The (Un)Deterministic Nature of Morphological Context

Kiril Ribarov

Institute of Formal and Applied Linguistics
Charles University
Malostranské náměstí 25
118 00 Prague
Czech Republic

e-mail: ribarov@ufal.mff.cuni.cz

Abstract

The aim of this paper is to contribute to the study of the context within natural language processing and to bring in aspects which, I believe, have a direct influence on the interpretation of the success rates and on a more successful design of language models. This work tries to formalize the (ir)regularities, dynamic characteristics, of context using techniques from the field of chaotic and non-linear systems. The observations are done on the problem of POS tagging.

1. Motivation

The natural language, within the Prague Linguistic School is a system of layers, where each layer by itself is a system with many relations and its own semantics, a dynamical system of core and peripheral units. According to the classical European traditions the analysis is a process of giving the form its meaning, which, when being projected to the stratified model of the language (Sgall, Hajičová, Panevová, 1986) is always between two adjacent layers ordered as follows: phonetic, morphonological, morphematic, syntactic, and tectogrammatical layer. So, in order to move from a 'lower' layer to an 'upper' one, we rely on the knowledge of the lower layers; the form from the lower layer gets its meaning on the way up. In that vein, context-based methods locate the additional knowledge in the context of an analysed form at a given layer.

Fig. 1 tries to summarise the results of many theoretical linguistic works which claim that (and give explanations why) ϵ is non-negative.

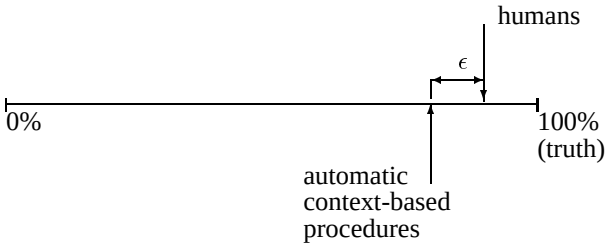

Figure 1: Is it ever possible to improve automatic context-based procedures?

What is the structure of the context then, and within it, how small ϵ can be, in other words, what the information richness of a given layer in our case the morphematical layer is?

Hence, this work would like to contribute to the study of the context, mainly by showing its importance and at the same time revealing some of its characteristics. The latter can be fruitful especially when building more successful language models and a better approach to coefficient's weights, first of all for statistical approaches. It is a fact that all of the modern methods dealing with natural language knowledge acquisition (language modeling, statistical as well as rule-based methods, analogy/memory based methods, systems based on grammars or any type of transducers/automata, neural networks) use the context as their 'world'.

In order to exemplify the context, one of the lowest layers has been selected, the morphematical layer. The data were taken from the Prague Dependency Treebank (PDT). The PDT is annotated by Czech morphological tags (positional structure, defined as a concatenation of 13 morphological categories); each morphological category corresponds to a single position. For further description on morphological tagging of Czech and on the description of the morphological tagset see (Hladká, 2000).

2. The Context within Non-Linear Dynamics

Motivated by (Ribarov, Sgall, 1998) and in order to support our analyses by a more formal background, modern techniques and procedures from the field of non-linear dynamics were used. Our study tries to reveal the (ir)regularities found within a context, projected on analysis of (ir)regularities inherent to the problem of rich tagging of Czech[1].

Let $T_{all} = (t_1, t_2, t_3, ..., t_i, ...)$ be tags excluded from $w_1|t_1 w_2|t_2 w_3|t_3 ... w_i|t_i ...$[2], where t_i is a morphological tag of the token w_i. Since w_i evolves in subsequent time intervals (as it is written or spoken), t_i has the same evolving characteristics.

Let $t_i \rightarrow x_i$ be a one-to-one mapping, where x_i is a member of [0,1]. Thus, each morphological tag has been substituted by its numerical identification, all of them being normalized on the interval [0,1]. So, the observed set,

[1] The Czech positional tag set consists of around 3,000 tags.

[2] taken from the PDT

serving as an input of further analyses, is the set $X = (x_1, x_2, x_3, ..., x_i, ...)$ extracted from the PDT, selected at random, and consists of 100,000 samples. I will refer to the set X also as to a signal.

In this study in the inside of X is of main interest: What kind of system generates such a signal? Is it a deterministic dynamical system? What is the character of the dynamics of such a system: is it periodic, or quasi-periodic, or (deterministic) chaotic, or of a different nature?

1. The study of the power spectra of the signal reveals periodic and quasiperiodic phenomena in the case when dominant frequencies are present with at least an instrumental width *2π/T*, where *T* is the length of the signal used (Eckmann, Ruelle, 1985). In the case of many frequencies, the couplings between the modes corresponding to distinct frequencies make the signal to be no more (quasi)periodic. The suspicion is that such a signal can be chaotic (Eckmann, Ruelle, 1985) - containing more or less broadened non-instrumentally sharp peaks, its spectrum is broadband and with noisy background. In our case (Fig. 2) and in a detail of it (Fig. 3), it is not possible to state dominant frequencies; we observe a broadband spectrum, distribution of peaks all over the spectrum and a probable presence of noise on the background. Hence, X is neither periodic nor quasiperiodic.

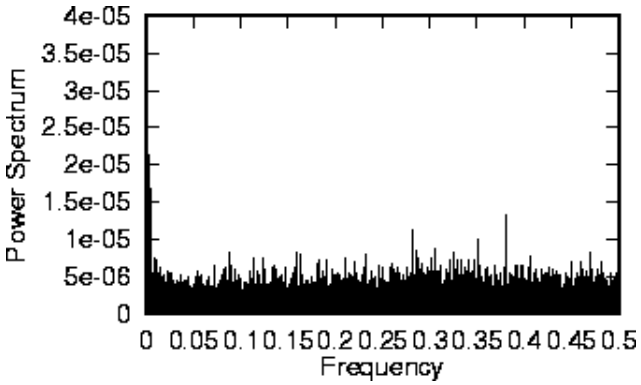

Figure 2: Full power spectrum

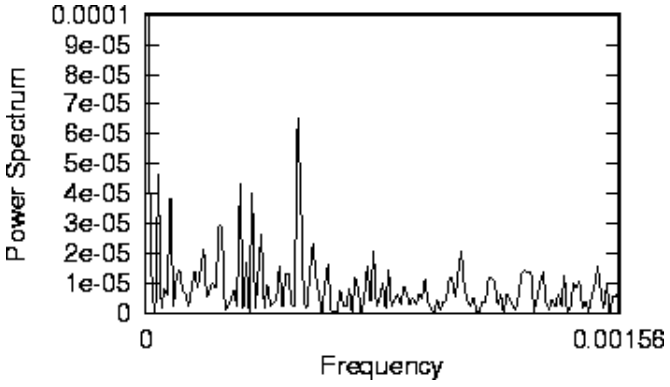

Figure 3: Detail of the power spectrum

2. To detect the power spectrum is not enough in order to support e.g. the statement that X is chaotic. The theory of the non-linear dynamical systems suggests a calculation of a dimension *d* of X, entropy (production of information within the system) and different characteristic exponents (Singular Value Decomposition, Lyapunov exponents)[3] These may serve in order to: (i) embed X in a suitable phase space, and (ii) make further conclusions upon the classification of X.

From X, using the methods of delays, m-dimensional vectors $y(i)$ were constructed, $y(i) = (x_i, x_{i+\tau}, ..., x_{i+(m-1)*\tau})$, where τ is a time delay.

The dimension *d* of X can be defined as an amount of information needed to specify X accurately. The idea is to construct N(ϵ) balls of radius ϵ in order to cover X. One way of establishing the dimension of phase space in which embedding of the system exists (embedding dimension *m*)[4] is by calculating the dimension *d* of X, that is the correlation dimension of X, as proposed by Grassberger and Procaccia (Grassberger, Procaccia, 1983):

$$dimX = \lim_{\epsilon \to 0} \frac{\log c(\epsilon, m)}{|\log \epsilon|}$$

$$c(\epsilon, m) = \frac{1}{N_{pairs}} \sum_{i=m}^{N} \sum_{k<j-\omega} \theta(\epsilon - |y_j - y_k|)$$

where N_{pairs} = (N-m+1)(N-m-ω +1)/2, θ is a step function and ω is Theiler window for excluding temporally correlated points.

3. The main problem in determining the correlation dimension is a choice of appropriate time delay between two subsequent points (determining the input vectors). The time lag was established from the autocorrelation function (being strictly periodic for periodic activity, or decays in time for the case of quasi-periodic or random processes). The time of the first zero of the autocorrelation function is supposed to be the optimal time delay τ (Fig. 4). For X, this function decays in time and has its first zero at i=233.

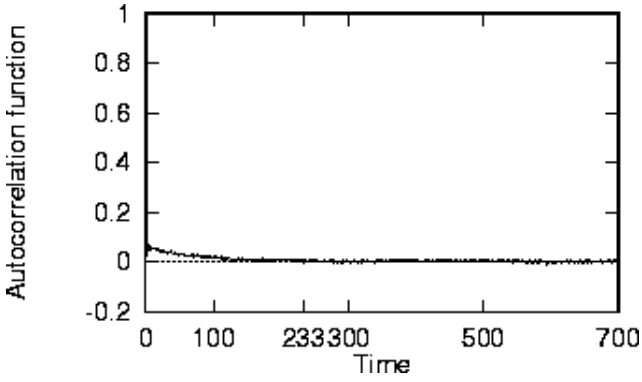

Figure 4: Autocorrelation function

[3]Calculations were provided using procedures from the TISEAN 2.0 package.

[4]According to the theorem of (Takens, 1981) and (Manne, 1981) if *d* is the dimensionality of the system, then $m \geq 2d + 1$ dimensional phase space is enough in order to provide its well embedding.

Another way to establish τ, is from the mutual information (Fig. 5), as its first minimum. As for X, the function decays in time, and it is difficult to use it for our purpose. Nevertheless, it supports the results from the autocorrelation function of a bigger τ[5].

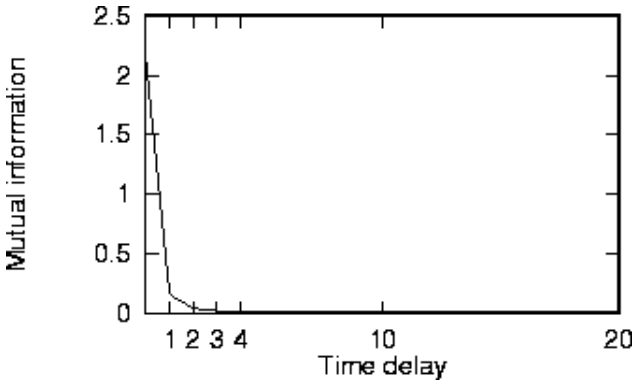

Figure 5: Mutual information

Fig. 6 shows, for each presupposed embedding dimension (tested for $m \geq 2$ and $m \leq 40$) the correlation sum on a log-log scale. Experimental determination of the dimension is done by determining the slope of the linear parts of the logarithm of the correlation sums dependent on the $\log \epsilon$. It is noticeable from Fig. 6 that it is difficult to locate a proper linear parts and study the convergence of their slopes. Further analysis at this point would lead to many speculations. Thus we tried another suggested method, the method of false nearest neighbors for a delay of 233 (Fig. 7). The embedding dimension results to be 12.

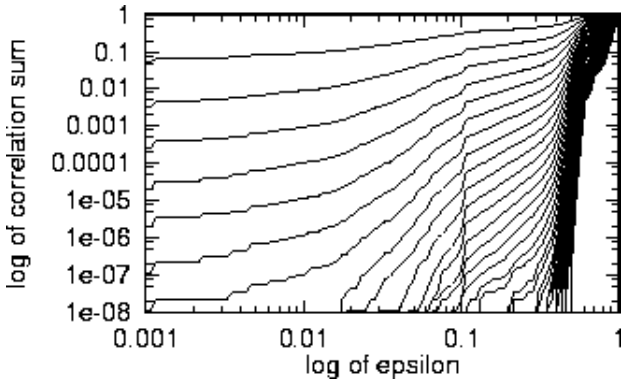

Figure 6: Correlation sums

4. In further revealing of the character of the signal a lower bound K of the Kolmogorov-Sinai entropy (sum of positive Lyapunov exponents, for which there is no safe way of calculation, thus they were not calculated) was calculated by a box counting approach[6].

[5]Different values of τ were tested as well.

[6]Let p_i is the probability to find the system state in box *i*, then the order *q* entropy is defined by the limit of small box size and large *m*:

$$\sum p_i^q \approx \epsilon^{-mh_2}$$

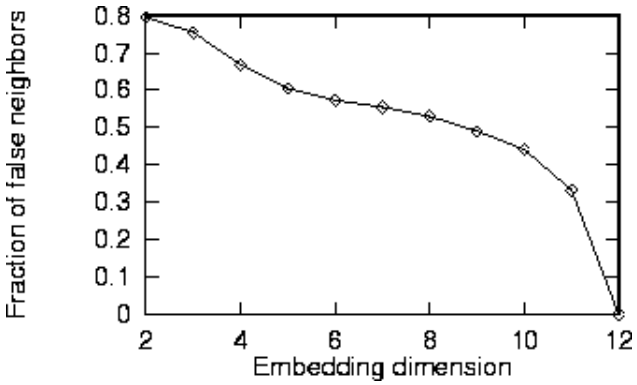

Figure 7: False nearest neighbors

For an ordered system K=0, for a random system K is infinite, while in a chaotic system K is constant and different from zero. Fig. 8 determines K to be constant and nonzero (it converges to a value approximately 11.5 as $\epsilon \to 0$ and $m \to \infty$).

5. In the attempt to plot X in 12-dimensional space, we tried to select the correct projection, from which the structure of the phase space occupied by X, reveals its properties. Singular Value Decomposition was used for this purpose the results of which show that X spreads broadly in the phase space without a 'special' preference towards any of the vectors of the orthogonal base of the 12-dimensional system.

Fig. 9 shows the 1-2 projection, marking the existence of frequently visited regions of the phase space (other projections share the same characteristics). It was not possible to present clear visibility of an attractor due to the fact that X spreads broadly in the phase space, with a lot of jumps between the noticeable regions presented.

3. Conclusion

The above analysis reveals that the data set X of tags has the following properties:

- The power spectrum does not contain distinguishable frequencies, it is broad band, and probably contains noise.
- Experiments carried out with different embedding dimension, with different time delays, show that proper capturing of X in a phase space is rather difficult. Presenting the results needed many supporting experiments (provided on the background), and different angles of view at X.
- The fact that the entropy is constant and greater than zero, supports that X is a result of a deterministic process.
- X seems to be highly non-linear.

.

- The embedding dimension as calculated by the False Neighbours procedure and observed as well on the log curves of the correlation sums, is established to be 12.
- The mutual information shows a rapid decrease (at 3).

In relation to the context, one may conclude that:

- The context contains rich but rather 'irregular' information.
- It will be very difficult (if at all possible) to find a recipe of what is the correct, i.e. the optimal context of a form. Selecting the context pattern can be based on the non-linear mapping parametrized for the specific application using the methods presented here.
- The relevant information is most likely to be found within a context of approximately 12 forms.
- A selective context (context patterns) should be applied in order to guarantee a higher success rate and better knowledge extraction.

4. Further considerations

Explaining any of the results needs extreme precautions. Modeling of the context pattern as a non-linear deterministic system is most likely the direction in which further studies should be made. For that purpose the non-linear analysis of given data sets (in our case the set X) needs further, more-in-detail analysis towards a possible attractor reconstruction, that can give the key of how to choose the contexts dynamically. We plan to consider various experiments on subtag levels of the positional morphological tags.

5. Acknowledgement

The results described herein have been obtained within the project No. 405/96/K214 sponzored by the Czech Grant Agency.

6. References

J.- P. Eckmann, D. Ruelle. *Ergodic Theory of Chaos and Strange Attractors*. Reviews in Modern Physics, 57, pp.617-656, 1985.

P. Grassberger and I. Procaccia. *Measuring the Strangeness of Strange Attractors*. Physica 9D, pp.189-208, 1983.

B. Hladká. *Czech Language Tagging*. PhD Thesis at the Faculty of Mathematics and Physics, Charles University, Prague, 2000.

R. Manne. *On the Dimension of the Compact Invariant Set of Certain Nonlinear Maps*. In Dynamical Systems and Turmulence, D.A. Rand, L.-S. Young, eds., Springer, Berlin, pp.230-242, 1981.

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

P. Sgall, E. Hajičová and J. Panevová. *The Meaning of the Sentence and its Semantic and Pragmatic Aspects*. Reidel, Dordrecht, Holland, 1986.

F. Takens. Detecting Strange Attractors in Turbulence. In *Dynamical Systems and Turbulence*, D.A. Rand, L-S. Young, eds., Springer, Berlin, pp.366-381, 1981.

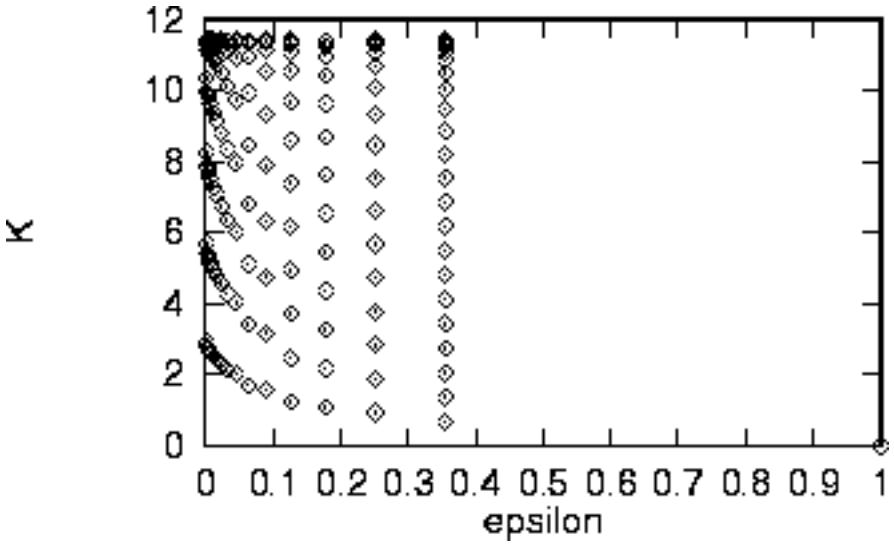

Figure 8: Lower limit of the Kolmogorov-Sinai entropy

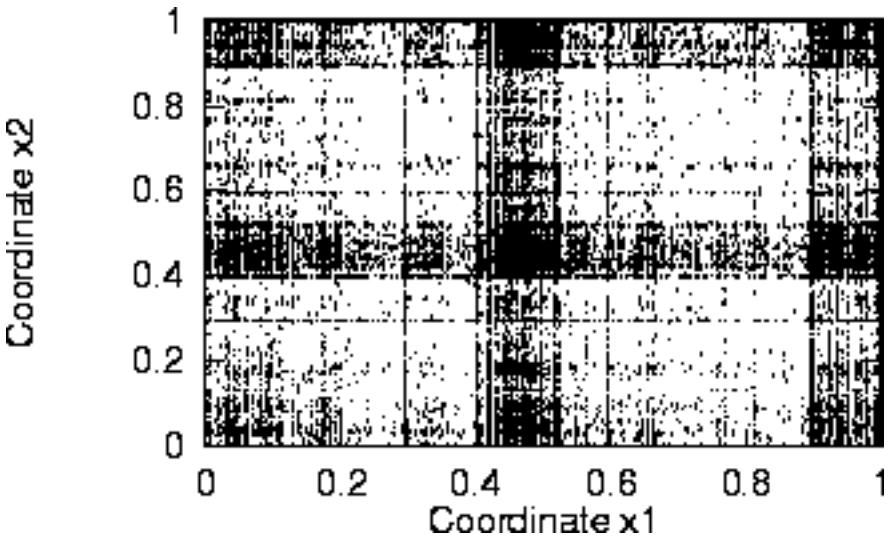

Figure 9: 1-2 projection from the 12-dimensional phase space