Searching for non-linearities in natural language

Kiril Ribarov and Otakar Smrž

Center for Computational Linguistics
Faculty of Mathematics and Physics
Charles University
{ribarov, smrz}@ckl.mff.cuni.cz

Inspired by wide range of applicability of what is commonly referred to as chaos theories, we explore the nature of energy series of a signal of human speech in the light of nonlinear dynamics. Using the TISEAN software package, analyses on various recordings of the language energy series were carried out (single speaker - different speeches; single speech - different speakers; dialogues; talkshows). Also correlated to other tenths of experiments conveyed on different linguistic inputs as written and morphologically analyzed texts, the presented experiment outputs (up to our knowledge, similar experiments have not been performed yet) reveal the complex and tricky nature of the language and are in favor of certain linguistic hypotheses. However, without further research, they do not encourage us to make explicit claims about the language signal such as dimension estimations (although probably possible) or attractor reconstruction.

Our main considerations include:

(a) a look into the stochastic nature of the language aiming towards reduction of the currently very large number of parameters present in language models based on Hidden Markov Models on language n-grams;

(b) visualization of the behavior of the language and revelation of what could possibly be behind the 'noisy' stream of sounds/letters/word-classes observed in our experiments; and last but not least

(c) presentation of a new type of signal to the community exploring natural non-linear phenomena.

Current inclusions of nonlinear dynamics in wider linguistic oriented studies can be found in:

- study of cognition in general (Thelen, Smith 1998),
- problems of evolution of natural language (Steels 2002),
- combinations of neural networks and language dynamics (Elman 1995),
- and others ranging from fractal structures in poetry texts, up to
- evolution modeling of development of words.

State-of-the-art tools currently applied in computational linguistic:

- statistical, based on Hidden Markov models
- maximum entropy
- various (probabilistic) formal grammars
- final state automata
- rule-based

The structure of the formal analysis of the langauge

1. The language as a dynamical system with its core and its perifery
2. The language in layers with two types of relations:
   - inter-layer (homomorphic) relations, and
   - layer-inherent relations.

It is very difficult to find out what is true about language, since what is observed is not what it always is.
The principle of least effort, expressed through the famous Zipf's law (later refined and brought closer to the language by Mandelbrot) does apply in various frequency studies like those of words or letters, or in the study of the number of sences and the plysemy of verbs, etc.

But, is there something else/more?

Written plain text: the language as a stream of characters evolving in time. O n c e u p o n a t i m e $l_1\ l_2\ l_3\ l_4 \ldots l_i \ldots l_n$

Ribarov, K. and Sgall, P. (1998). The Micro and the Macro of Linguistic Description. In *ELSNET in Wonderland Proceedings*, pp. 95-99, ELSNET.

Written text annotated with POS: the language as a stream of POS tags. Mary saw Jim with a telescope N V N P D N

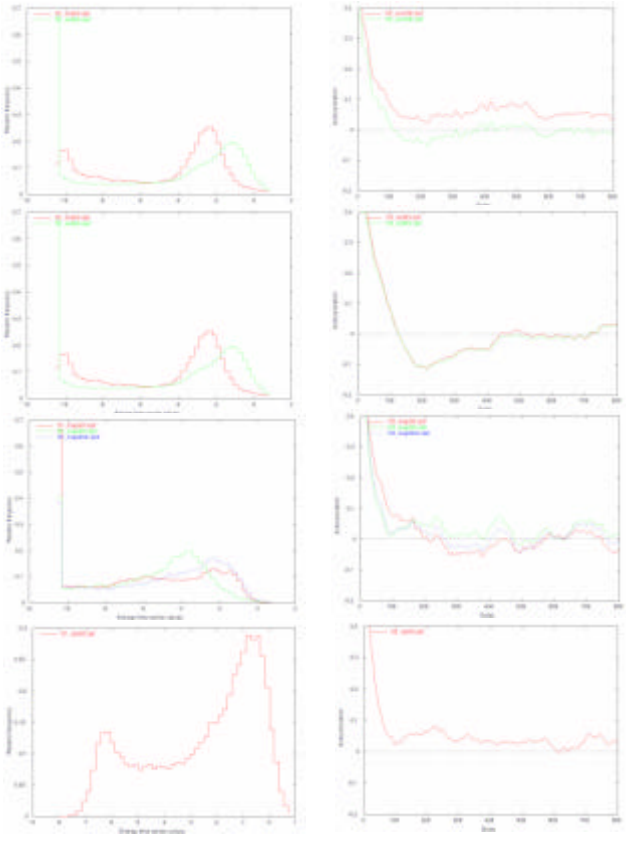
Ribarov, K. (2000). The (Un)Deterministic Nature of Morphological Context. In *Proceedings of the Second International Conference on Language Resources and Evaluation*, vol. III, pp. 1743-1747. Athens, Greece.

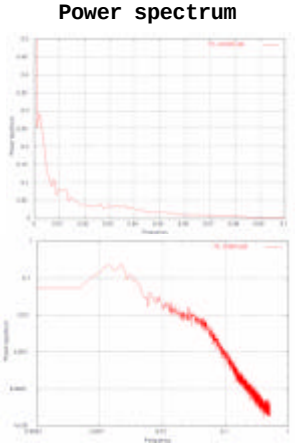
Similar studies of various levels are possible, where each level has different type of organization.
The separable language levels as assumed in theory are: phonetic, morphonological, morphematic, syntactic (surface and deep).

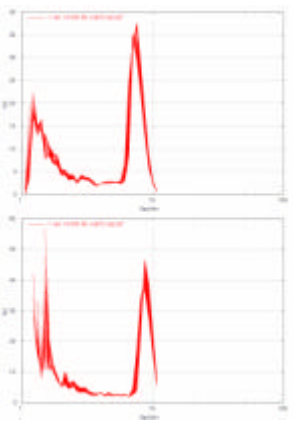
The spoken text energy series and its characteristics

Histograms | **Autocorrelation** | **Power spectrum** | **D2**

Some of the experiments
(all of them with qualitatively similar results)







Poincare section

The linguistics classifies the language as a system, being dynamic, sensitive to changes, orderred but still irregular. But, could we view it as dynamic, non-linear, or chaotic, or ? system?