$^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$
Word frequency sig. indep. predictor for early measures.

Modeling Results

Parsing model: unlexicalized PCFG

Correlations (r)

	first	skip	gaze	total	length	freq
length	$.365^{***}$	$-.436^{***}$	$.335^{***}$	$.279^{***}$		
freq	$-.227^{*}$	$.303^{**}$	$-.197^{*}$	$-.169^{*}$	$-.556$	
prob	$-.091$	$.142$	$-.059^{**}$	$-.052^{*}$	$-.229$	$.685$

Unlexicalized PCFG probability sig. indep. predictor for later measures.

Modeling Results

Parsing model: lexicalized PCFG (Charniak 2000)

Correlations (r)

	first	skip	gaze	total	length	freq
length	.365***	−.436***	.335***	.279***		
freq	−.227	.303*	−.197	−.169	−.556	
prob	−.260***	.289*	−.245***	−.207***	−.519	.576

Lexicalized PCFG probability sig. indep. predictor for early and later measures.

Discussion

- Word frequency sig. predictor of early measures (first fixation, skipping rate):
  - ⇒ well-know effect from the eye-tracking literature;
- Unlexicalized PCFG probabilities sig. indep. predictors of later measures (gaze duration, total time):
  - ⇒ models syntactic information that goes beyond word frequency and influences later processing stages;
- Lexicalized PCFG probability sig. indep. predictor of both early and late measures; higher correlation than unlex. PCFG:
  - ⇒ combines lexical and syntactic information; models syntactic information more accurately than PCFG.

Summary

- Robust probabilistic parsing models were tested as models of human parsing;
- these models share the accuracy, broad coverage, and efficiency of the human parser;
- all models predict garden variety eye-tracking data, but differ as to whether they predict early or later measures;
- the most successful parsing model was a lexicalized PCFG, which is able predict both early and later measures.

Part II

Models of Text Processing

From Sentence to Text

- We have successfully modeled the processing of *individual sentences* using probabilistic models.
- Can the probabilistic approach be extended to text, i.e., connected *sequences of sentences?*
- Use notions from *information theory* to formalize the relationship between processing effort for sentences and processing effort for text.

From Sentence to Text

Entropy Rate Principle

Speakers produce language whose entropy rate is on average constant (Genzel and Charniak 2002, 2003; G&C).

Motivation:

- information theory: most efficient way of transmitting information through a noisy channel is at a constant rate;
- if human communication has evolved to be optimal, then humans produce text and speech with constant entropy;
- some evidence for speech (Aylett 1999).

Entropy Rate Principle

Applying the Entropy Rate Principle (ERP) to text:

- entropy is constant, but the amount of context available to the speaker increases with increasing sentence position;
- *prediction:* if we measure entropy out of context (i.e., based on the probability of isolated sentences), then entropy should increase with sentence position;
- G&C show that this is true for both function and content words, and for a range of languages and genres;
- entropy can be estimated using a language model or a probabilistic parser.

Entropy Rate Principle

Sentences in context:

1. a a a a a a a
2. b b b b b b b
3. c c c c c c c
4. d d d d d d d
5. e e e e e e e

Sentences out of context:

a a a a a a a $H = 0.4$

e e e e e e e $H = 0.7$

Predictions for Human Language Processing

Out-of-context predictions

- out-of-context entropy increases with sentence position; tested extensively by G&C (replicated in Exp. 1);
- out-of-context processing effort increases with sentence position;
- reading time as an index of processing effort;
- *prediction:* out of context reading time correlated with sentence position (Exp. 3).

Predictions for Human Language Processing

In-context predictions

- in-context entropy on average the same for all sentences;
- *prediction:* in-context reading time not correlated with sentence position (Exp. 2).
- processing effort increases with entropy;
- reading time as an index of processing effort;
- *prediction:* reading time correlated with entropy (Exp. 2).

Experiment 1: Method

Replication of G&C's results:

- use Wall Street Journal corpus (1M words), divided into training and test set;
- treat each article as a separate text; compute sentence position by counting from beginning of text (1–149).
- compute per-word entropy computed using an *n*-gram language model:

$$\hat{H}(X) = -\frac{1}{|X|} \sum_{x_i \in X} \log P(x_i | x_{i-(n-1)} \dots x_{i-1})$$

Extension of G&C's results:

- correlation on raw data or on binned data (avg. by position);
- baseline model: sentence length $|X|$.

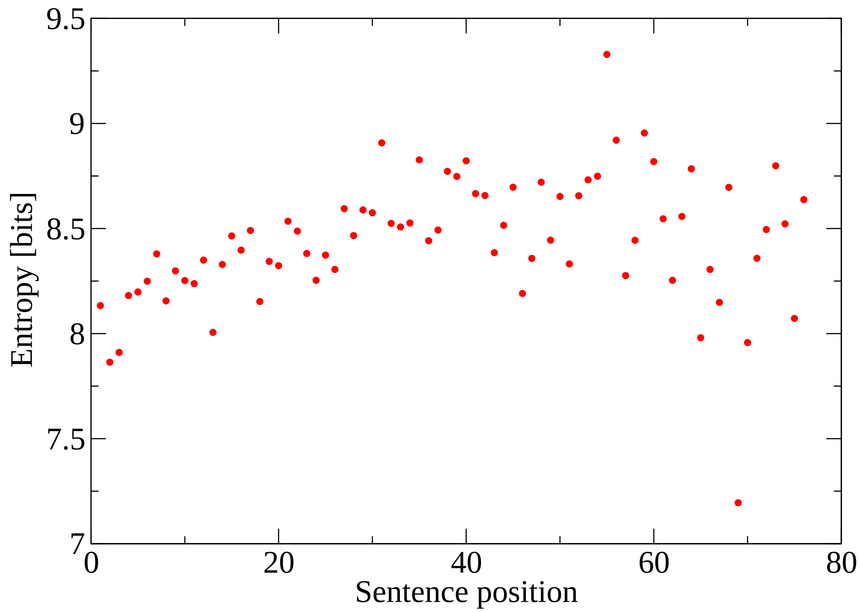
Results

Correlation of sentence entropy and sentence position ($c = 25$):

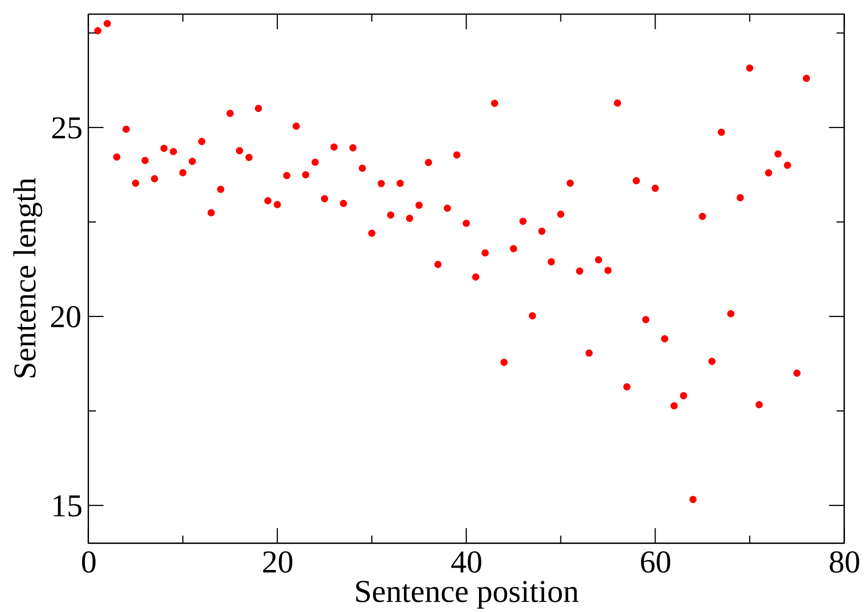
	Binned data	Raw data
Entropy 3-gram	0.6387^{**}	0.0598^{**}
Sentence length	-0.4607^{*}	-0.0635^{**}

Significance level: $^{*}p < .05$, $^{**}p < .01$

Results



Correlation of sentence entropy and sentence pos. (binned data)



Correlation of sentence length and sentence pos. (binned data)

Results

We need to disconfound entropy and sentence length.

Compute correlation of entropy and sentence length with sentence position, with the other factor partialled out ($c = 25$):

Binned data	Binned data	Raw data
Entropy 3-gram	0.6708**	0.0784**
Sentence length	−0.7435**	−0.0983**

Discussion

- Results of Exp. 1 confirm G&C's main finding: entropy increases with sentence position;
- *however:* sign. negative correlation between sentence position and sentence length: longer sentences tend to occur earlier in the text;
- further analyses show that entropy rate is a *significant independent predictor,* even if sentence length is controlled for;
- G&Cs effect holds even without binning: important for evaluation against human data (binning not allowed).

Aims of Experiment 2

This experiment tests the psycholinguistic predictions of the ERP *in context:*

- entropy predicted to correlate with processing effort;
- test this using a corpus of newspaper text annotated with eye-tracking data;
- eye-tracking measures of reading time reflect processing effort for words and sentences;
- sentences position predicted not to correlate with processing effort for in-context sentences.

Method

- *Test set:* Embra eye-tracking corpus (McDonald and Shillcock 2003); 2,262 words of text from UK newspapers;
- regression used to control confounding factors: word length, word frequency (Lorch and Myers 1990);
- *training and development set:* broadsheet newspaper section of the BNC; training: 6.7M words, development: 0.7M words;
- sentence position: 1–24 in test set, 1–206 in development set;
- entropy computed as in Experiment 1;

Results

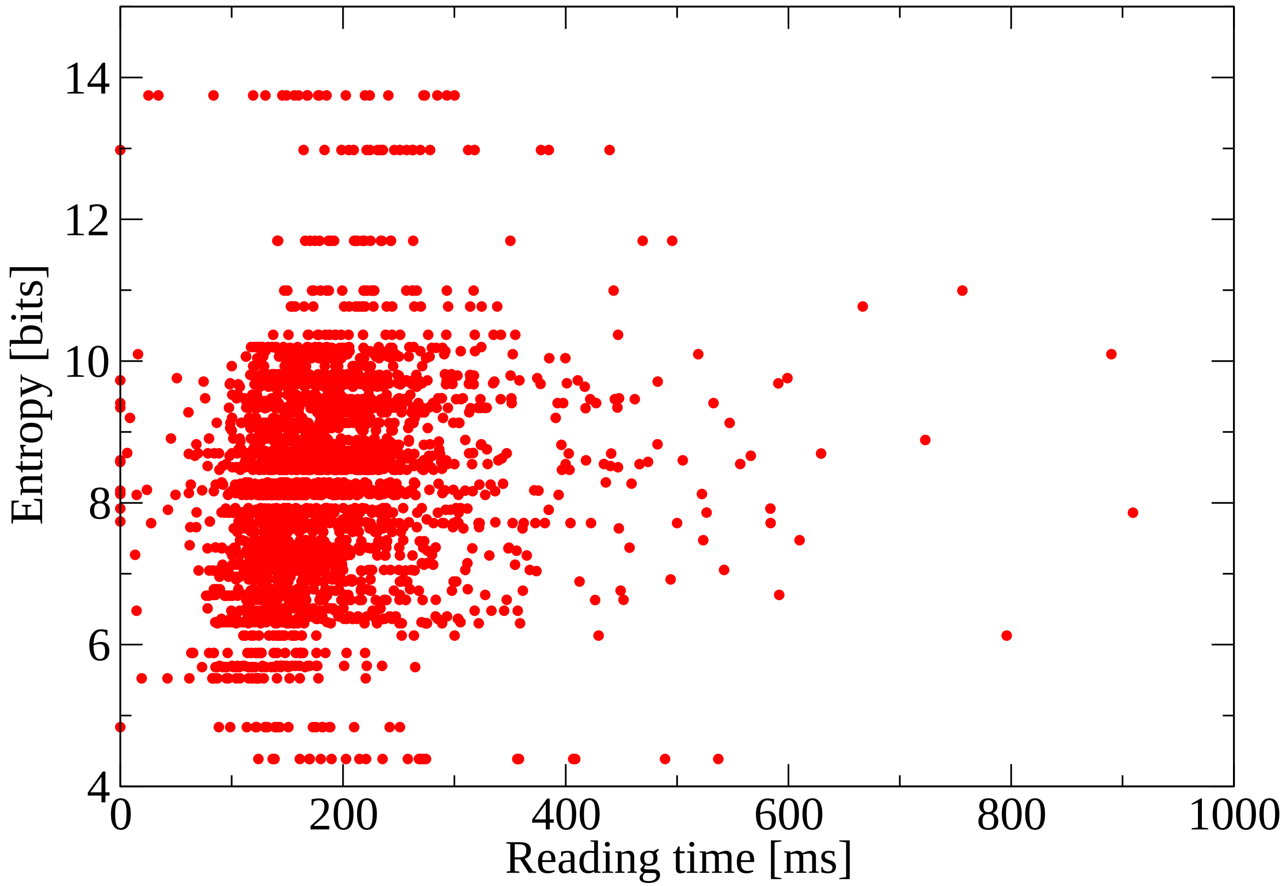
Correlation of entropy and position on the Embra corpus:

	Binned data	Raw data
Entropy 3-gram	-0.5512^{**}	-0.1674
Sentence length	0.3902	0.0885

Correlation of reading times with entropy and sentence position:

Entropy 3-gram	0.1646^{**}
Sentence position	-0.0266

Results



Correlation of reading time and sentence position

Discussion

- Results on Embra corpus show sign. correlations between entropy and sentence position;
- sign. correlation between entropy and reading time (with word length and frequency partialled out);
- confirms ERP assumption: *sentences with higher entropy are harder to process;*
- no sign. correlation between position and reading time;
- confirms ERP prediction: *entropy constant in connected text.*

Aims of Experiment 3

This experiment further investigates the psycholinguistic predictions of the ERP *out of context:*

- entropy predicted to correlate with processing effort;
- test this using out-of-context sentences;
- self-paced reading time reflects processing effort for words and sentences;
- sentences position predicted to correlate with processing effort for out-of-context sentences.

Method

- 60 sentences samples randomly from Embra corpus; 5 sentences each for positions 1–12;
- sentences presented out of context in random order; 24 filler sentences interspersed;
- 32 native speakers read the sentences using a word-by-word self-paced reading paradigm;
- measure of processing effort: total reading time for a sentence, normalized by sentence length;
- regression used to control confounding factors: word length, word frequency (as in Exp. 2);
- entropy computed as in Exp. 1 and 2;

Results

Correlation of entropy and position in the stimulus set:

	Binned data	Raw data
Entropy 3-gram	0.1201	−0.0366
Sentence length	−0.1023	−0.0464

Correlation of reading times with entropy and sentence position:

Entropy 3-gram	0.0523
Sentence position	0.0504**

Discussion

- Sign. correlation between sentence position and reading time for sentences presented out of context;
- confirms ERP prediction: *out-of-context entropy increases with sentence position.*
- however: no sign. correlation between entropy and sentence position; no sign. correlation between entropy and reading time;
- probably due to small data set compared to Exp. 2.

Summary

- Probabilistic models extended from sentence to text using *entropy rate principle;*
- confirmed *in-context* predictions of ERP using reading time data for connected text:
  - ⇒ correlation between entropy and processing effort (i.e., reading time);
  - ⇒ no correlation between position and processing effort;
- confirmed *out-of-context* predictions of ERP using reading time data for isolated sentences:
  - ⇒ correlation between sentence position and processing effort.

Future Work

Sentence Processing Models

- *Incrementality:* test parsing models on a word-by-word basis; requires an incremental parser (and probability model);
- *Measures:* replace sentence probability with more realistic measure of processing difficulty (probability ratio, Jurafsky 1996, or entropy, Hale 2003);
- *Garden paths:* show that the model not only works for garden variety, but also for garden path sentences.

Text Processing Models

- *Integration:* combine with sentence processing models; use prob. grammars instead of language models.

References

Aylett, Matthew. 1999. Stochastic suprasegmentals: Relationships between redundancy, prosodic structure and syllabic duration. In *Proceedings of the 14th International Congress of Phonetic Sciences*. San Francisco.

Charniak, Eugene. 1997. Statistical parsing with a context-free grammar and word statistics. In *Proceedings of the 14th National Conference on Artificial Intelligence*. AAAI Press, Cambridge, MA, pages 598–603.

Charniak, Eugene. 2000. A maximum-entropy-inspired parser. In *Proceedings of the 1st Conference of the North American Chapter of the Association for Computational Linguistics*. Seattle, WA, pages 132–139.

Collins, Michael. 1997. Three generative, lexicalised models for statistical parsing. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Conference of the European Chapter of the Association for Computational Linguistics*. Madrid, pages 16–23.

Crocker, Matthew W. and Thorsten Brants. 2000. Wide-coverage probabilistic sentence processing. *Journal of Psycholinguistic Research* 29(6):647–669.

Genzel, Dmitriy and Eugene Charniak. 2002. Entropy rate constancy in text. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, pages 199–206.

Genzel, Dmitriy and Eugene Charniak. 2003. Variation of entropy and parse trees of sentences as a function of the sentence number. In Michael Collins and Mark Steedman, editors, *Proceedings of the Conference on Empirical Methods in Natural*