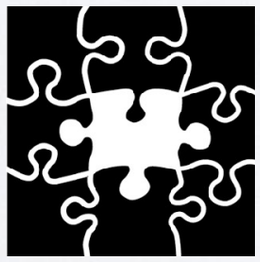
LTG Seminar 14 March 2022

Efficient Strategies of Language Production: An Information-Theoretic Analysis

Mario Giulianelli

Institute for Logic, Language and Computation
University of Amsterdam

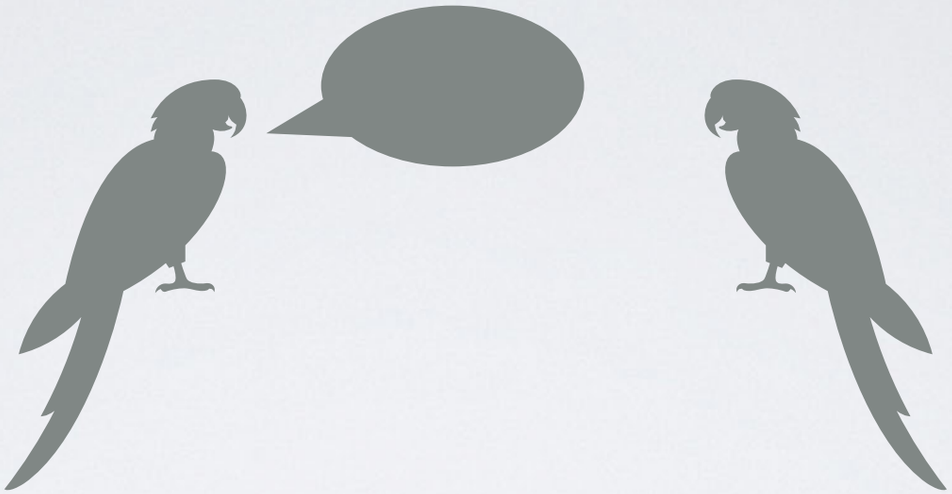
m.giulianelli@uva.nl | glnmario.github.io



Efficient Strategies of Language Production: An Information-Theoretic Analysis

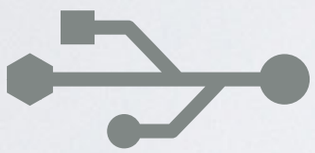
- Three studies, with Raquel Fernández and Arabella Sinclair
- An information-theoretic model of language production
- A hypothesis: humans follow efficient strategies of information transmission, leading to near-optimal collaborative effort
- Utterance surprisal (information density) as a proxy of effort, estimated with an autoregressive transformer language model

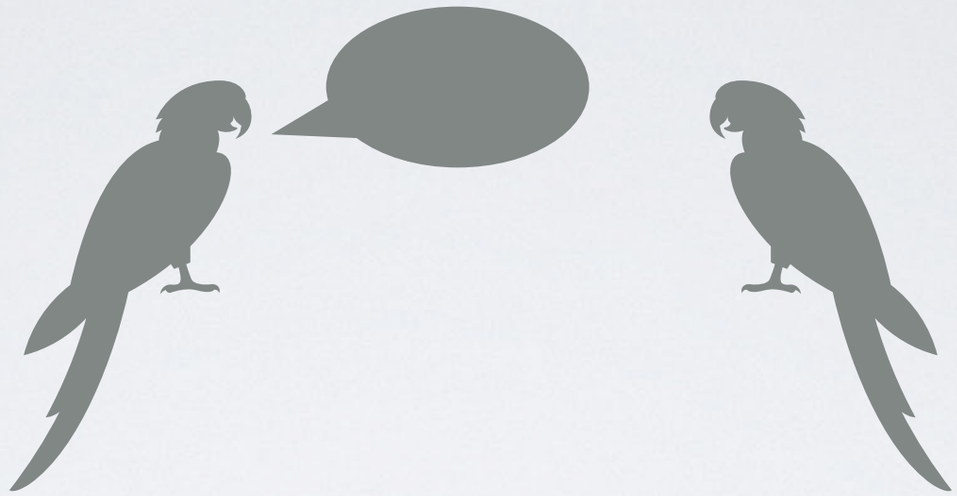
Collaborative effort in language production



Collaborative effort in language production

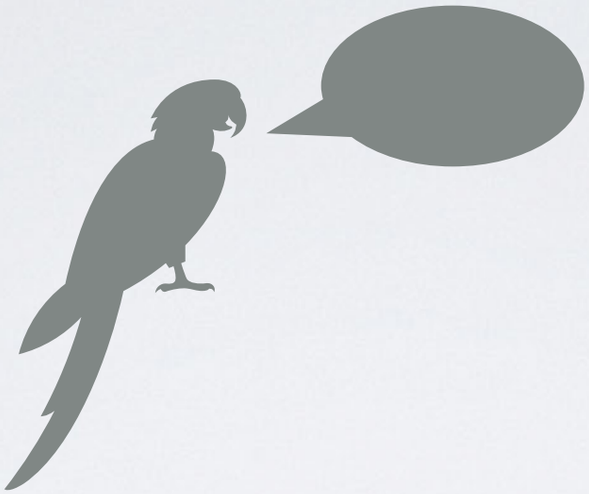
PRODUCTION
EFFORT





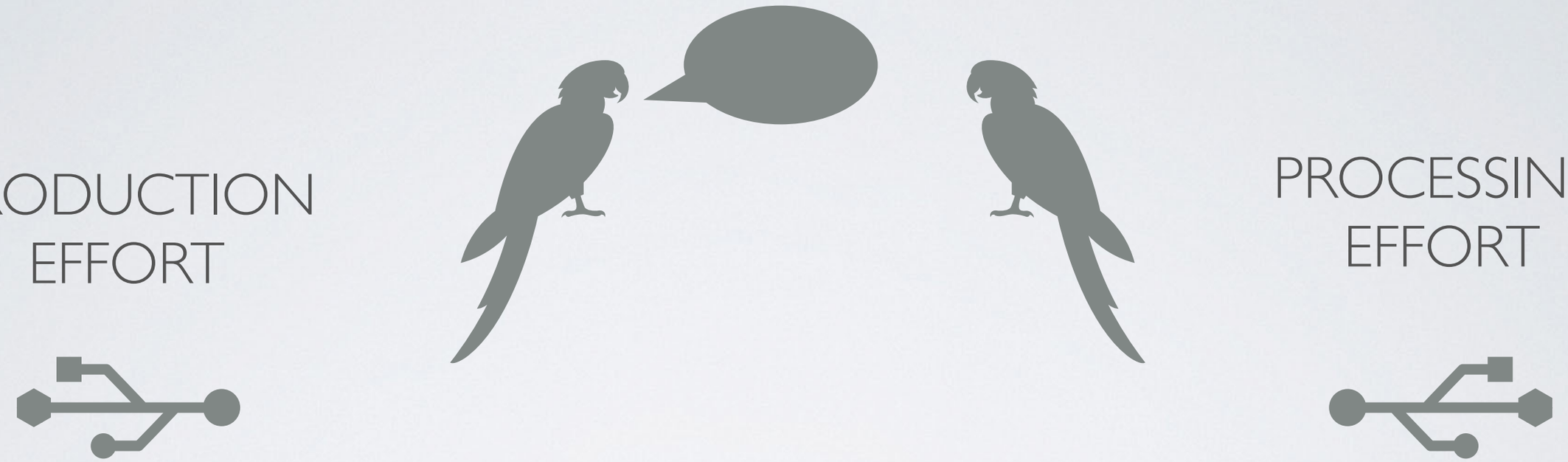
Collaborative effort in language production

PRODUCTION EFFORT



PROCESSING EFFORT

Collaborative effort in language production



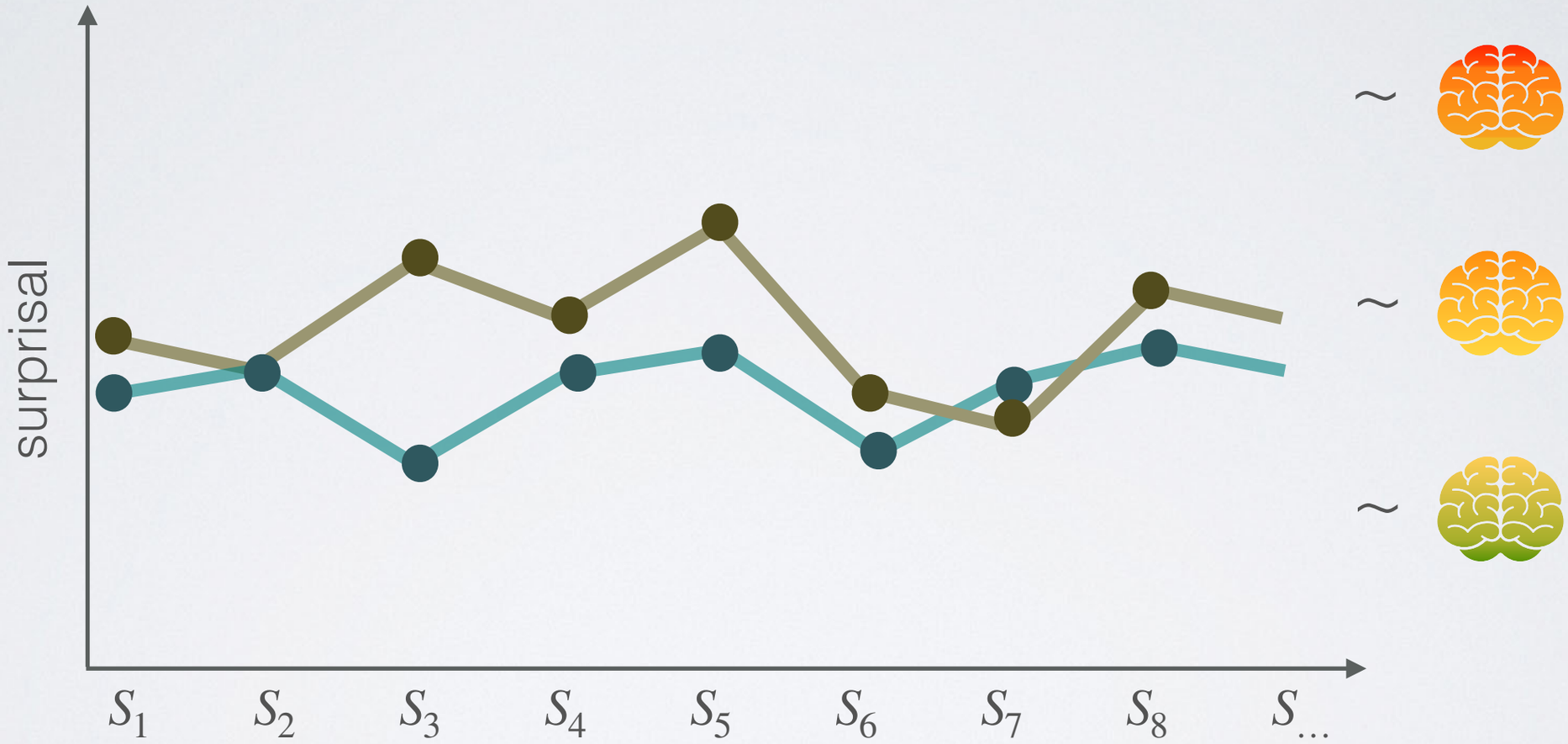
PRODUCTION EFFORT

PROCESSING EFFORT

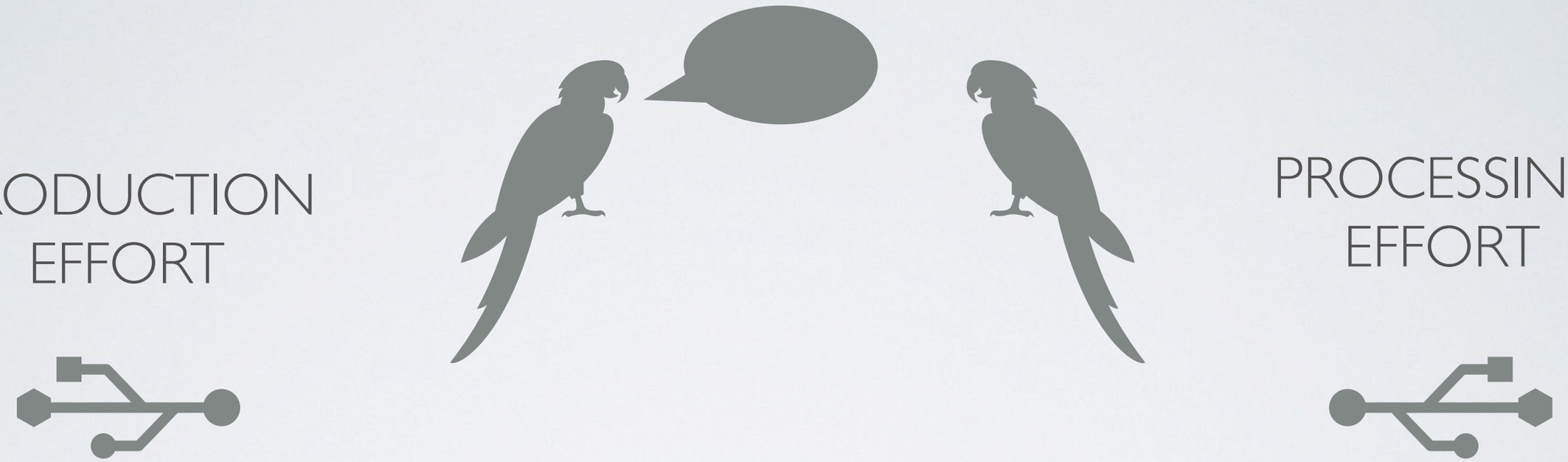
Unpredictability

Surprisal

Collaborative effort in language production



Collaborative effort in language production



PRODUCTION EFFORT

PROCESSING EFFORT

Unpredictability

Surprisal

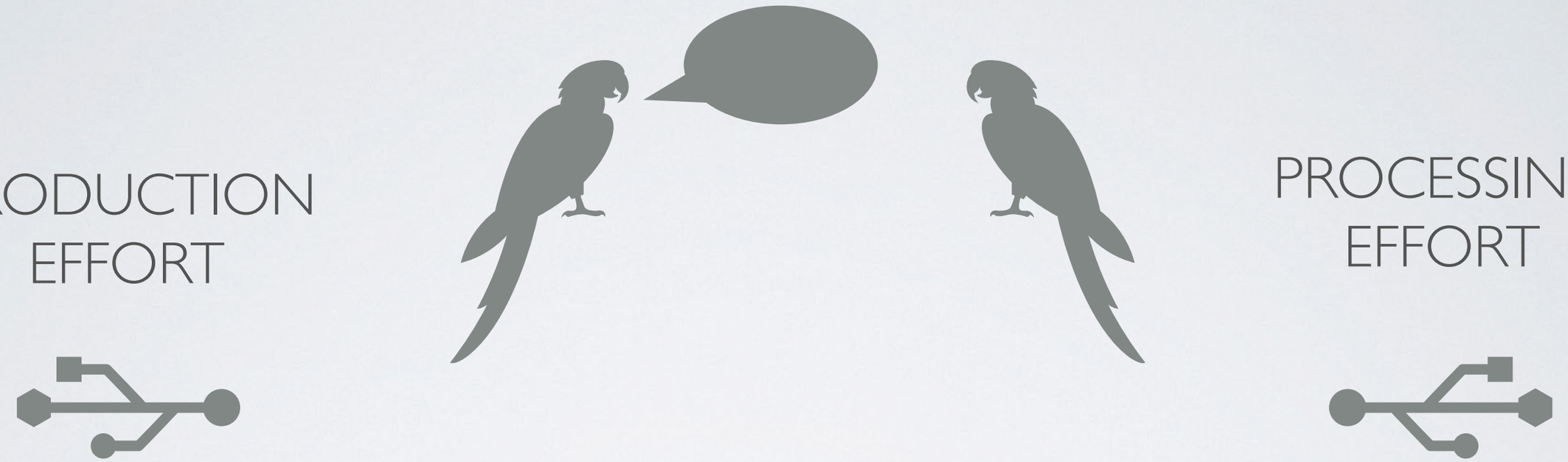
phoneme duration, syntactic structures, turn taking

perception, reading, sentence interpretation

Aylett and Turk, 2004, 2006; Bell et al., 2003; Demberg et al., 2012; Jaeger and Levy, 2007; Frank and Jaeger, 2008; Jaeger, 2010; Dethlefs et al., 2016

Jelinek et al., 1975; Clayards et al., 2008; Keller, 2004; Demberg and Keller, 2008; Levy et al., 2009; Levy, 2008; Gibson et al., 2013

Collaborative effort in language production



PRODUCTION EFFORT

PROCESSING EFFORT

Unpredictability

Surprisal

Information content

Information density

Information Theory (Shannon, 1948)

Information-theoretic measures

Information-theoretic measures

$H(S)$ Out-of-context surprisal

$$-\log_2 P(S)$$

Amcore Financial Inc. said it agreed to acquire Central of Illinois Inc. in a stock swap.

Shareholders of Central, a bank holding company based in Sterling, will receive Amcore stock equal to 10 times Central's 1989 earnings, Amcore said.

For the first nine months of 1989, Central earned $2 million.

Information-theoretic measures

$H(S \mid C)$ In-context surprisal

$$-\log_2 P(S \mid C)$$

Amcore Financial Inc. said it agreed to acquire Central of Illinois Inc. in a stock swap.

Shareholders of Central, a bank holding company based in Sterling, will receive Amcore stock equal to 10 times Central's 1989 earnings, Amcore said.

For the first nine months of 1989, Central earned $2 million.

Information-theoretic measures

$I(S;C)$ Context informativeness

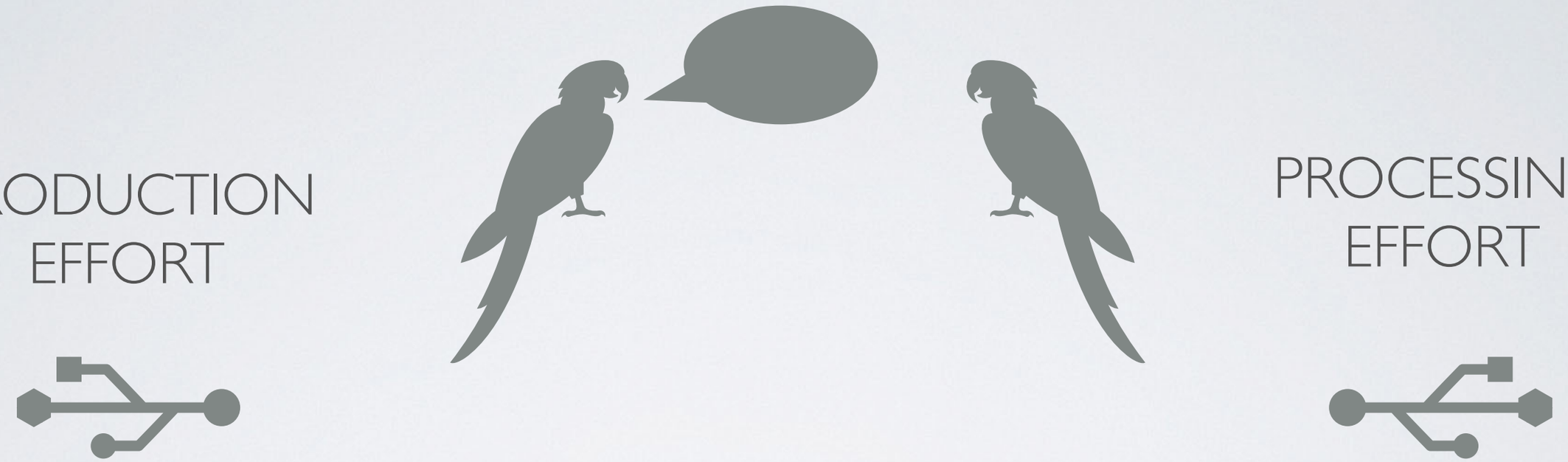
$$H(S) - H(S \mid C) = -\log_2 P(S) + \log_2 P(S \mid C)$$

Amcore Financial Inc. said it agreed to acquire Central of Illinois Inc. in a stock swap.

Shareholders of Central, a bank holding company based in Sterling, will receive Amcore stock equal to 10 times Central's 1989 earnings, Amcore said.

For the first nine months of 1989, Central earned $2 million.

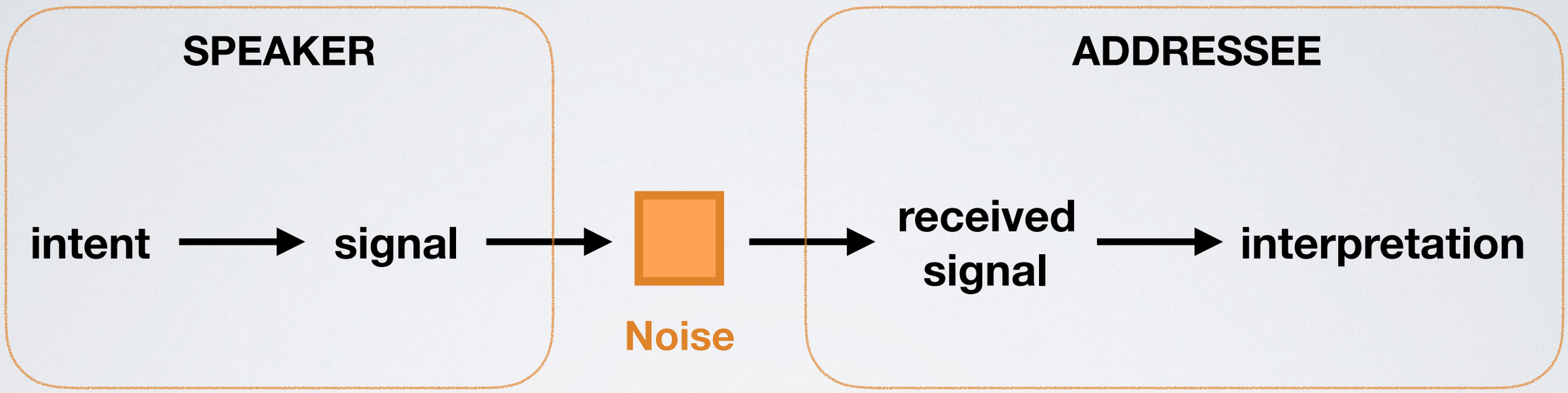
Collaborative effort in language production



Surprisal

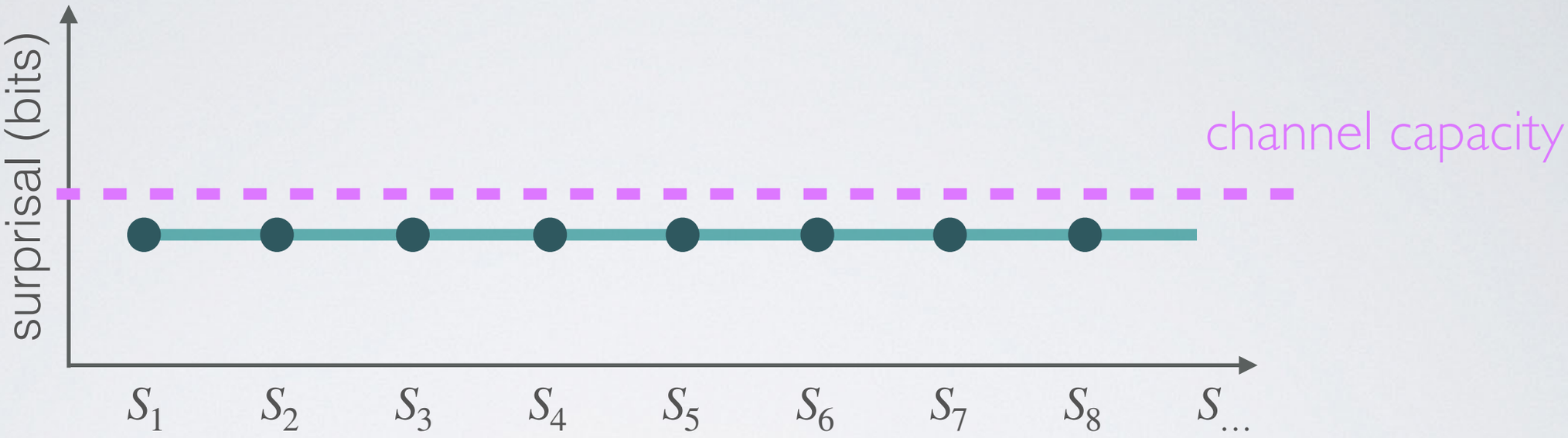
Strategies of language production

Strategies of language production

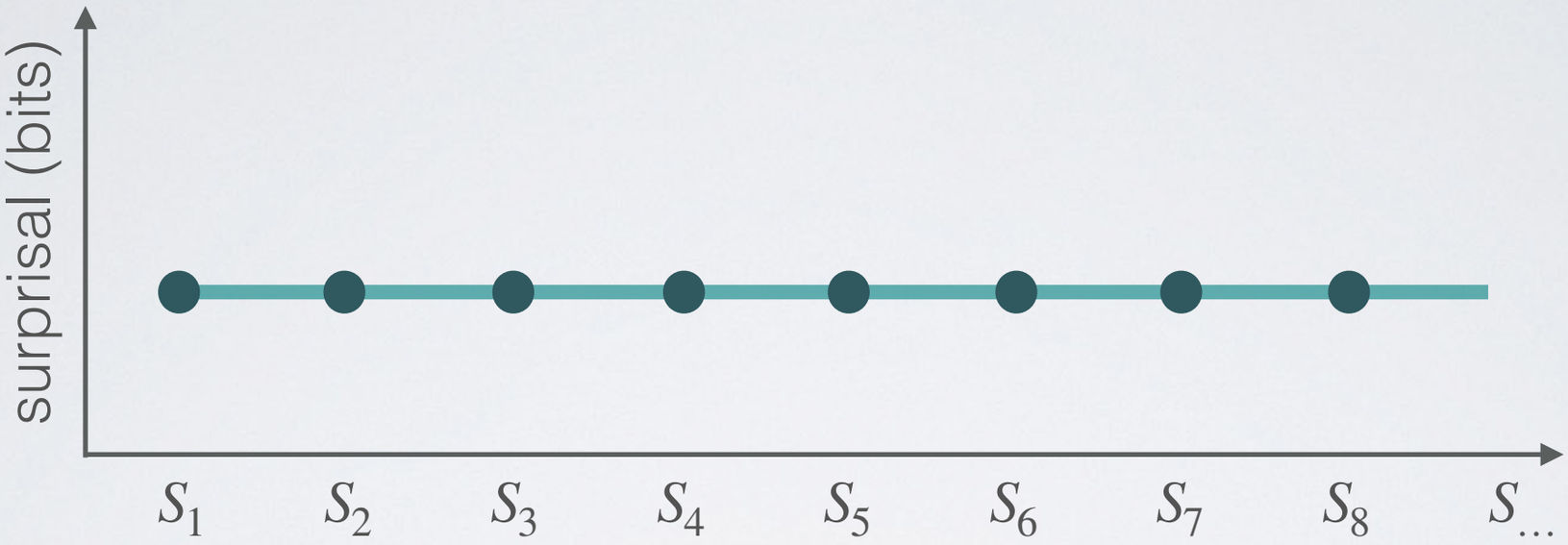


For any given degree of noise, it is possible to communicate nearly error-free up to a *maximum rate* through the channel.
(Shannon, 1948)

Strategies of language production



Strategies of language production



Constancy Rate Principle

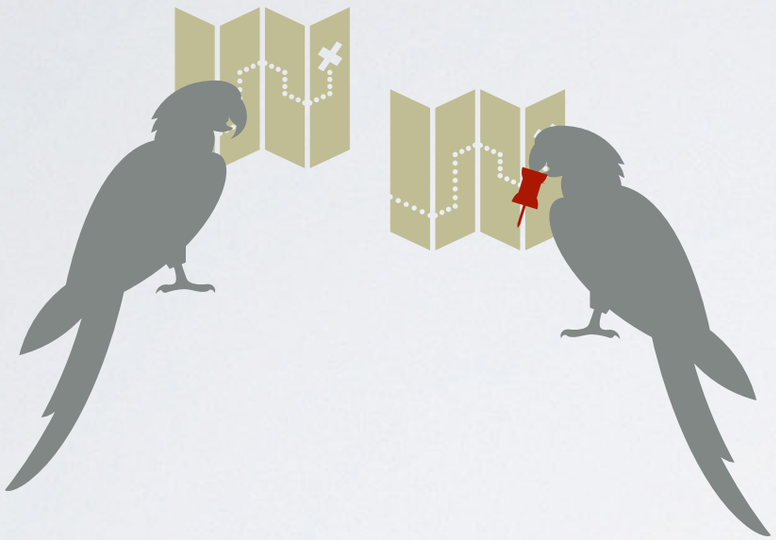
(Genzel & Charniak, 2002)

Mario Giulianelli, Arabella Sinclair, Raquel Fernández - EMNLP 2021

Study 1

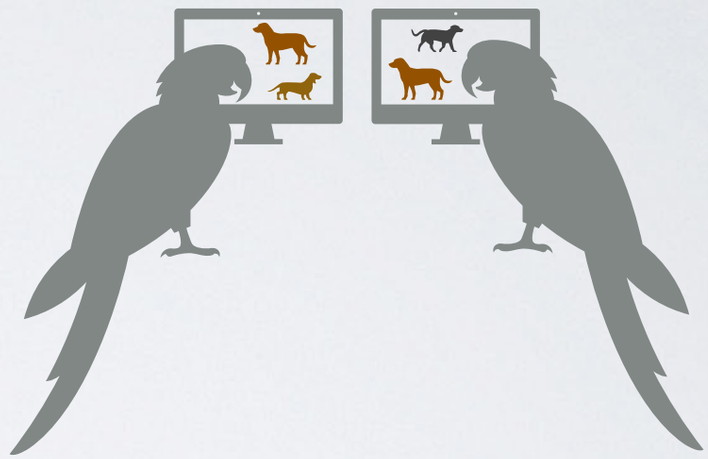
Does the Constancy Rate Principle hold in task-oriented dialogues?

Task-oriented dialogue corpora



MapTask
instruction giving
and following

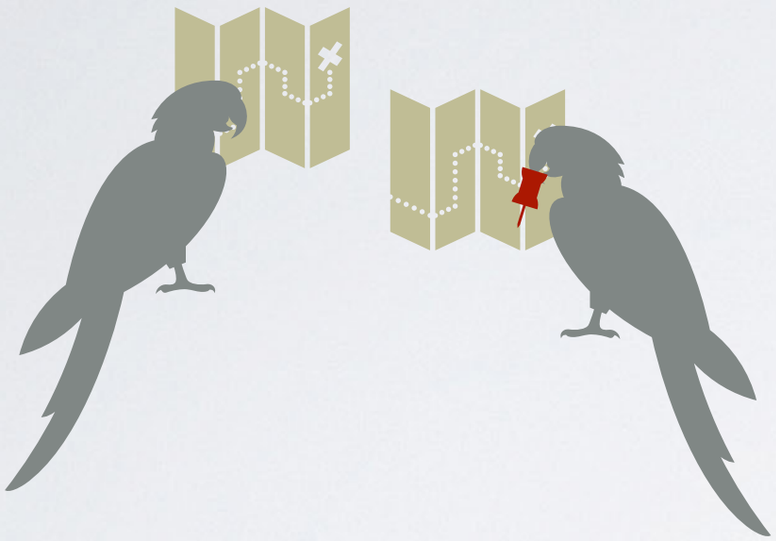
(Anderson et al., 1991)



PhotoBook
written cooperative
reference game

(Haber et al., 2019)

Task-oriented dialogue corpora



MapTask
instruction giving
and following

(Anderson et al., 1991)

Transaction 2

G	okay
G	go down
F	down towards missionary camp
G	aye
G	just right down uh-huh
F	right
F	i'm i've stopped right above it
G	then
G	you've what
F	i'm right above the missionary camp

Transaction 3

G	just go right down to round down to […]
F	right
G	okay
F	so go round it
G	uh-huh
G	round to the left uh the left yeah
F	keep going down
G	uh-huh
G	just stop where you when you reach […]

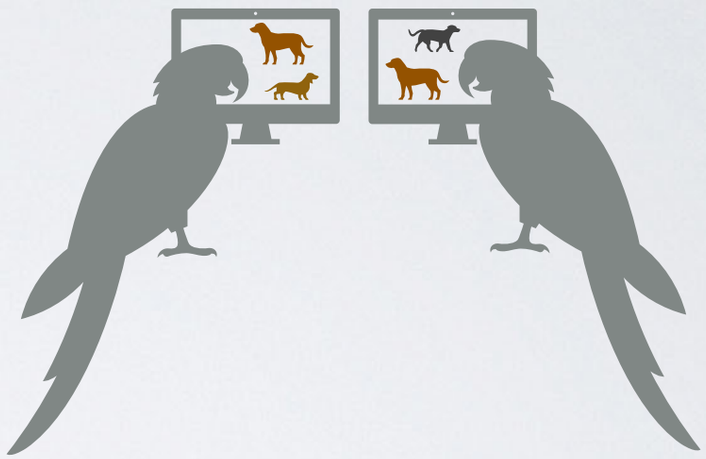
Task-oriented dialogue corpora

Round 3

B	pink sweater woman and man w/umbrella
A	yes
B	statue man with umbrella
A	y
B	guy in black suit with 2 plaid blue umbrellas
A	no

Round 4

B	guy in black suit with 2 plaid blue umbrellas
A	no
B	lady in pink and guy with umbrella
A	no
B	statue/umbralla again
A	no



PhotoBook
written cooperative reference game
(Haber et al., 2019)

Task-oriented dialogue corpora

Round 3

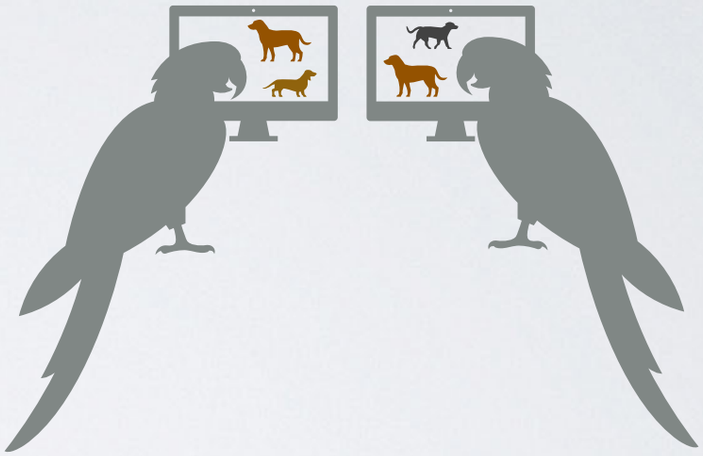
B	pink sweater woman and man w/umbrella
A	yes
B	statue man with umbrella
A	y
B	guy in black suit with 2 plaid blue umbrellas
A	no

Round 4

B	guy in black suit with 2 plaid blue umbrellas
A	no
B	lady in pink and guy with umbrella
A	no
B	statue/umbralla again
A	no

Reference chains

A	man eating slice of pizza
B	last one for me is guy with pizza
A	pizza eater
B	pizza

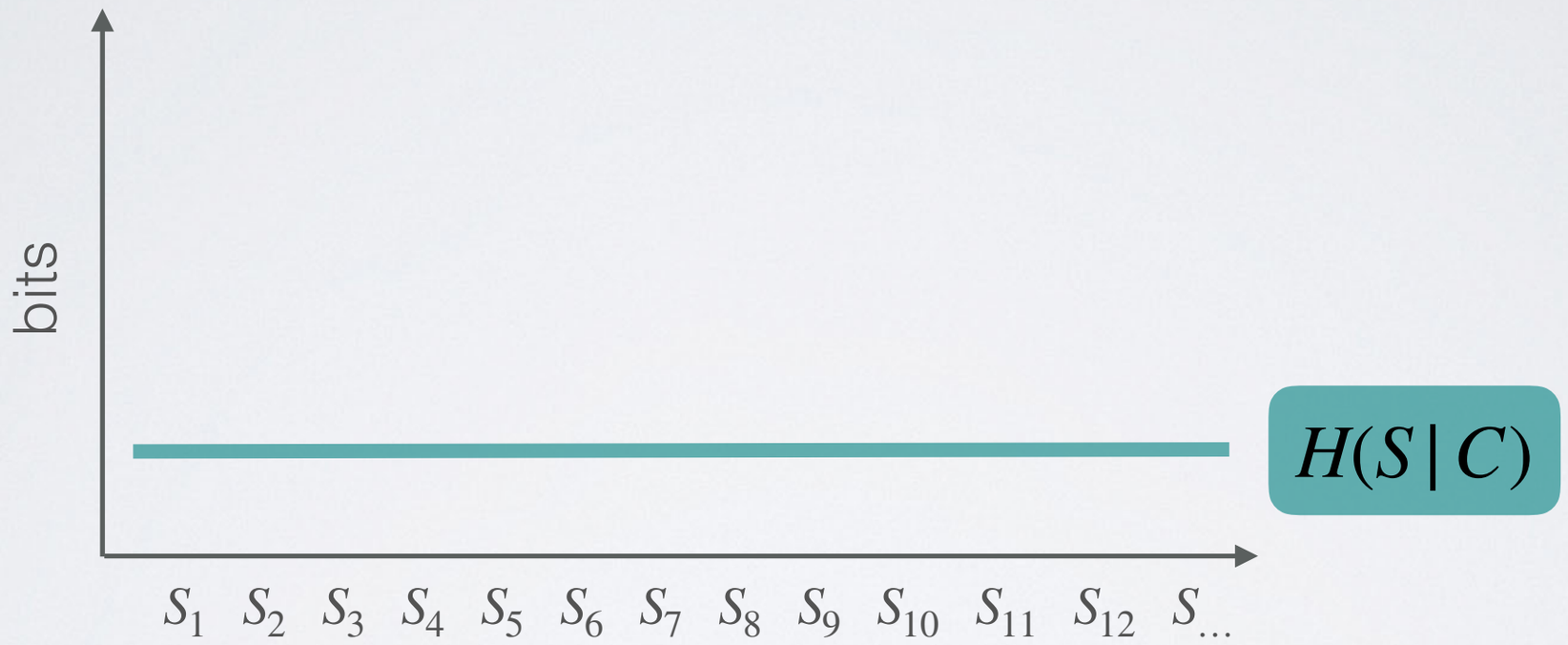


PhotoBook
written cooperative reference game
(Haber et al., 2019)

Reference chain dataset
(Takmaz et al., 2020)

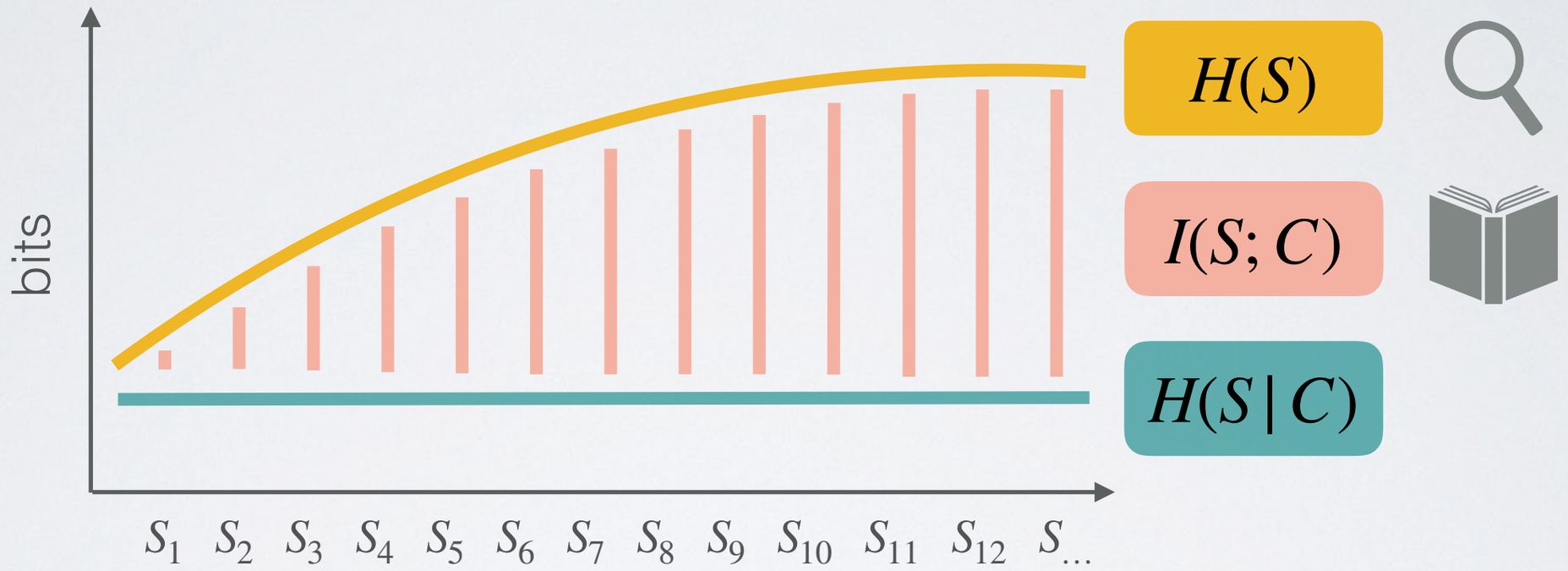
Constancy Rate Principle

(Genzel & Charniak, 2002)



Constancy Rate Principle

(Genzel & Charniak, 2002)

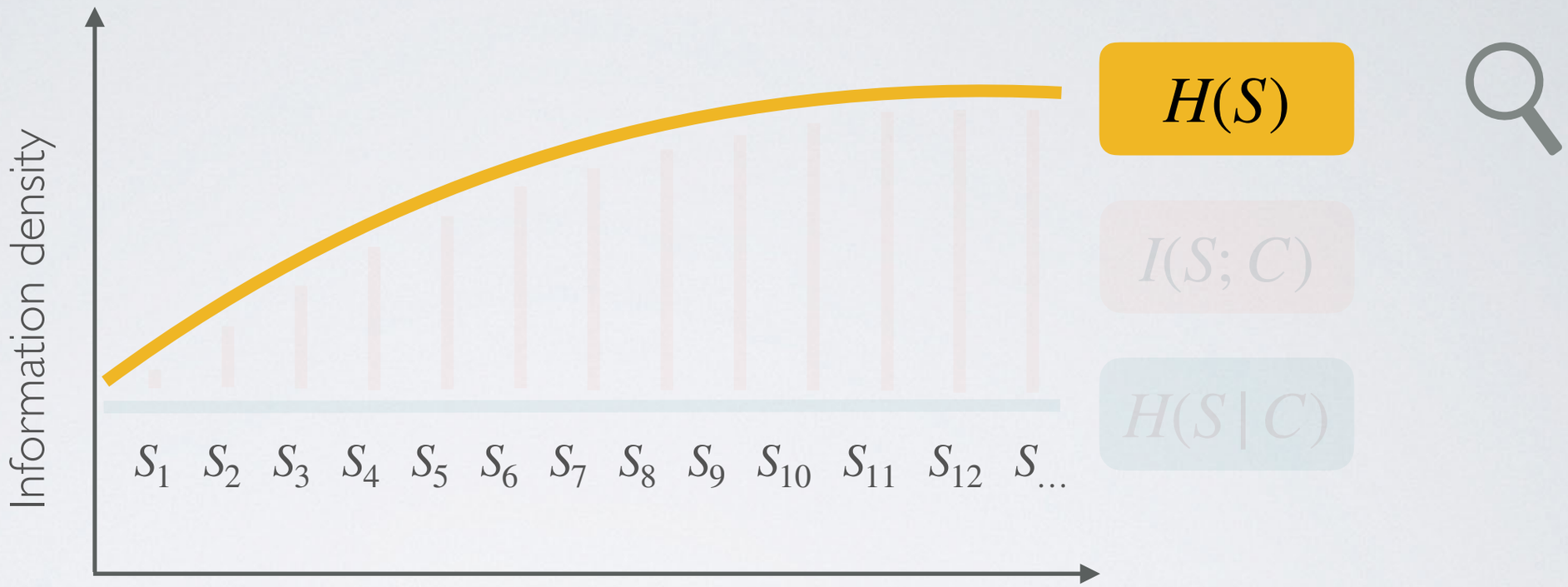


Assume: as discourse develops, context informativeness increases

Measure: whether out-of-context surprisal increases

$$H(S|C) \equiv H(S) - I(S;C)$$

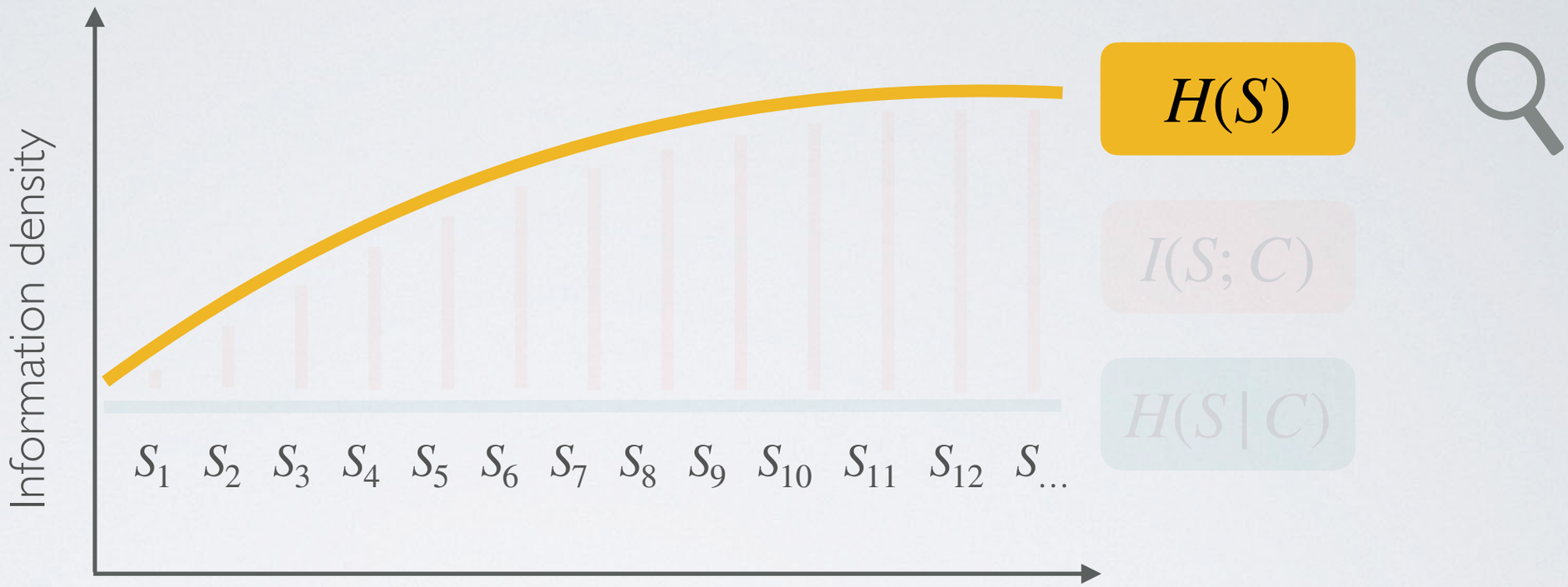
Constancy Rate Principle



We consider the Constancy Rate Principle to hold

if $H(S_i)$ increases with dialogue turn i

Estimates of surprisal

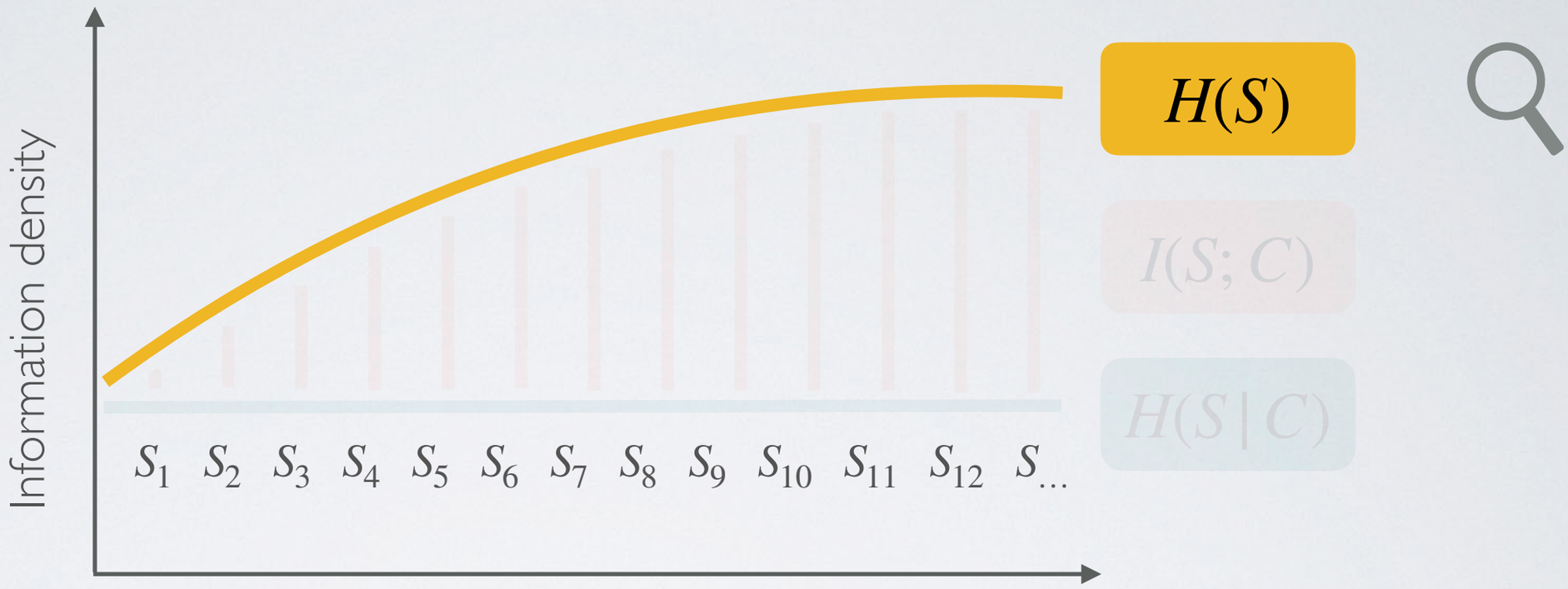


$$H(S) = -\log_2 P(S) = -\frac{1}{|S|}\sum_{w_i \in S} \log_2 P(w_i \mid w_1, \ldots, w_{i-1})$$

$P(w_i \mid \ldots)$ estimates obtained with GPT-2 fine-tuned on 70% of each target corpus

(30% held-out for analysis)

Linear Mixed Effect Model



$H(S)$

$I(S;C)$

$H(S \mid C)$

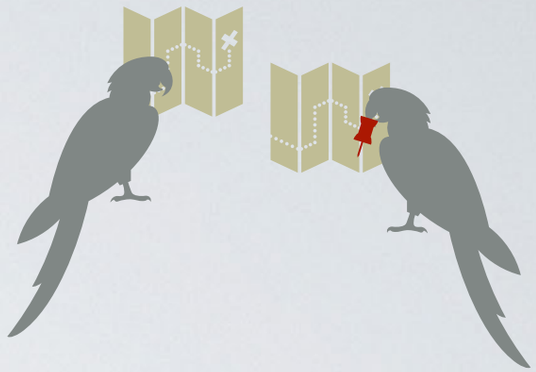
length-normalised
(Xu & Reitter, 2018)

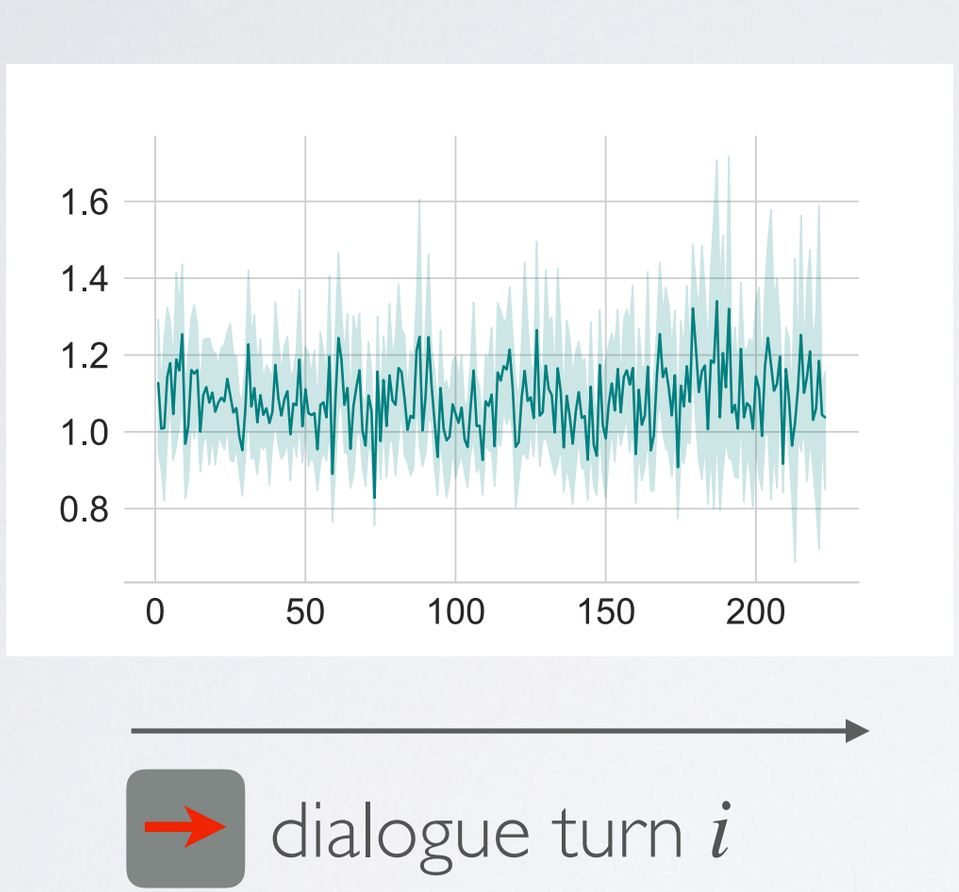
$H(S)$ ~ 1 + log dialogue turn + (1 + log dialogue turn | dialogue_id)

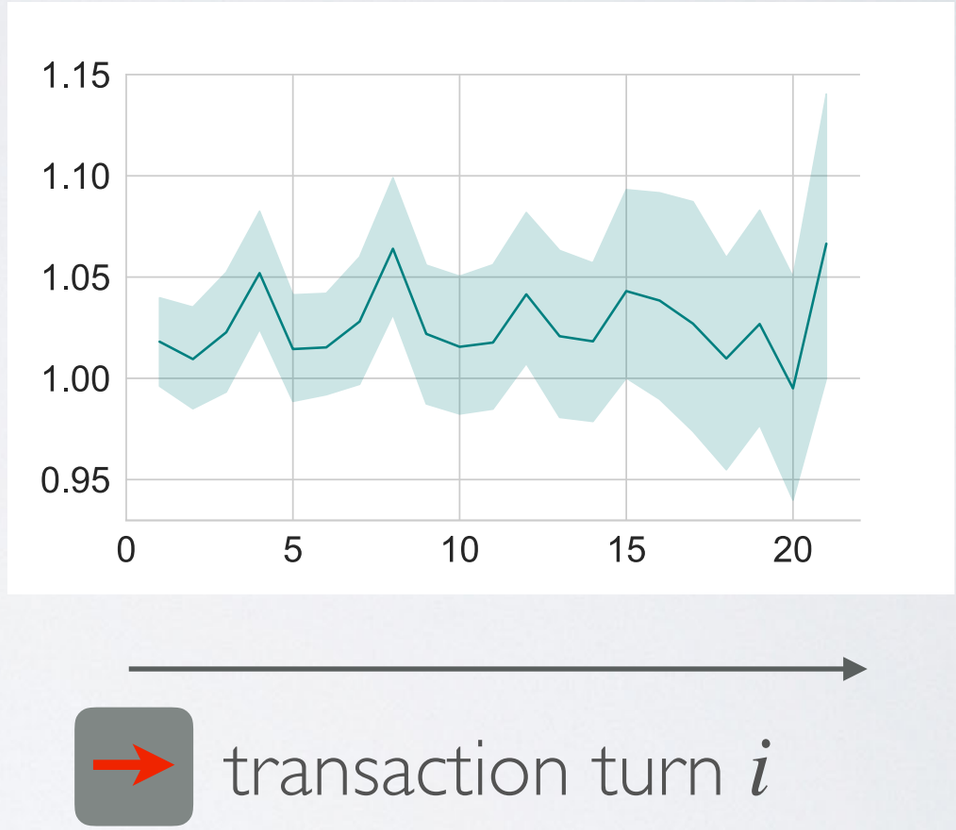
fixed effect

random intercept and slope
grouped by dyad

Results: MapTask

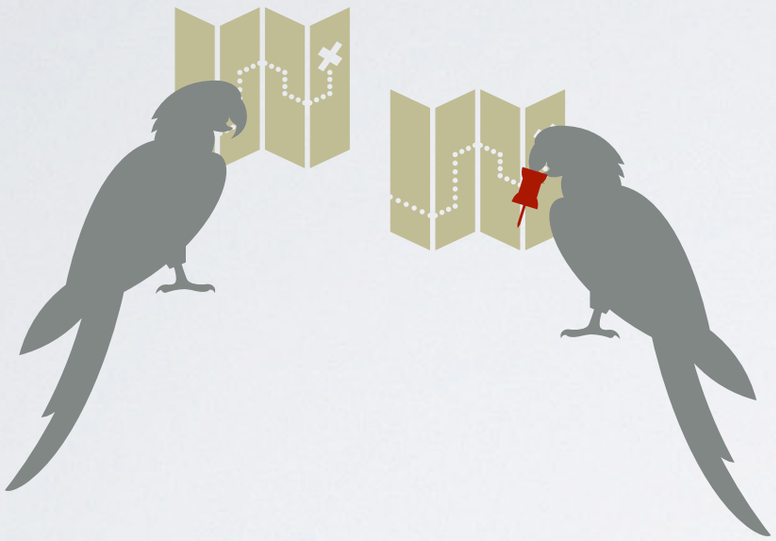






$H(S)$ does not increase
over an entire dialogue and over a transaction

Information transmission acts



MapTask
instruction giving
and following
(Anderson et al., 1991)

Transaction 2

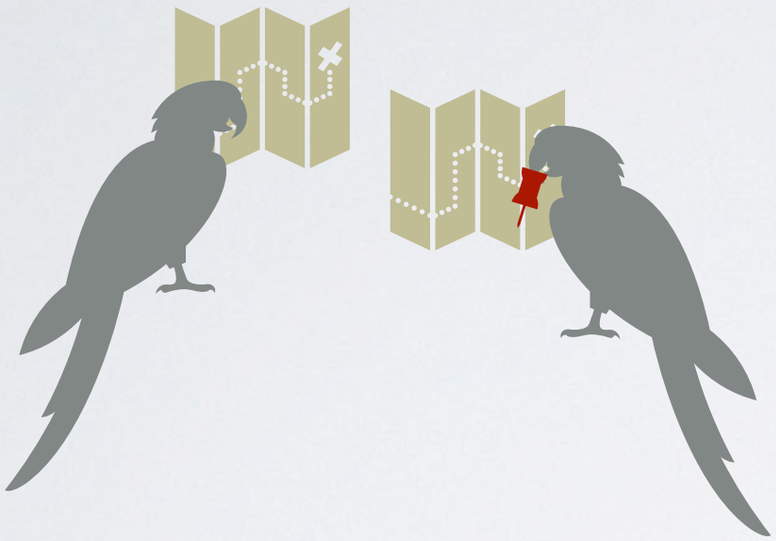
G	okay
G	go down
F	down towards missionary camp
G	aye
G	just right down uh-huh
F	right
F	i'm i've stopped right above it
G	then
G	you've what
F	i'm right above the missionary camp

Transaction 3

G	just go right down to round down to […]
F	right
G	okay
F	so go round it
G	uh-huh
G	round to the left uh the left yeah
F	keep going down
G	uh-huh
G	just stop where you when you reach […]

Information transmission acts

(no backchannels and grounding acts: *okay, mmhmm*)



MapTask

instruction giving and following

(Anderson et al., 1991)

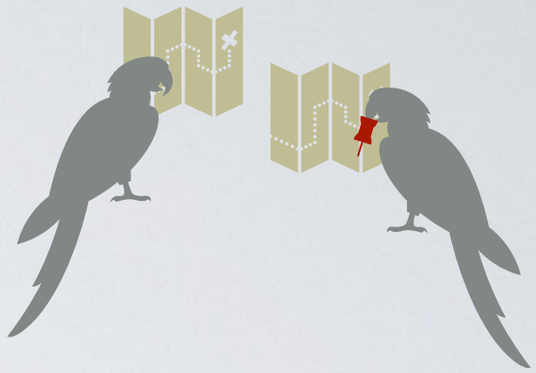
Transaction 2

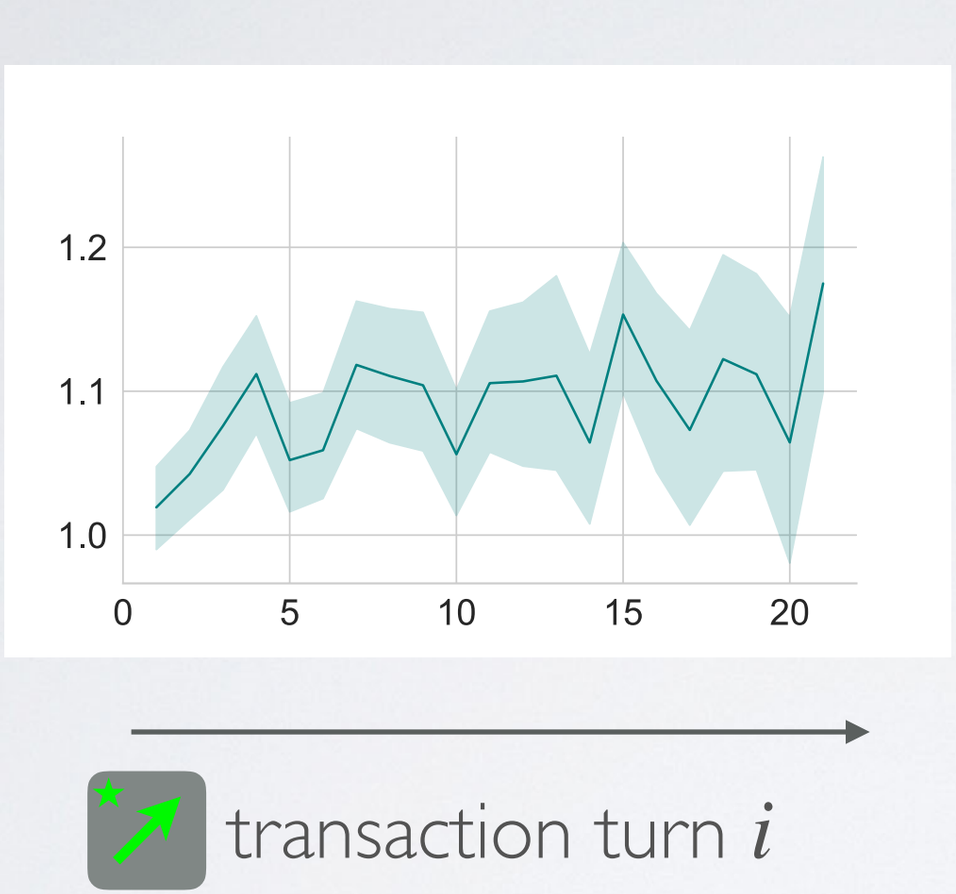
G	okay
G	go down
F	down towards missionary camp
G	aye
G	just right down uh-huh
F	right
F	i'm i've stopped right above it
G	then
G	you've what
F	i'm right above the missionary camp

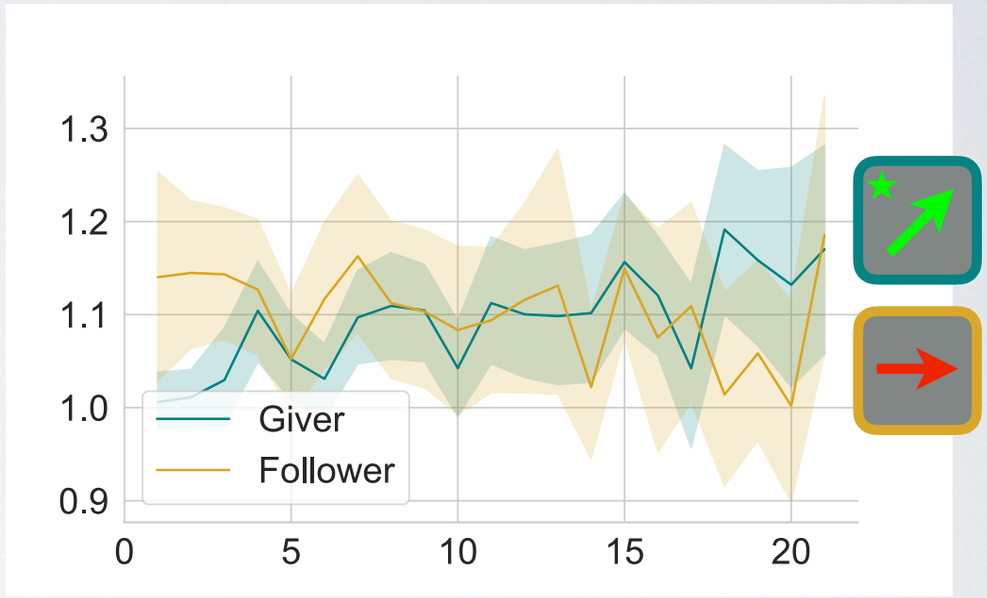
Transaction 3

G	just go right down to round down to […]
F	right
G	okay
F	so go round it
G	uh-huh
G	round to the left uh the left yeah
F	keep going down
G	uh-huh
G	just stop where you when you reach […]

Results: MapTask

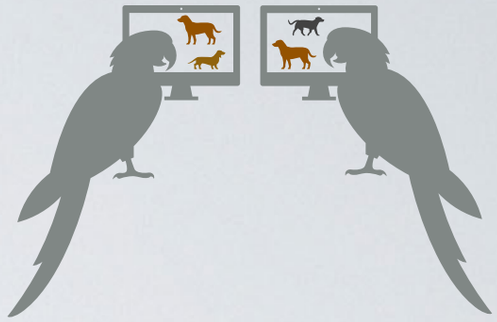






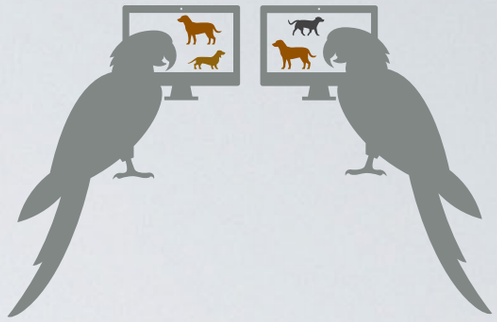
$H(S)$ increases
in the information-transmission dialogue acts of transactions
(no backchannels and grounding acts: *okay, mmhmm*)

Results: PhotoBook



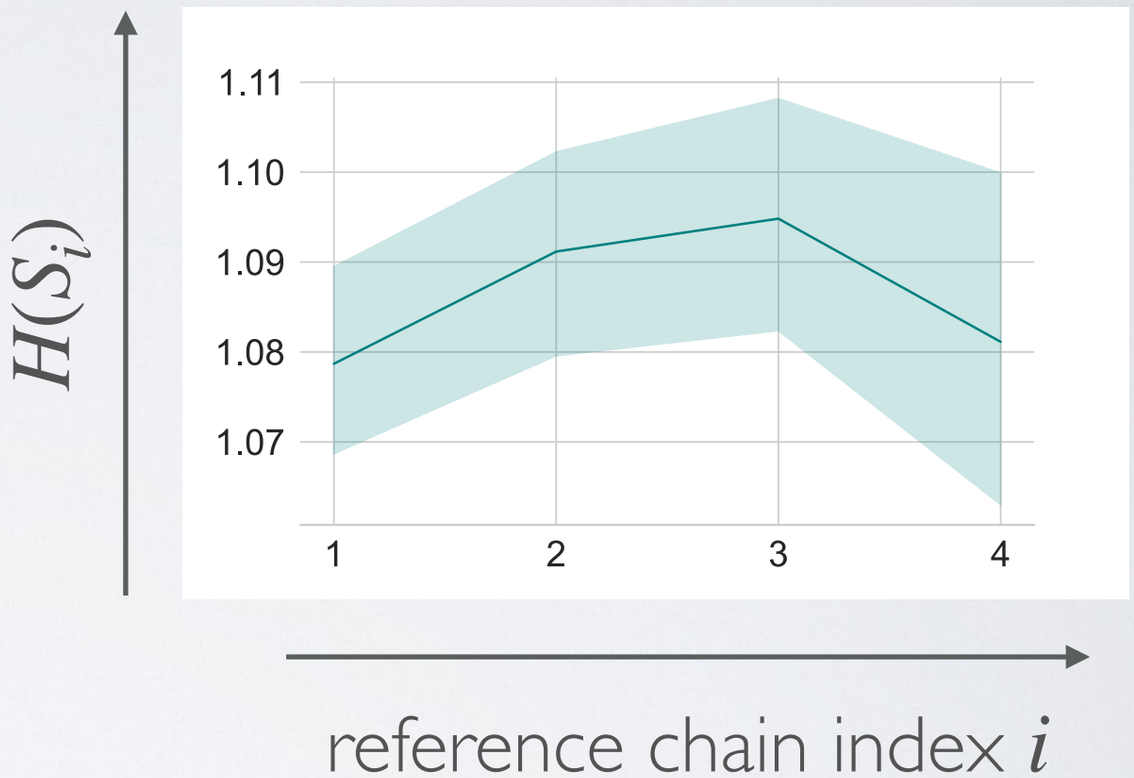
$H(S)$ increases
over an entire dialogue and over a reference chain

Results: PhotoBook



$H(S_i)$

man eating slice of pizza	0.69
last one for me is guy with pizza	0.78
pizza eater	0.91
pizza	0.67



Reduction + information compression

Study 1 - Summary

The Constancy Rate Principle holds
within more **topically and referentially coherent contextual units** of task-oriented dialogues:

MapTask information transmission acts in transactions

PhotoBook dialogues and reference chains

Strategies of information transmission

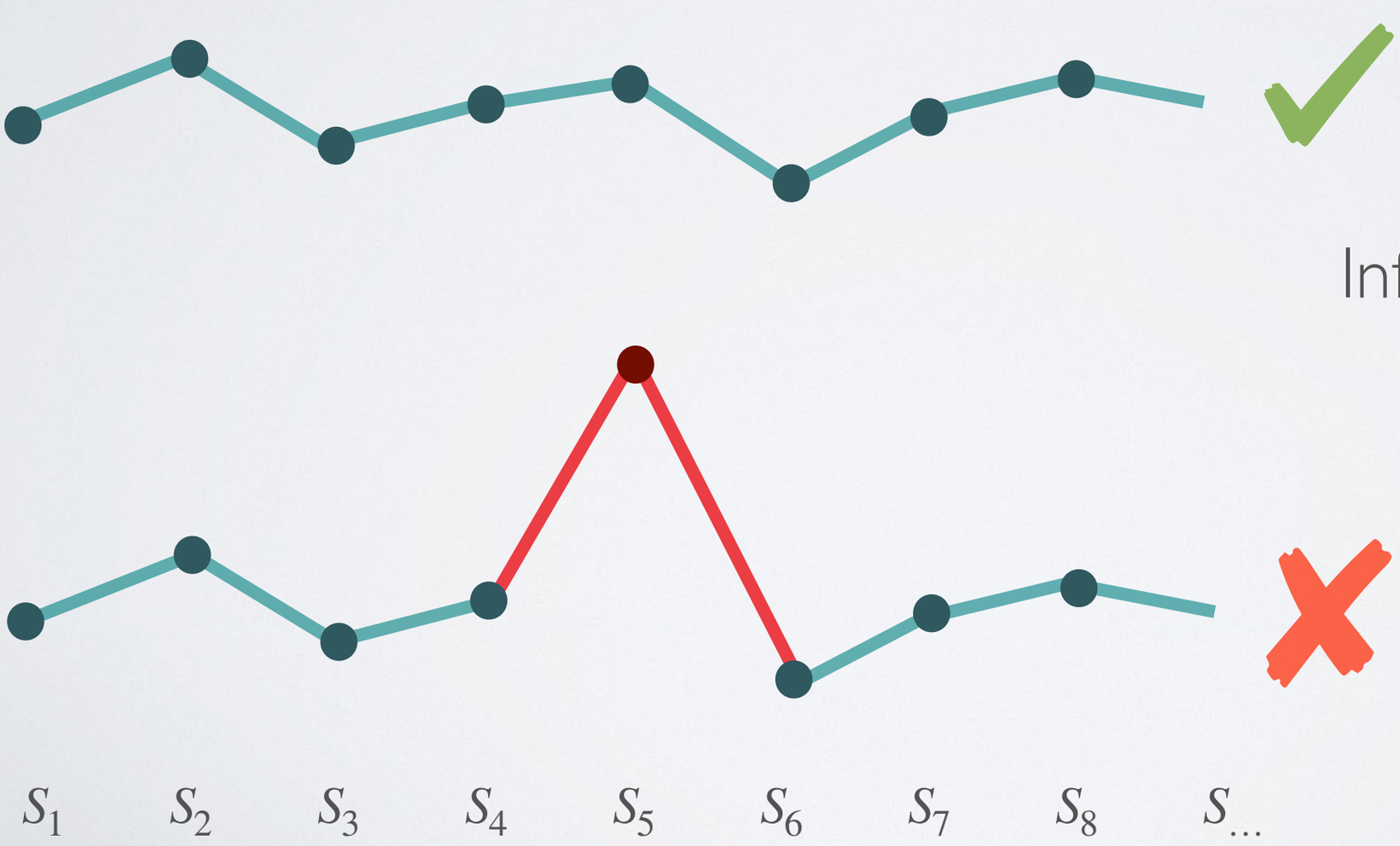
Constancy Rate Principle

(Genzel & Charniak, 2002)

Strategies of information transmission

Entropy Rate Constancy

(Genzel & Charniak, 2002)



Uniform Information Density

(Aylett & Turk, 2004; Jaeger & Levy, 2007; Levy, 2010)

Mario Giulianelli and Raquel Fernández - CoNLL 2021

Study 2

Are surprisal patterns better described as constant (CRP) or as uniform (UID)?

Mario Giulianelli and Raquel Fernández - CoNLL 2021

Study 2

Are surprisal patterns better described as constant (CRP) or as uniform (UID)?

Here, we measure surprisal as a function of discourse context.

Estimates of surprisal (revised)

$$H(S) = -\log_2 P(S) = -\frac{1}{|S|}\sum_{w_i \in S} \log_2 P(w_i \mid w_1, \ldots, w_{i-1})$$

Estimates of surprisal (revised)

$$H(S) = -\log_2 P(S) = -\frac{1}{|S|}\sum_{w_i \in S} \log_2 P(w_i \mid w_1, \ldots, w_{i-1})$$

$$H(S \mid C) = -\log_2 P(S \mid C) = -\frac{1}{|S|}\sum_{w_i \in S} \log_2 P(w_i \mid w_1, \ldots, w_{i-1}, C)$$

Estimates of surprisal (revised)

$$H(S) = -\log_2 P(S) = -\frac{1}{|S|}\sum_{w_i \in S} \log_2 P(w_i \mid w_1, \ldots, w_{i-1})$$

$$H(S \mid C) = -\log_2 P(S \mid C) = -\frac{1}{|S|}\sum_{w_i \in S} \log_2 P(w_i \mid w_1, \ldots, w_{i-1}, C)$$

$$I(S; C) \equiv H(S) - H(S \mid C) = -\log_2 P(S) + \log_2 P(S \mid C)$$

Estimates of surprisal (revised)

$$H(S) = -\log_2 P(S) = -\frac{1}{|S|}\sum_{w_i \in S} \log_2 P(w_i \mid w_1, \ldots, w_{i-1})$$

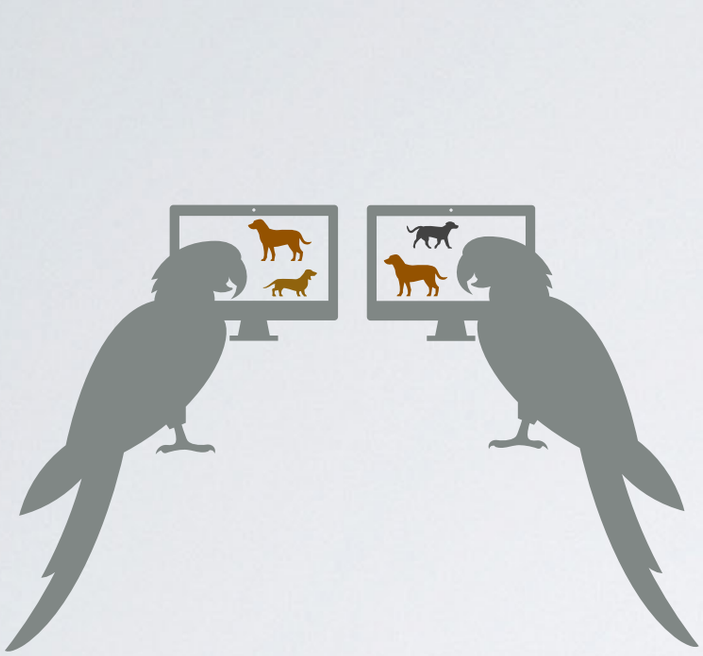
$$H(S \mid C) = -\log_2 P(S \mid C) = -\frac{1}{|S|}\sum_{w_i \in S} \log_2 P(w_i \mid w_1, \ldots, w_{i-1}, C)$$

$$I(S; C) \equiv H(S) - H(S \mid C) = -\log_2 P(S) + \log_2 P(S \mid C)$$

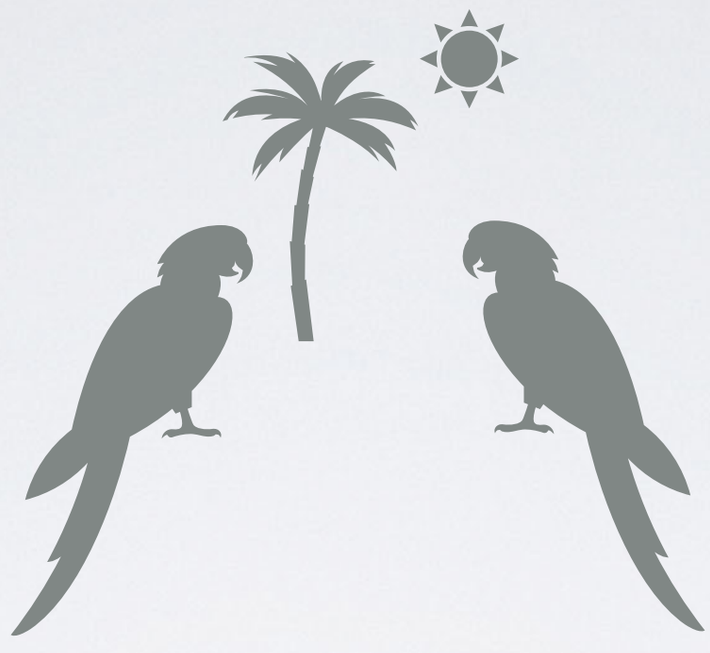
$P(w_i \mid \ldots)$ estimates obtained with GPT-2
fine-tuned on 70% of each target corpus

(30% held-out for analysis)

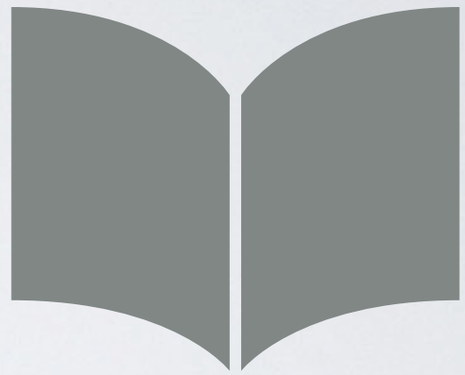
Experimental data



PhotoBook
task-oriented
written dialogues
(Haber et al., 2019)

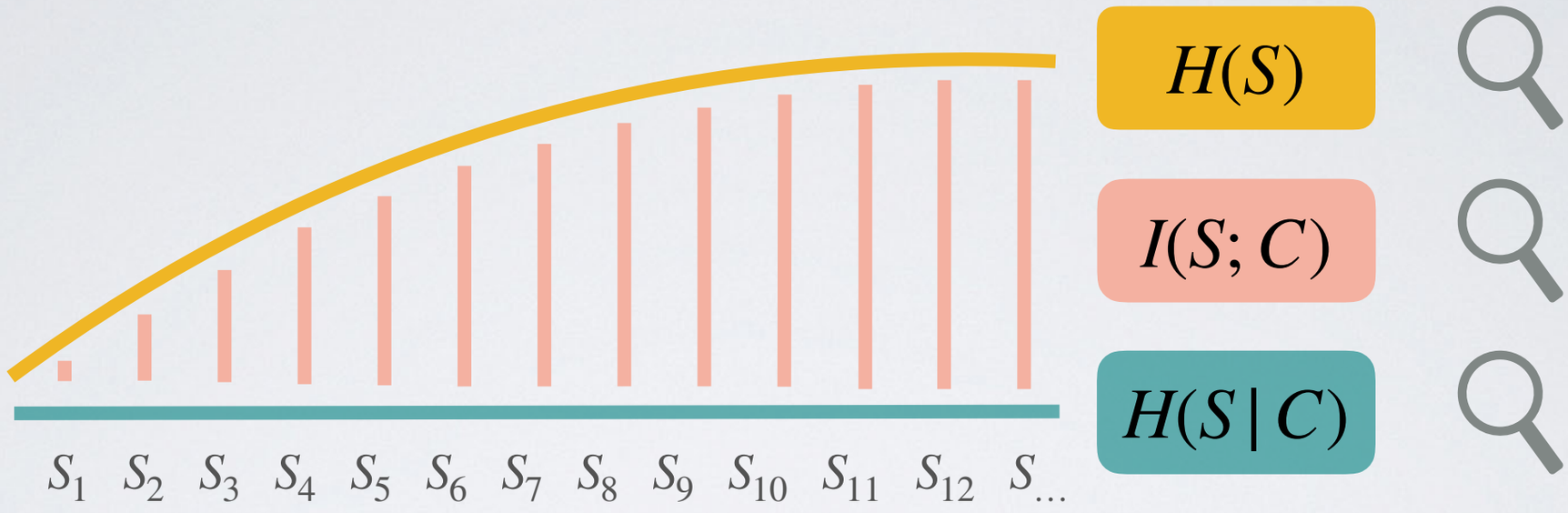


Spoken BNC
open domain
dialogues
(Love et al., 2017)

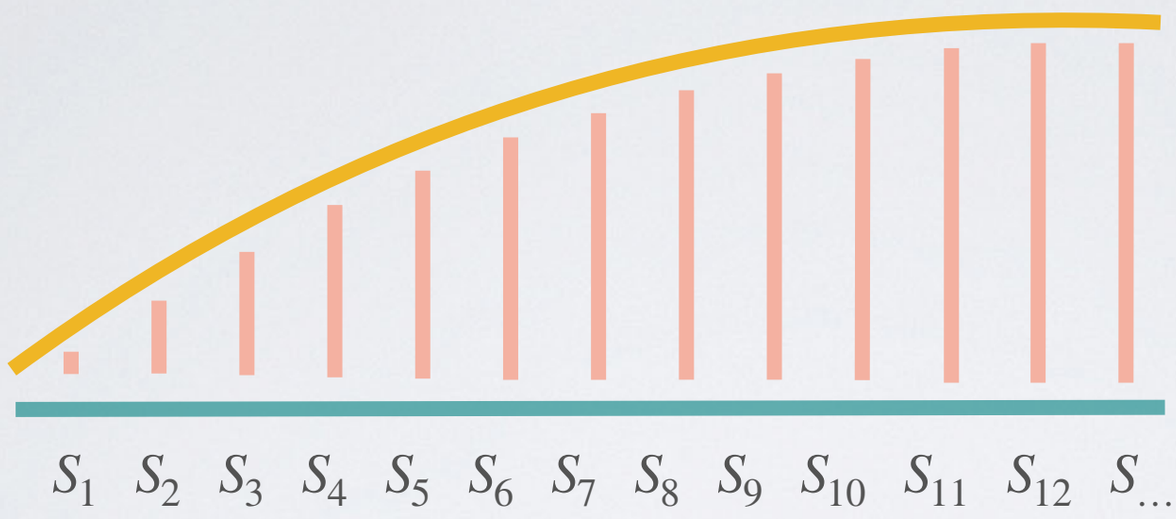


Penn Treebank
newspaper
articles
(Mitchell et al. 1999)

Can surprisal be described as constant?



Can surprisal be described as constant?



$H(S)$

$I(S;C)$

$H(S \mid C)$

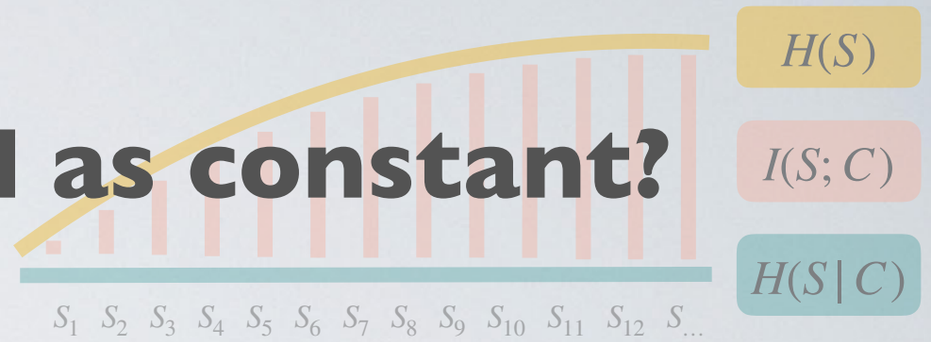
$H(S)$ ~ fixed effects

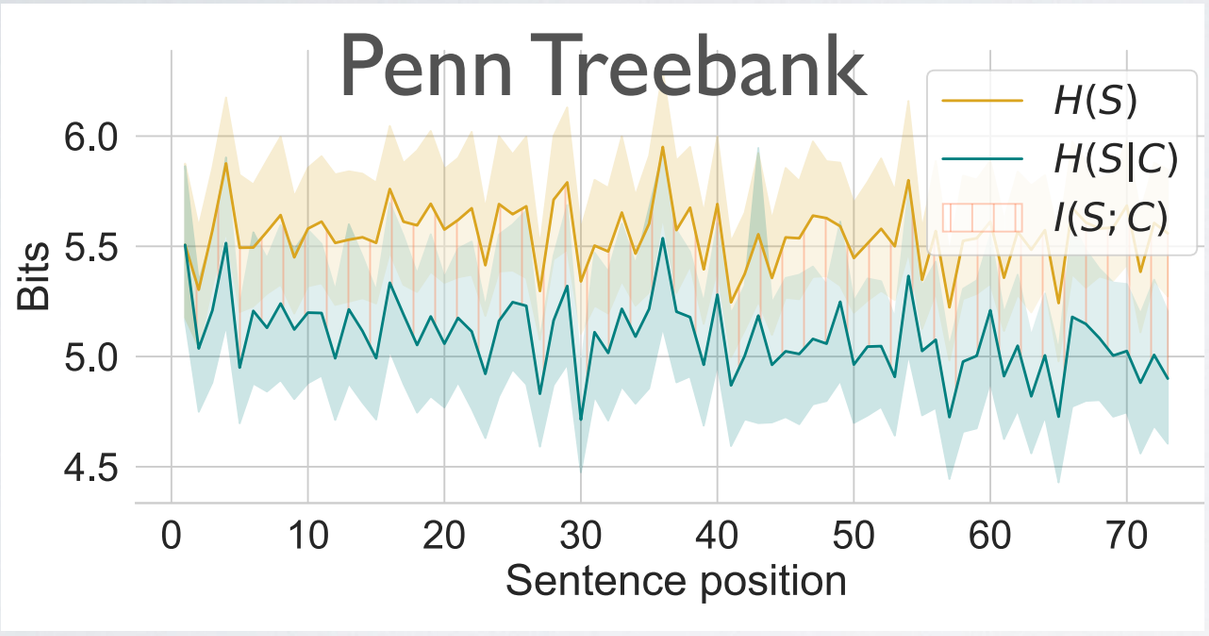
$I(S;C)$ ~ 1 + log sentence position + log sentence length

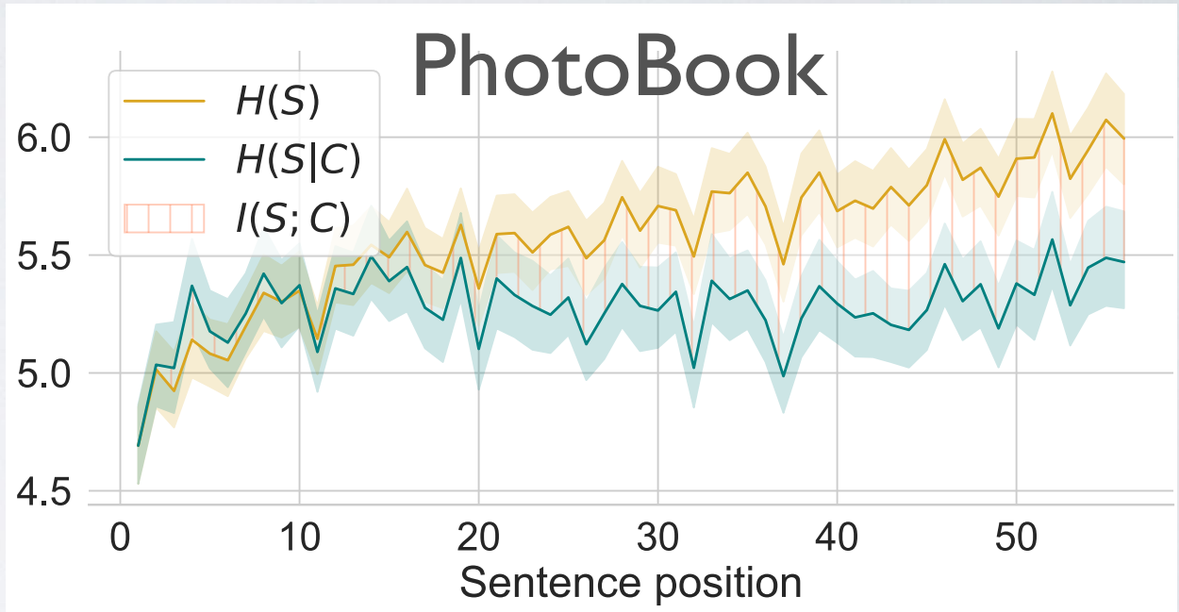
+ (1 + log sentence position + log sentence length | doc_id)

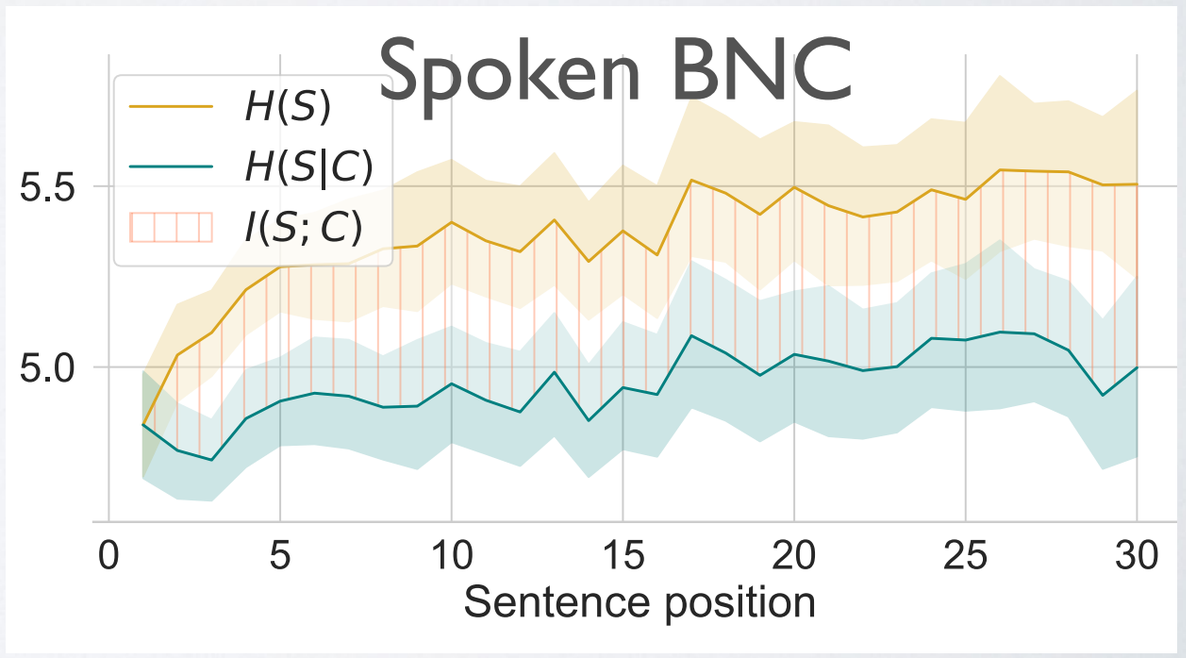
$H(S \mid C)$ ~ random intercept and slope
grouped by text / dialogue

Can surprisal be described as constant?

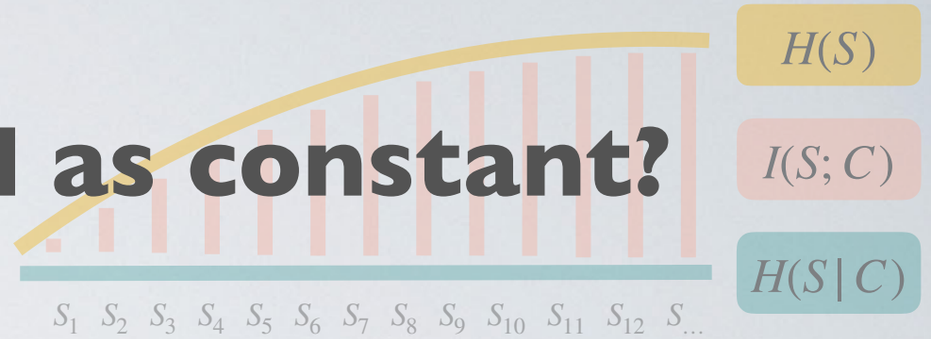


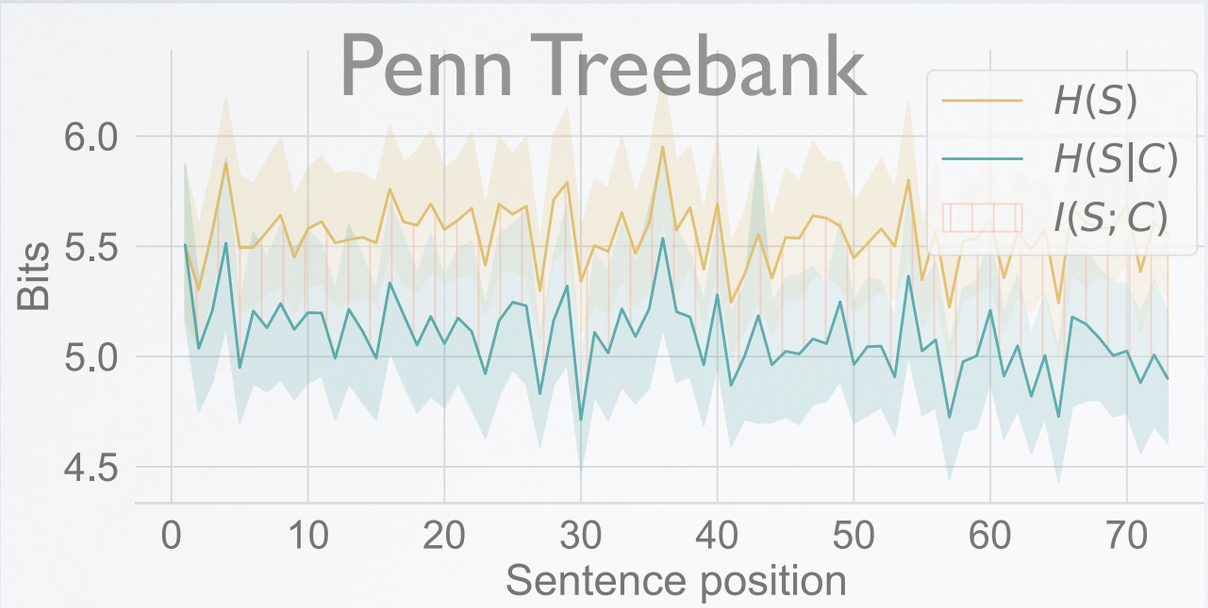


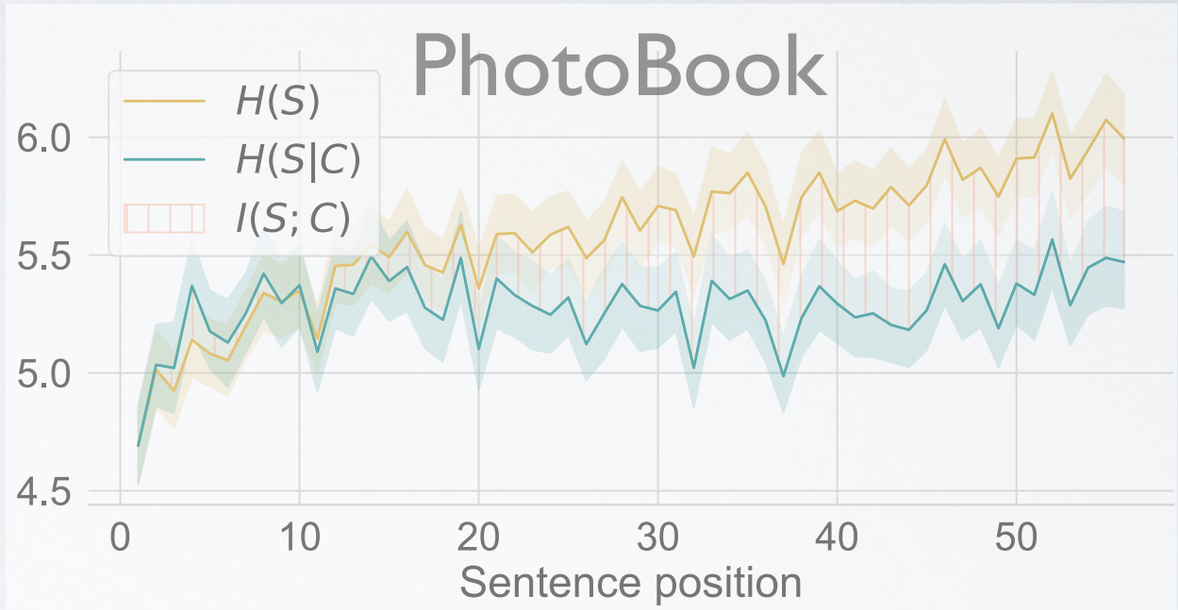


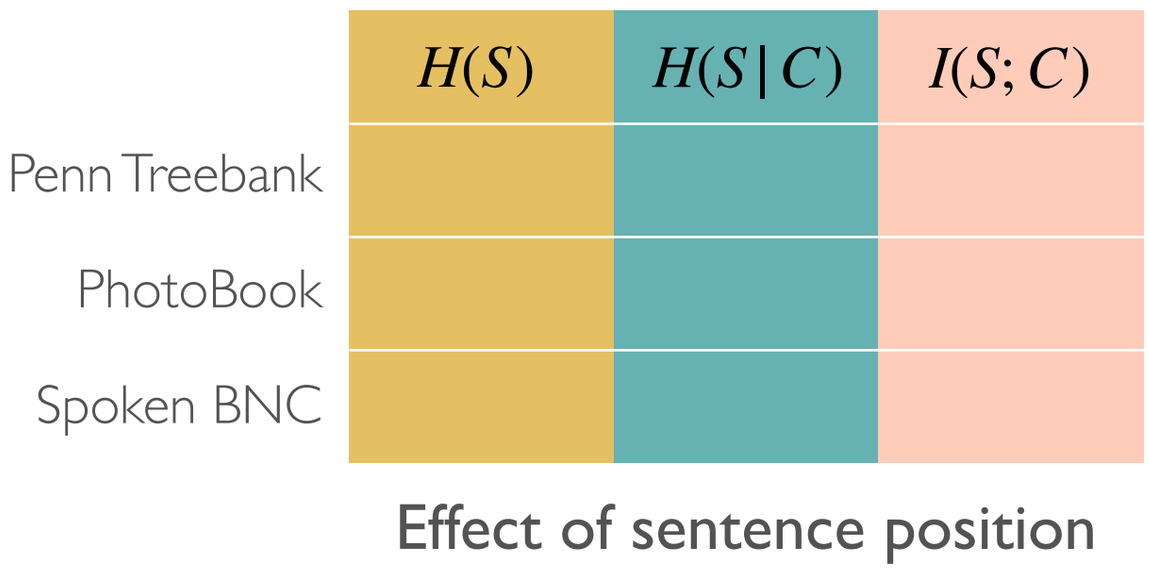


Can surprisal be described as constant?

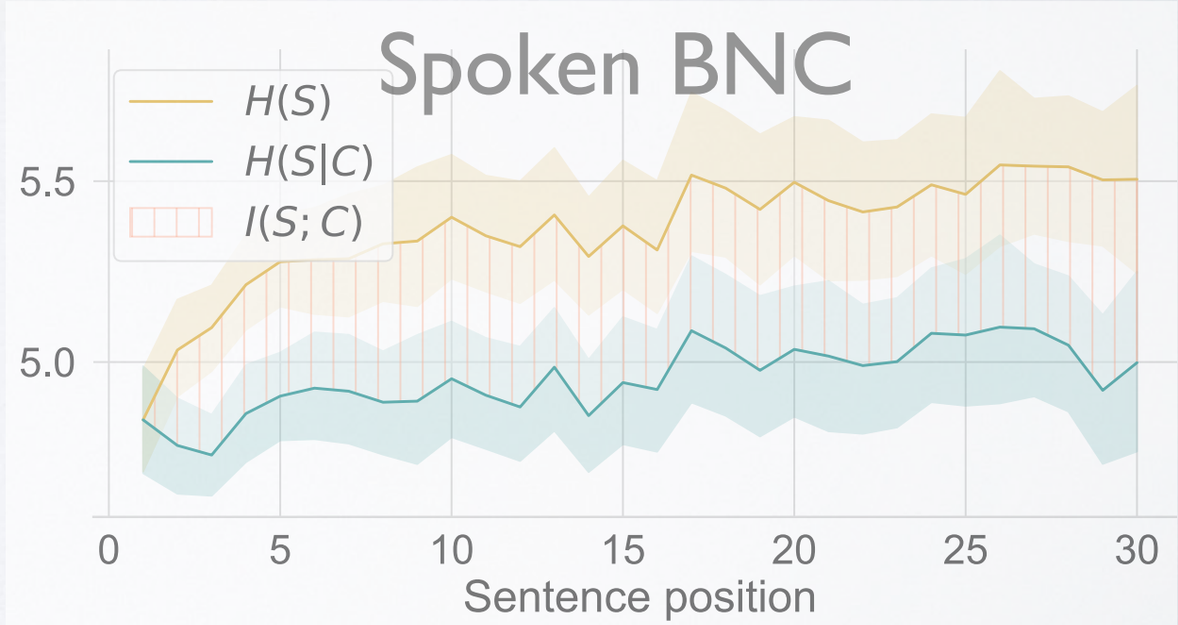




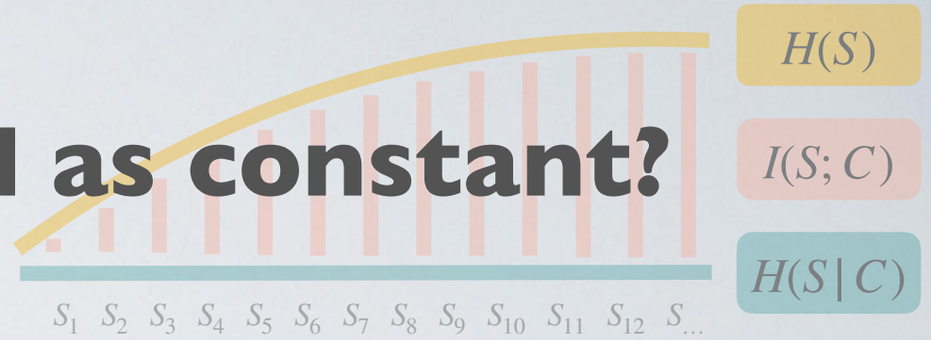


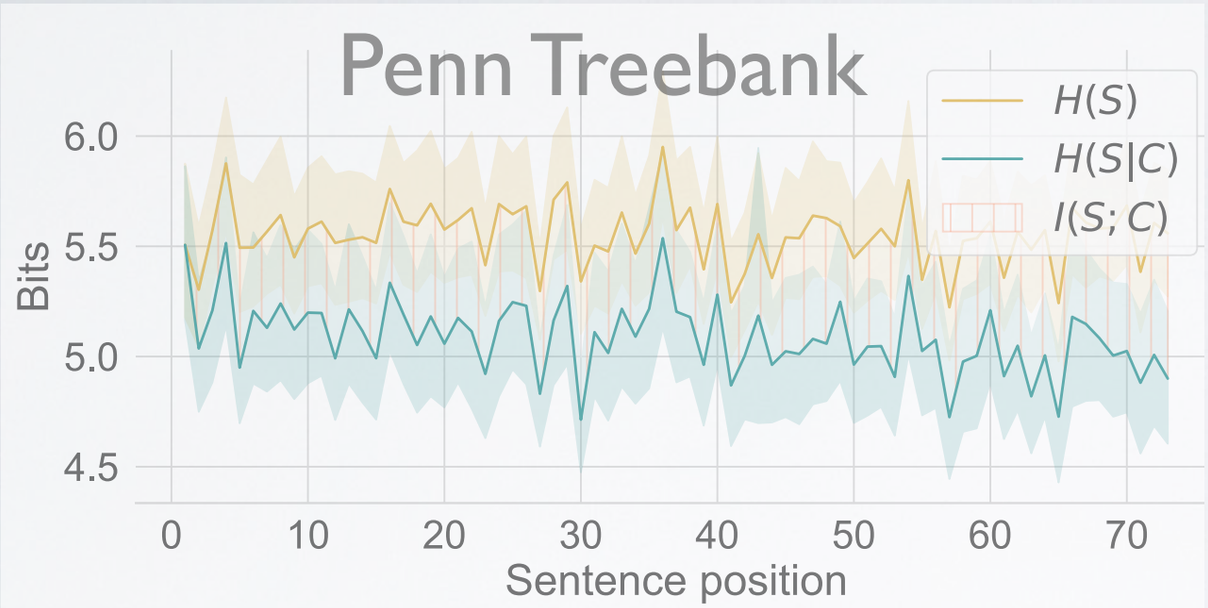


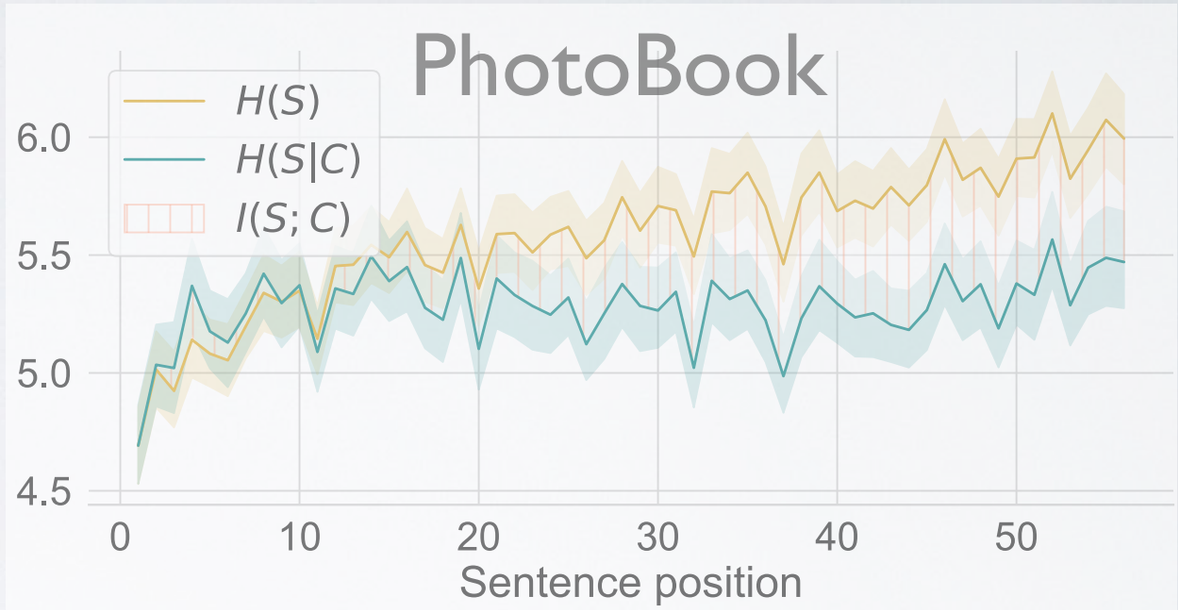
Effect of sentence position

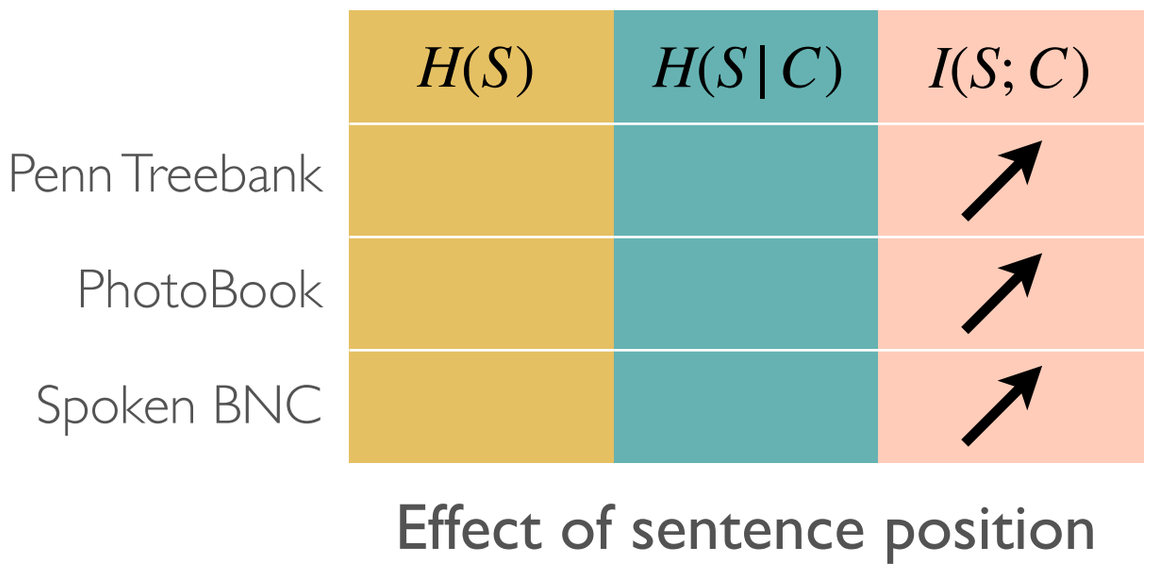


Can surprisal be described as constant?

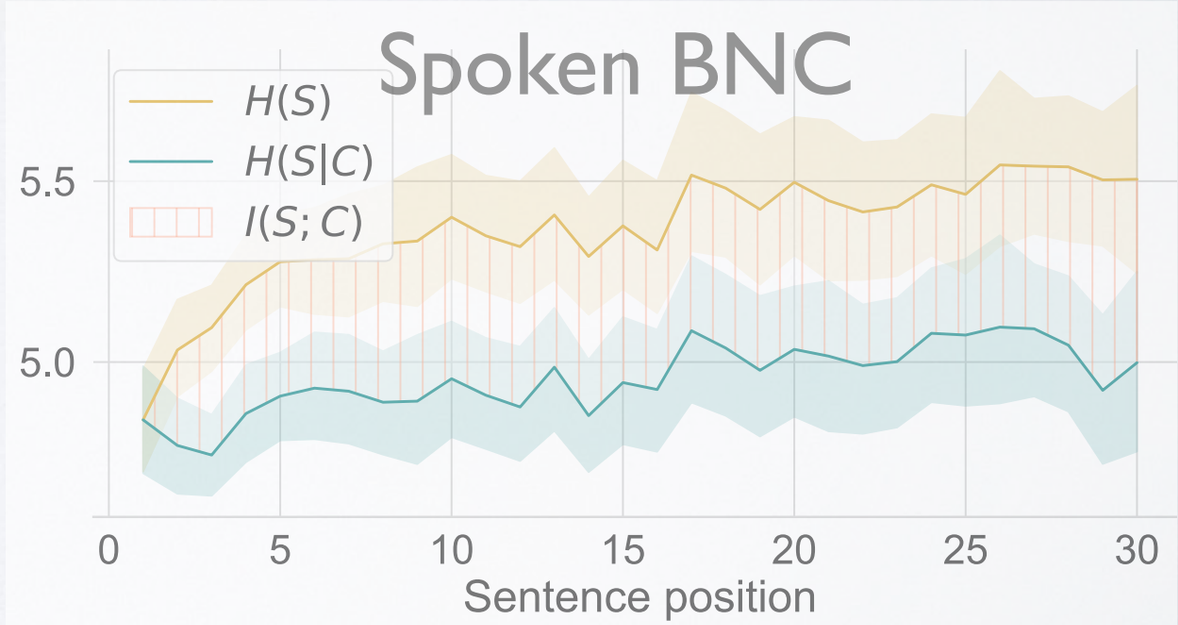




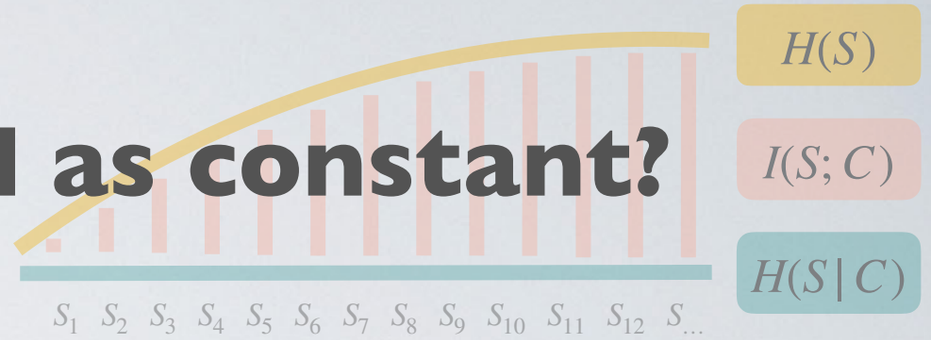


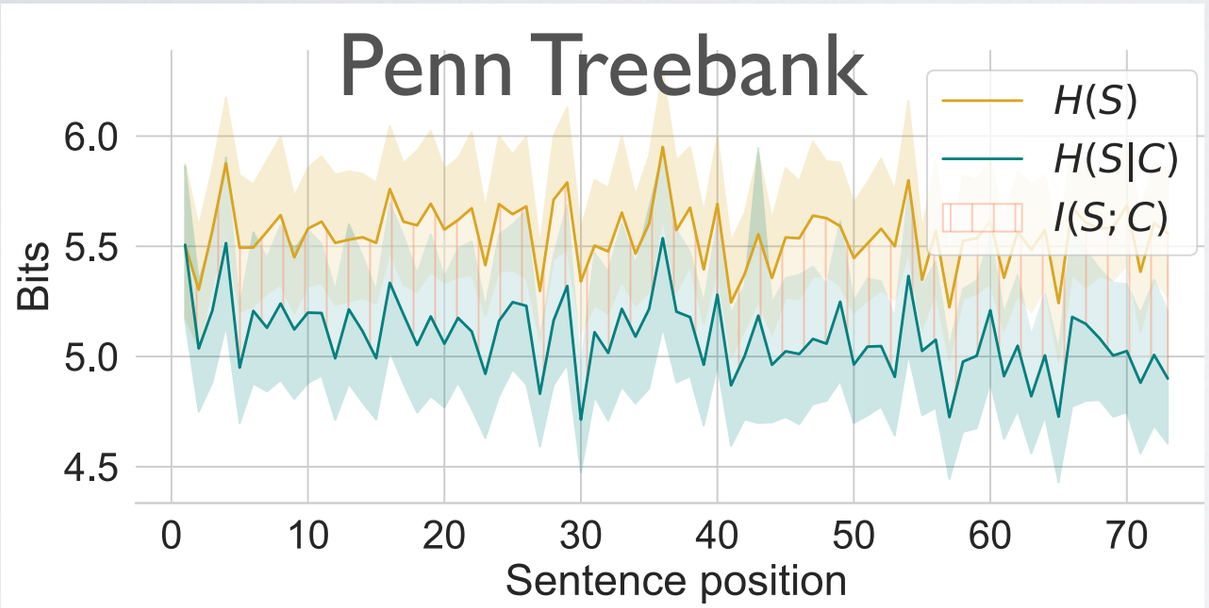


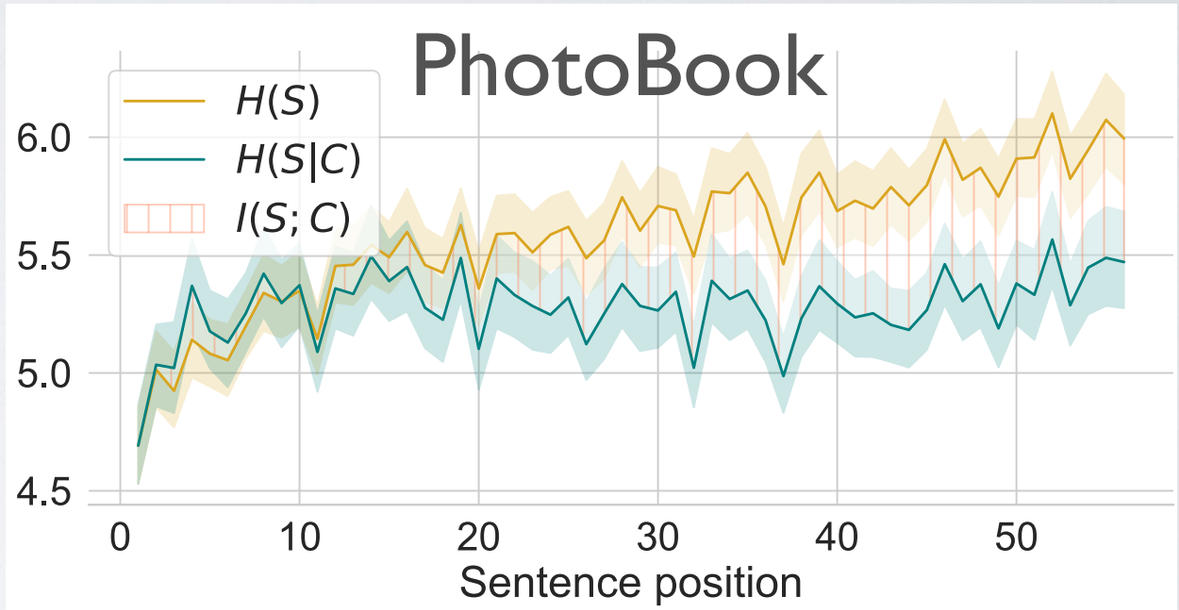
Effect of sentence position

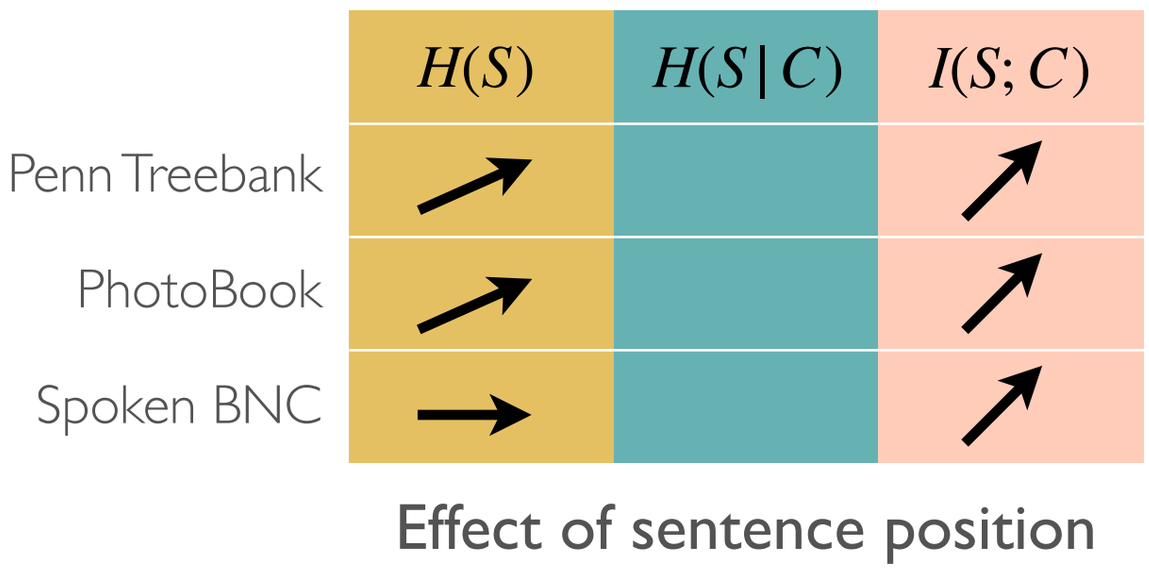


Can surprisal be described as constant?

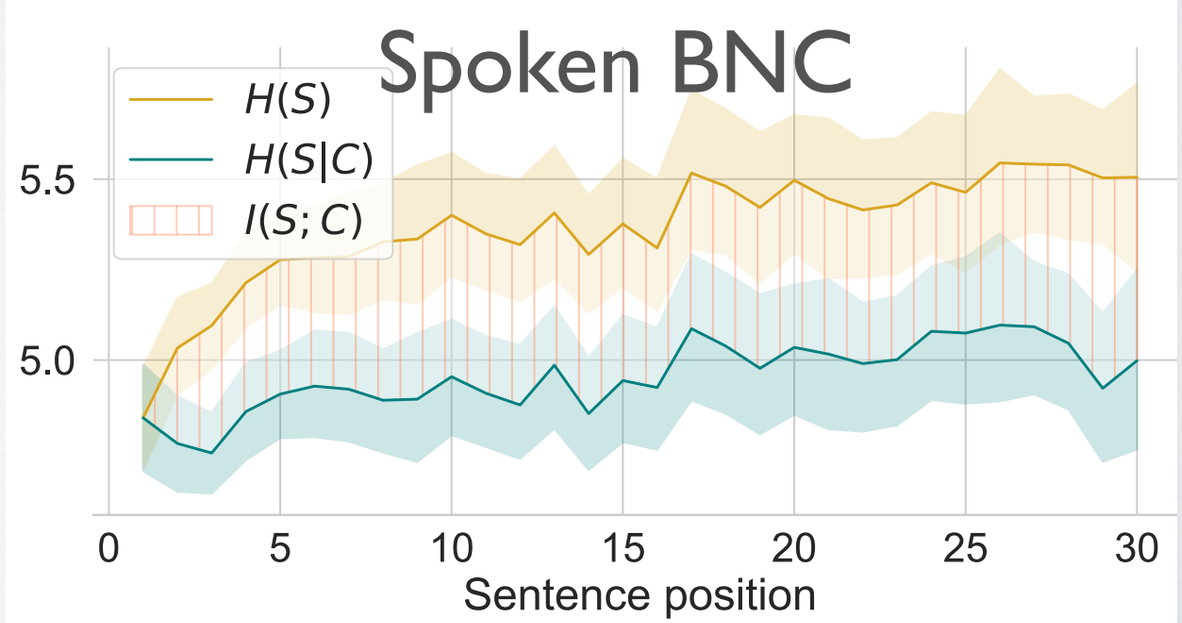




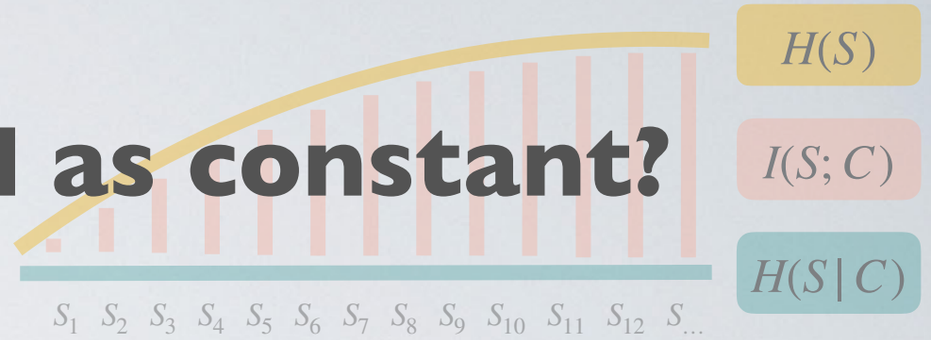


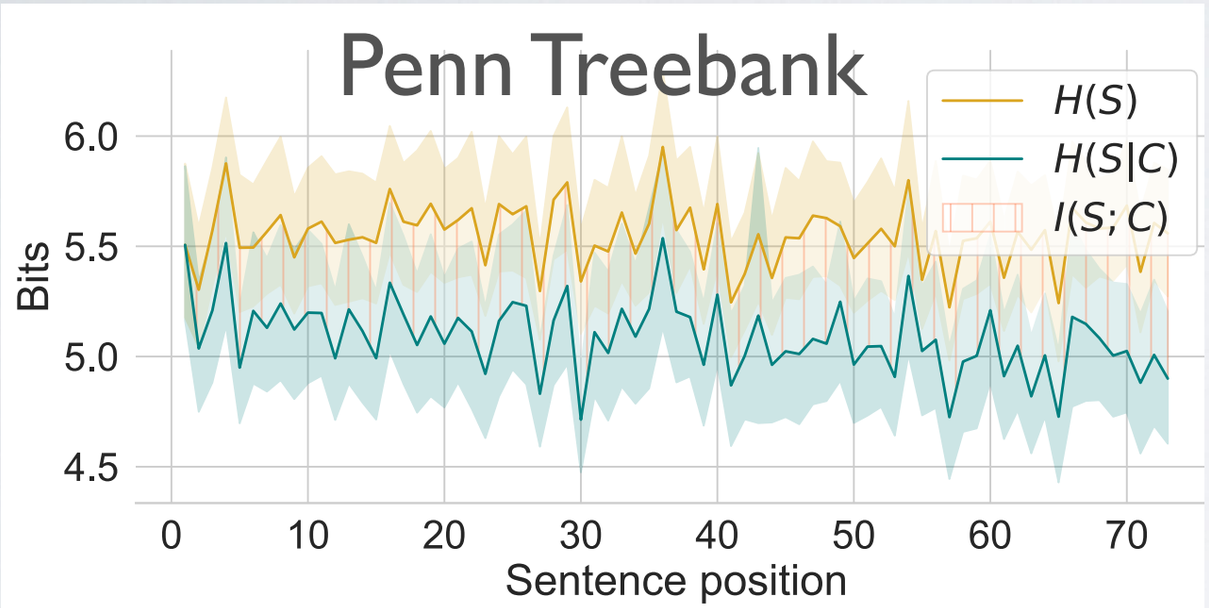


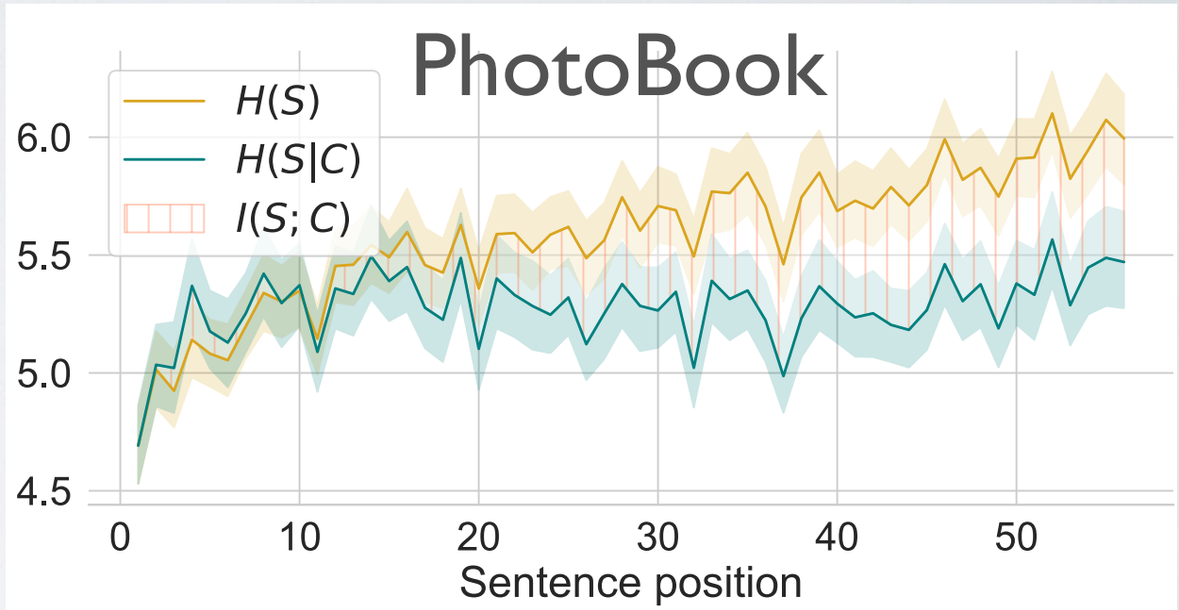
Effect of sentence position

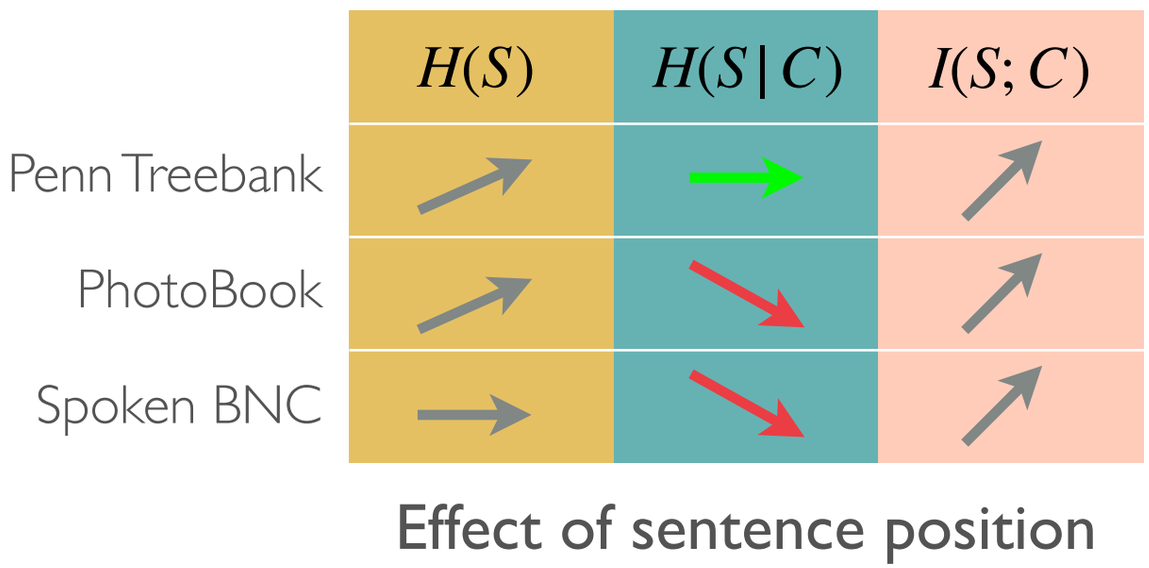


Can surprisal be described as constant?

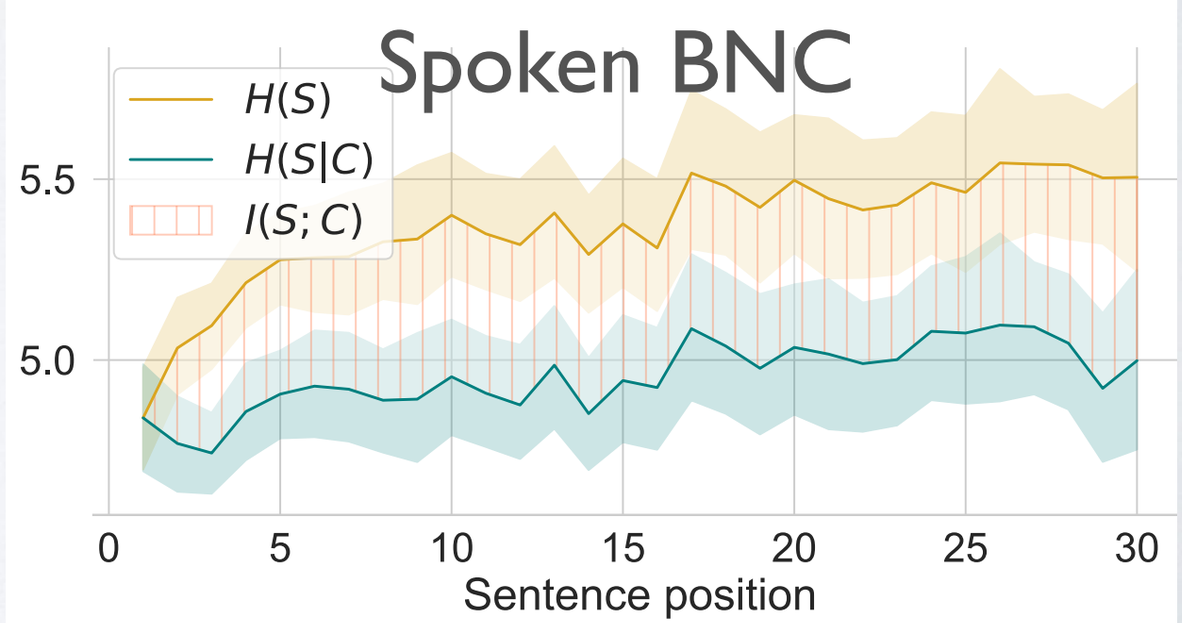








Effect of sentence position



Can surprisal be described as uniform?

Can surprisal be described as uniform?

Criteria of uniformity (Collins, 2014)

Global centrality

$$-\frac{1}{N}\sum_{i=1}^{N}\left(H\left(S_i \mid C_i\right)-\mu\right)^2$$

Local predictability

$$-\frac{1}{N}\sum_{i=2}^{N}\left(H\left(S_i \mid C_i\right)-H\left(S_{i-1} \mid C_{i-1}\right)\right)^2$$

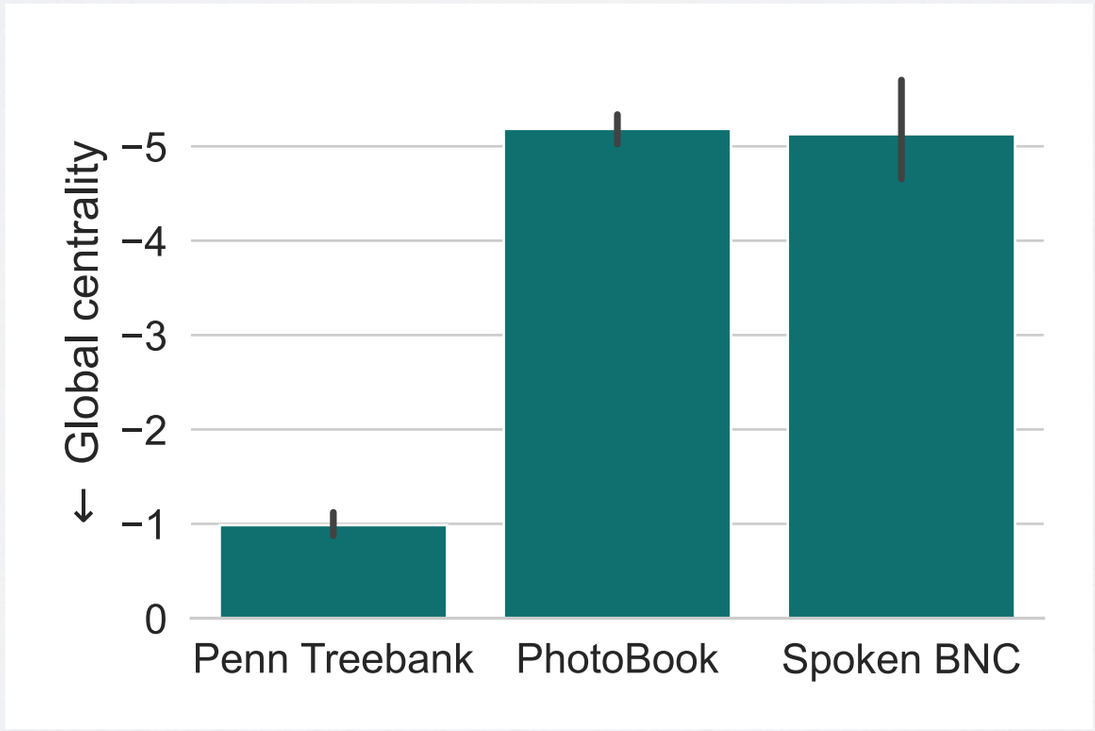
N number of sentences in the text / dialogue

μ average information content in the text / dialogue

Can surprisal be described as uniform?

Global centrality

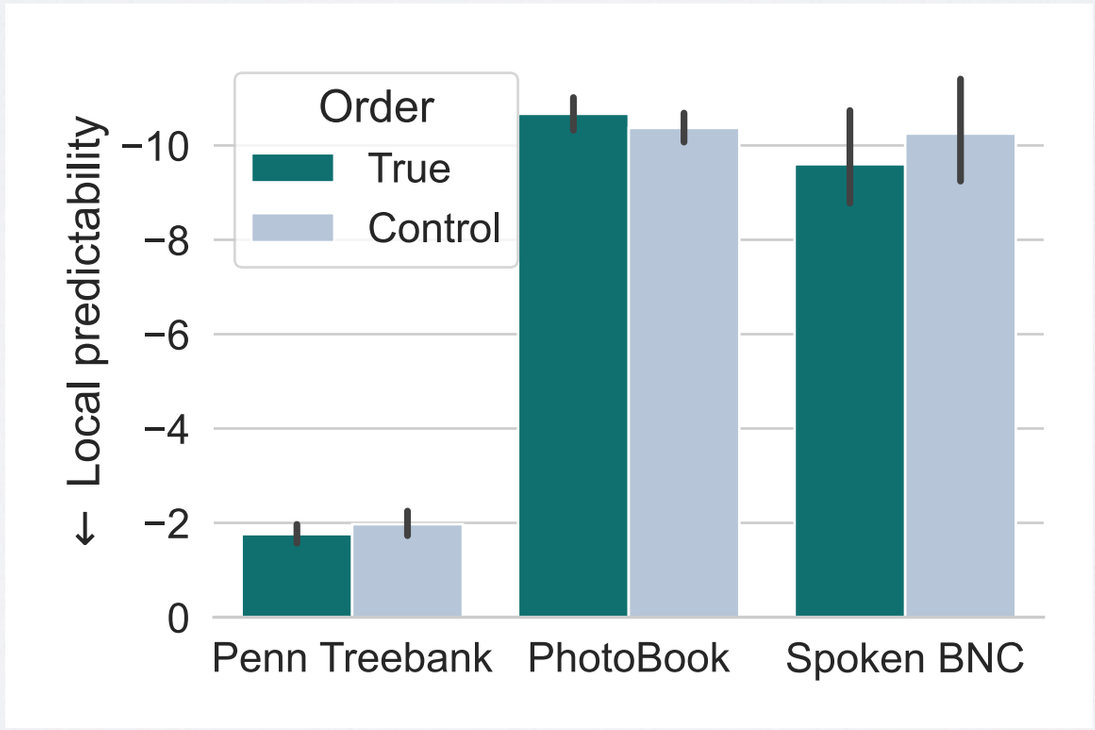
$$-\frac{1}{N}\sum_{i=1}^{N}\left(H\left(S_i \mid C_i\right)-\mu\right)^2$$



Can surprisal be described as uniform?

Local predictability

$$-\frac{1}{N}\sum_{i=2}^{N}\left(H\left(S_i \mid C_i\right) - H\left(S_{i-1} \mid C_{i-1}\right)\right)^2$$



Study 2 - Summary

We have used Transformers language models to obtain surprisal estimates for sentences **within their discourse context**.

In newspaper articles surprisal remains stable,
as predicted by the Constancy Rate Principle.

Surprisal decreases in dialogues:
spoken open domain and written task-oriented dialogues.

Global uniformity is a more faithful criterion than local uniformity.

In dialogues, surprisal is stable within
topically and referentially coherent contextual units…

yet it decreases overall, throughout the entire dialogue

In dialogues, surprisal is stable within
topically and referentially coherent contextual units…

yet it decreases overall, throughout the entire dialogue

Any other efficient production strategies at play?

Study 3

Does construction repetition reduce surprisal? Is it an efficient strategy?

Wrapping up

Information-theoretic model of communication to describe decision making in language production.

Surprisal-based measures of collaborative effort, estimated via an autoregressive neural language model.

Humans follow **efficient strategies**, leading to near-optimal collaborative effort:

- ★ surprisal decreases throughout dialogues (unlike in written texts)
- ★ topically and referentially coherent contextual units show stable (globally uniform) levels of surprisal
- ★ repetitions of non-topical and non-referential expressions (constructions) reduces surprisal

What's next?

Apply this analysis to machine-generated utterances, to **evaluate** adherence to human efficiency principles.

Engineer better **training and decoding objectives** for NLG systems, to bring their production strategies closer to those followed by humans.

Open questions

How to transform information-theoretic measures into evaluation metrics?

How to incorporate them into the training and decoding objectives of NLG systems?

How to improve our model of communication?

- continual adaptation (without catastrophic forgetting)
- partner-specificity
- separate production and processing effort
- include other factors (lexical access, conversational goals, …)

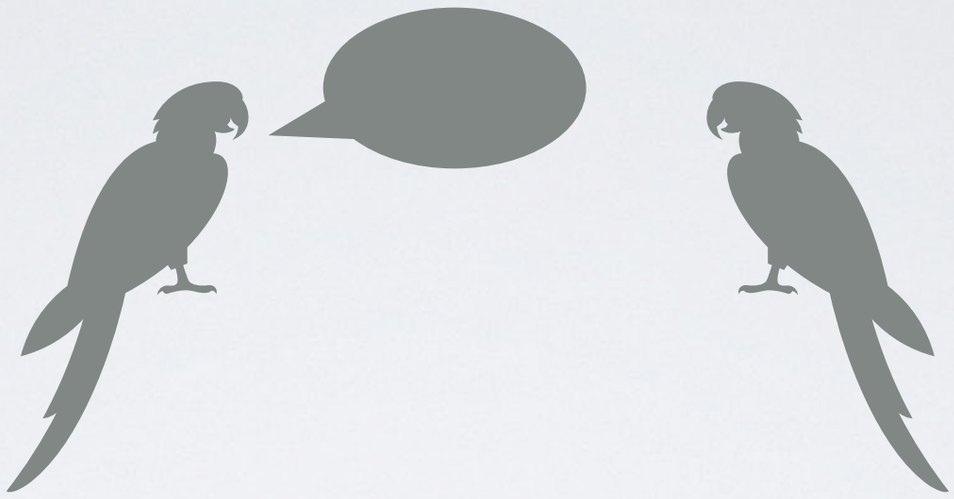
Can we find evidence for efficiency principles in other languages?

APPENDIX

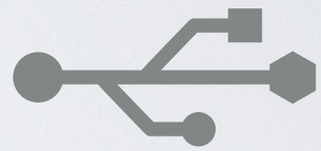
Strategies of language production

- Grice's maxims and the cooperative principle
- Neo-Gricean principles (Horn and Ward, Levinson)
- Relevance theory (Sperber and Wilson)
- Conversation analysis (Sacks, Schegloff, Jefferson)
- Least collaborative effort (Clark)
- Rational Speech Act framework (Frank and Goodman)

Collaborative effort in language production



PROCESSING EFFORT



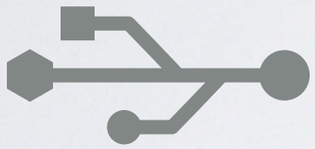
Amcore Financial Inc. said it agreed to acquire Central of Illinois Inc. in a stock swap.

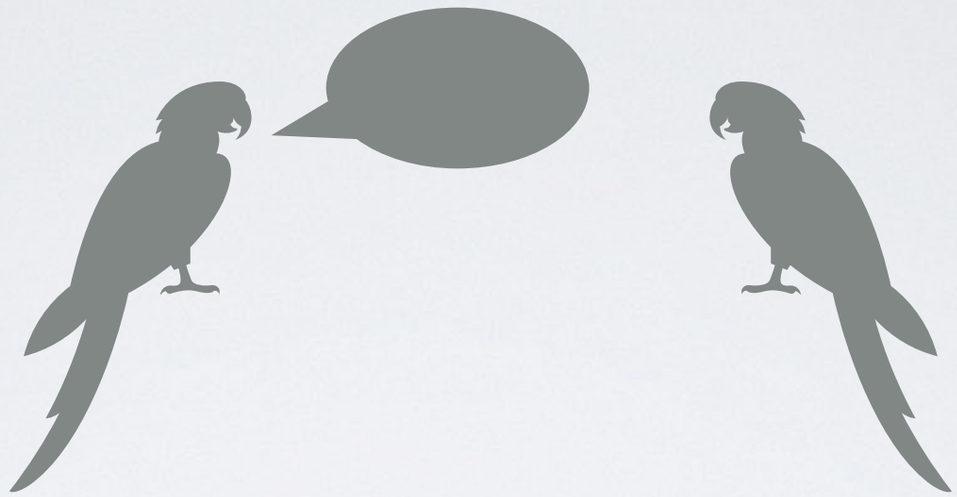
Shareholders of Central, a bank holding company based in Sterling, will receive Amcore stock equal to 10 times Central's 1989 earnings, Amcore said.

For the first nine months of 1989, Central of Illinois Inc., a bank holding company based in Sterling, earned $2 million.

Collaborative effort in language production

PRODUCTION EFFORT





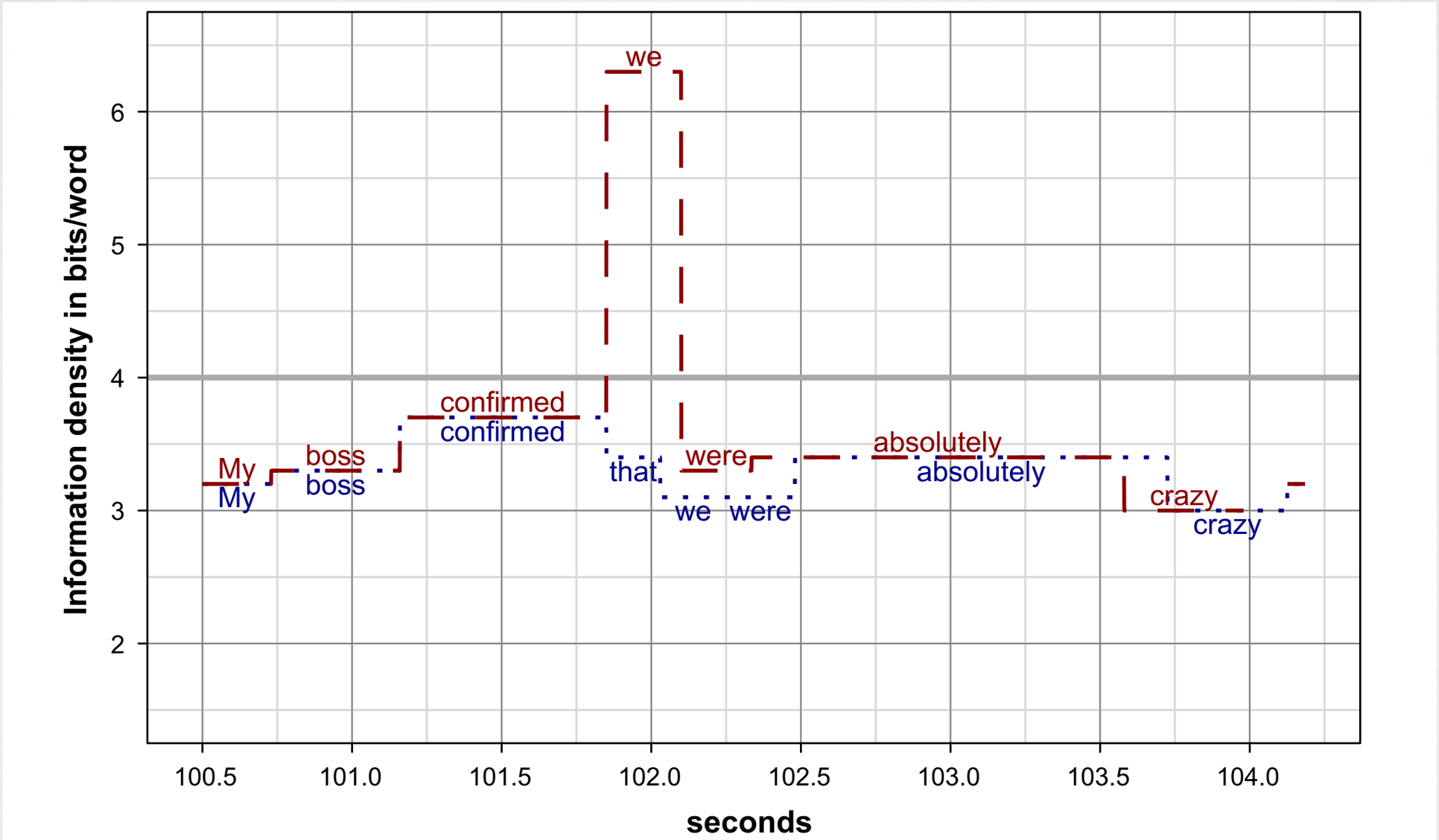
Amcore Financial Inc. said it agreed to acquire Central of Illinois Inc. in a stock swap.

Shareholders of Central, a bank holding company based in Sterling, will receive Amcore stock equal to 10 times Central's 1989 earnings, Amcore said.

For the first nine months of 1989, Central earned $2 million.

Uniform Information Density

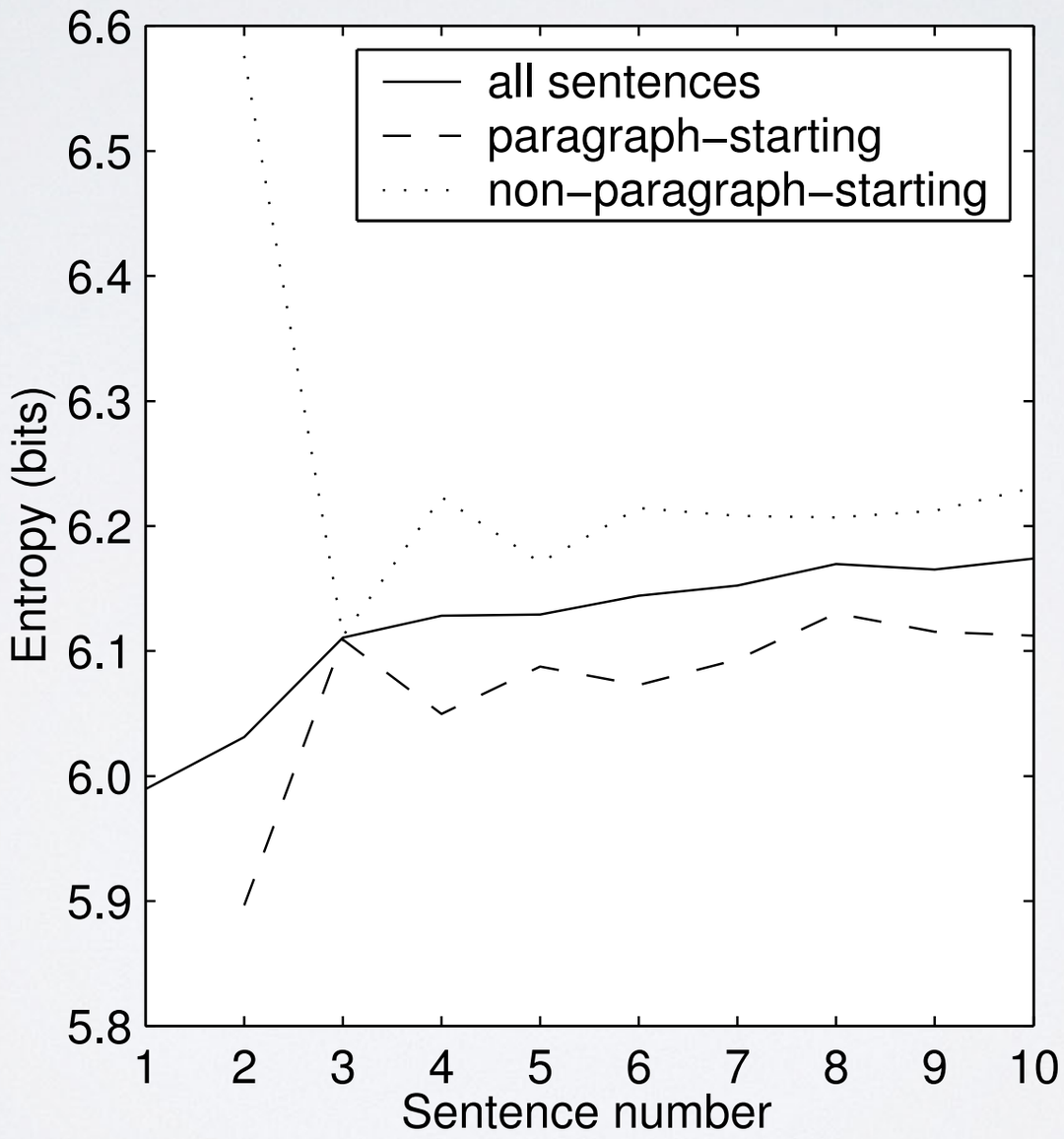
Example 1: Syntax
Complementiser *that*-mentioning (e.g., Jaeger , 2010)



(Jaeger , 2010; Figure 1a)

Uniform Information Density

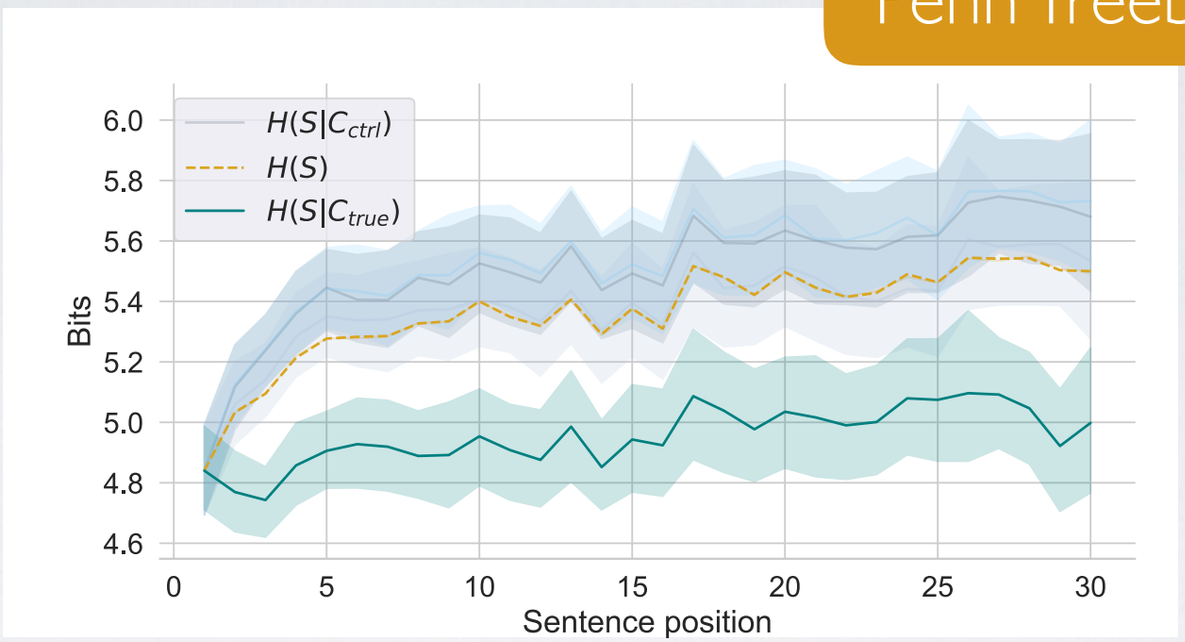
Example 2: Discourse
Entropy rate constancy (Genzel & Charniak, 2002, 2003)



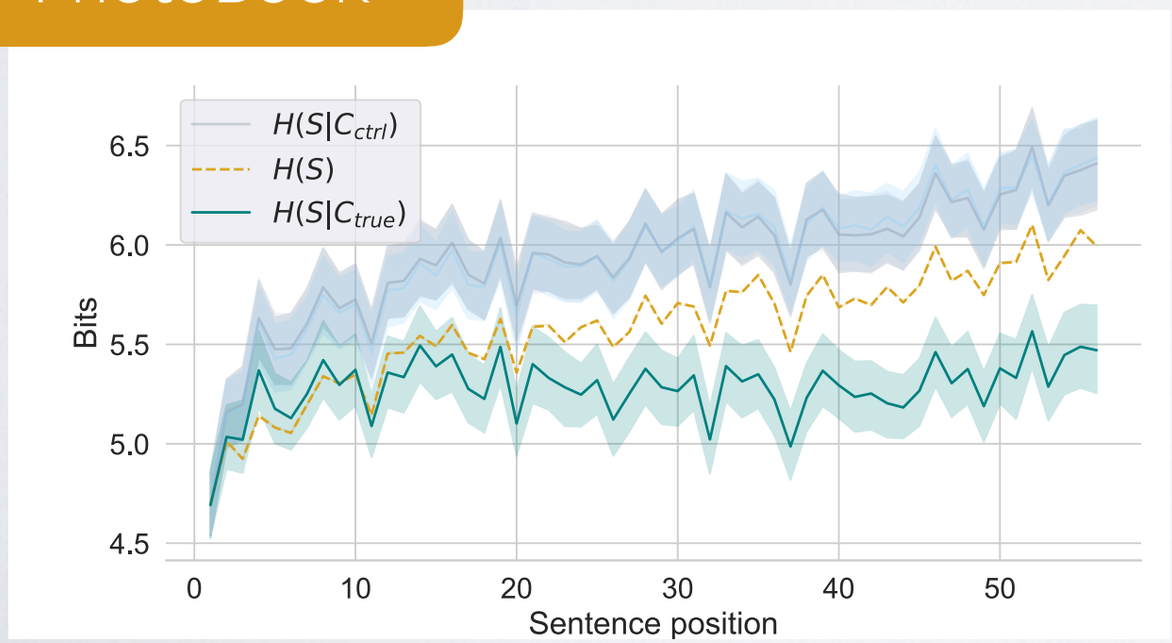
(Genzel & Charniak, 2003; Figure 1)

Control runs

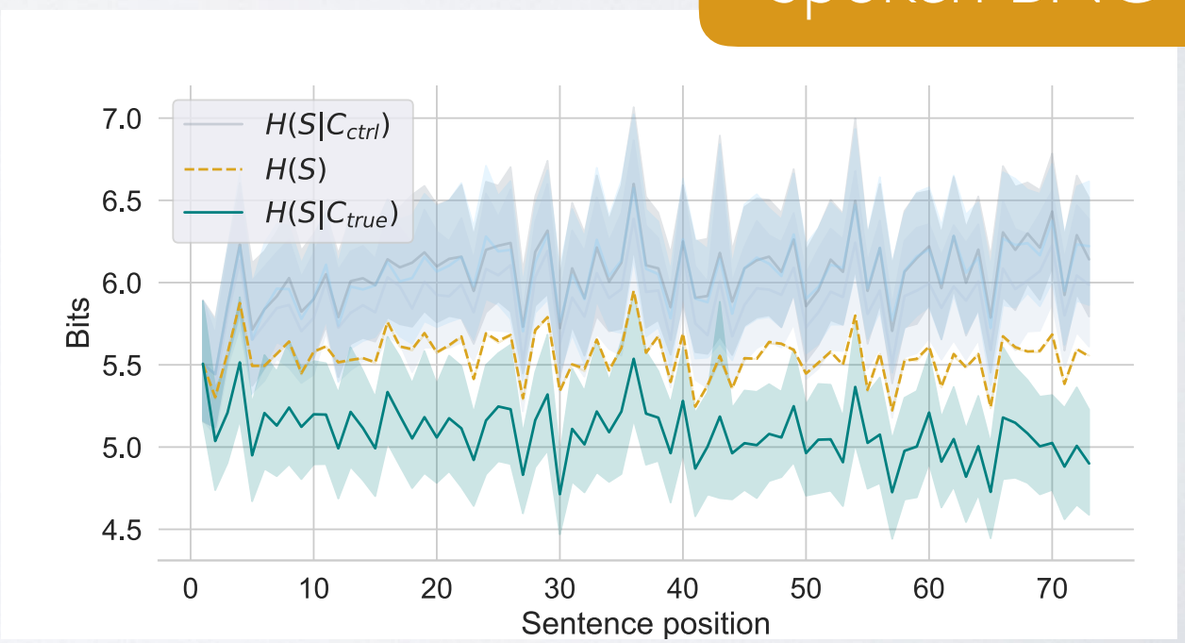
Penn Treebank



PhotoBook

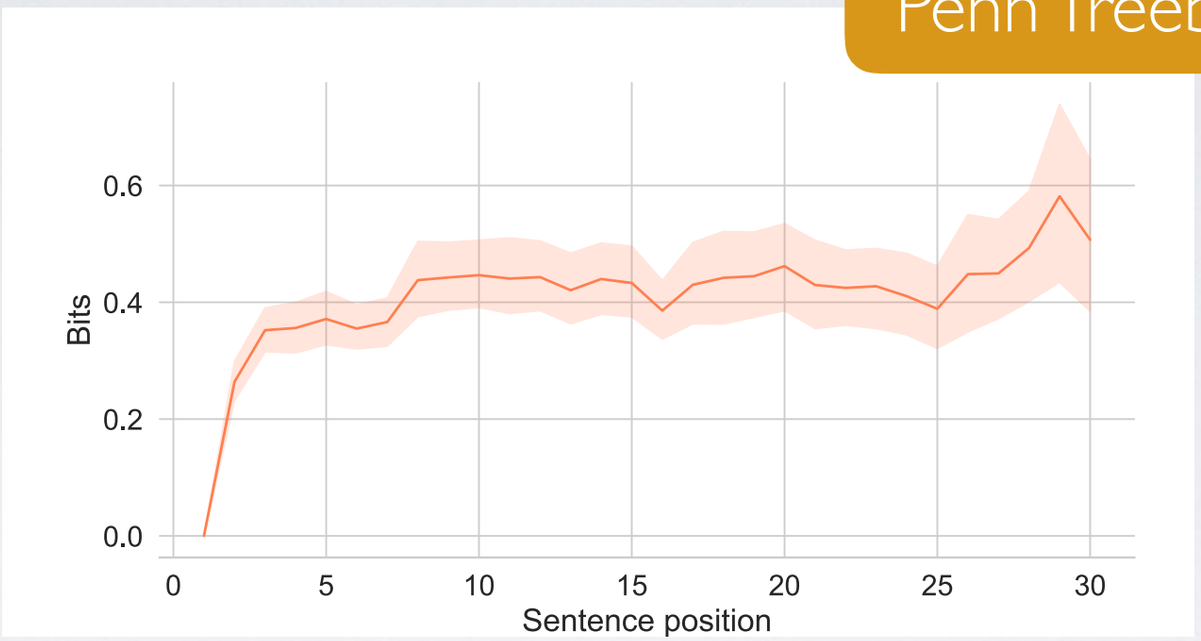


Spoken BNC

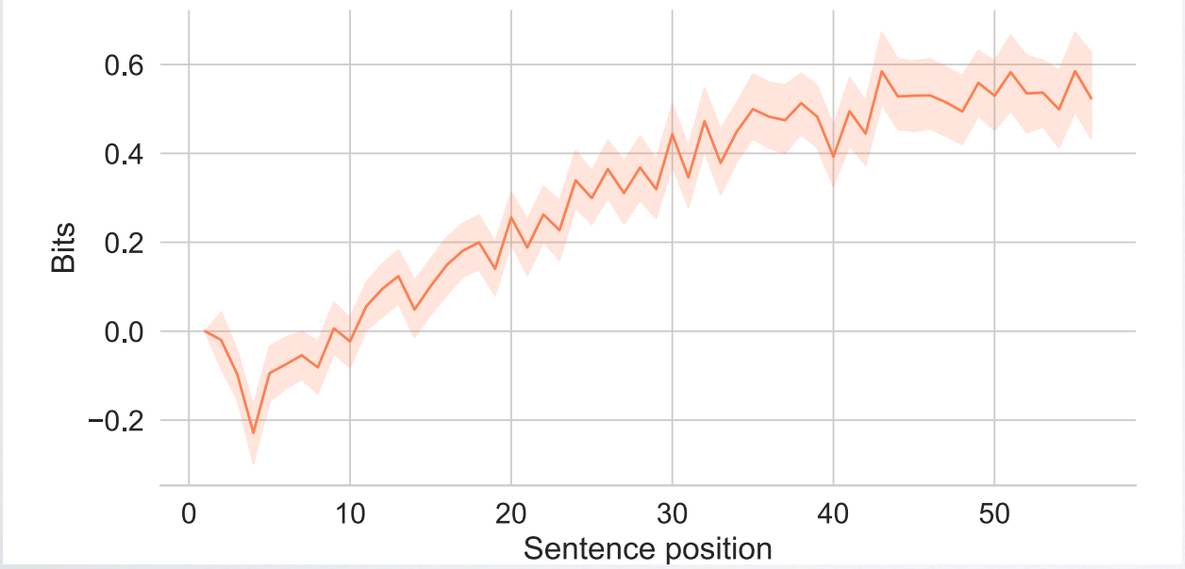


Context informativeness

Penn Treebank



PhotoBook



Spoken BNC