Abstract. We investigated the informational structure of written texts (also in the form of speech transcriptions) using Recurrence Quantification Analysis (RQA). RQA technique provides a quantitative description of text sequences at the orthographic level in terms of structuring, and may be useful for a variety of linguistics-related studies. We used RQA to measure differences in linguistic samples from different subjects. They were divided in subgroups based on personality and culture differences. We used RQA and KRQA (Cross Recurrence) to measure the coupling and synchronization during the conversation (semiotic interaction) of different subjects. We discuss results both for the improvement of methodology and some general implications for neuro-cognitive science.

INTRODUCTION

Claude E. Shannon posed the foundations of modern studies on the informational structure of texts and speech [1] His less famous work on the prediction and entropy of printed English [2] is a source for inspiring new research.

While we could place Shannon’s contributions at the micro-semiotic level George K. Zipf [3] was mostly working at the meso-semiotic level (words). He described two laws of word frequencies:

1. the Rank-Frequency Law: when $n$ tokens in a text are sorted by decreasing frequency and a rank number is assigned to each token. The plot of $\log$ (frequency) (y axis) versus $\log$ (rank) (x axis) approximates a straight line of slope $n^{-1}$.
2. The Number-Frequency Law: being $n$ the frequency of a token, the plot of $\log$ (n) (y axis) versus $\log$ (number of tokens with frequency n) (x axis) approximates a straight line of slope $-0.5$.

Recent studies used different approaches Schenkel [4], for example, studied long range correlation in various human writings by mapping them into a simple 1d random walk model. This approach allowed to obtain better quality scaling data than the traditional power spectrum methods. Optimally written computer programs seem to have the highest scaling exponent, i.e., close to that of ideal 1/f noise. Literature, on the other hand, leads to smaller exponents. The Bible, however, has the strongest correlation among the Roman letter writings examined.

Amit [5] studied the long-range correlations in various translations of the Bible by mapping them onto a one dimensional random walk model. He concluded that an
exponent $\alpha$, which characterizes the algebraic decay law of these correlations depends on both language and coding.

Maturana & Varela [6], from another perspective, suggested that structural coupling is a process which occurs when two structurally plastic systems repeatedly perturb one another’s structure in a non-destructive way over a period of time. This leads to the development of structural ‘fit’ between the systems. There is an intimate relationship between this process and the emergence of ‘appropriate’ behaviour from the interplay between interacting systems, because the structure of a system determines its responses to perturbing environmental events. Maturana stressed this dynamical approach in semiotic terms: “Language is a manner of living together in a flow of coordinations of coordinations of consensual behaviours or doings that arises in a history of living in the collaboration of doing things together” [7]. It is a co-evolutionary perspective.

This dynamical systems’ approach leads, naturally to control & synchronization in complex systems. The history of synchronization goes back to the 17th century when the famous Dutch scientist Christiaan Huygens (1673) reported on his observation of synchronization of two pendulum clocks. Systematic study of this phenomenon, experimental as well as theoretical, was started by Edward Appleton (1922) and Balthasar van der Pol (1927). They showed that the frequency of a triode generator can be entrained, or synchronized, by a weak external signal with slightly different frequency. These studies were of high practical importance because such generators became basic elements of radio communication systems. Mutual synchronization of two weakly nonlinear oscillators was analytically treated by Mayer (1935) and Gaponov (1936). An important step was done by Stratonovich (1958) who developed a theory of external synchronization of a weakly nonlinear oscillator in the presence of random noise.

Pecora & Carroll [8], Ott, Grebogi & Yorke [9] and Pyragas [10] opened a new reliable way to contemporary research on control and synchronization. A comprehensive framework has been defined by Boccaletti et al. [11, 12] and Pikovsky, Rosemblum & Kurths [13].

**FRAMEWORK**

The dynamical systems’ approach to language is defining the new trend in linguistics and semiotics [14, 15, 16]. Following this theory, lexicon consists of regions of state space within a semiotic system. Grammar consists of the dynamics (attractors and repellers) which constrain dynamics in that space. Representations are highly context-sensitive, continuously varied and probabilistic.

Objects of mental representation are trajectories through mental space rather than things which are constructed. Representations are not abstract symbols but rather regions of state space. Rules are not operations on symbols but rather embedded in the dynamics of the system: dynamics which permits movement from certain regions to others while making other transitions difficult.

Between the different methodologies available in dynamical systems theory we found Recurrence Quantification Analysis – RQA particularly useful for our scopes. RQA was first introduced in physics by Eckmann, Kamphorst & Ruelle [17]. Later, Webber and Zbilut [18] enhanced the technique by defining nonlinear variables that were
found to be diagnostically useful in the quantitative assessment of time series structure in fields ranging from molecular dynamics to physiology, see also Giuliani & Manetti [18].

A recurrence plot is a 2-dimensional $N \times N$ pattern of points where $N$ is the number of embedding vectors obtained from the delay coordinates of the input signal. From the occurrence of lines parallel to the diagonal in the recurrence plot it can be seen how fast neighboured trajectories diverge in phase space. Therefore, the average length of these lines is a measure of the reciprocal of the largest positive Lyapunov exponent. The main RQA parameters are:

- $n# \text{REC} =$ Total number of recurrent points in upper triangular area.
- $n\% \text{REC} =$ Percent recurrence $= \#\text{RECURS} / \text{triangular area}$. 
- $n\% \text{DET} =$ Percent determinism $= \#\text{recurrent points forming upward diagonal lines} / \#\text{RECURS}$ (set equal to $-1$ if $\#\text{RECURS} = 0$).

Recurrence plots help revealing phase transitions and instationarities. Visible rectangular block structures with a higher density of points in the recurrence plot indicate phase transitions within the signal.

**STRUCTURE**

In a preliminary work [20] we investigated the informational structure of written texts (also in the form of speech transcriptions) using Recurrence Quantification Analysis (RQA).

Mathematical language analysis has enjoyed varying degrees of success but the use of orthography as a basis for computational studies of human languages has experienced much less enthusiasm owing to the view that orthography actually represents a superficial amalgam of phonological changes and conventions (an opinion which has had little objective confirmation).

In order to evaluate this view, a reliable metric is needed to compare different texts (orthographic sequences) corresponding to speech samples. This implies a technique able to identify and measure the amount of prosodic structuring of a given text. Important requirements the technique must meet include: 1) relative invariance with respect to the original language of the texts; 2) consistency with already known and recognizable prosodic structures; 3) maximal dependence upon dynamical (order dependent) features of texts and a relative independence upon statistical (order independent) features. Furthermore, the chosen technique must be independent of distributional assumptions.

The preliminary experimental evidence suggests that recurrence quantification analysis (RQA) meets these requirements.

**Linguistic Data Sets.**

1. Italian poems of the 20th century (set 1): 9 Italian poems of the last century were chosen with the aim of maximizing the range of variation of the degree of prosodic structuring. The texts include poems with a minimum level of redundancy, as well as very structured poems using long repetitive motifs.
2. American (English) poems of the XXth century (set 2): this set is made up of 11 American poems chosen with the same criteria adopted for set 1.
3. Swedish contemporary poems and their correspondent Italian translations (set 3): this set is made up of 12 Swedish poems together with their Italian translations. It is
important to note that the translation was a poetical one, i.e. the translator had the goal to reproduce the prosodic structure of the original poem, so as to provide a check for the language invariance of the method.

4. Italian speech samples (set 4): transcriptions of the speech samples of 7 people with different cultural backgrounds (3000-4000 letter lengths).

5. American (English) speech samples (set 5): transcriptions of the speech samples of 10 university students. The recorded speech samples correspond to periods of time in which the individuals are freely and fluently speaking with no external constraints and after a period of adaptation to the presence of the recorder.

As pertains to this work, it is sufficient to explain the 2 basic descriptors: percent recurrence (REC) and percent determinism (DET). REC is the percentage of recurrent pairs of points in the above described distance matrix: a pair is considered recurrent if the distance between the elements is lower than a predetermined cut-off. In our case, given the symbolic character of the studied series, the cut-off was set to 0, and only totally super-imposable epochs (text segments) in the embedding matrix are considered recurrent. This choice makes the analysis completely independent from the chosen code. DET is the percentage of recurrent points that appear in sequence forming diagonal line structures in the distance matrix. DET corresponds to the existence of patches of recurrent behavior in the studied series.

![Figure 1](image.png)

**FIGURE 1.** Plot of structure distribution of all described linguistic samples: a combined analysis of American poems (AMP), Italian poems (ITP), American speech (AMS), and Italian speech (ITS).

The figure above shows a combined analysis of American poems (AMP), Italian poems (ITP), American speech (AMS), and Italian speech (ITS). It is worth noting the common scaling of the texts along a linear relationship between REC and DET, \( r = \)
0.87, p < 0.001). This common scaling points to a possibility of using the position on the REC-DET plane as a simple numerical index of the relative complexity of a text.

In our application of this methodology, texts were considered as continuous streams of letters with no interruptions, and letters were coded in alphabetical order. The embedding dimension was set to 3 in order to situate the window of analysis at the word level. Different choices of embedding dimension (4 and 5 dimensions) gave coinciding results (data not shown). The choice of a 3-dimensional embedding maximizes the sensitivity of the method by increasing the number of scored recurrences without disturbance from low-level statistical features of language (asymmetrical distribution of couples of letters typical of each language). A recurrence is scored whenever 2 identical 3 letter patterns are recognized along the sequence. The departure of the analyzed texts from random sequences was evaluated by comparing the real sequences with their randomly shuffled counterparts. The shuffling procedure keeps invariant the statistical features of the studied sequence so as to evaluate the amount of pure dynamical (i.e. phasic) information present in the text at the orthographic level.

**Overview of results:**

- **Sets 1 and 2 (Italian poems):** it is worth noting the disposition of poems going from maximum complexity (left-lower corner of the graph) to very recurrent, deterministic structures (right-superior corner). The complex end is made up of poems with no discernible rhyme or rhythmic structure, while the low complexity end is occupied by texts with highly repetitive motifs. When poems were shuffled, the randomized series with RQA profiles were hardly discernible from the most complex texts and very far apart from the highly structured poems. All the poems, when shuffled, gave very similar results [Mean (REC) = 0.35, SD = 0.01; Mean (DET = 8.00; SD = 0.32)], and were different from their shuffled counterparts with the only exceptions being SAB1 and CAPRI1. The high correlation between REC and DET variables (Pearson's r = 0.90, p < 0.01) highlights a common scaling of poems in the REC-DET plane pointing to a common structural design of texts, which was further confirmed by the next analyses. There is common general structure for the poems of different languages. Reading the texts makes apparent how the two poems are characterized by the same general structural feature; i.e., the iteration of a common pattern organizing the text. 3) The extremely high redundancy of Doctor Suess' poem is immediately recognizable and probably represents a real maximum of structuring degree as well as RF1. CAPRI1 and SAB1, given their contiguity with random sequences, probably constitute a real minimum.

- **Set 3 (Swedish poems and Italian translation):** the analysis of this data set allowed us to investigate the issue of the recognition of a given prosodic structure across two different languages. The demonstration of the possibility to recognize the same structure (a poem) in terms of RQA parameters when translated into another language, is both a stringent test of the language invariance of the technique, as well as of the consistency of the dynamical descriptors. The obvious linguistic differences between Swedish and Italian languages makes the test more cogent. The Swedish texts were transliterated using the Italian 21 letter standard alphabet (ch for k, and so on) maintaining the same phonetic character of the original. There are statistically significant correlations between the Swedish and Italian versions (r = 0.85 and r = 0.90 for REC and DET respectively, p < 0.01). It is worth noting that the technique
was able to discover the prosodic linkage between the texts without being affected by the obvious morphological differences between Italian and Swedish languages.

- **Set 4 (Italian speech samples):** human speech samples (SP) fit very well in the REC-DET plane individuated by poems maintaining the strong relation of the two descriptors ($r = 0.89, p < 0.001$). This suggests that speech samples have a general structure similar to poems. While the prosodic structure of the poems is characterized by very short sequences, the same general structure in speech samples involves more lengthy texts (3000-4000 letters). This fact makes the poems the ideal reference standard for speech samples because they give condensed (and thus easily understandable) specimens of the prosodic structure representative of long speech samples. Speech samples are significantly different from their shuffled counterparts thus highlighting an orthographic structuring.

- **Set 5 (American speech samples):** there is a correlation between the 2 descriptors ($r = 0.88, p < 0.001$), confirming the fit of spoken language samples into the space spanned by poems. The homogeneity of speakers is reflected by the tight distribution of samples, and the American and Italian samples have a similar scaling: [American Mean (REC) = 0.416, SD = 0.03, Mean (DET) = 20.00, SD = 1.36; Italian Mean (REC) = 0.676, SD= 0.06; Mean (DET) = 27.94, SD = 2.66]. Nevertheless the 2 groups of speakers were significantly different for both REC and DET mean values (t-test corrected for variance non-homogeneity: $p < 0.0001$) pointing to a greater complexity (lesser degree of structuring) of American speech. This result is probably linked to a higher cultural background of the American sample (students) with respect to the Italian one (psychiatric patients), but further investigation is needed to resolve this point. American speakers displayed less variance than the Italian ones, demonstrating a statistical significance for REC (F-test for homogeneity of variances: $F = 4.69, p < 0.04$), pointing to the possibility of using the method to discriminate sets of speakers.

**DIFFERENCES**

RQA technique provides a reliable quantitative description of text sequences at the orthographic level in terms of structuring, and may be useful for a variety of linguistics-related studies.

We used RQA to measure differences in linguistic samples from different subjects. They were divided in subgroups based on personality and culture differences. The question was: Can we differentiate different people’s informational coding of speech, basing these differences on culture, personality or psychopathology?

The sample was subdivided in:

1. 50 healthy subjects speech transcriptions (Italian and American students);
2. 18 pathological subjects speech transcriptions (severe psychopathologies i.e. schizophrenia, borderline and major affective disorders);
3. 20 pathological subjects writings (pathologies as above).
In looking at data we compared the DET8 for speech and writing (for normal vs psych) with a REC8 used as a covariate. With the analysis we found that DET PSYC speech VS NORMAL to be significant at $p=.003$ and DET PSYC writing to be $p=.10$. We find this interesting, as it suggests that with speech there is a free-flowing production of ideas which is then constrained by the writing—even here with a $p=.10$ this is close to significant indicating even here where one is forced to choose a symbolic expression perhaps not as free as regular speech.

The determinism at 8 is important since we presume a high number of degrees of freedom in language. Also the DET obviously is a determinism: thus it might seem a psychiatric population may more easily hide abnormalities in writing but not in spoken language.

In our first research we had seen that while written texts were disposed in a wide area of the plot, speech transcriptions were crowded in a small area almost in the middle. So, in order to zoom in this area we should focus on higher degrees of freedom than we did with written texts (which were differentiable at about 3 degrees of freedom). Presumed that DET and REC 8 referred to embedding, with speech there is more noise, which would require higher embeddings. We might think of all the various place holder words-phrases-sounds. As spoken discourse is less linear the more high is embedding we are exposed to randomness, but we might be able to detect determinism at meso-macro levels.

The need to place our analysis at higher embeddings than the previous one on written texts generates a need for much longer (about 10 times) samples. Leaving behind the difficulty of collecting such pure samples (monologues of about 1 hour), this might mean that a structure in a monologue becomes differentiable at a level higher than the meso one: between word and phrase. This is the macro level (phrase, narration).
COUPLINGS

We used RQA and KRQA (Cross Recurrence) also to measure the coupling and synchronization during the conversation (semiotic interaction) of different subjects. We measured a series of samples derived from a natural conversation and a clinical conversation. While the first one is supposed to be free the second one is supposed to be finalized.

FIGURE 3-1 and 3-2: both represent a semiotic interaction each dot being a turn in conversation. In 3-a white and black dots are for friend A or B; in 3-b black dots are for patient P and white dots for therapist T.

Sample 1 represents a natural dialogue between two friends: they are talking about some general topic as “the meaning of life”. While in the beginning of conversation they seem quite well synchronized, as they are going to approach the end of conversation and separation they tend to desynchronize.

Sample 2. A clinical dialogue between a patient and a psychiatrist. As this is a first interview they didn’t know each other before. It is evident how synchronization develops during conversation, including some crossing.

Transcriptions have been coded and analyzed at the orthographic level using RQA, and every dot in plots represent a turn in conversation. These data are taken at embedding 3: this means that synchronization is active at a micro-meso level. A level which implies resonances in rhythms, phonetics and redundancies of words, more than in narrative macro structures. This might also mean that coupling and synchronization start in non-conscious ways, before the self consciousness implied in narrative structures is well established. It might be stressed that before should be considered both in logical and chronological dimensions.

To highlight differences in structure of each dynamic interaction we used also Cross Recurrence KRQA which compares the dynamical behaviour of two interacting time series, which are simultaneously embedded in the same phase space [21].
Values for comparing dynamics in KRQA can be summarized:

<table>
<thead>
<tr>
<th></th>
<th>NATURAL</th>
<th>CLINICAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recurrence</td>
<td>0.139</td>
<td>0.123</td>
</tr>
<tr>
<td>Determinism</td>
<td>41.15</td>
<td>36.33</td>
</tr>
<tr>
<td>Maxline</td>
<td>10</td>
<td>9</td>
</tr>
<tr>
<td>Entropy</td>
<td>1.47</td>
<td>1.44</td>
</tr>
</tbody>
</table>

**CONCLUSIONS**

While our work was focused at the microscopic level, the influence of mesoscopic (word level) and macroscopic (phrase and narrative level) factors on the observed synchronization is open for hypotheses.

Our research can be considered complimentary with a study by Shockley et al [22] in which interpersonal coordination during conversation was based on recurrence strategies, to evaluate the shared activity between 2 postural time series in reconstructed phase space. Pairs of participants were found to share more locations in phase space (greater recurrence) in conditions where they were conversing with one another to solve a puzzle task than in conditions in which they conversed with others. The trajectories of pairs of participants also showed less divergence when they conversed with each other than when they conversed with others as well. Objective evidence of interpersonal coordination of postural sway in the context of a cooperative verbal task.

In a study on speech and rhythmic behavior Port et al. [23] found that animals and humans exhibit many kinds of behavior where the frequencies of gestures are related by small integer ratios (like 1:1, 2:1 or 3:1). Does speech exhibit timing similarities to other kinds of rhythmic or periodic behavior by our species? People exhibit many cyclic behaviors: walking, waving an object, repetitive reaching, finger drumming, scratching an itch or dancing. We suspect that many properties of these skills will be found in speech as well.
PERSPECTIVES

All these findings are related to interpersonal and intrapersonal embodied dynamics. Measures of the influence of these variables could be of great interest for example in psychometrics, which is now entirely based on inferential and non-direct measures. A differentiation of cultural, personality and psychopathology based on direct measures of semiotic samples could be of theoretical and clinical value.

Our findings in the synchronization of conversation dynamics can be relevant for the general issue of structural coupling of psychobiological organizations. Implications are related with psycho-chrono-biology research, learning environments and the clinical field.

This might have impact on neuro-cognitive dynamical diseases treatment. As dynamical diseases are considered to be based on a sudden qualitative change in the temporal pattern of physiological variables.

Data on synchronization suggest that the activity of some form of psychobiological gravitation can be evident also in semiotic dynamics (besides the well established research on biological oscillators).

These isomorphisms between biological and cognitive structures are blurring old Cartesian boundaries.

ACKNOWLEDGEMENTS

We are grateful for contributions from Chuck Webber, Marco Luccioni, Marianna Mazza, Kimberly Walters and Stefano Boccaletti.

REFERENCES


93