Chapter 4:
Entropy

Section 4.1.
The Shannon-Game

2

Informal definition of Entropy

Entropy:

Amount of information contained in a message (after removing all possible redundancy)

(number of bits that the message has after compression)

3

Complete the Sentence

Th_r_ _s _nly _n_ w_y t_ f_ll _n th_ v_w_ls _n th_s s_nt_nc_

There is only one way to fill in the vowels in this sentence

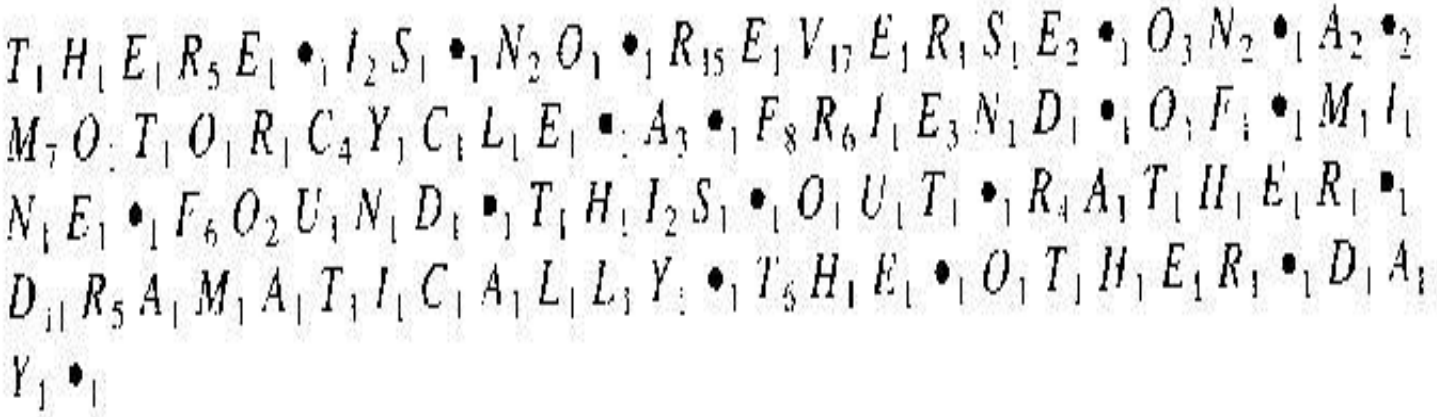
4

Entropy of a Language: Shannon's Approach

1. Show somebody the beginning of a text
2. Ask him/her to guess the next letter
3. Count the number of trials

5

The Shannon Game



Play it yourself at:
http://www.math.ucsd.edu/~crypto/java/ENTROPY/

6

Cover and Kings Variation to the Shannon Game

1. The test person gets a certain amount of money
2. Show the person the beginning of the text
3. Ask the person to bet on the next letter
4. Observe how the fortune of the person evolves

IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. IT-24, NO. 4, JULY 1978

A Convergent Gambling Estimate of the Entropy of English

THOMAS M. COVER, FELLOW, IEEE, AND ROGER C. KING, STUDENT MEMBER, IEEE

7

Example of Results

TABLE I

EXPERIMENTAL RESULTS ON ESTIMATING ENTROPY OF ENGLISH USING SEQUENCE OF 75 SYMBOLS FROM *Jefferson the Virginian*

Subject	Capital Achieved	Resultant Entropy Estimate
1	1.50×10^{78}	1.29 bits/sym
2	1.46×10^{76}	1.38
3	3.36×10^{75}	1.41
4	2.37×10^{73}	1.51
5	6.45×10^{71}	1.57
6	3.22×10^{71}	1.59
7	2.30×10^{70}	1.64
8	4.00×10^{70}	1.67
9	2.21×10^{69}	1.68
10	9.63×10^{68}	1.70
11	3.88×10^{67}	1.76
12	3.60×10^{64}	1.90

Note the difference between the subjects

8

The Entropy of Natural Languages

TABLE 4-3 ESTIMATES OF THE ENTROPY OF NATURAL LANGUAGES

Language	Size of alphabet	Letter models with order: −1	0	1	2	3	7	11	≥100	Word model	Source
From statistical analysis of text											
English	26	4.70	4.14	3.56	3.3					2.62	Shannon (1951)
	26+1	4.75	4.03	3.32	3.1					2.14	
English	26	4.70	4.12							1.65	Barnard (1955)
French	26	4.70	3.98							3.02	
German	26	4.70	4.10							1.08	
Spanish	26	4.70	4.02							1.97	
English	26+1	4.75	4.09	3.23	2.85	2.66	2.43	2.40			Newman and Waugh (1960)
Samoan	16+1	4.09	3.40	2.68	2.40	2.28	2.16	2.14			
Russian	35+1	5.17	4.55	3.44	2.95	2.72	2.45	2.40			
Portugese	26?	4.70?	3.92	3.51	3.15						Manfrino (1970)
Tamil	30	4.91	4.34								Siromoney (1963)
Kannada	49	5.61	4.55								Rajagopalan (1965)
Telugu	53	5.73	4.59	3.09							Balasubrahmanyam and Siromoney (1968)
Arabic	32	5.00	4.21	3.77	2.49						Wanas et al. (1976)
Chinese	4700	12.20	9.63								Wong and Poon (1976)
From experiments with subjects' best guesses											
English	26+1	4.75									
upper bound (smoothed)			4.0	3.4	3.0	2.6	2.1	1.9	1.3		Shannon (1951)
lower bound (smoothed)			3.2	2.5	2.1	1.8	1.2	1.1	0.6		
English	26+1	4.75				2.2	1.8	1.8	1.7		Jamison and Jamison (1968)
Italian	26+1	4.75				2.9	2.6	2.8	3.0		
Italian*	26+1	4.75				3.4	3.1	3.3	3.8		
French*	26+1	4.75				3.5	2.8	2.9	3.2		
From experiments with subjects using gambling											
English	26+1	4.75							1.25		Cover and King (1978)
Malay	26+1	4.75							1.32		Tan (1981)

Gambling experiment gives more accurate results than simple guessing

9

Formal Definition of Entropy

$$H(V) = \mathbb{E}[-\log(p(V))]$$

$$= \sum_{w_i \in V} - p(w_i) \log(p(w_i))$$

Note: if you want the "unit" of the entropy to be "bit" you have to use the log to the basis 2

10

Example

Vocabulary with two words:

V={a,b}

P(a)=p

P(b)=1-p

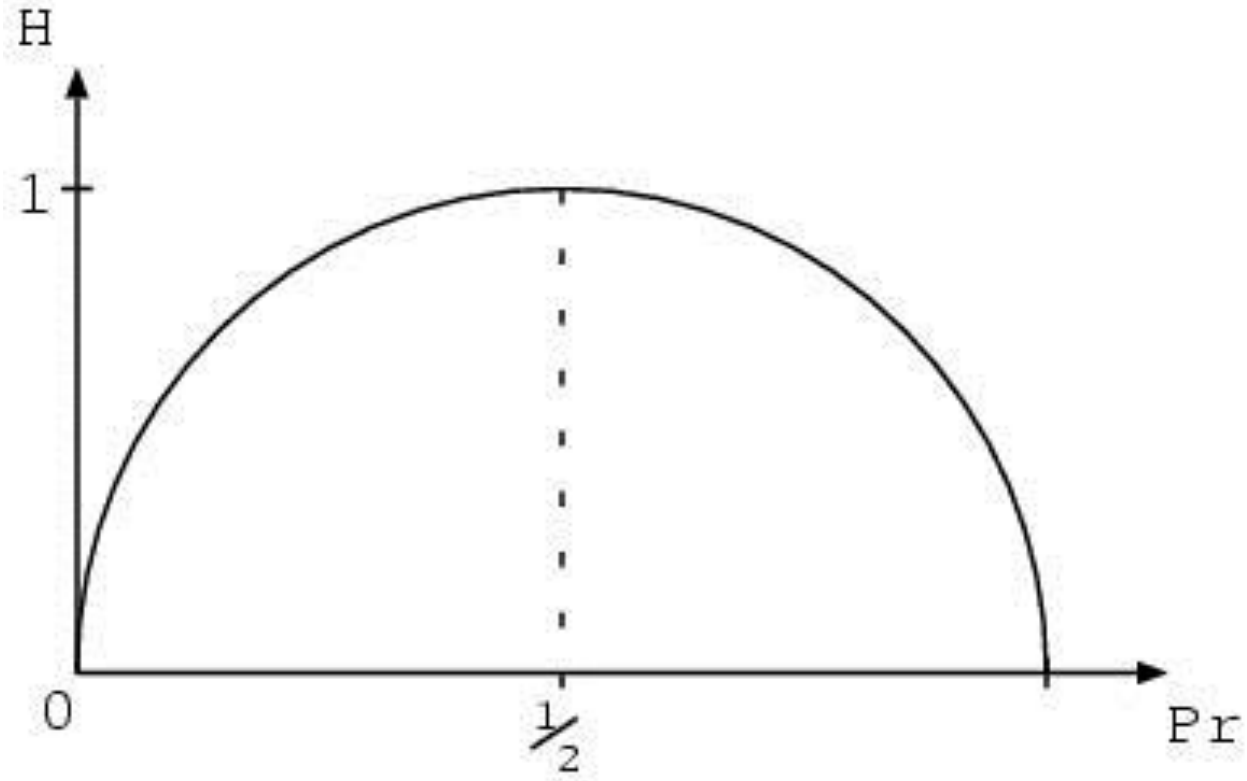
H=-p log p – (1-p) log(1-p)

p=0 $\mapsto$ H=0

p=1 $\mapsto$ H=0

11

Entropy H=-p log p – (1-p) log(1-p)



12

Example 2

Vocabulary of W words w_i with uniform distribution $p(w_i)=1/W$

$$H = \sum_{i=1}^{W} - p(w_i)\log(p(w_i)) = \sum_{i=1}^{W} - \frac{1}{W}\log(\frac{1}{W})$$

$$= -W\frac{1}{W}\log(\frac{1}{W}) \quad = -\log(\frac{1}{W}) = \log(W)$$

Entropy for uniform distribution: log of the number of symbols

13

Theorem: Entropy is never negative

Theorem $H(V) \geq 0$

Proof:

Idea of proof: use that the log is a concave function

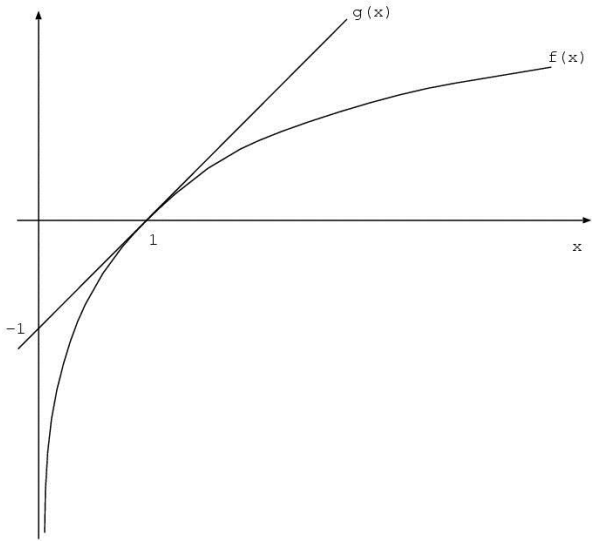
14

Proof: entropy is not negative

Define auxiliary functions

f(x)=log(x)

g(x)=x-1



$$\mapsto \quad g(x) \geq f(x) \quad \forall x > 0$$

$$\mapsto \quad x - 1 \geq \log(x) \quad \forall x > 0$$

$$\mapsto \quad \boxed{-\log(x) \geq 1 - x \quad \forall x > 0}$$

15

Proof: entropy is not negative

$$H = \sum_{i=1}^{W} - p(w_i) \log(p(w_i))$$

Use

$$-\log(x) \geq 1 - x \quad \forall x > 0$$

$$\geq \sum_{i=1}^{W} \underbrace{p(w_i)}_{\geq 0} \underbrace{(1 - p(w_i))}_{\geq 0} \geq 0$$

16

Definition 2-Word Entropy

$$H(W,W) = -\sum_{w_1,w_2} p(w_1,w_2)\log(\,p(w_1,w_2))$$

Similarly higher order entropies can be defined

17

Example for 2-Word Entropy

- Let w_1 and w_2 statistically independent

$$p(w_1, w_2) = p(w_1)\,p(w_2)$$

$$H(W,W) = -\sum_{w_1,w_2} p(w_1, w_2)\log(\,p(w_1, w_2))$$

$$= -\sum_{w_1,w_2} p(w_1)\,p(w_2)\log(\,p(w_1)\,p(w_2))$$

$$= -\sum_{w_1,w_2} p(w_1)\,p(w_2)[\log\,p(w_1) + \log\,p(w_2)]$$

18

Example for 2-Word Entropy

$$= -\sum_{w_1, w_2} p(w_1)\, p(w_2)[\log p(w_1) + \log p(w_2)]$$

$$= -\sum_{w_1, w_2} p(w_1)\, p(w_2) \log p(w_1) - \sum_{w_1, w_2} p(w_1)\, p(w_2) \log p(w_2)$$

$$= -\sum_{w_1} p(w_1) \log p(w_1) - \sum_{w_2} p(w_2) \log p(w_2)$$

$$= 2H(W)$$

$$H(W, W) = 2H(W)$$

for statistical independence

19

Definition: Kullback-Leibler (KL) Divergence

$$D(p \parallel q) = \sum_i p_i \log \frac{p_i}{q_i}$$

Remark:

-p_i is a short hand for $p(w_i)$

-q_i is a short hand for $q(w_i)$

-p_i and q_i are different distributions

-$D(p\|p)=0$

-$D(p\|q)$ is **not** symmetric ($D(p\|q)$ is not equal to $D(q\|p)$)

20

Kullback-Leibler (KL) Divergence and Entropy

$$D(p \,\|\, q) = \sum_i p_i \log \frac{p_i}{q_i}$$

$$= -\sum_i p_i \log q_i + \sum_i p_i \log p_i$$

$$= E_p[-\log(q)] - E_p[-\log(p)]$$

$$= H(p,q) - H(p)$$

Cross Entropy

Entropy

We will see later: KL divergence measures how many more bits need to be transmitted if you have a mismatch in the model of the language

21

Kullback-Leibler (KL) Divergence: Example

$$q_A = 1/2; q_B = 1/4; q_C = 1/4;$$
$$p_A = 1/3; p_B = 1/3; p_C = 1/3;$$

See white board

22

Theorem: KL-Divergence is non-negative

$$D(p \| q) \geq 0$$

Proof:

$$D(p \| q) = -\sum_i p_i \log \frac{q_i}{p_i}$$

$$\geq \sum_i p_i \left(1 - \frac{q_i}{p_i}\right)$$

$$\geq \sum_i p_i - \sum_i q_i = 1 - 1 = 0$$

Use

$$-\log(x) \geq 1 - x \quad \forall x > 0$$

23

Theorem: H(W,W) bounded by 2H(W)

Theorem $H(W,W) \leq 2H(W)$

Proof: $D\big(p(w_1,w_2) \,\|\, p(w_1)p(w_2)\big)$

$$= \sum_{w_1,w_2} p(w_1,w_2) \log \frac{p(w_1,w_2)}{p(w_1)p(w_2)}$$

$$= \sum_{w_1,w_2} p(w_1,w_2) \log p(w_1,w_2) - \sum_{w_1,w_2} p(w_1,w_2) \log p(w_1) - \sum_{w_1,w_2} p(w_1,w_2) \log p(w_2)$$

24

Theorem: H(W,W) bounded by 2H(W)

$$= \sum_{w_1,w_2} p(w_1,w_2)\log p(w_1,w_2) - \sum_{w_1,w_2} p(w_1,w_2)\log p(w_1) - \sum_{w_1,w_2} p(w_1,w_2)\log p(w_2)$$

$$= \sum_{w_1,w_2} p(w_1,w_2)\log p(w_1,w_2) - \sum_{w_1} p(w_1)\log p(w_1) - \sum_{w_2} p(w_2)\log p(w_2)$$

$$= -H(W,W) + H(W) + H(W) \geq 0$$

Hence:

$$\boxed{H(W,W) \leq 2H(W)}$$

25

Section 4.2.
Text Compression

26

Text Compression 19th Century (Admiralty Shutter Telegraph, 1825)

Design of Telegraph

Lines in southern England

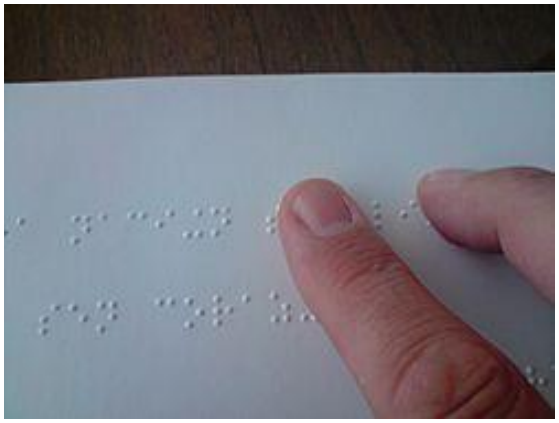
- Visual transmission of a message using 6 bit encoding
- Very limited bandwidth makes good compression necessary

Source: http://www.douglas-self.com/MUSEUM/COMMS/telegraf/telegraf.htm

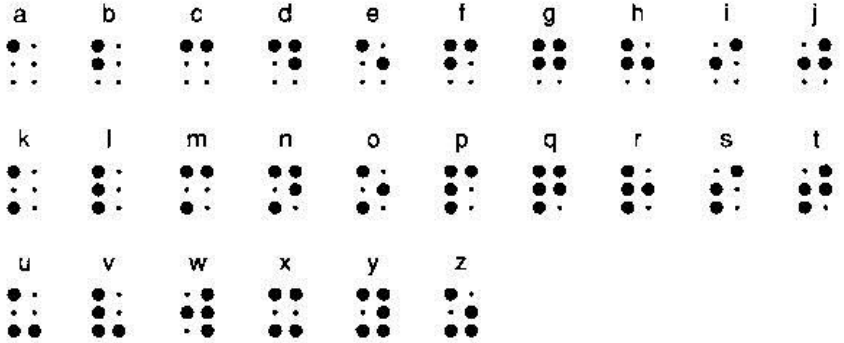
27

Braille-Letters: tactile writing system for the blind

Example of Text



Encoding

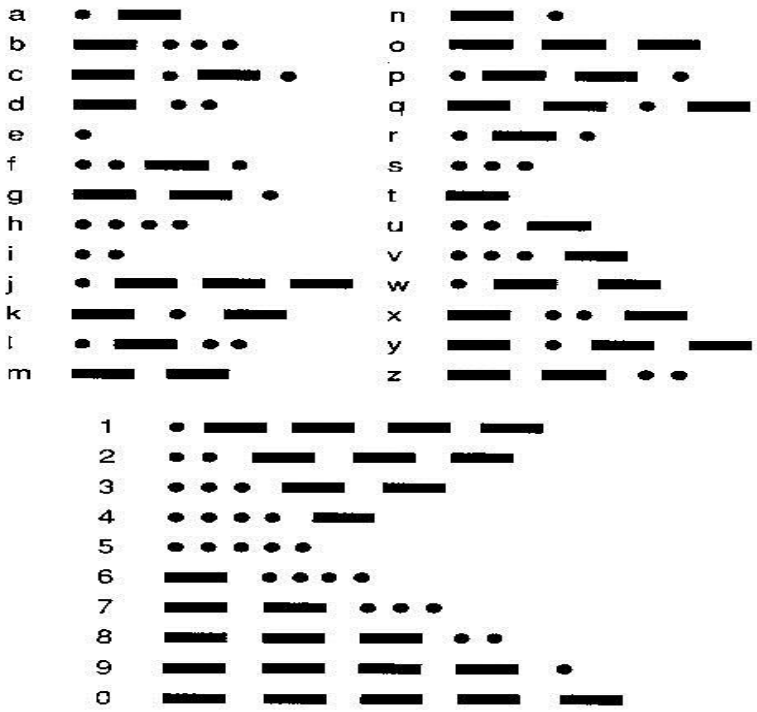


Source: Wikipedia

28

Morse Code

a method of transmitting text information
as a series of log and short tones



Note: frequent characters
Have short code words

29

Definition: Code

A code $C(w_i)$ is a mapping of words w_i to finite strings of a D-nary alphabet

Example: suppose you want to code the most frequent English words “the”, “and”, “of” and “he” a possible binary code would be:

C(“the”)=0
C(“and”)=10
C(“of”)=110
C(“he”)=111

30

Definition: Code

C(“the”)=0
C(“and”)=10
C(“of”)=110
C(“he”)=111

What is the decoded message corresponding to:

10101101110

What makes a code uniquely decodable?

31

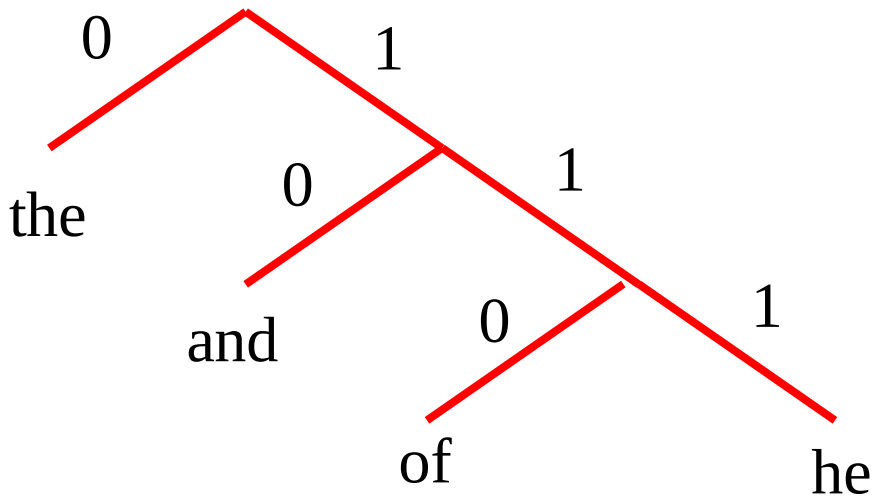
Definition: Prefix Code

- No code word is prefix of another code word
- Organize code as a tree

Example for encoding:

C(“the”)=0
C(“and”)=10
C(“of”)=110
C(“he”)=111

Corresponding tree:



32

Length Distribution of Example Code

Length l_i: number of letters of D-nary alphabet

Word	Code	Length l_i
“the”	0	1
“and”	10	2
“of”	110	3
“he”	111	3

Observe:

$$\sum_{i=1}^{4} 2^{-l_i}$$

$$= \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{8}$$

$$= 1$$

33

Theorem: Kraft's Inequality

For every prefix code with code word length l_1 to l_m the following inequality holds true

$$\sum_{i=1}^{m} D^{-l_i} \leq 1$$

For each length distribution that satisfies this inequality there is a prefix code

D: base of the code (e.g. D=2 for binary code)
M: number of code words

34

Optimal length distribution

The expected length (per word) of a text after coding is

$$L = \sum_i l_i \, p(w_i)$$

The optimal length of code words $C(w_i)$ is

$$\boxed{l_i = -\log_D p(w_i)}$$

This choice minimizes L

35

Proof: Optimal length distribution

The expected length of a text after coding is

$$L = \sum_i l_i \, p(w_i)$$

with

$$\sum_{i=1}^{m} D^{-l_i} \leq 1$$

36

Proof: Optimal length distribution

Optimize a modified function

$$L^* = \sum_i l_i p(w_i) + \lambda\left(\sum_i D^{-l_i} - 1\right)$$

λ :
Lagrange multiplier
(to be determined later)

Derived from constraint

$$\sum_{i=1}^{m} D^{-l_i} \leq 1$$

37

Proof: Optimal length distribution

Optimize a modified function

$$L^* = \sum_i l_i p(w_i) + \lambda\left(\sum_i D^{-l_i} - 1\right)$$

Calculate first derivative

$$\frac{\partial L^*}{\partial l_j} = p(w_j) + \lambda \frac{\partial D^{-l_j}}{\partial l_j} = 0$$

38

Proof: Optimal length distribution

$$\frac{\partial L^*}{\partial l_j} = p(w_j) + \lambda \frac{\partial D^{-l_j}}{\partial l_j} = 0$$

solve for derivative

$$\frac{\partial D^{-l_j}}{\partial l_j} = \frac{\partial}{\partial l_j} e^{-l_j \log(D)} = -\log(D) e^{-l_j \log(D)} = -\log(D) D^{-l_j}$$

insert back

$$\frac{\partial L^*}{\partial l_j} = p(w_j) - \lambda \log(D) D^{-l_j} = 0$$

solve

$$D^{-l_j} = \frac{p(w_j)}{\lambda \log(D)}$$

39

Proof: Optimal length distribution

Goal: determine λ

Insert $D^{-l_i} = \frac{p(w_i)}{\lambda \log(D)}$ into $\sum_{i=1}^{m} D^{-l_i} = 1$

Equation to solve:

$$1 = \sum_{i=1}^{m} \frac{p(w_i)}{\lambda \log(D)} = \frac{1}{\lambda \log(D)} \quad \Rightarrow \quad \lambda = \frac{1}{\log(D)}$$

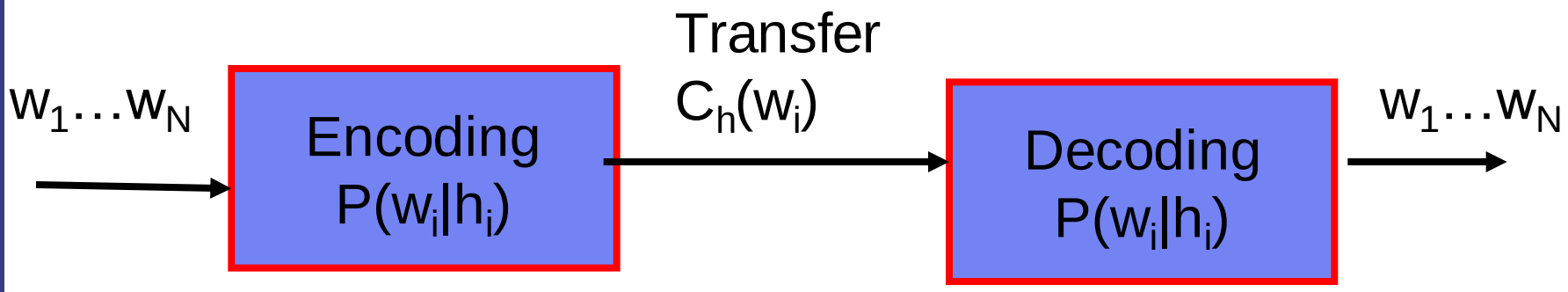
Insert λ into first equation:

$$p(w_i) = D^{-l_i}$$

$$l_i = -\frac{\log p(w_i)}{\log D} = -\log_D p(w_i)$$

40

Text Compression and Perplexity



- Code words $C_h(w_i)$ depend on history
- Length of code words $C_h(w_i)$: $l_h(w_i) = - \log_D P(w_i|h_i)$

Question: what is the length of the compressed message?

41

Length of Compressed Message

$$L_{total} = \sum_{w,h} N(h,w) l_h(w) = N \sum_{w,h} f(h,w)\left(-\log_D P(w|h)\right)$$

$$= -N \sum_{w,h} f(h,w) \frac{\log P(w|h)}{\log D} \quad = \frac{N}{\log D}\left(-\sum_{w,h} f(h,w) \log P(w|h)\right)$$

$$= \frac{N}{\log D} \log PP$$

$$L_{total} = \frac{N}{\log D} \log PP$$

Length of compressed text is proportional to logarithm of perplexity

42

Limits of the optimal Code

Theorem (entropy-bound):

Let

$l_i = \text{int}(-\log_D(p(w_i))+1)$

then

$$H_D(W) \leq L < H_D(W)+1$$

and there is a prefix-code for l_i

Proof: direct calculation

43

Compression with a mismatched code

Theorem: The probabilities of words w_i of a text are p_i, however, it is compressed based on q_i with l_i=-log q_i

Then the length of the compressed text is

$$H(p)+D(p||q)\leq L<H(p)+D(p||q)+1$$

44

Section 3.3.
Long Range Dependencies and Conditional Entropy

45

Two Alternatives

- Correlation function
- Conditional Entropy

46

Analyzing Long Range Dependencies

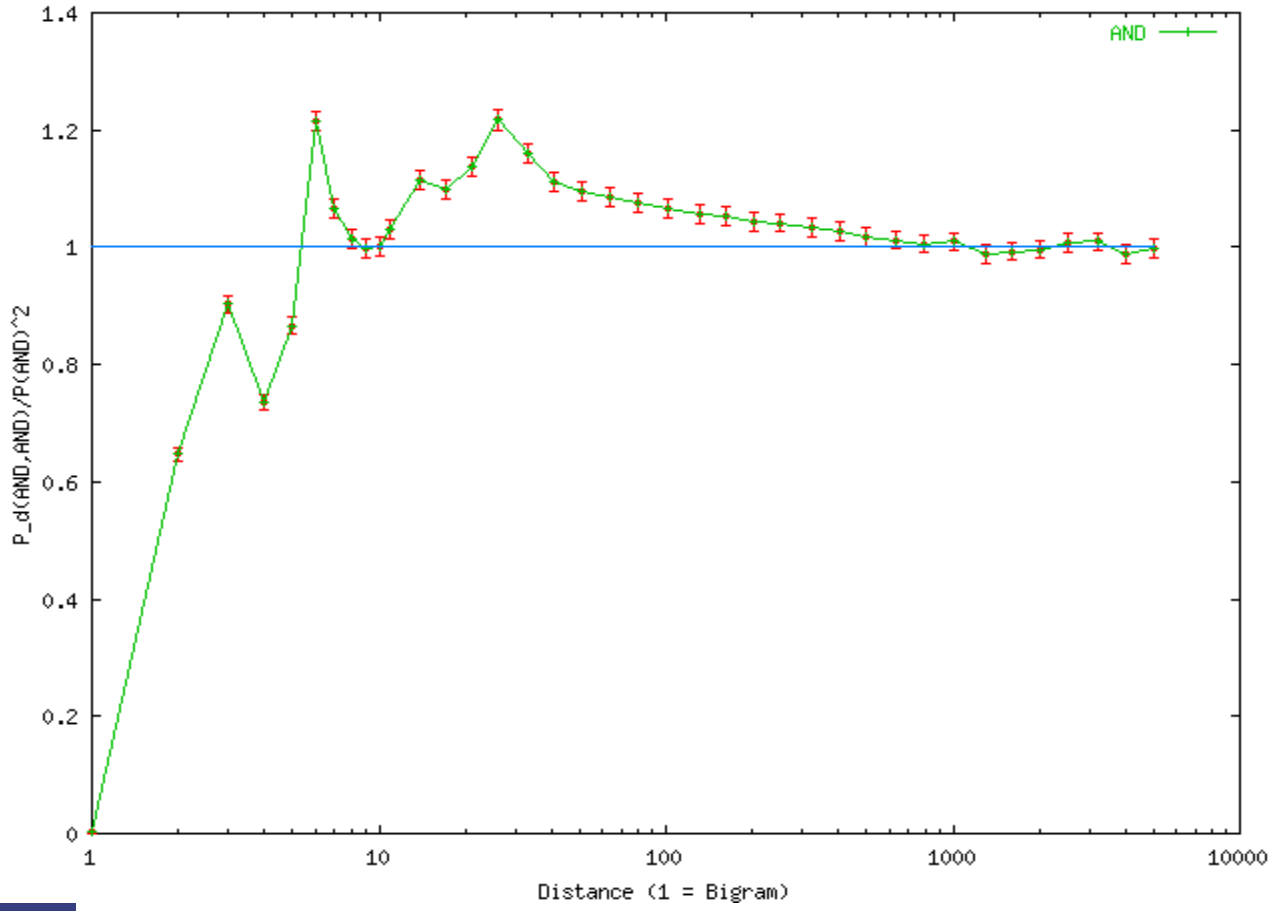
- Correlation Function:

$$c_d(w) = \frac{P_d(w\ w)}{P(w)^2}$$

  - d: distance between two observations of word w
  - Statistical independence: c(w)=1
  - Measures correlation

47

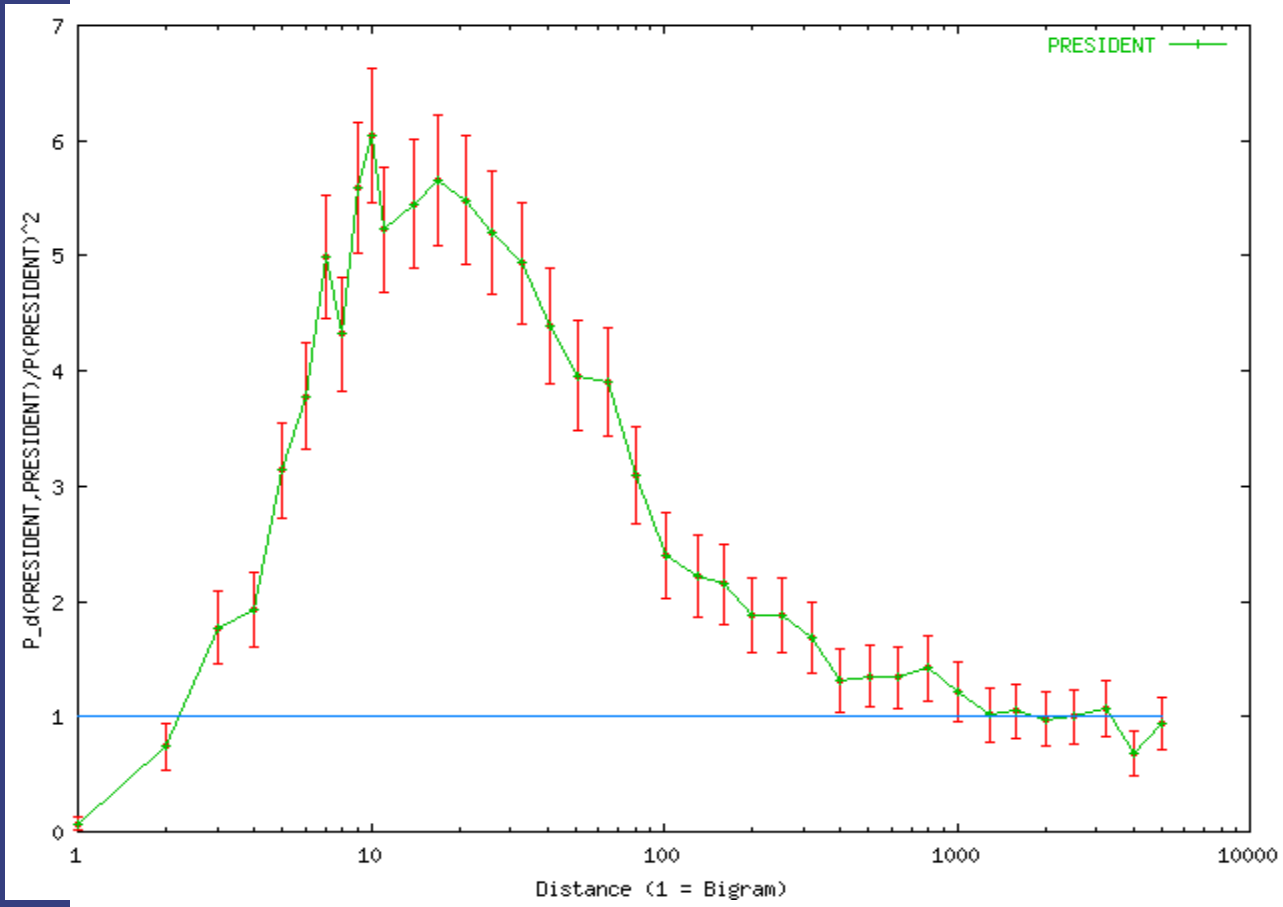
Correlation Function „and”



Only weak short range dependencies

48

Correlation Function „President”

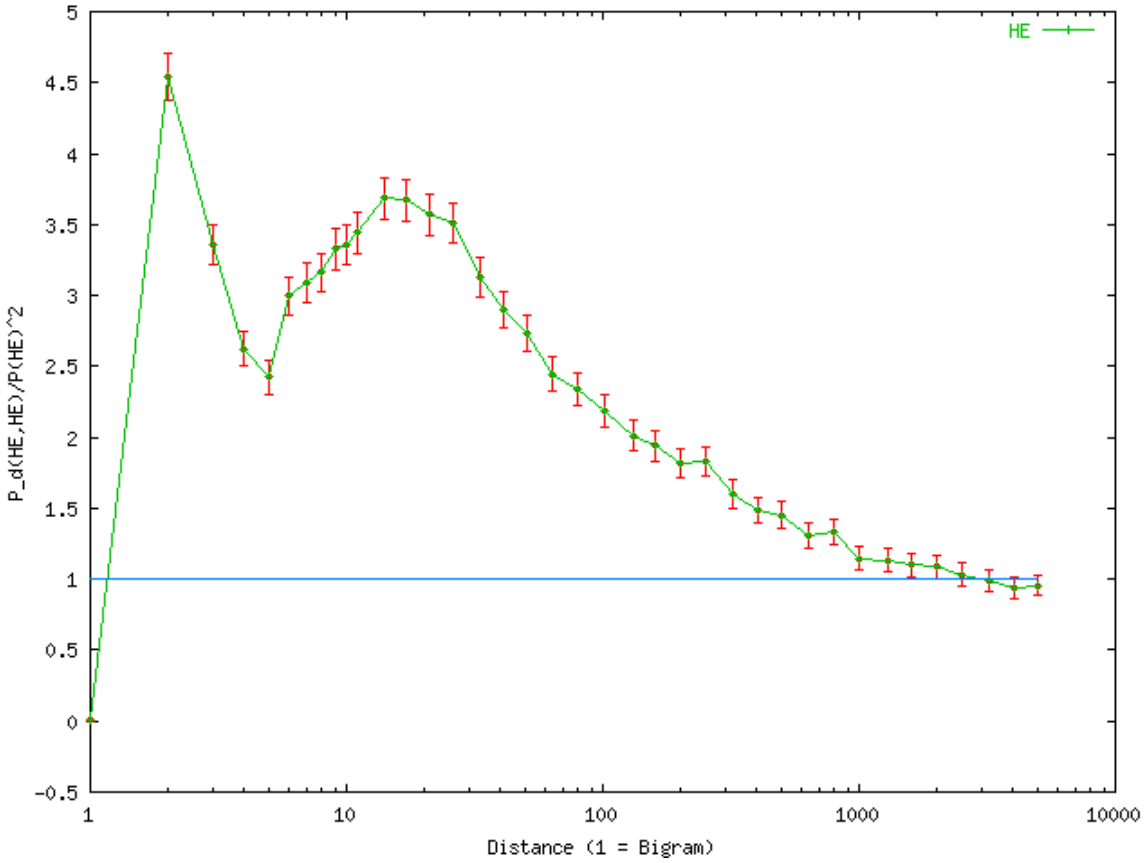


-Long range (semantic) dependency

-Decay of correlations after about 1000 words

49

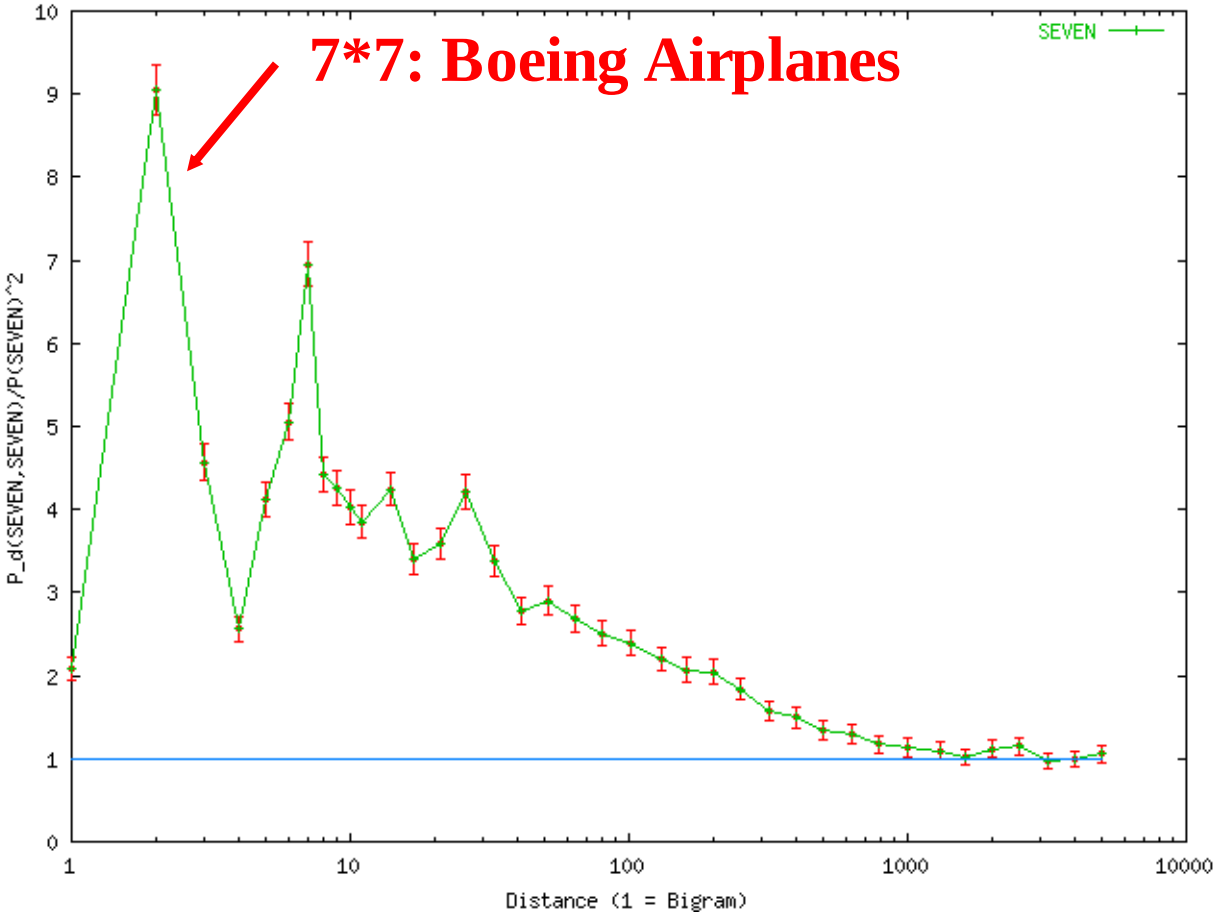
Correlation Function „he“



Short- and Long Range Dependencies

50

Correlation Function „seven”



51

Conditional Entropy

Definition

$$h_n = -\sum_{w_1 \ldots w_{n+1}} P(w_1 \ldots w_{n+1}) \log P(w_{n+1} | w_1 \ldots w_n)$$

Study examples to understand meaning of conditional entropy

52

Example I: independence of words (Bernoulli Sequence)

Assume $P(w_1 \ldots w_{n+1}) = \prod_{i=1}^{n+1} P(w_i)$

$$h_n = -\sum_{w_1 \ldots . w_{n+1}} P(w_1 \ldots . w_{n+1}) \log P(w_{n+1} | w_1 \ldots . w_n)$$

$$= -\sum_{w_1 \ldots . w_{n+1}} \left(\prod_{i=1}^{n+1} P(w_i) \right) \log P(w_{n+1}) \quad = -\sum_{w_{n+1}} P(w_{n+1}) \log P(w_{n+1})$$

$= H(W)$ $\mapsto$conditional entropy is independent of length

$\mapsto$conditional entropy equals one word entropy

53

Example II: periodic sequence

- After p symbols, the sequence repeats itself
- Example of sequence with period p=3
  - ABCABCABC…

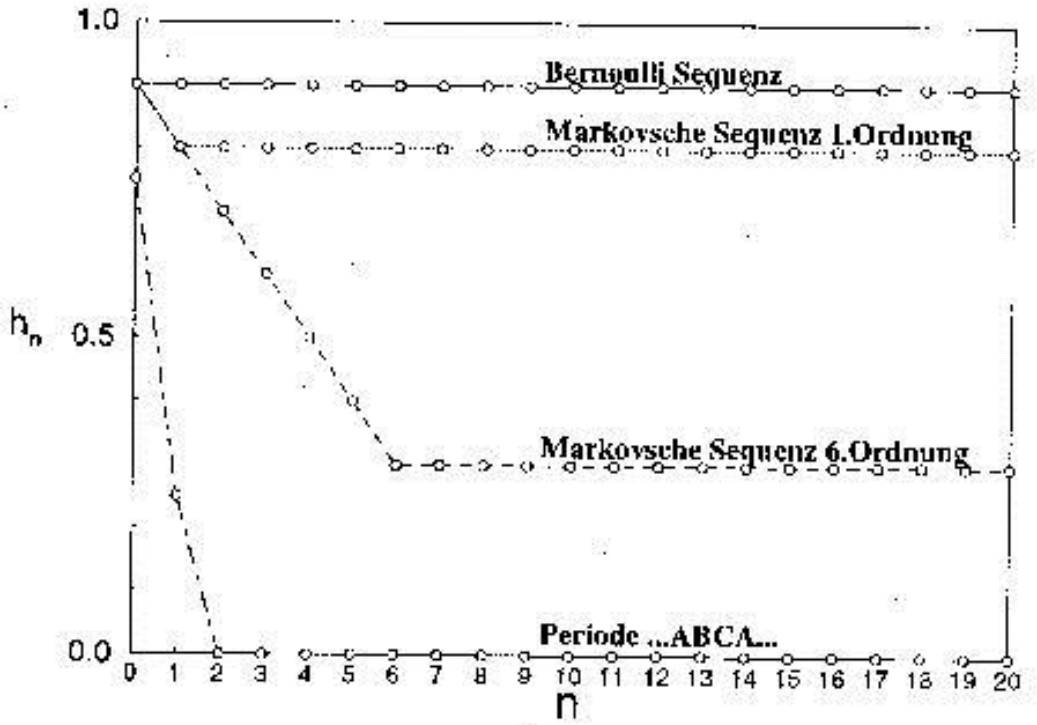
After one period p we have $P(w_{p+1} \mid w_1 \ldots w_p) = 1$

In general $P(w_{p+1+i} \mid w_1 \ldots w_{p+i}) = 1 \quad \text{for all } i \geq 1$

$\Rightarrow \quad h_n = 0 \quad \text{for } n \geq p$

54

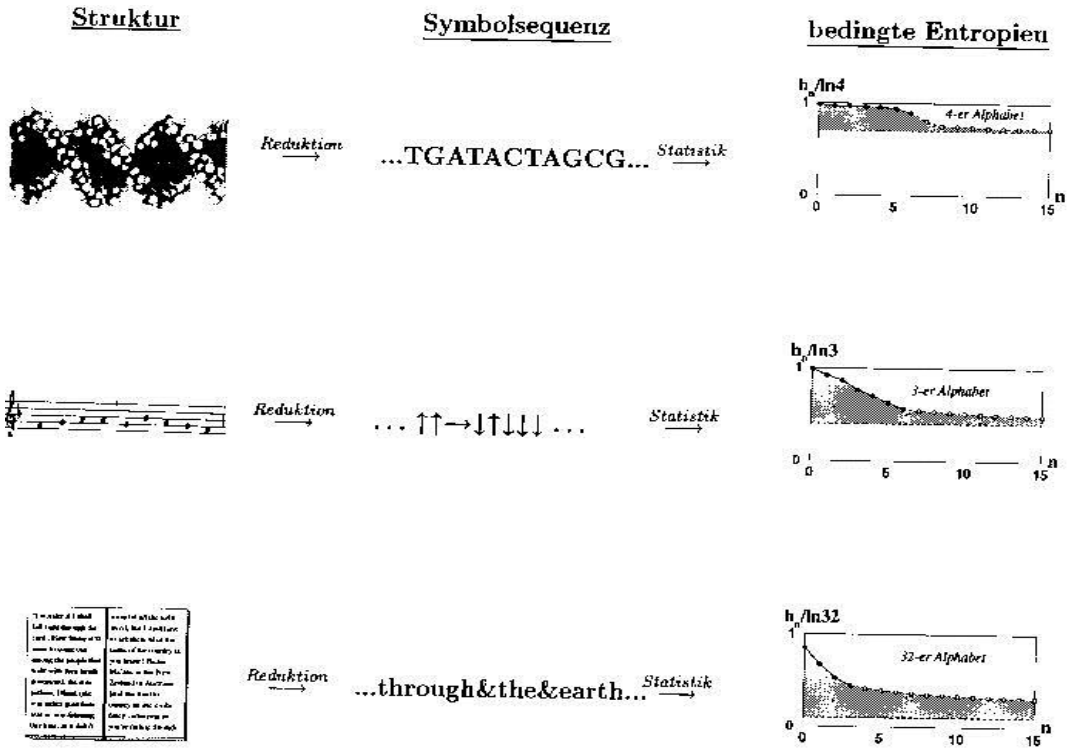
Long-Range Dependencies: Typical Sequences



Prototypical artificial cases

55

Conditional Entropy: Other “Languages”



From: Ebeling

56

Summary

- Entropy measures the information content of a message in bits
- It is related to the compressibility of text
- Probabilities of symbols determine optimal code length
- Lower perplexity $\mapsto$ better compression
- Long range dependencies in language:
  - Correlation function and conditional entropy

57