Information Theory And Language

Romain Brasselet, SISSA

09/07/15

Reprinted with corrections from *The Bell System Technical Journal,*
Vol. 27, pp. 379–423, 623–656, July, October, 1948.

A Mathematical Theory of Communication

By C. E. SHANNON

INTRODUCTION

THE recent development of various methods of modulation such as PCM and PPM which exchange bandwidth for signal-to-noise ratio has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important papers of Nyquist[1] and Hartley[2] on this subject. In the present paper we will extend the theory to include a number of new factors, in particular the effect of noise in the channel, and the savings possible due to the statistical structure of the original message and due to the nature of the final destination of the information.

The fundamental problem of communication is that of reproducing at one point either exactly or ap-

Framework of Information Theory

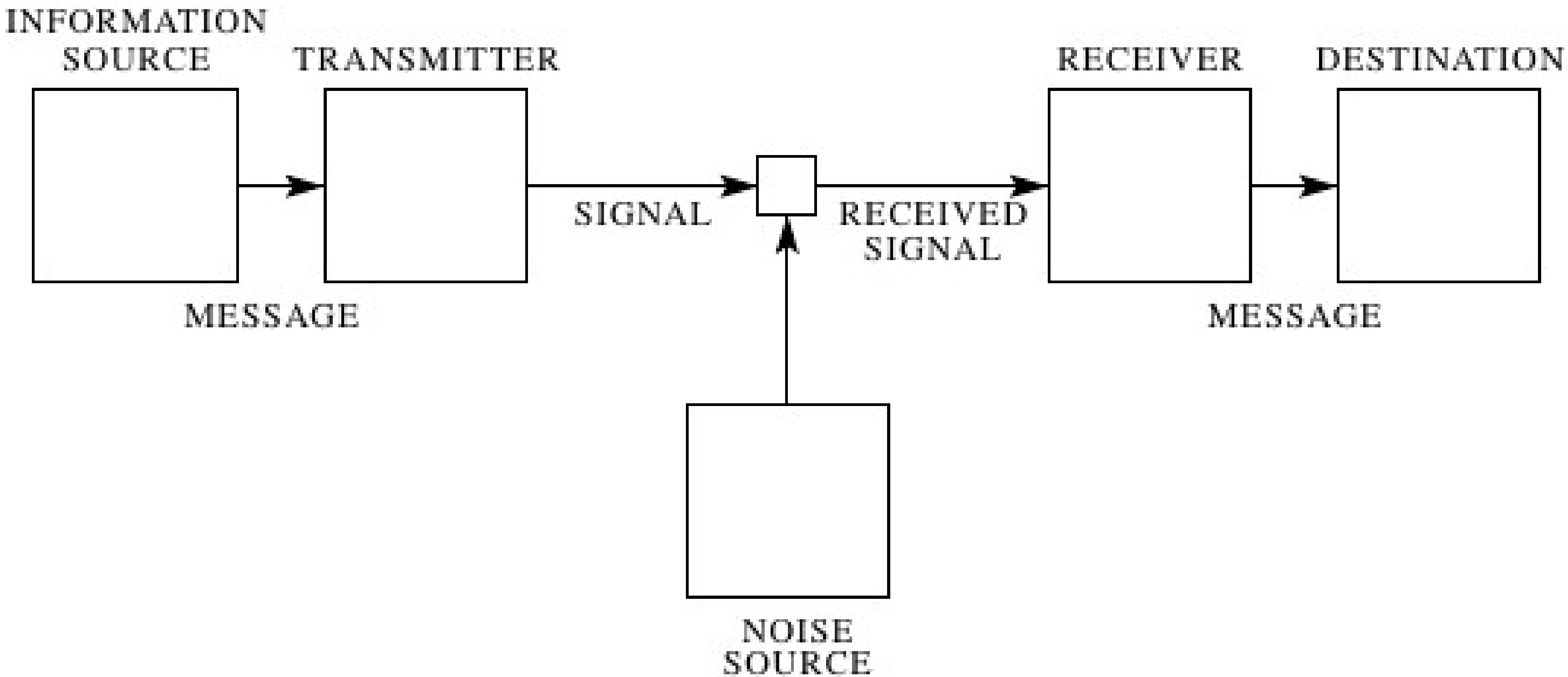
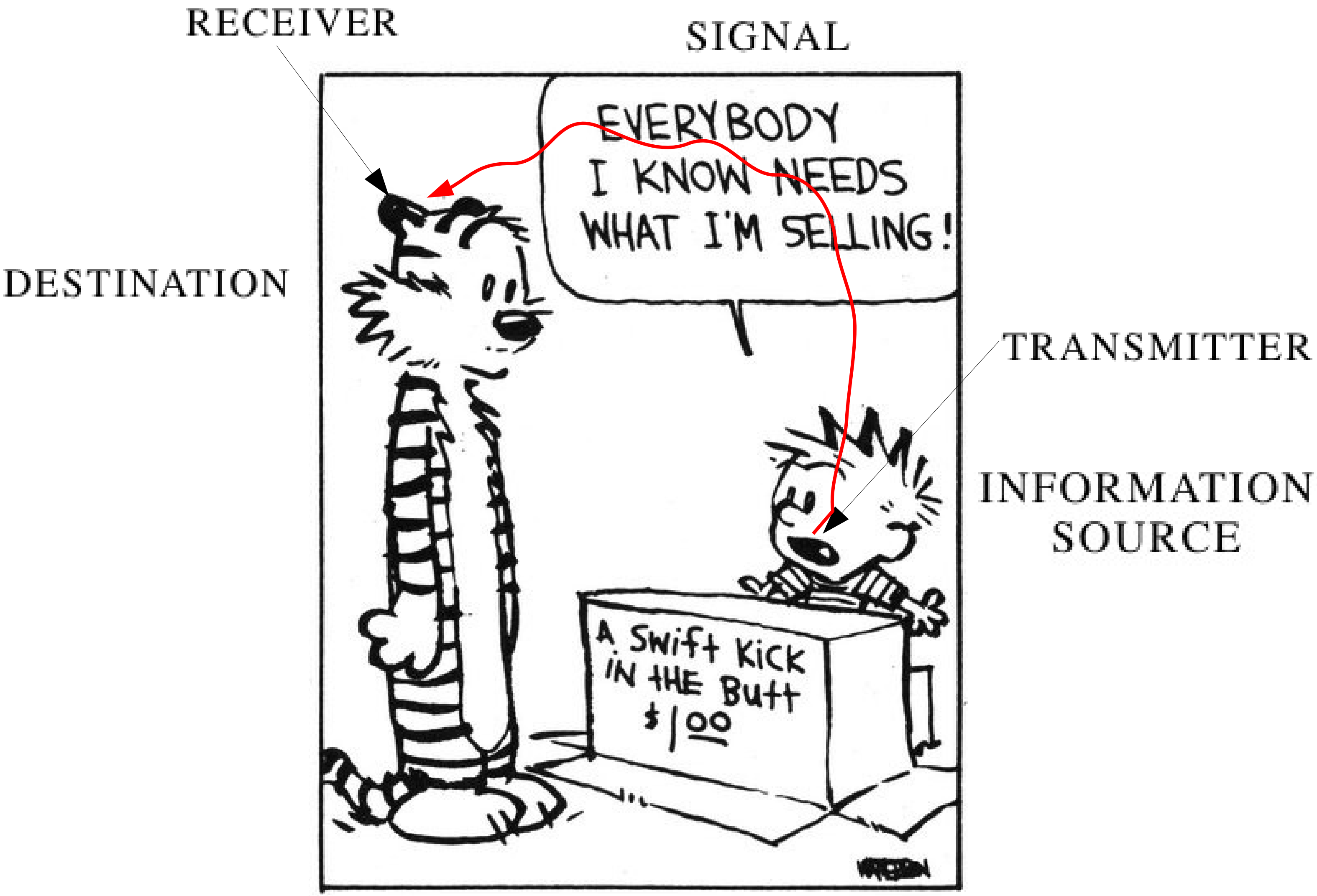
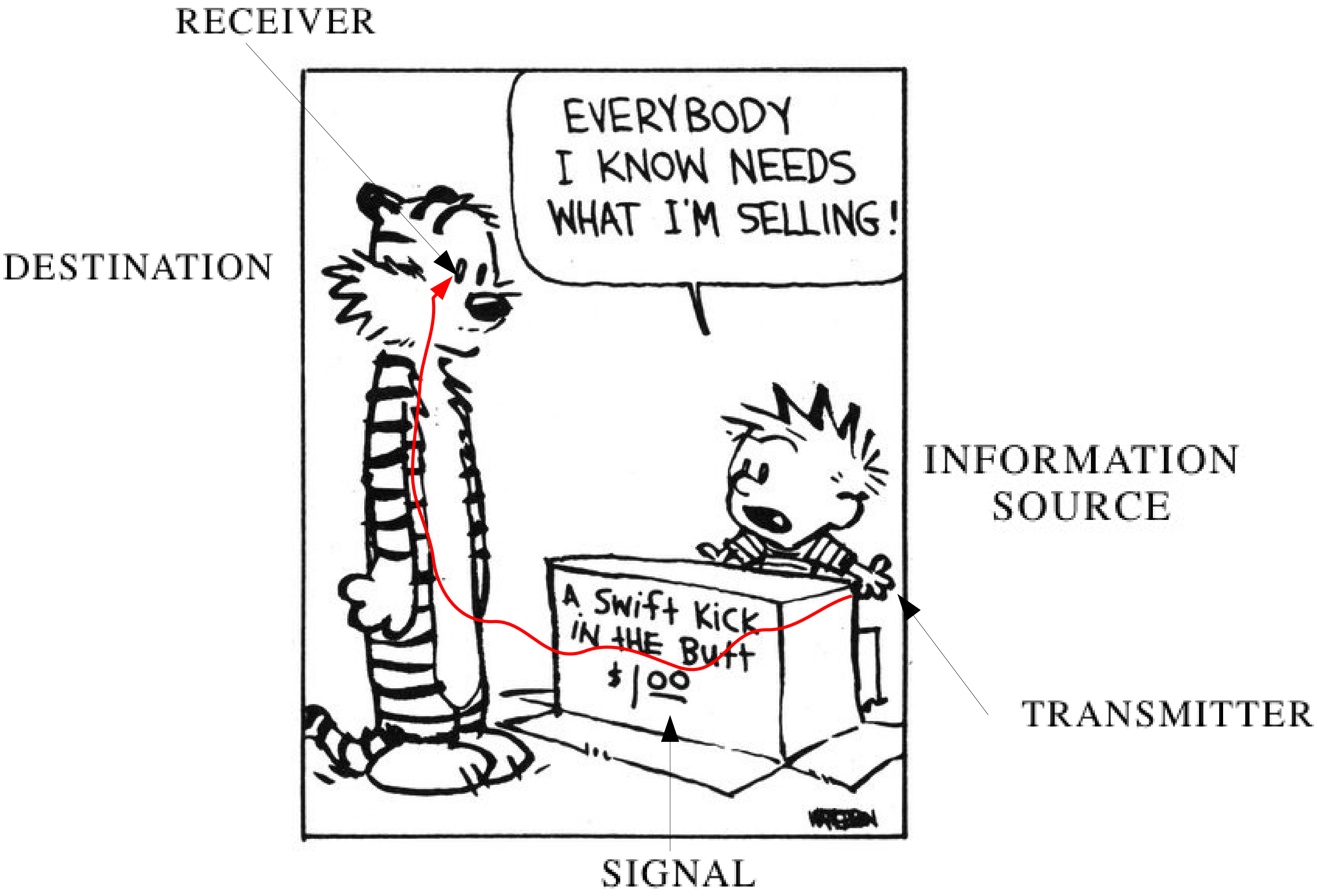


Fig. 1—Schematic diagram of a general communication system.

Exemplified



Exemplified



Entropy and redundancy of written language

Space of letters

h n r

w

j c k e

q

a y

x p

l z m

v g

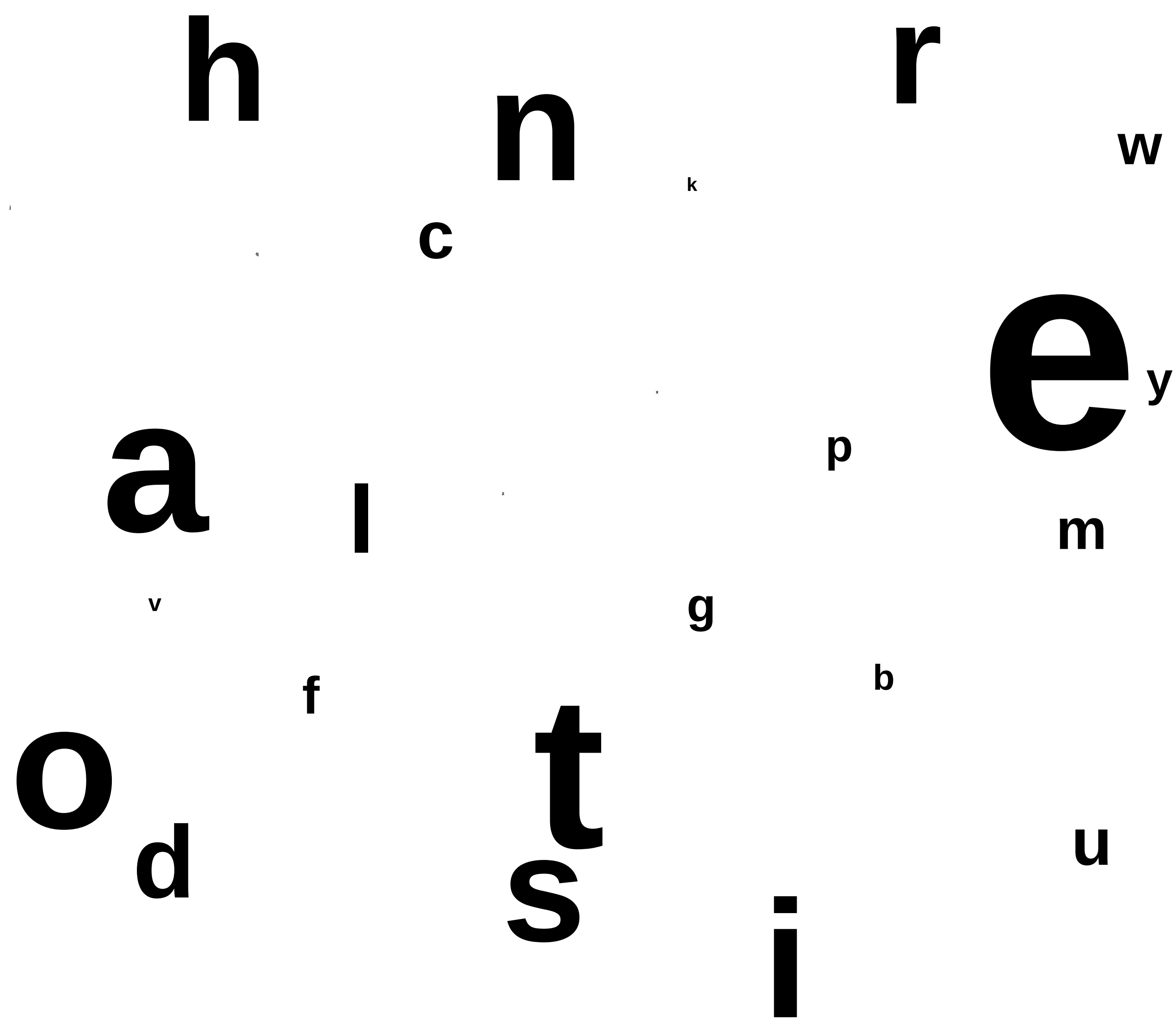
o f t b

d s u

i

→ no correlation analysis

Probabilities



Entropy as a measure of uncertainty and information

$$X = \{x_i\}_{1 \leq i \leq N} \longrightarrow \{p_i\}_{1 \leq i \leq N}$$

$$surprise(x_i) = -\log p(x_i)$$

$$H(X) = < -\log p(x_i) >_i$$

$$H(X) = -\sum_i p(x_i) \log p(x_i)$$

Entropy of a coin

$$H = -(p \log p + q \log q)$$

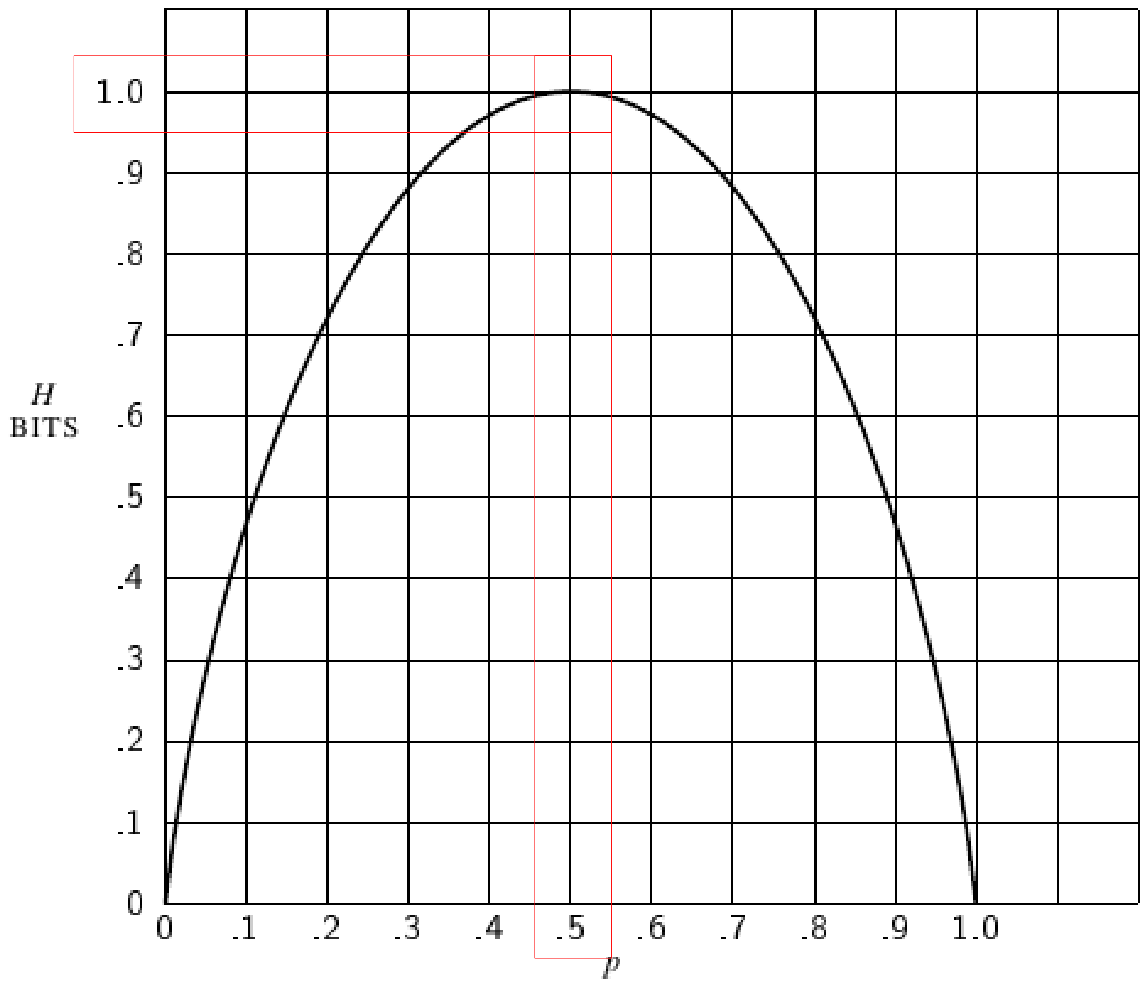
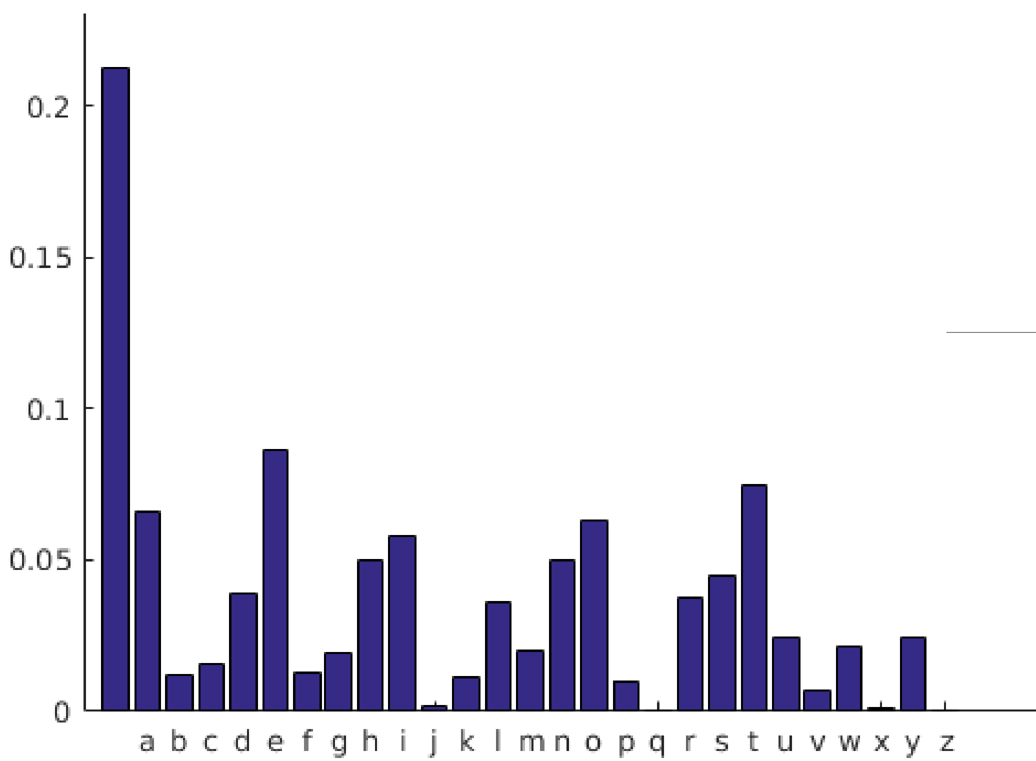


Fig. 7—Entropy in the case of two possibilities with probabilities p and $(1-p)$.

Entropy of language?

$$p(x_i) = \frac{1}{N} \longrightarrow H(X) = \log_2 N = 4.75 bits$$



$$\longrightarrow H(X) = 4.08 bits$$

Redundancy

$$R = 1 - \frac{H(X)}{H_{equi}}$$

$$H(X) = 4.08bits \longrightarrow R = 14\%$$

Redundancy ~ structure

Language has structure and therefore is redundant.

Conditional probabilities

if you really want to hear about it the first thing youll probably want to know is where i was born

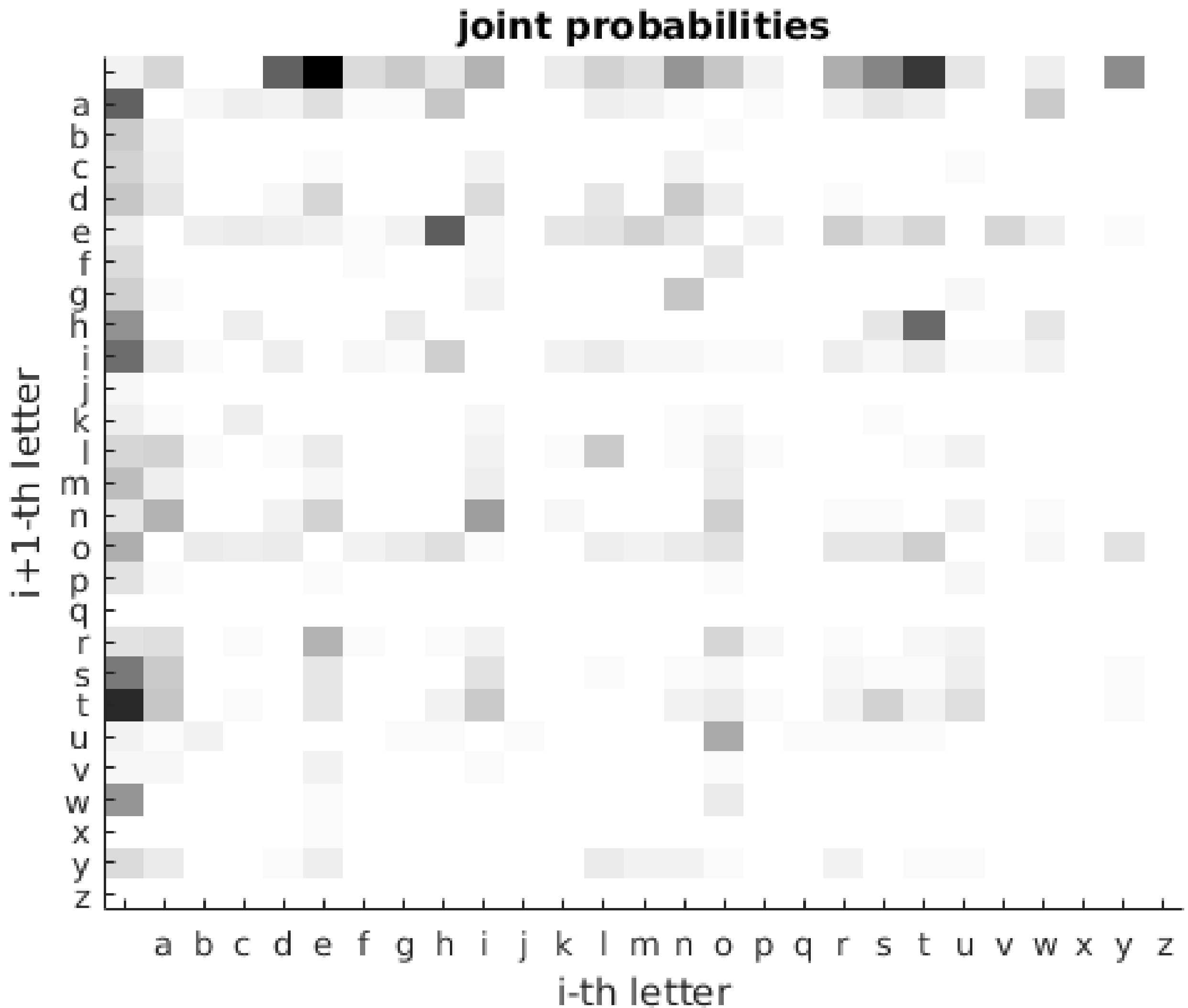
$$p(x^n|x^{n-1}) \neq p(x)$$

$$p(x^n|x^{n-1}, x^{n-2}...) \neq p(x)$$

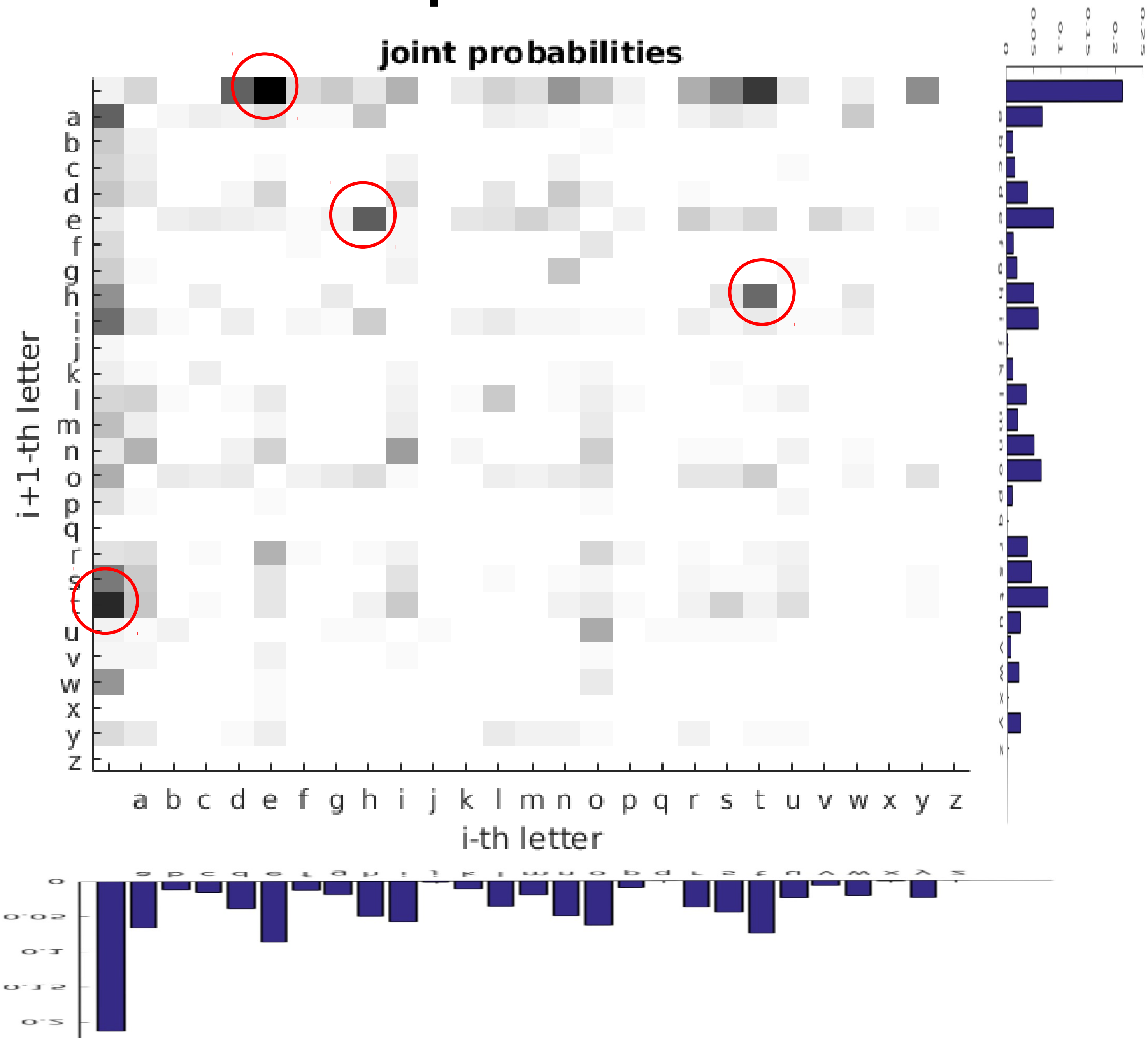
Catcher in the Rye

if you really want to hear about it the first thing you ll probably want to know is where i was born an what my lousy childhood was like and how my parents were occupied and all before they had me and all that david copperfield kind of crap but i don t feel like going into it if you want to know the truth in the first place that stuff bores me and in the second place my parents would have about two hemorrhages apiece if i told anything pretty personal about them they re quite touchy about anything like that especially my father they re nice and all i m not saying that but they re also touchy as hell besides i m not going to tell you my whole goddam autobiography or anything i ll just tell you about this madman stuff that happened to me around last christmas just before i got pretty run down and had to come out here and take it easy i mean that s all i told db about and he s my brother and all he s in hollywood that isn t too far from this crumby place and he comes over and visits me...

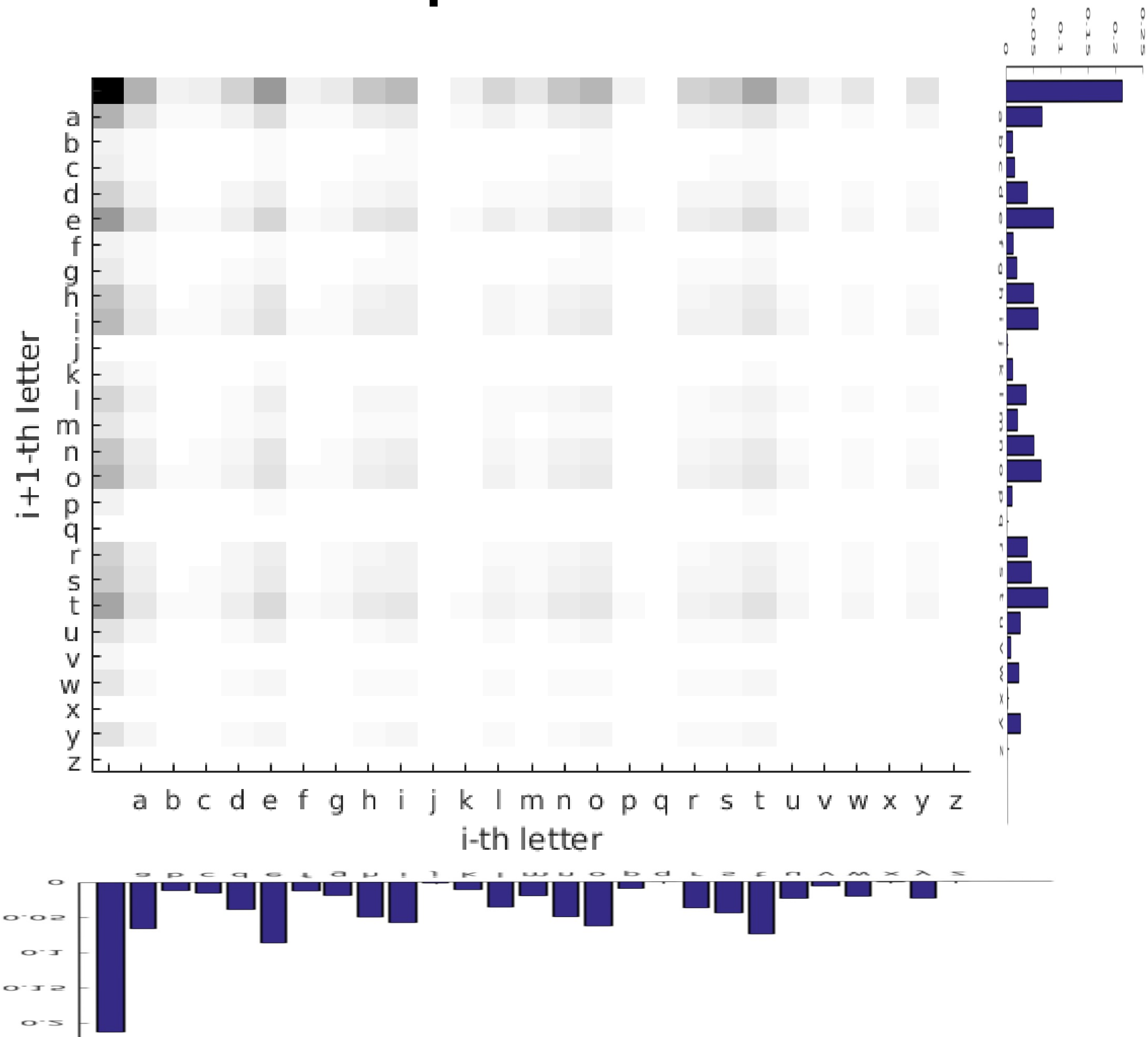
Joint probabilities



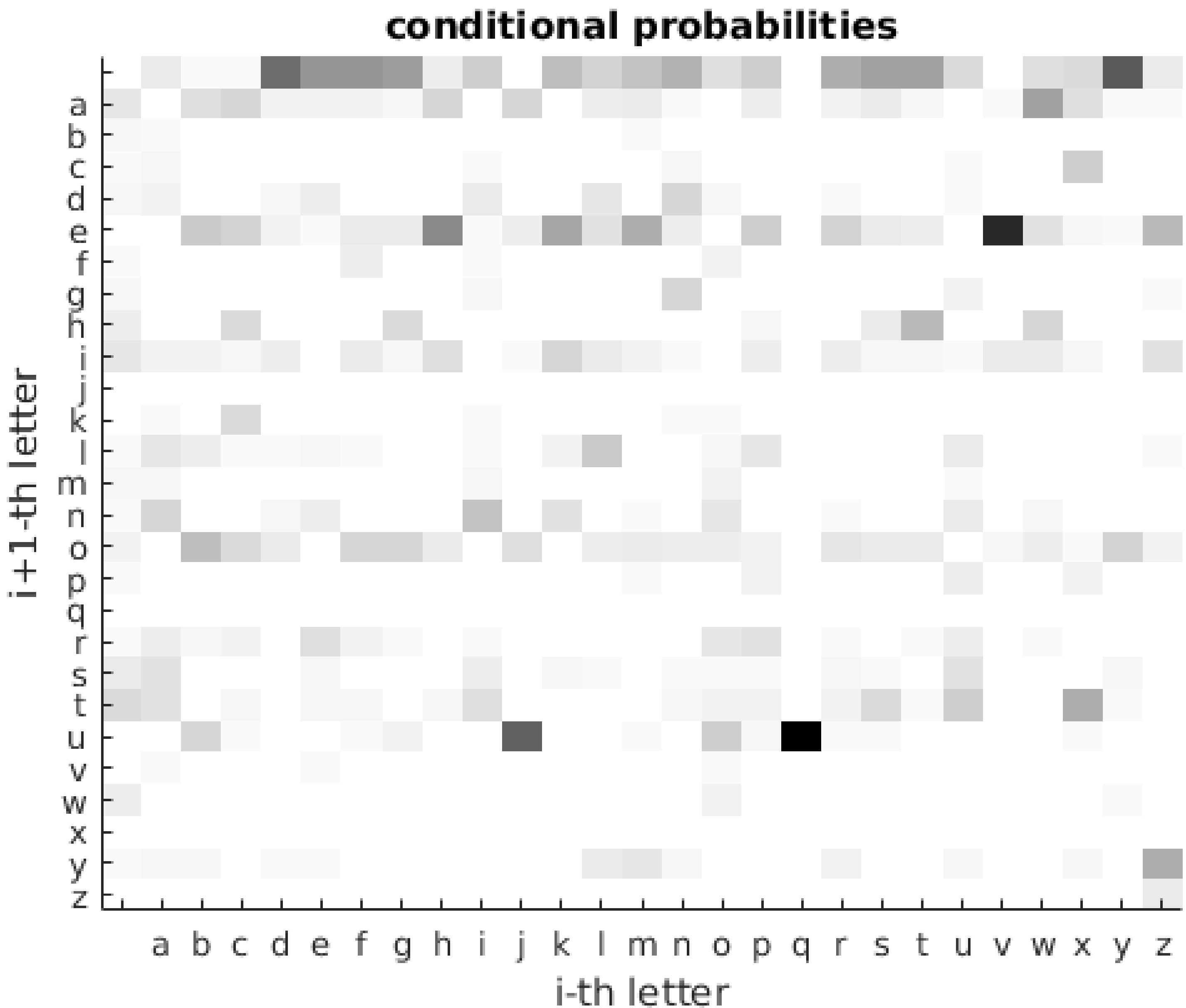
Joint probabilities



Joint probabilities



Conditional probabilities

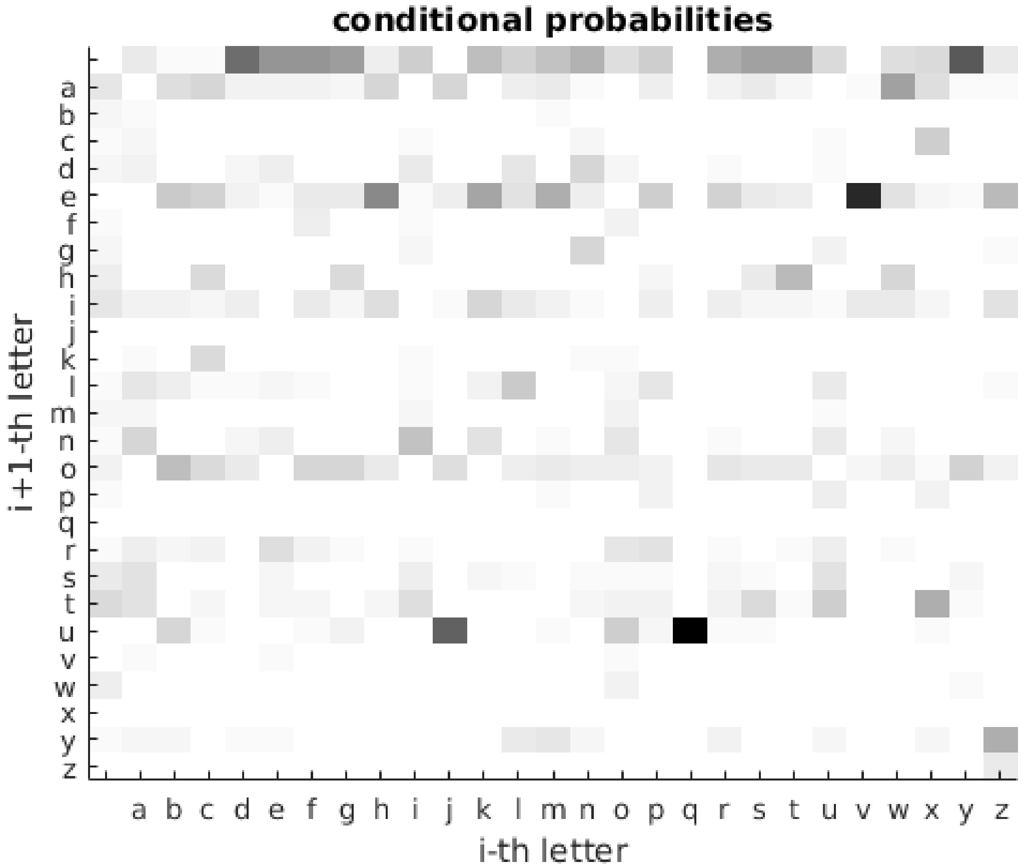


Conditional entropy

$$H(X^n|X^{n-1}) =$$

$$-\sum_{x^{n-1}} p(x^{n-1}) \sum_{x^n} p(x^n|x^{n-1}) \log p(x^n|x^{n-1})$$

Conditional probabilities



$$\longrightarrow H(X) = 3.28 bits$$

$$R = 31\%$$

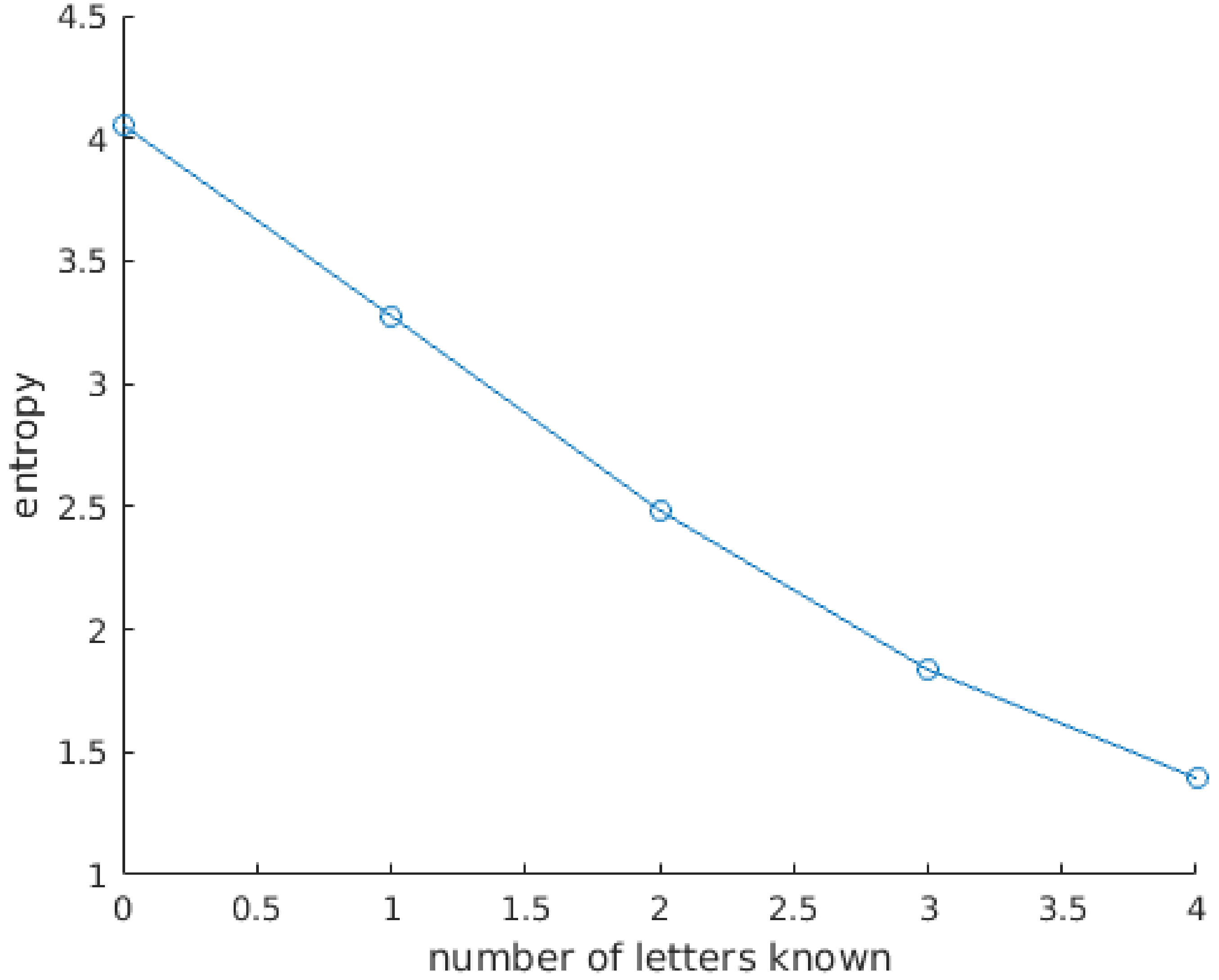
Conditional entropy

$$H(X^n|X^{n-1}) =$$

$$-\sum_{x^{n-1}} p(x^{n-1}) \sum_{x^n} p(x^n|x^{n-1}) \log p(x^n|x^{n-1})$$

$$H(X^n|X^{n-1}, X^{n-2}) = -\sum_{x^{n-2}, x^{n-1}} p(x^{n-2}, x^{n-1})...$$

Entropies


entropy
number of letters known
4.5
4
3.5
3
2.5
2
1.5
1
0
0.5
1
1.5
2
2.5
3
3.5
4

Generation of sentences

myig ohi lunnh p mtoswers h oc llwdn cdsieal tihd r hhhicggnd   w daeasereeoynth
0 iar iehttiomlmele dazoo toede orhsiuee adfatc tfku u uahtd lk tninnorn ena tod oof
tualm lletnsth qiiwoetli s esd t

2

4

Generation of sentences

myig ohi lunnh p mtoswers h oc llwdn cdsieal tihd r hhhicggnd   w daeasereeoynth
0 iar iehttiomlmele dazoo toede orhsiuee adfatc tfku u uahtd lk tninnorn ena tod oof
tualm lletnsth qiiwoetli s esd t

the chat agodding ancid nier ove m fen hin aftelee diall or ando an s jusea pen he not
onting whame the new a sup everse mides he it inee s have ve way i wit she my wit
2 kictle th cradlay to fave sorriven thembeets bally heintice goddamearobvin onsted i
loozencey got hating bon the ater hell the bouldiew hat king ught mid her a pread ing
yout did hand he teeng like hels and peng abou

4

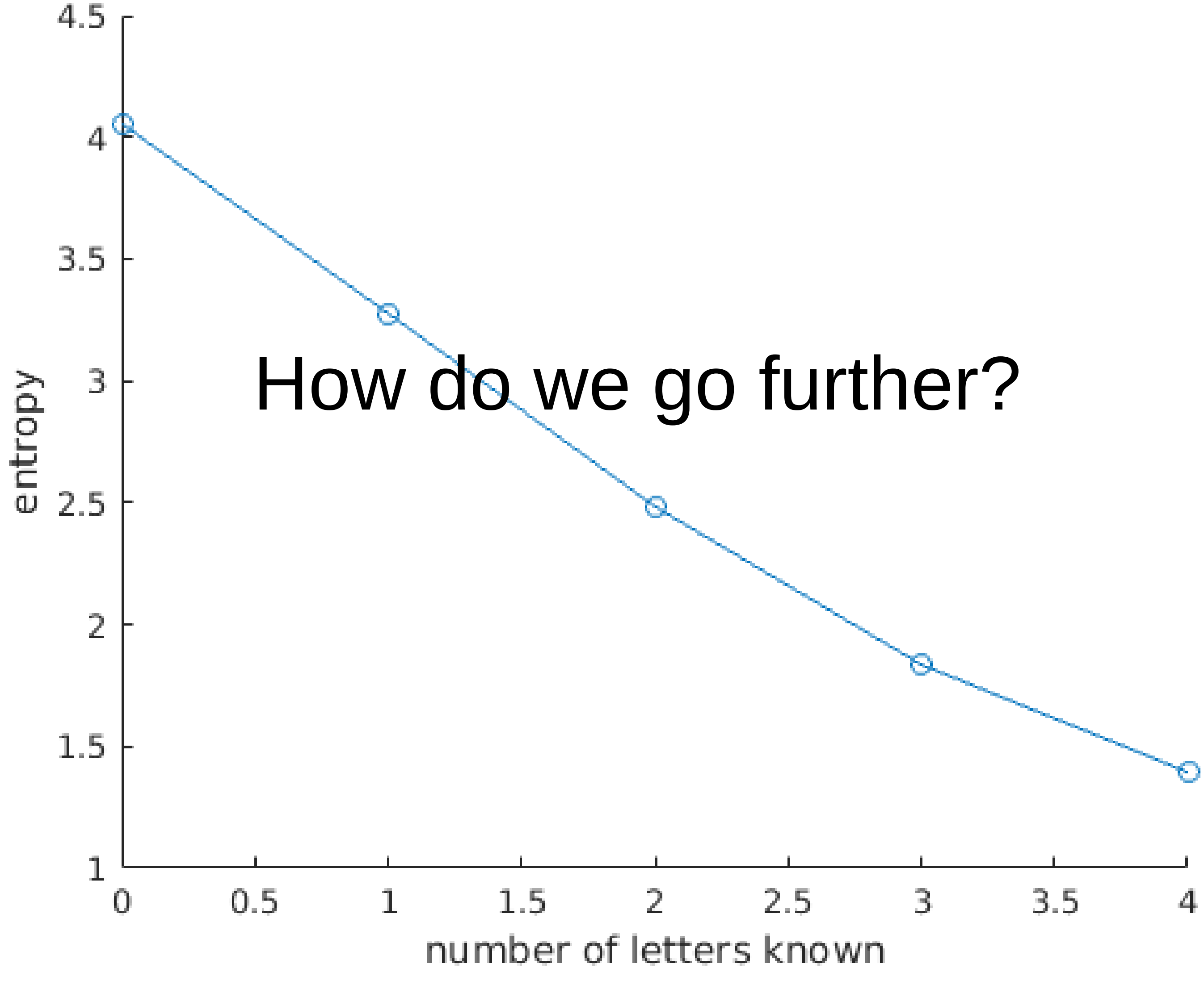
Generation of sentences

myig ohi lunnh p mtoswers h oc llwdn cdsieal tihd r hhhicggnd   w daeasereeoynth
0 iar iehttiomlmele dazoo toede orhsiuee adfatc tfku u uahtd lk tninnorn ena tod oof
tualm lletnsth qiiwoetli s esd t

the chat agodding ancid nier ove m fen hin aftelee diall or ando an s jusea pen he not
onting whame the new a sup everse mides he it inee s have ve way i wit she my wit
2 kictle th cradlay to fave sorriven thembeets bally heintice goddamearobvin onsted i
loozencey got hating bon the ater hell the bouldiew hat king ught mid her a pread ing
yout did hand he teeng like hels and peng abou

the crumby bar when i got him except giving out gear her and running teachests at
pretty were this guts i could hartzell over man keep you re you happened about a
4 *handshaking her i have one of stuff they probably hurt sort of my hardy up at the was*
the d even he hardly guy   right and parents were s goddam hound none comed and
that we got booth

Entropies



How do we go further?

Prediction and Entropy of Printed English

By C. E. SHANNON

(*Manuscript Received Sept. 15, 1950*)

A new method of estimating the entropy and redundancy of a language is described. This method exploits the knowledge of the language statistics possessed by those who speak the language, and depends on experimental results in prediction of the next letter when the preceding text is known. Results of experiments in prediction are given, and some properties of an ideal predictor are developed.

Shannon's guessing game 1

(1) THE ROOM WAS NOT VERY LIGHT A SMALL OBLONG

(2) ----ROO------NOT-V-----I------SM----OBL----

(1) READING LAMP ON THE DESK SHED GLOW ON

(2) REA----------O------D----SHED-GLO--O--

(1) POLISHED WOOD BUT LESS ON THE SHABBY RED CARPET

(2) P-L-S-----O---BU--L-S--O------SH-----RE--C------

Shannon's guessing game 2

(1) T H E R E I S N O R E V E R S E O N A M O T O R C Y C L E A
(2) 1 1 1 5 1 1 2 1 1 2 1 1 15 1 17 1 1 1 2 1 3 2 1 2 2 7 1 1 1 1 4 1 1 1 1 1 3 1

(1) F R I E N D O F M I N E F O U N D T H I S O U T
(2) 8 6 1 3 1 1 1 1 1 1 1 1 1 1 1 6 2 1 1 1 1 1 1 2 1 1 1 1 1 1

(1) R A T H E R D R A M A T I C A L L Y T H E O T H E R D A Y
(2) 4 1 1 1 1 1 1 11 5 1 1 1 1 1 1 1 1 1 1 1 6 1 1 1 1 1 1 1 1 1 1 1 1 (9)

Shannon's guessing game 2

(1) T H E R E I S N O R E V E R S E O N A M O T O R C Y C L E A
(2) 1 1 1 5 1 1 2 1 1 2 1 1 15 1 17 1 1 1 2 1 3 2 1 2 2 7 1 1 1 1 4 1 1 1 1 1 3 1
(1) F R I E N D O F M I N E F O U N D T H I S O U T
(2) 8 6 1 3 1 1 1 1 1 1 1 1 1 1 1 6 2 1 1 1 1 1 1 2 1 1 1 1 1 1
(1) R A T H E R D R A M A T I C A L L Y T H E O T H E R D A Y
(2) 4 1 1 1 1 1 1 1 1 5 1 1 1 1 1 1 1 1 1 1 1 6 1 1 1 1 1 1 1 1 1 1 1 1 1 1   (9)

what letter? → what guess?

$$p(1) = \sum_{Ngrams} p(Ngram) \max_j p(j|Ngram)$$

$$p(2) = \sum_{Ngrams} p(Ngram) \max_j^2 p(j|Ngram)$$

$$H(X) = -\sum_i p(x_i) \log p(x_i)$$

Guessing game

TABLE I

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	100
1	18.2	29.2	36	47	51	58	48	66	66	67	62	58	66	72	60	80
2	10.7	14.8	20	18	13	19	17	15	13	10	9	14	9	6	18	7
3	8.6	10.0	12	14	8	5	3	5	9	4	7	7	4	9	5	
4	6.7	8.6	7	3	4	1	4	4	4	4	5	6	4	3	5	3
5	6.5	7.1	1	1	3	4	3	6	1	6	5	2	3			4
6	5.8	5.5	4	5	2	3	2			1	4	2	3	4	1	2
7	5.6	4.5	3	3	2	2	8		1	1	1	4	1		4	1
8	5.2	3.6	2	2	1	1	2	1	1	1	1		2	1	3	
9	5.0	3.0	4		5	1	4		2	1	1	2		1		1
10	4.3	2.6	2	1	3		3	1					2			
11	3.1	2.2	2	2	2	1			1	3		1	1	2	1	
12	2.8	1.9	4		2	1	1	1			2	1	1		1	1
13	2.4	1.5	1	1	1	1	1	1	1	1		1	1			
14	2.3	1.2		1			1					1				1
15	2.1	1.0	1	1							1	1	1			
16	2.0	.9					1			1					1	
17	1.6	.7	1		2	1	1				1		2	2		
18	1.6	.5													1	
19	1.6	.4			1	1			1		1					
20	1.3	.3		1		1	1									
21	1.2	.2														
22	.8	.1														
23	.3	.1														
24	.1	.0														
25	.1															
26	.1															
27	.1															

Entropy of written english

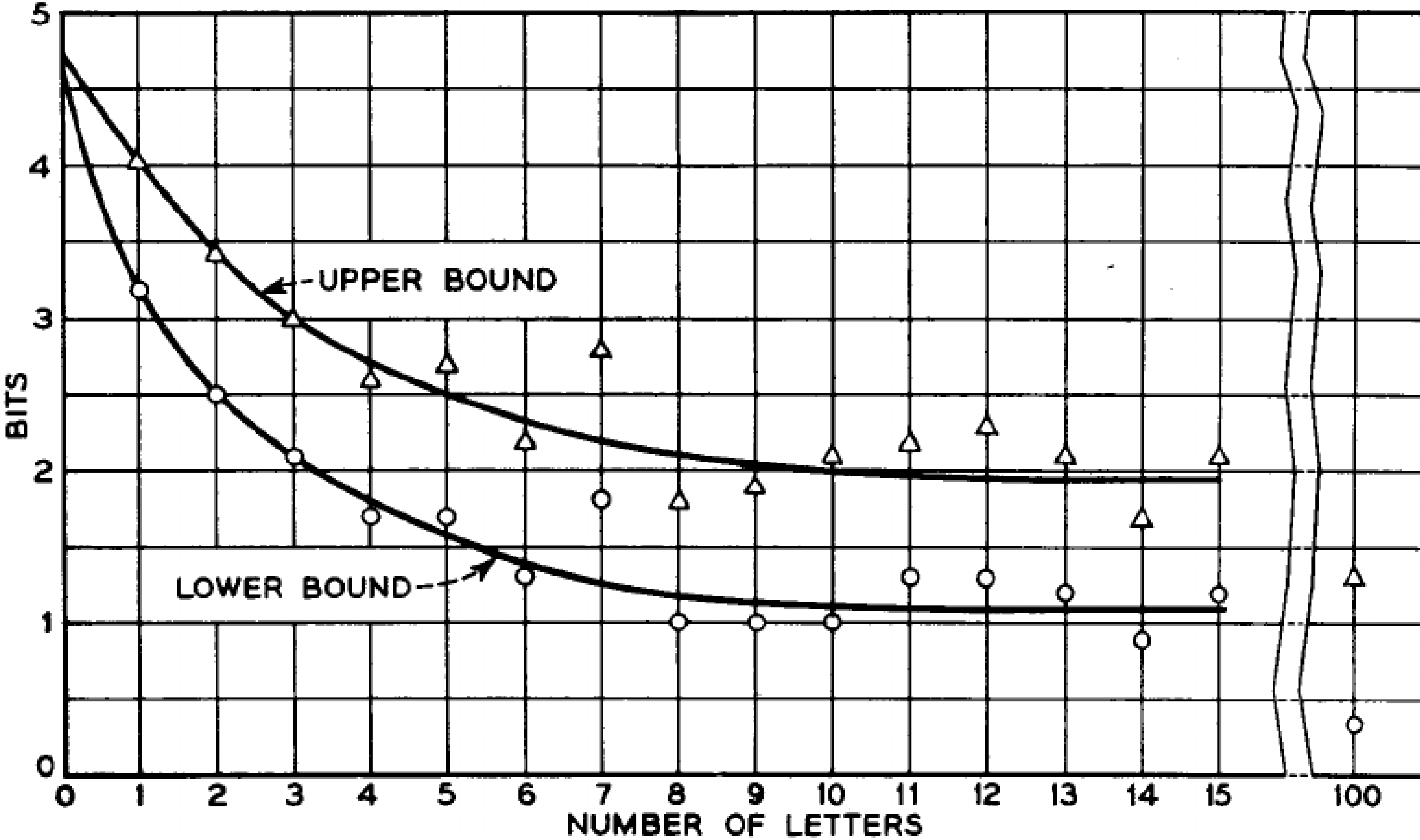


Fig. 4—Upper and lower experimental bounds for the entropy of 27-letter English.

However...

The entropy of the code depends on the writer.

be realized that English is generated by many sourccs, and each source has its own characteristic entropy. The operational meaning of entropy is clear. It is the minimum expected number of bits/symbol necessary for the characterization of the text. A gambling approach will yield an

The guessing game depends on the knowledge of the reader.

Thus an intelligent well-educated gambler will do better than a gambler untrained in quantitative thinking who is relatively unfamiliar with the language. Nonetheless, it will be true that there is an upper bound on how well a gambler can do. If there were no such bound, then the true entropy of the creative process of the writer would be zero and his writing totally predictable. This upper bound

Cover and King, IEEE transactions on Information Theory 1978

Source coding theorem

Importance of redundancy

- Redundancy is a measure of how efficiently symbols are used.
- It is a sign of structure in the language.
- It reduces communication rate but increases predictability.
- Redundancy allows us to reconstruct noisy signals:
  "Turn phat mufic down"

- We can see language as a compromise between information and redundancy.

Zipf's law

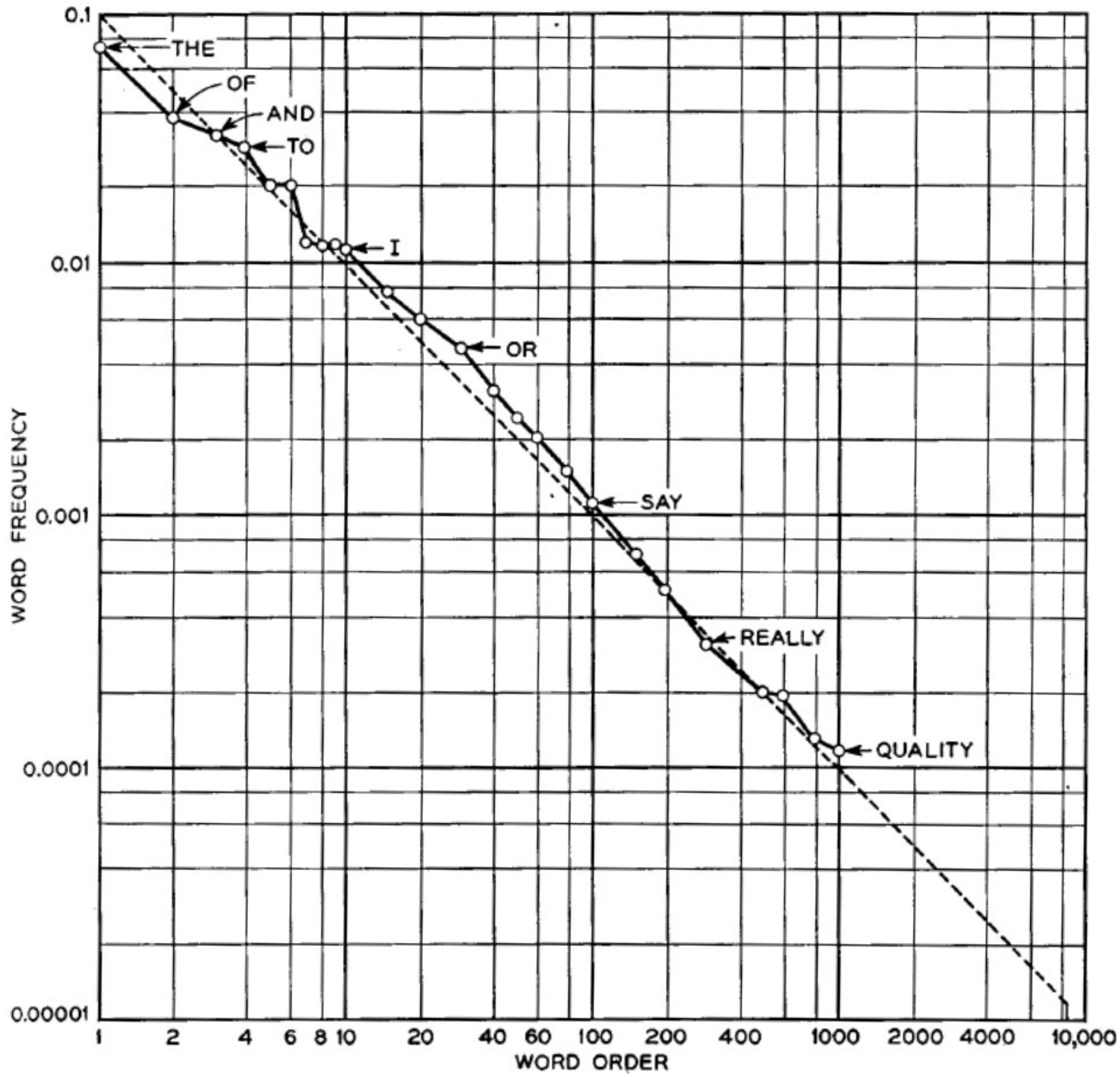


Fig. 1—Relative frequency against rank for English words.

$$p_n = \frac{0.1}{n}$$

$$\sum_1^N \frac{0.1}{n} = 1$$

$$N = 12370$$

Word entropy

$$H(W) = -\sum_n p_n \log p_n = 9.83 bits$$

Universal Entropy of Word Ordering Across Linguistic Families

Marcelo A. Montemurro[1]*, Damián H. Zanette[2]

1 The University of Manchester, Manchester, United Kingdom, **2** Consejo Nacional de Investigaciones Científicas y Técnicas, Centro Atómico Bariloche and Instituto Balseiro, San Carlos de Bariloche, Argentina

Abstract

Background: The language faculty is probably the most distinctive feature of our species, and endows us with a unique ability to exchange highly structured information. In written language, information is encoded by the concatenation of basic symbols under grammatical and semantic constraints. As is also the case in other natural information carriers, the resulting symbolic sequences show a delicate balance between order and disorder. That balance is determined by the interplay between the diversity of symbols and by their specific ordering in the sequences. Here we used entropy to quantify the contribution of different organizational levels to the overall statistical structure of language.

Methodology/Principal Findings: We computed a relative entropy measure to quantify the degree of ordering in word sequences from languages belonging to several linguistic families. While a direct estimation of the overall entropy of language yielded values that varied for the different families considered, the relative entropy quantifying word ordering presented an almost constant value for all those families.

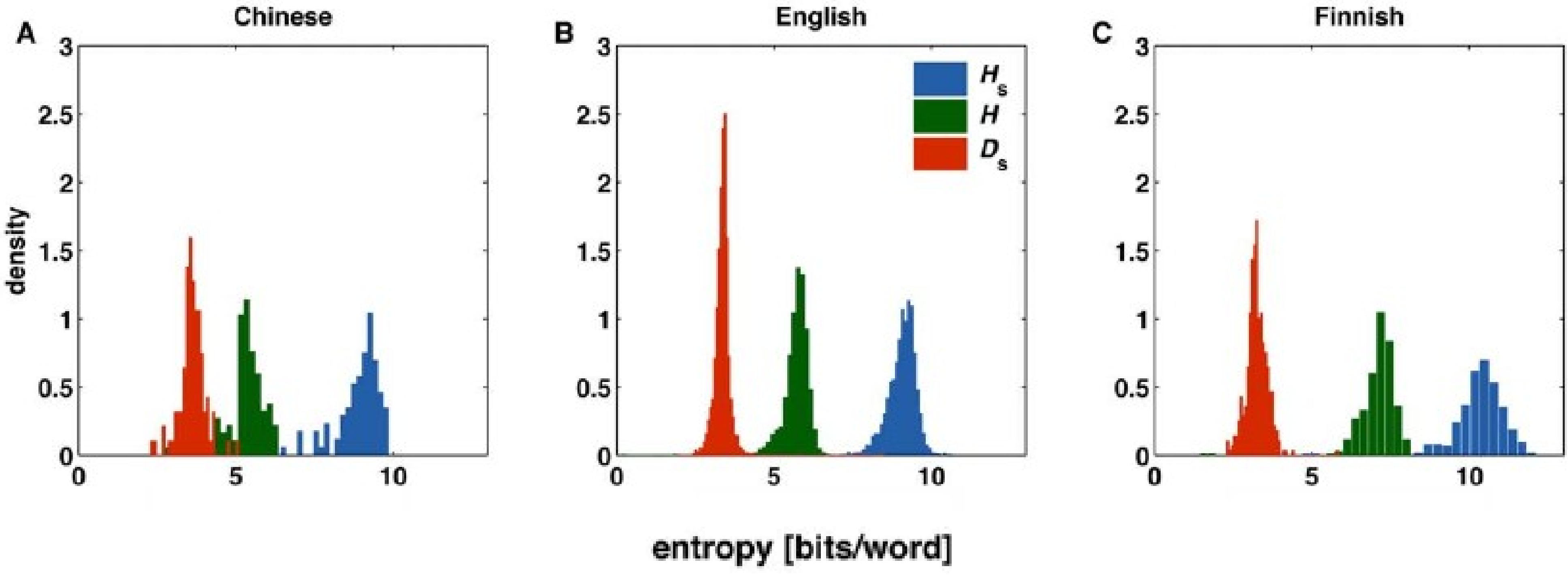
Conclusions/Significance: Our results indicate that despite the differences in the structure and vocabulary of the languages analyzed, the impact of word ordering in the structure of language is a statistical linguistic universal.

Citation: Montemurro MA, Zanette DH (2011) Universal Entropy of Word Ordering Across Linguistic Families. PLoS ONE 6(5): e19875. doi:10.1371/journal.pone.0019875

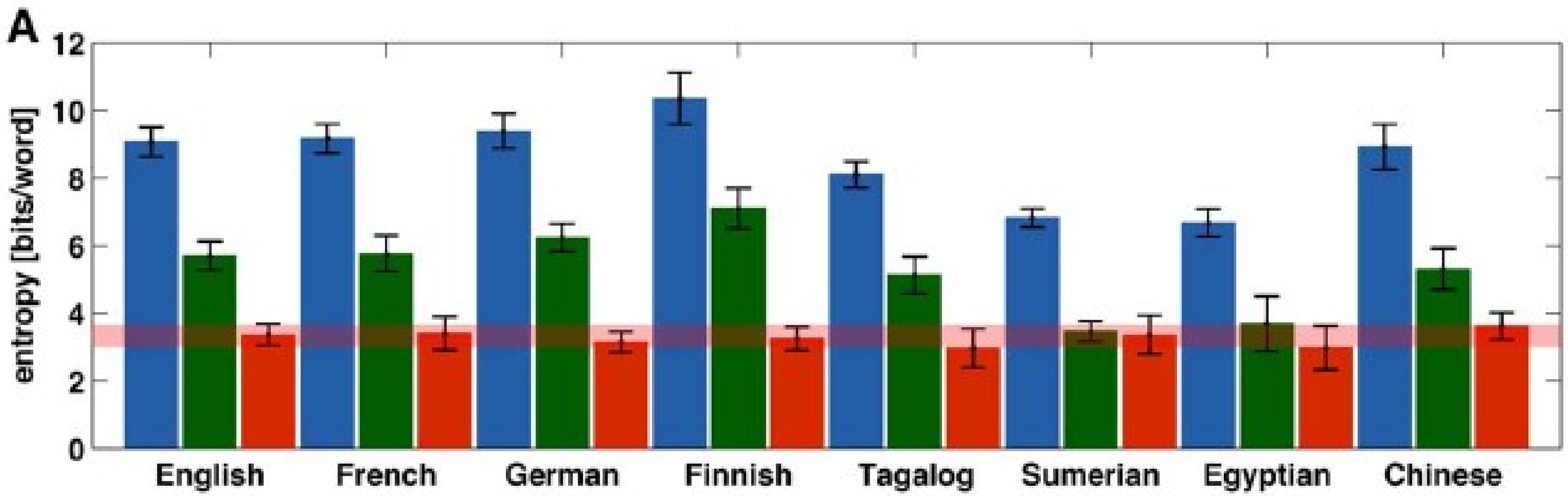
Editor: Michael Breakspear, Queensland Institute of Medical Research, Australia

Received August 24, 2010; **Accepted** April 19, 2011; **Published** May 13, 2011

Entropy of word ordering



Entropy of word ordering



Model of word formation

J. theor. Biol. (2000) **205,** 147–159
doi:10.1006/jtbi.2000.2053, available online at http://www.idealibrary.com on IDEAL®

Language Evolution and Information Theory

JOSHUA B. PLOTKIN* AND MARTIN A. NOWAK

Institut for Advanced Study, Princeton, NJ 08540, *U.S.A.*

(*Received on* 21 *October* 1999, *Accepted in revised form* 24 *March* 2000)

This paper places models of language evolution within the framework of information theory. We study how signals become associated with meaning. If there is a probability of mistaking signals for each other, then evolution leads to an error limit: increasing the number of signals does not increase the fitness of a language beyond a certain limit. This error limit can be overcome by word formation: a linear increase of the word length leads to an exponential increase of the maximum fitness. We develop a general model of word formation and demonstrate the connection between the error limit and Shannon's noisy coding theorem.

Model

Consider a population of individuals who can communicate via signals.
Signals may include gestures, facial expressions, or spoken sounds.

Each individual is described by an active matrix P and a passive matrix Q.

The entry P denotes that the probability that the individual, as a speaker, will refer to object i by using signal j.
The entry Q denotes the probability that the individual, as a listener, will interpret signal j as referring to object i.

Model

Q'

P

$$F(L, L') = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{m} P_{ij} Q'_{ji} + P'_{ij} Q_{ji}$$

Model

Q

U

P

$$F(L, L) = \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{m} P_{ij} U_{jk} Q_{ki}$$

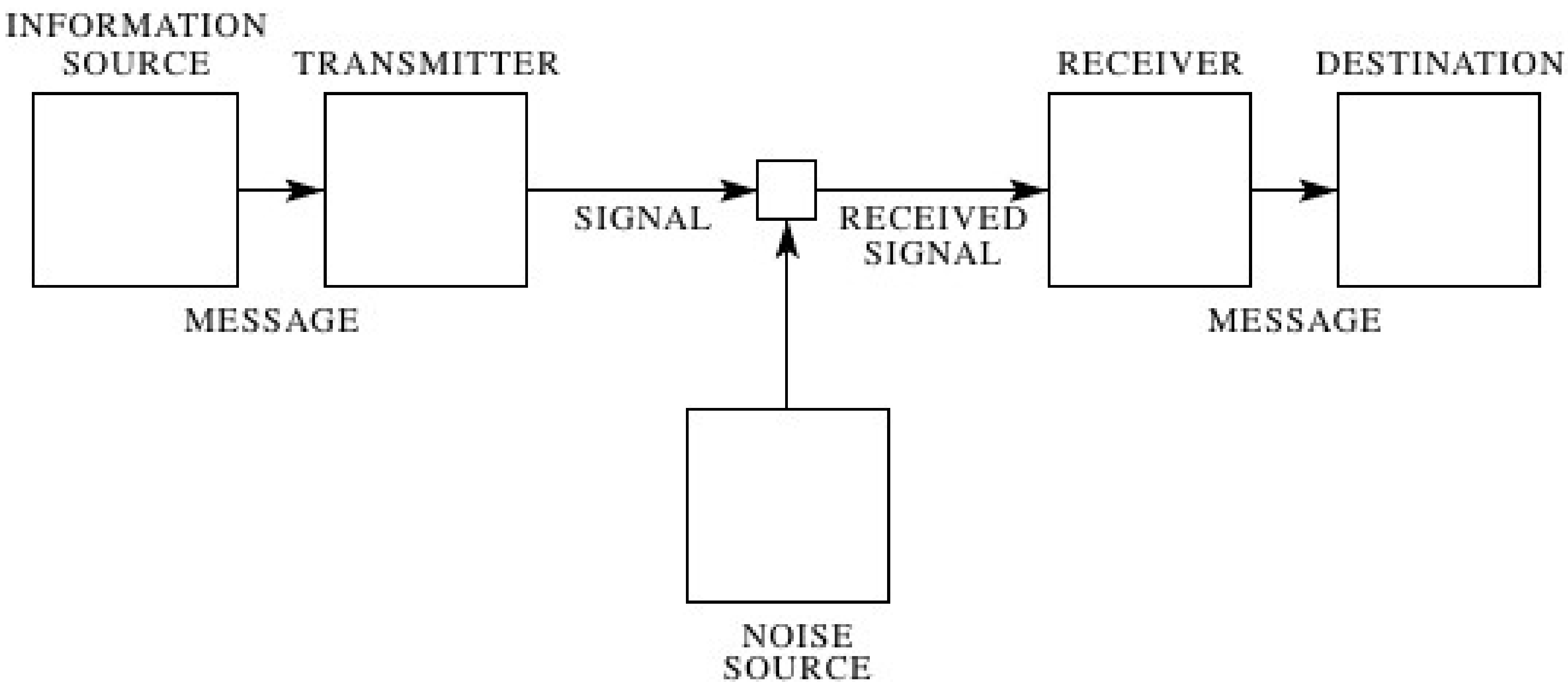


Fig. 1 — Schematic diagram of a general communication system.

Noise/confusion

- Languages whose basic signals consist of m phonemes.
- The words of the language are all l-phonemes long.
- The probability of confusion between words is defined by the product of the probability of confusion of their phonemes.

$$U_{\alpha\beta} = \prod_{k=1}^{l} V_{\alpha^{(k)}\beta^{(k)}}$$

where $\alpha^{(k)}$ denotes the k-th phoneme of word α.

$$\begin{aligned} F(L, L') &= \sum_{i=1}^{n} \sum_{\alpha \in \Phi^l} \sum_{\beta \in \Phi^l} P_{w_i \alpha} U_{\alpha\beta} Q_{\beta w_i} \\ &= \sum_{i=1}^{n} \sum_{\alpha \in \Phi^l} P_{w_i \alpha} \sum_{\beta \in \Phi^l} Q_{\beta w_i} \prod_{k=1}^{l} V_{\alpha^{(k)}\beta^{(k)}} \end{aligned}$$

Where $\Phi = \{\phi_1, \ldots, \phi_m\}$ are the phonemes.

P emitter matrix

$P =$

	aa	am	ap	ma	mm	mp	pa	pm	pp
Mother	0	0	0	$1-2\varepsilon$	ε	0	ε	0	0
Food	0	0	0	ε	$1-2\varepsilon$	0	ε	0	0
Father	0	0	0	ε	ε	0	$1-2\varepsilon$	0	0

Miller & Nicely 1955

TABLE V. Confusion matrix for $S/N = +6$ db and frequency response of 200–6500 cps.

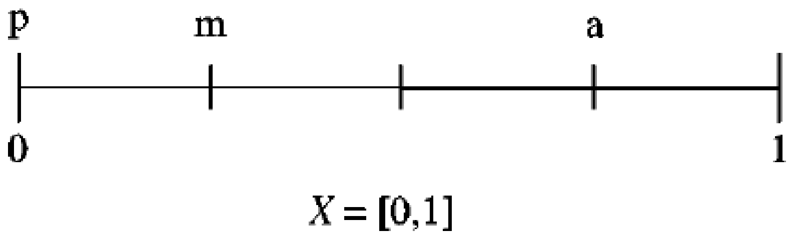
	p	*t*	*k*	*f*	*θ*	*s*	*ʃ*	*b*	*d*	*g*	*v*	*ð*	*z*	*ʒ*	*m*	*n*
p	162	10	55	5	3							1				
t	8	270	14													
k	38	6	171	1												
f	5	1	2	207	57			3			1					
θ	5	1	2	71	142	3					2	2				
s		1		1	7	232	2			1						
ʃ						1	239									
b				1	2			214			31	12				
d									206	14		9	1	2		
g								11	64	194		4	2	1		
v				1	1			14		2	205	39	5			1
ð								2		4	55	179	22	2		
z									3	10	2	20	198	3		
ʒ									3	4			2	215		
m															217	3
n									1						2	285

Miller & Nicely 1955

TABLE III. Confusion matrix for $S/N = -6$ db and frequency response of 200–6500 cps.

	p	*t*	*k*	*f*	*θ*	*s*	*ʃ*	*b*	*d*	*g*	*v*	*ð*	*z*	*ʒ*	*m*	*n*
p	80	43	64	17	14	6	2	1	1		1	1			2	
t	71	84	55	5	9	3	8	1				1	2		2	3
k	66	76	107	12	8	9	4					1			1	
f	18	12	9	175	48	11	1	7	2	1	2	2				
θ	19	17	16	104	64	32	7	5	4	5	6	4	5			
s	8	5	4	23	39	107	45	4	2	3	1	1	3	2		1
ʃ	1	6	3	4	6	29	195		3							1
b	1			5	4	4		136	10	9	47	16	6	1	5	4
d							8	5	80	45	11	20	20	26	1	
g					2			3	63	66	3	19	37	56		3
v				2		2		48	5	5	145	45	12		4	
ð					6			31	6	17	86	58	21	5	6	4
z					1	1	1	7	20	27	16	28	94	44		1
ʒ								1	26	18	3	8	45	129		2
m	1							4			4	1	3		177	46
n					4			1	5	2		7	1	6	47	163

U noise matrix



$$s_{\text{a,a}} = s_{\text{m,m}} = s_{\text{p,p}} = 1$$

$$s_{\text{a,m}} = s_{\text{m,a}} \approx 0.08,$$

$$s_{\text{a,p}} = s_{\text{p,a}} \approx 0.02,$$

$$s_{\text{m,p}} = s_{\text{p,m}} \approx 0.29.$$

$V =$

	a	m	p
a	0.90	0.07	0.02
m	0.06	0.73	0.21
p	0.02	0.22	0.76

$U =$

	aa	am	ap	ma	mm	mp	pa	pm	pp
ma	0.05	0.00	0.00	0.66	0.05	0.02	0.19	0.02	0.00
mm	0.00	0.04	0.01	0.04	0.53	0.15	0.01	0.15	0.04
pa	0.02	0.00	0.00	0.20	0.02	0.00	0.69	0.06	0.02

What is the optimal Q passive matrix?

First, a guess:
a listener should interpret perceived output word w as object i with a probability which equals the probability that, when trying to communicate object i, the perceived output would be w.

Q receiver matrix

	Mother	Food	Father
aa	$0.73 - 1.2\varepsilon$	$0.05 + 0.85\varepsilon$	$0.22 + 0.34\varepsilon$
am	$0.09 + 0.73\varepsilon$	$0.89 - 1.65\varepsilon$	$0.03 + 0.92\varepsilon$
ap	$0.09 + 0.73\varepsilon$	$0.88 - 1.65\varepsilon$	$0.03 + 0.92\varepsilon$
ma	$0.73 - 1.2\varepsilon$	$0.05 + 0.85\varepsilon$	$0.22 + 0.34\varepsilon$
$Q =$ mm	$0.09 + 0.73\varepsilon$	$0.88 - 1.65\varepsilon$	$0.03 + 0.92\varepsilon$
mp	$0.09 + 0.73\varepsilon$	$0.88 - 1.65\varepsilon$	$0.03 + 092\varepsilon$
pa	$0.21 + 0.36\varepsilon$	$0.01 + 0.96\varepsilon$	$0.77 - 1.3\varepsilon$
pm	$0.07 + 0.79\varepsilon$	$0.68 - 1.04\varepsilon$	$0.25 + 0.25\varepsilon$
pp	$0.07 + 0.79\varepsilon$	$0.68 - 1.04\varepsilon$	$0.25 + 0.25\varepsilon.$

Fitness as a function of noise

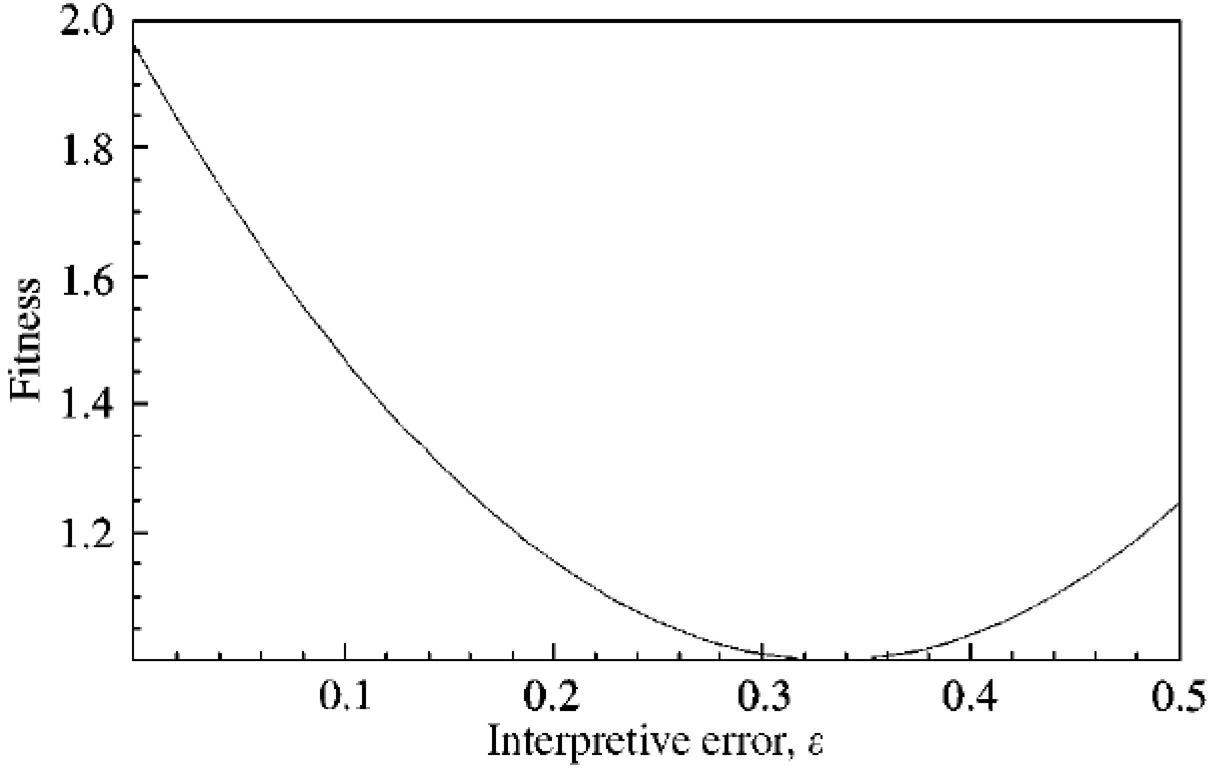


FIG. 3. Graph of the language fitness $F(L, L)$ obtained as a function of ε. The parameter ε measures the amount of interpretive error in the language. The fitness of the language L is maximized when there is no chance for misinterpretation ($\varepsilon = 0$).

Maximum likelihood Q matrix

$Q^{ML} =$

	Mother	Food	Father
aa	1	0	0
am	0	1	0
ap	0	1	0
ma	1	0	0
mm	0	1	0
mp	0	1	0
pa	0	0	1
pm	0	1	0
pp	0	1	0

.

Fitness as a function of noise

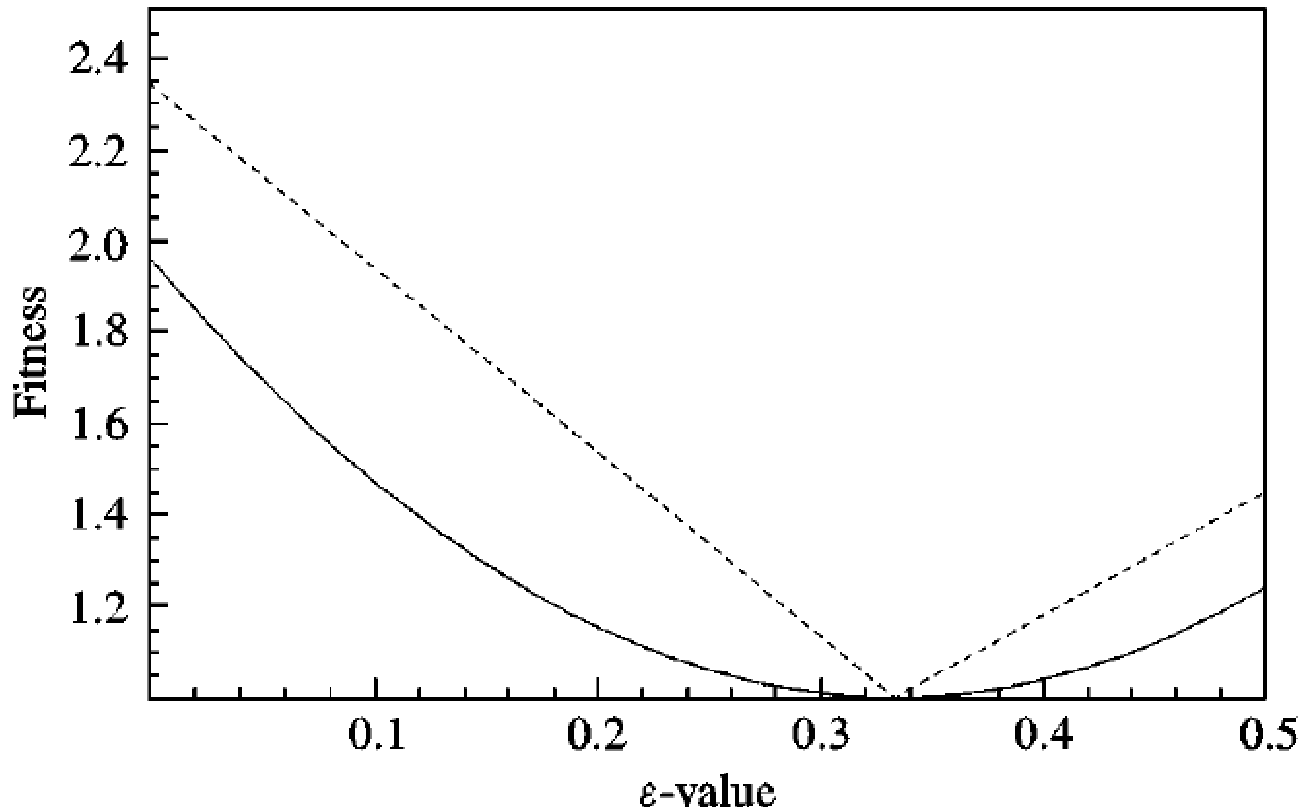


FIG. 4. Graph of the language fitness $F(L, L)$ obtained via the non-deterministic decoder Q as opposed to the deterministic, maximum-likelihood decoder Q^{ML}. A language is always better served by the maximum-likelihood decoder. Thus, we expect that languages should evolve towards maximum-likelihood decoding. (- - - -) Shannon decoder (Q^{ML}); (——) non-deterministic decoder (Q).

Word formation

is it possible, by increasing the word length *l*, to increase a language's payoff without bound? In light of the error limit, this inquiry addresses a fundamental question regarding the adaptive benefits of word formation.

Theorem

Theorem 4.1 (Shannon, 1948). *If a discrete memoryless channel V has capacity $C > 0$ and R is any positive quantity with $R < C$, then there exists a sequence of codes $(\mathfrak{C}_n | 1 \leqslant n < \infty)$ such that*
(a) $\mathfrak{C}_n$ *has* $2^{\lfloor Rn \rfloor}$ *codewords of length* $l = n$,
(b) *the error probability satisfies* $e(\mathfrak{C}_n) \leqslant A\mathrm{e}^{-Bn}$,
where the constants A and B depend only on the channel V and on R.

where $$e(\mathfrak{C}) = \frac{1}{n} \sum_{i=1}^{n} \Pr(\text{error in communication} \mid \text{codeword } w_i \text{ is transmitted})$$

$$F(L, L) = |\mathfrak{C}|(1 - e(\mathfrak{C}))$$

Theorem

Theorem 4.1 (Shannon, 1948). *If a discrete memoryless channel V has capacity $C > 0$ and R is any positive quantity with $R < C$, then there exists a sequence of codes* $(\mathfrak{C}_n | 1 \leqslant n < \infty)$ *such that*
(a) $\mathfrak{C}_n$ *has* $2^{\lfloor Rn \rfloor}$ *codewords of length* $l = n$,
(b) *the error probability satisfies* $e(\mathfrak{C}_n) \leqslant A\mathrm{e}^{-Bn}$, *where the constants A and B depend only on the channel V and on R.*

Theorem 5.1 (word formation). *Given a phoneme error-matrix V (with non-zero capacity), there exists a sequence of languages L_n with linearly increasing word length and exponentially increasing fitness.*

Word formation

$$C(V) \approx 0.7988$$

Thus, since $C(V) > 0$, Shannon's theorem indeed applies to our explicit example. In particular, Shannon's theorem guarantees a sequence of languages L_n, $n = 1, 2, 3, \ldots$, each with a lexicon of $2^{\lfloor 0.79n \rfloor}$ words of length $l = n$, with exponentially increasing fitnesses.

Of course, in reality, words don't grow arbitrarily longer.
But they still permit a decrease in the error rate.

Evolution of syntax

The evolution of syntactic communication

Martin A. Nowak*, **Joshua B. Plotkin*** & **Vincent A. A. Jansen†**

** Institute for Advanced Study, Princeton, New Jersey 08540, USA*
† School of Biological Sciences, Royal Holloway, University of London, Egham Surrey, TW20 0EX UK

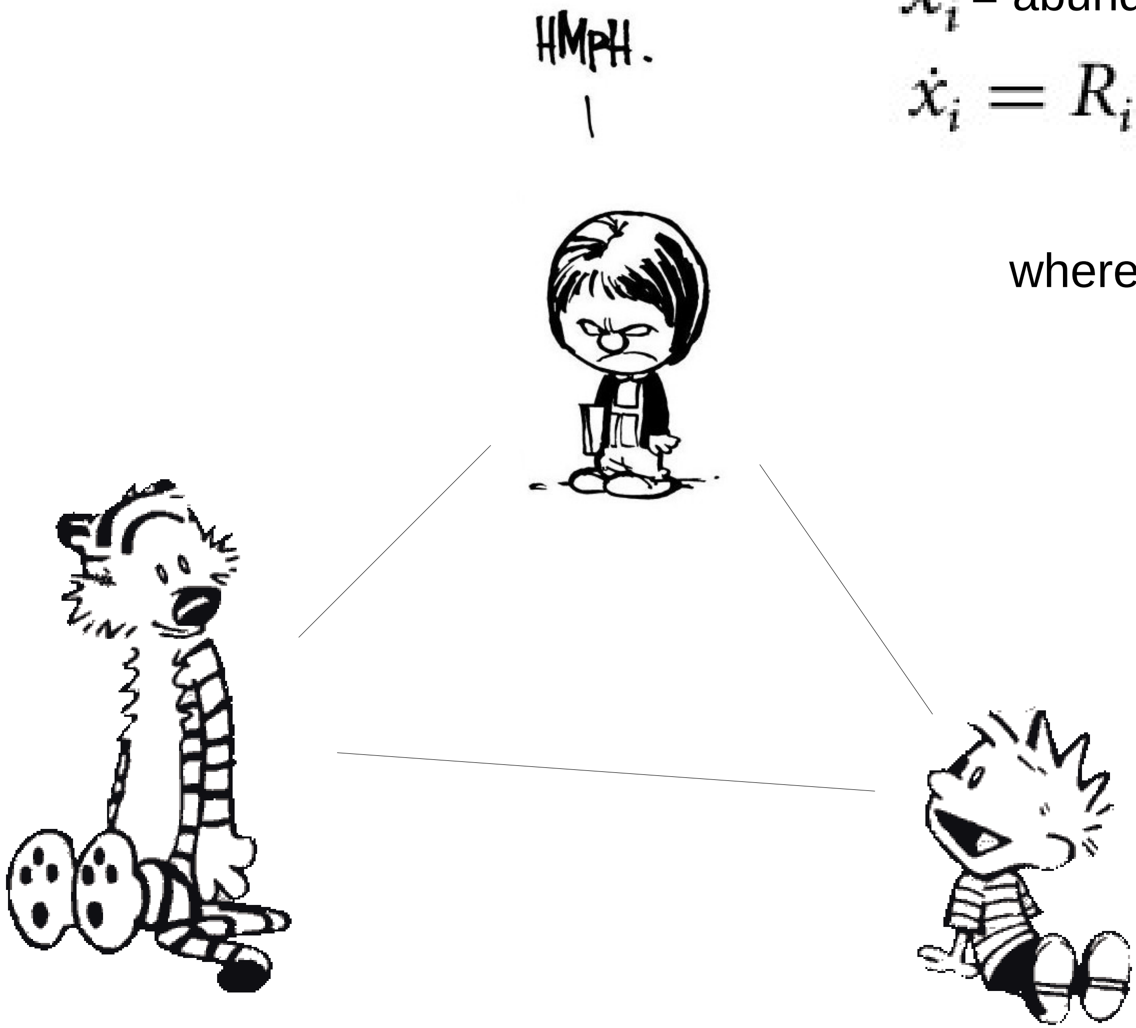
Animal communication is typically non-syntactic, which means that signals refer to whole situations[1–7]. Human language is syntactic, and signals consist of discrete components that have their own meaning[8]. Syntax is a prerequisite for taking advantage of combinatorics, that is, "making infinite use of finite means"[9–11]. The vast expressive power of human language would be impossible without syntax, and the transition from non-syntactic to syntactic communication was an essential step in the evolution of human language[12–16]. We aim to understand the evolutionary dynamics of this transition and to analyse how natural selection can guide it. Here we present a model for the population dynamics of language evolution, define the basic reproductive ratio of words and calculate the maximum size of a lexicon. Syntax allows larger repertoires and the possibility to formulate messages that have not been learned beforehand. Nevertheless, according to our model natural selection can only favour the emergence of syntax if the number of required signals exceeds a threshold value. This result might explain why only humans evolved syntactic communication and hence complex language.

Word learning

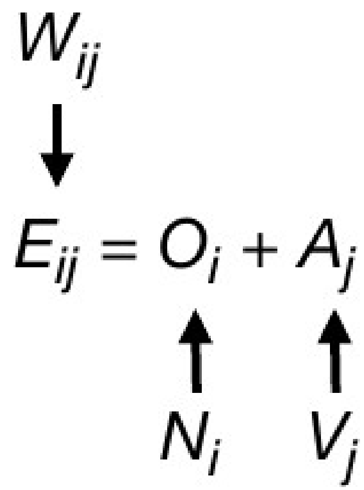
x_i = abundance of word in population

$$\dot{x}_i = R_i x_i (1 - x_i) - x_i$$

where $R_i = bq\phi_i$,



To syntax or not to syntax?



To syntax or not to syntax?

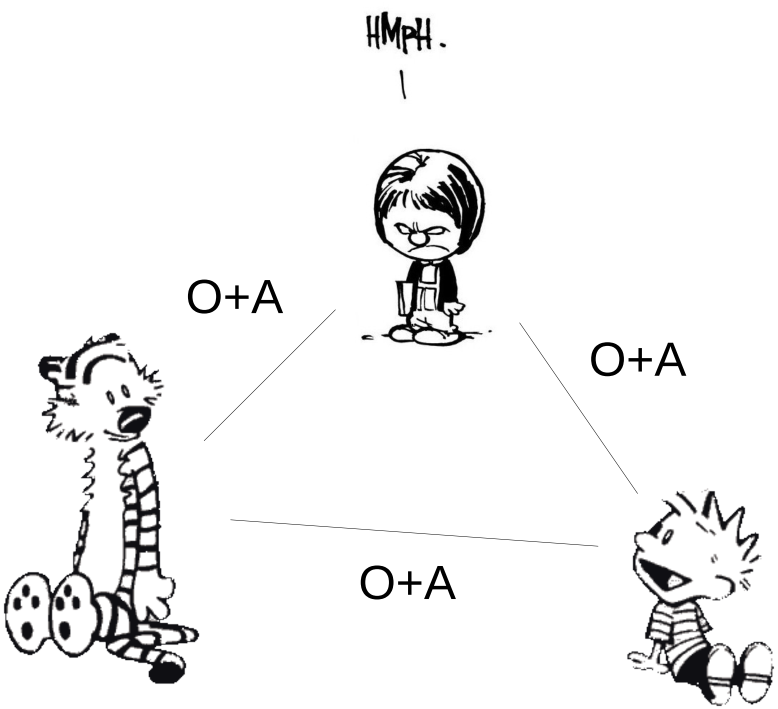
$$W_{ij}$$

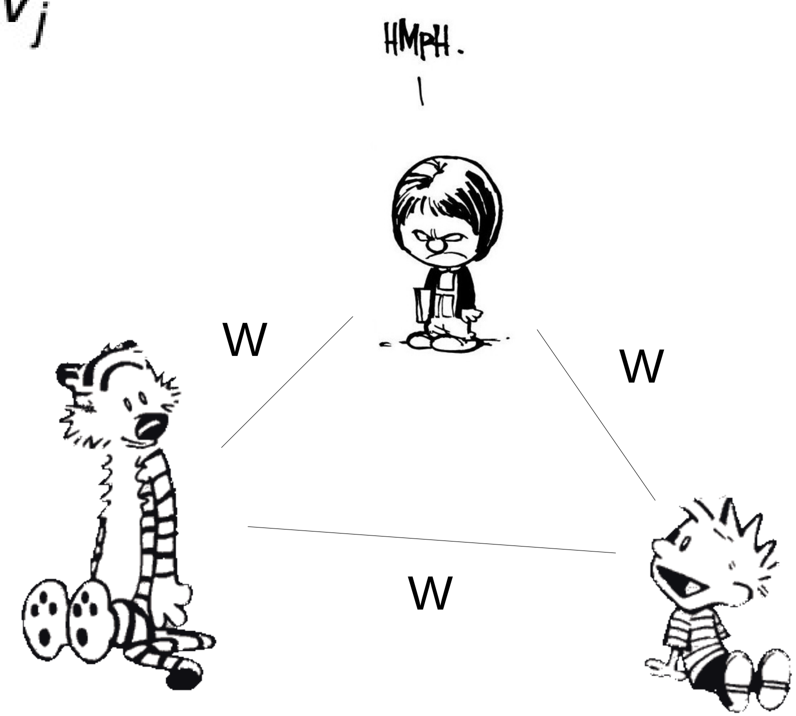
$$\downarrow$$

$$E_{ij} = O_i + A_j$$

$$\uparrow \quad \uparrow$$

$$N_i \quad V_j$$





$$\dot{x}(N_i V_j) = R(N_i)x(N_i)[x(V_j) - x(N_i V_j)] + R(V_j)x(V_j)[x(N_i) - x(N_i V_j)] - x(N_i V_j)$$