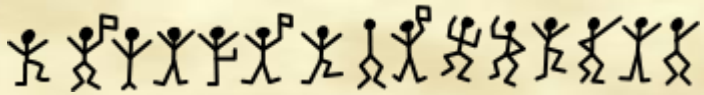
LANGUAGE Analysis

Simple Statistics

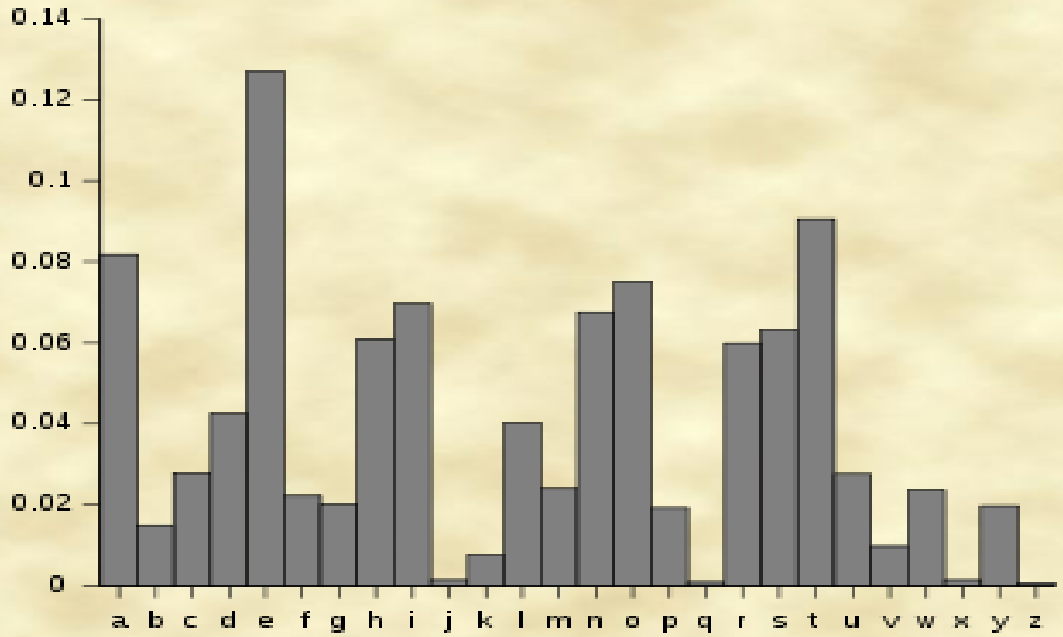
- Letter frequency
- Sherlock Holmes : “**The Adventure of the Dancing Men**”



- Frequency analysis

Frequency Analysis

Letter distribution



Beyond lexical analysis

- Correlations
- Frequency Time series
- Length time series

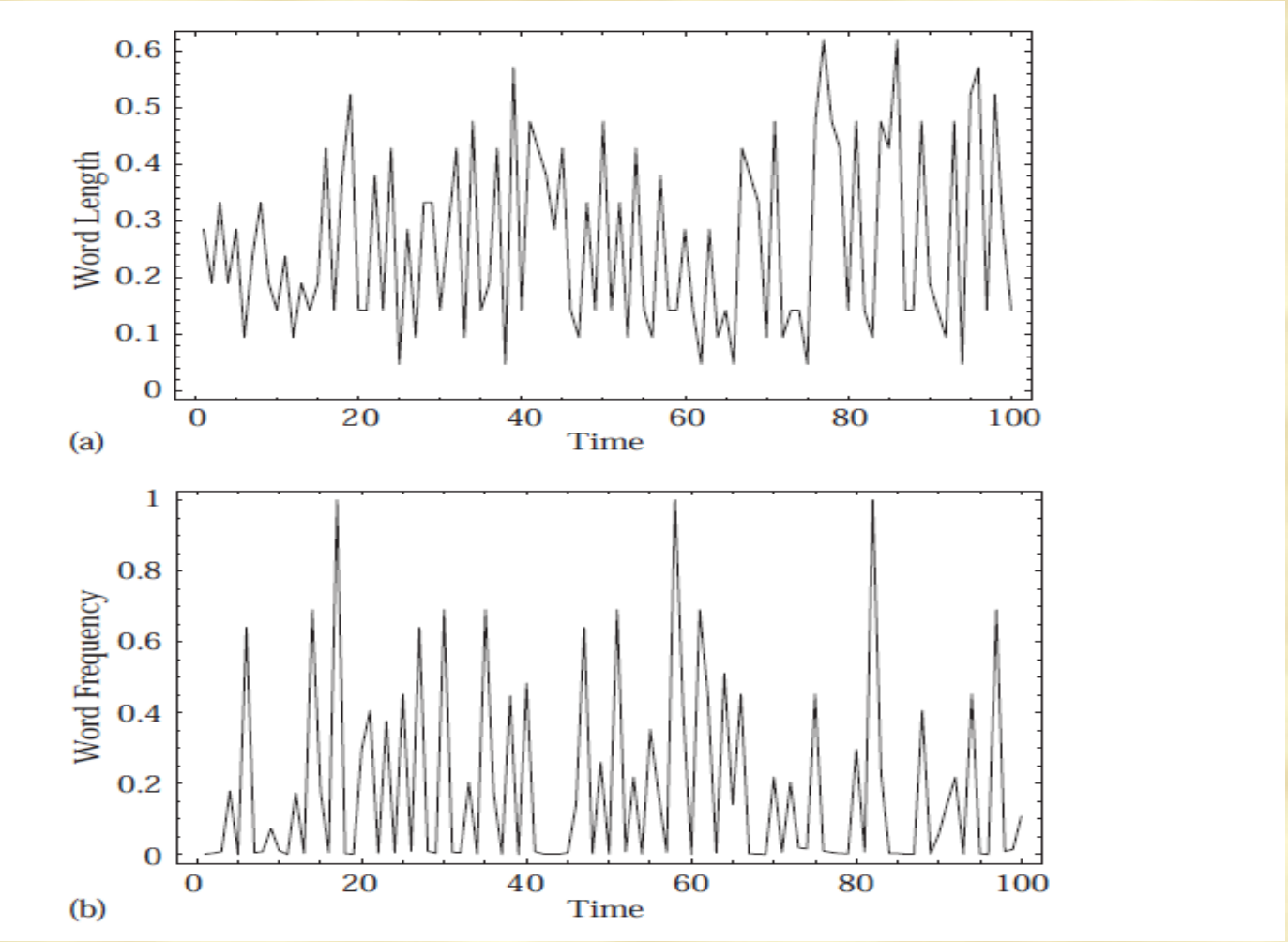
Methods

- Detrended Fluctuations Analysis
- Grassberger-Procaccia Analysis

Methods to create series

- LTS- Length Time Series
- FTS- Frequency Time Series

Series


Word Length
0.6
0.5
0.4
0.3
0.2
0.1
0
0
20
40
60
80
100
Time
(a)
Word Frequency
1
0.8
0.6
0.4
0.2
0
0
20
40
60
80
100
Time
(b)

Long Range Corelations

Do long words follow long words?

Do long words folow small words?

Or is it Random?

Long Range Corelations-DFA

- DFA compares the (detrented) series fluctuations with the fluctuations of a random walk.
- Natural languages are uncoralated.
- Computer code has long range corelations.

Long Range Corelations

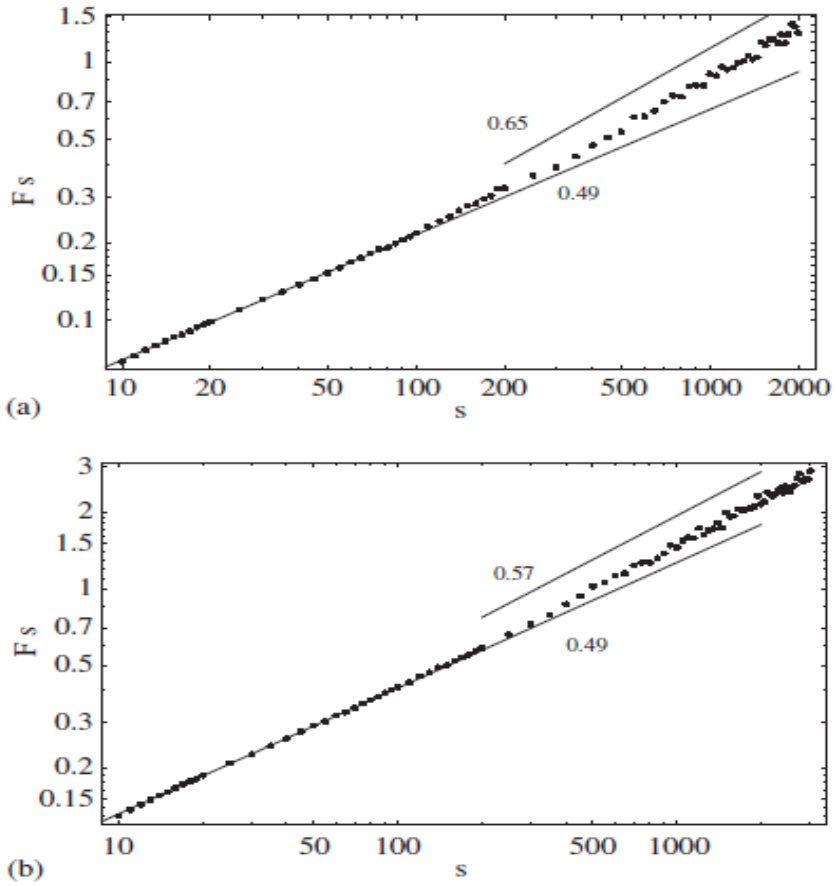


Fig. 3. (a) Plot of the DFA3 Fluctuation function $F(s)$ vs. s of the Length Time Series for an English document (A Christmas Carol by Dickens). The length of the series is 28 713 words. Again, a power-law behavior is observed and again the slope is equal to 0.48 very close to that of a signal with no or short range correlations. (b) Same analysis for the Frequency Time Series for the same English document. The initial slope here is 0.49, almost identical to that of an uncorrelated signal.

Long Range Corelations

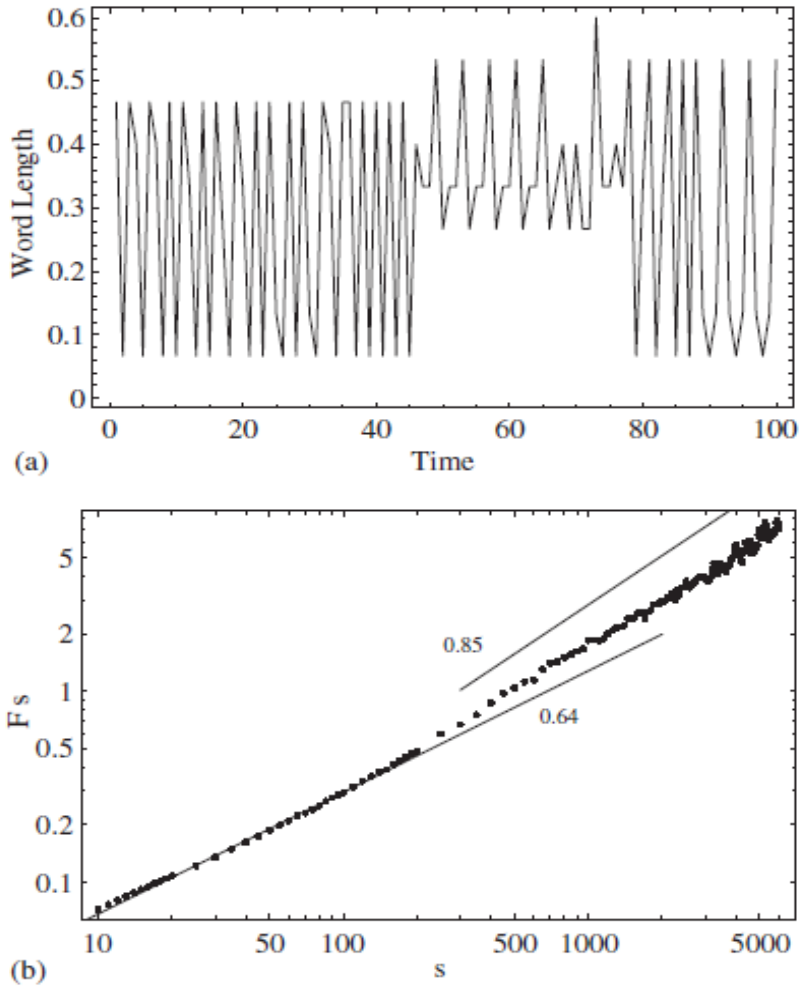


Fig. 4. (a) A Length Time Series for the Linux Kernel, and (b) DFA3 for the Linux Kernel. The signal seems to exhibit long-range correlations as the slope of the straight line is equal to 0.64.

GP Analysis

Time Series of a single variable contains all info of the dynamics

We can determine the dimensionality of the phase space.

Climate has a low dimensionality phase space (5 variables)

Language is an infinite dimensional system.

GP Analysis

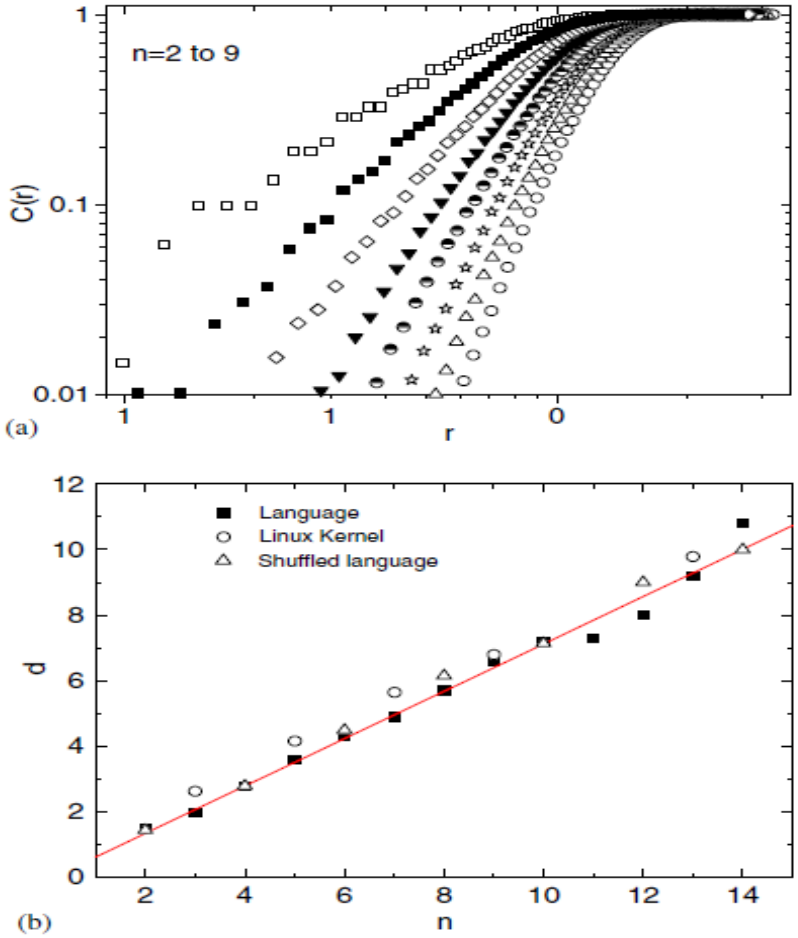


Fig. 5. (a) Double logarithmic plot of the correlation integral $C(r)$ for several embedding dimensionalities n. The exponents d are calculated from the slope of the straight line segments. (b) Plot of exponent d versus the dimensionalities n. *Rectangles*: Greek language document. *Circles*: Linux Kernel. The straight line has a slope equal to 0.69 and shows no saturation. *Triangles*: Shuffled Greek language document.