Chaotic dynamics and forbidden words

J.H.B. Deane and D.J. Jefferies
School of Electronic Engineering, Information Technology and Mathematics,
University of Surrey,
Guildford GU2 7XH,
United Kingdom
e-mail: J.Deane/D.Jefferies@eim.surrey.ac.uk
http://www.ee.surrey.ac.uk/Personal/J.Deane/ /D.Jefferies/

Abstract. We report results on the symbol sequences generated by a simple rule operating on a continuous chaotic sequence. We propose a complexity measure arising out of this work and discuss applications to natural language analysis.

1. Introduction

We examine in this paper the sequences $S = \{s_1, s_2 \ldots\}$ of symbols, chosen from a finite alphabet, according to rule operating on a continuous state variable of a possibly chaotic system (Wiggins, 1998). In particular, we study the distribution of substrings of S of length L, which we refer to as words. We choose two examples of potentially chaotic systems: (i) the logistic map

$$x_{n+1} = M(x_n) = ax_n(1 - x_n) \tag{1}$$

which is a discrete time system; and (ii) a sinusoidally driven particle moving in an infinite periodic force

$$\frac{\mathrm{d}^2x}{\mathrm{d}t^2} + K\frac{\mathrm{d}x}{\mathrm{d}t} + \Omega^2 F(x) = A\sin\omega t \tag{2}$$

where the periodic force $F(x) = x - 2\lfloor(1 + x)/2\rfloor$, an odd sawtooth waveform with period equal to two. This is a continuous time system.

In case (i) the rule for turning a real x_n into a discrete symbol s_n is simply

$$s_n = \Delta(x_n) = \begin{cases} 0 & 0 \le x_n < \frac{1}{2} \\ 1 & \text{otherwise} \end{cases} \tag{3}$$

In case (ii) the rule is $s_n = 0$ if the n-th transition between one well and the next takes place in the direction of decreasing x, with $s_n = 1$ otherwise. The sequence S may in this case be finite (for instance, for small A — the particle just oscillates within a single well). However, for present purposes, we only need to know that there exist parameter sets for (2) for which the particle hops from well to well in a chaotic manner; it is such sequences that we study here.

We are interested in answering the following questions:

- Can all possible sequences S occur? If not, what governs which sequences do occur?

- How are the S that do occur arranged?
- Is there any distinction between the S produced in cases (i) and (ii)?

2. Case (i): the logistic map

2.1. $a = 4$, ergodic behaviour

It is shown in, for instance (Schuster, 1984), that for $a = 4$ the logistic map is ergodic, which implies that for all i and almost all initial conditions x_0, (a) s_i and s_{i+1} behave as if statistically independent; and (b) s_i is equally likely to be 0 or 1. These observations arise because, for $a = 4$, the logistic map is topologically conjugate to the tent map

$$x \mapsto \begin{cases} 2x & \text{for } 0 \le x < 1/2 \\ 2(1-x) & \text{for } 1/2 \le x \le 1 \end{cases}$$

which provably has these properties. This in turn implies that, for $a = 4$, the logistic map can produce *any* sequence S of the symbols 0 and 1; which sequence is actually observed depends only on the initial value x_0. (The Bernoulli shift $x \mapsto 2x \bmod 1$ also has this property, in a somewhat more obvious way.) Since the two symbols 0 and 1 are equally likely for almost all x_0, it is also true that each two-bit word (00, 01, 10, 11) occurs with equal probability, as does each 3-, 4-... bit word. The autocorrelation function

$$A(k) = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} s_i s_{i+k}$$

should be $1/2$ for $k = 0$ and $1/4$ otherwise, if the s_i are independent. All these properties can be verified experimentally for some large N; results for $N = 2^{20}$ and for the particular orbit with $x_0 = 0.77$ are

$$P(0) = 0.4997, \quad P(1) = 0.5003$$

$$P(00) = 0.2500, \quad P(01) = 0.2503, \quad P(10) = 0.2490, \quad P(11) = 0.2507$$

The autocorrelation function, $A(k)$, is shown in figure 1, for various values of a.

2.2. $a < 4$, chaotic behaviour

The situation is considerably more complicated when $a < 4$ and such that the sequence $\{x_0, x_1, x_2, \ldots\}$ is chaotic. To illustrate this, we now look at the case $a = 3.91$ in some detail. For this value of a,

- $0.0859955625 \le x_n \le 0.9775$ for all n, provided that x_0 lies in this range: since we are only interested in *asymptotic* behaviour, we always chose an initial condition in this range, or, equivalently, pre-iterate a few dozen times;
- The Lyapunov exponent λ, defined as

$$\lambda(x_0) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \ln |f'(x_i)|$$

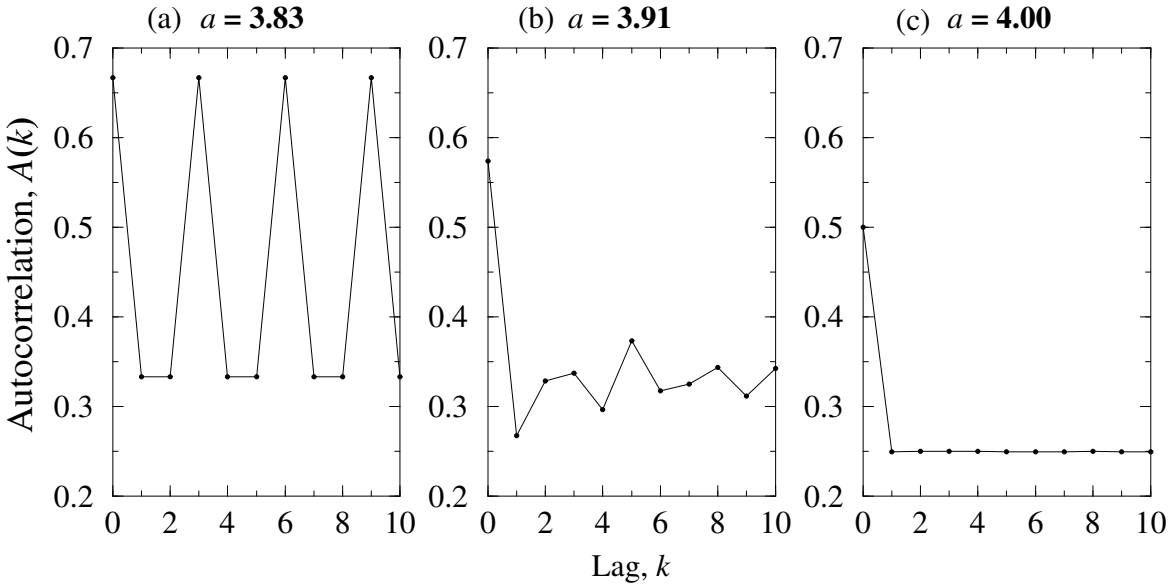

Figure 1: *The autocorrelation function of S for three different values of a. (a) $a = 3.83$ (period-3 behaviour); (b) $a = 3.91$ (chaotic) and (c) $a = 4.00$ (ergodic). The rate of decay in case (b) is a measure of the rate at which x_{n+k} becomes independent of x_n.*

for a one-dimensional map f, is a measure of the average stretching apart of two initially close points under the action of the mapping. In this case, $\lambda \approx 0.492$;

- There is dependence within S, that is, x_{n+k} depends on x_n for some $k > 1$;
- Not all possible sequences are produced.

We now consider the occurrence or not of L-bit words, for $L = 1 - 10$. The 1–5-bit words that occur are shown in figure 2. Both 1-bit words are seen, although not with equal probability: $P(0) \approx 0.425$, $P(1) \approx 0.575$. All 2-bit words also occur and the probabilities are $P(00) \approx 0.119$, $P(01) \approx 0.306$, $P(10) \approx 0.306$ and $P(11) \approx 0.268$. (These probabilities were obtained by observing one particular orbit.) Note that these cannot be estimated from $P(0)$ and $P(1)$ since the sequence is not statistically independent (by contrast with the ergodic case $a = 4$) — figure 1(b) shows that dependence lasts for at least ten timesteps. For $L = 3$, the situation is different again, since only 7 out of the 8 possible 3-bit words occur: the word 000 is never seen, and we prove this below. All 4-bit words are seen except the three (0000, 0001, 1000) that contain the 'forbidden' string 000. For $L = 5$ a new forbidden word, 00110, appears. Further new forbidden words appear at $L = 6, 7$ and 10.

The above information was extracted from a sequence S of length 2^{20}. To determine forbidden words of length L, each adjacent subsequence was checked, *i.e.* $s_1 - s_L$, $s_2 - s_{L+1}$ *etc.*

The table below summarises the forbidden words with $L \leq 10$, that appear for a selection of values of a. Blank entries mean that no new forbidden words appear for that value of L. Note that the forbidden words for $a = 3.91$ and 3.96 are the same, at least for $L \leq 10$.

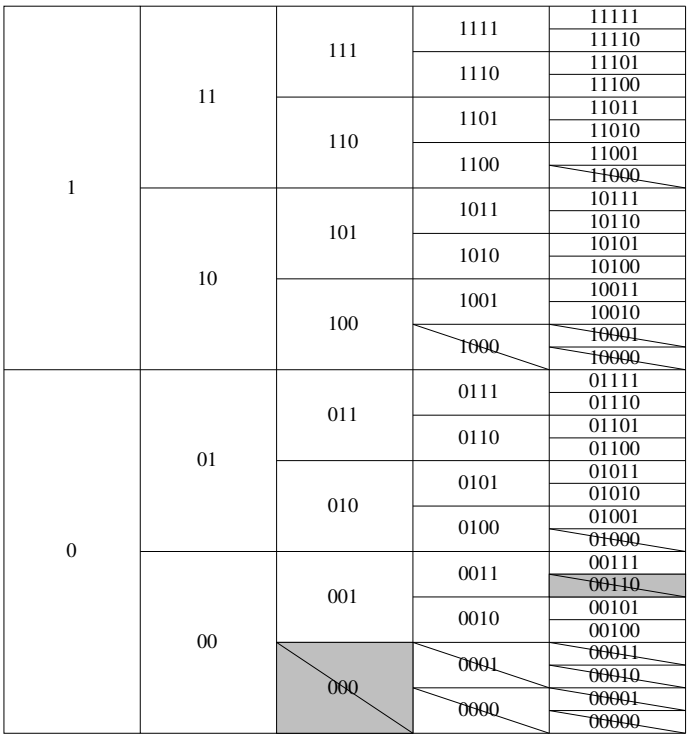

Figure 2: *The words of lengths 1–5 produced from the logistic map with* $a = 3.91$. *The shaded boxes contain new* — i.e. *not predictable from previous* — *forbidden words; the crossed out boxes contain words which are not seen, which fact can be deduced from previous forbidden words.*

	Forbidden words							
a	$L = 3$	4	5	6	7	8	9	10
3.90	000	0011						0010100101
3.91	000		00110	001111	0011101			0011100100
3.93	000		00110		0011110			
3.95	000			001100				
3.96	000		00110	001111	0011101			0011100100
3.97		0000	00011		0001010	00010111	000101101	
3.99		0000						

2.2.1. Proof of forbidden words

The results in the above table were all obtained 'experimentally', *i.e.*, by producing a sequence S of 2^{20} symbols and observing which words did not appear. This heuristic method suggests that the forbidden words may never appear, but in reality only shows that they have low probability. Thus, we now take a specific case, the fact that 000 is apparently forbidden when $a = 3.90$, and *prove* this to be the case. We need to show that, if $x_0 > 1/2, x_1 < 1/2$ and $x_2 < 1/2$, then x_3 must be $> 1/2$, giving the sequence $\Delta(x_0), \Delta(x_1), \Delta(x_2), \Delta(x_3) = 1001$.

We refer to figure 3, which is a plot of $M(x)$ for $a = 3.9$. We first observe that all iterates starting from $X_l \leq x_0 \leq X_u$ lie in the range $[X_l, X_u]$, which is not true for iterates starting from $0 \leq x_0 \leq 1$. The proof relies on the fact that initial conditions

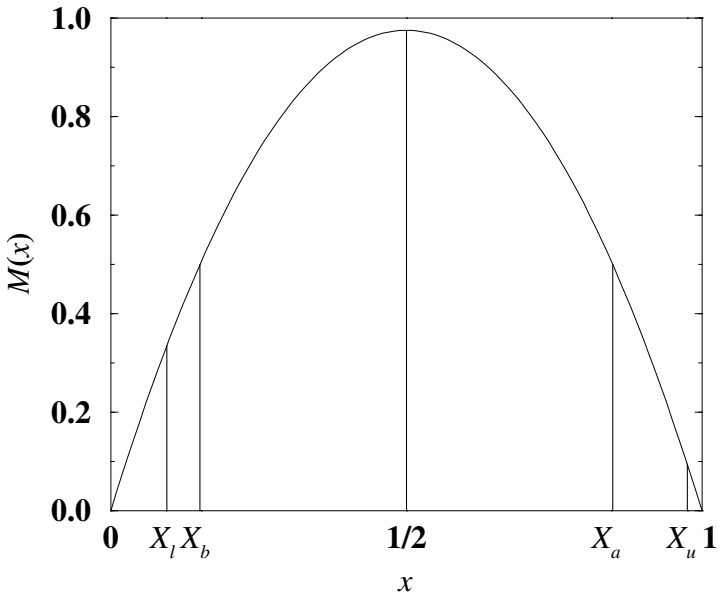

Figure 3: *A plot of $M(x)$ with $a = 3.9$, used in the proof that 000 is a forbidden word. For this value of a, $X_u = M(1/2)$, $X_l = M(X_u)$, $M(X_a) = 1/2$ with $X_a > 1/2$, and $M(X_b) = 1/2$ with $X_b < 1/2$.*

x_0 were chosen from $X_l \leq x_0 \leq X_u$, or sufficient pre-iteration was carried out first to put the initial value in this range—remember, we are only interested in the asymptotic behaviour. For $a = 39/10$, $X_u = 39/40$ and $X_l = 1521/16000$.

The first transition is $1 \to 0$. In order for this to obtain, we must have $X_a < x_0 \leq X_u$, where $M(X_a) = 1/2$ and $X_a > 1/2$. With this restriction, M is a one-to-one mapping from $X_a < x_0 \leq X_u$ to $X_l \leq x_1 < 1/2$, with $X_l = M(X_u)$. Solving the appropriate quadratic equation gives $X_a = 1/2 + \sqrt{741}/78$.

We require the second transition to be $0 \to 0$. For this to occur, x_1 must be further restricted to $X_l < x_1 \leq X_b$, where $M(X_b) = 1/2$ and $X_b < 1/2$, giving $X_b = 1/2 - \sqrt{741}/78$. With this restriction, $x_2 = M(x_1)$ is confined to $M(X_l) \leq x_2 < 1/2$.

Hence, finally, $x_3 = M(x_2)$ is restricted to $M(M(X_l)) \leq x_3 < M(1/2)$, which, for $a = 3.9$ gives $\sim 0.8695 \leq x_3 < 39/40$. Thus

If $a = 3.9$ and x_0 is such that s_0, s_1, s_2 is 100, then $s_3 = 1$.

This proves that 000 is a forbidden word when $a = 3.9$. Proof of 'forbiddenness' of all other words could in principle be carried out in the same way, although the longer the word, the longer the proof, and each case needs to be dealt with separately.

2.3. $a < 4$, period-p behaviour

This case is easily treated. If S is periodic with period p, then there are a maximum of p words of length L for any L. For $L > p$, almost all words are forbidden.

2.4. The number of allowed words

Figure 4 shows the base-2 logarithm of the number of allowed words, N_a, as a function of L, *i.e.* the number of words which do not contain forbidden strings. In the ergodic case the plot has unity slope since $N_a = 2^L$ — all words are present. In the period-3 case illustrated, the plot has zero gradient for $L > 3$: the number of words is constant as

L increases. In the chaotic case the gradient is somewhere between these two extremes. We propose that the slope of this graph is a useful measure of complexity of the sequence S, and of the underlying dynamics of the continuous nonlinear system producing it.

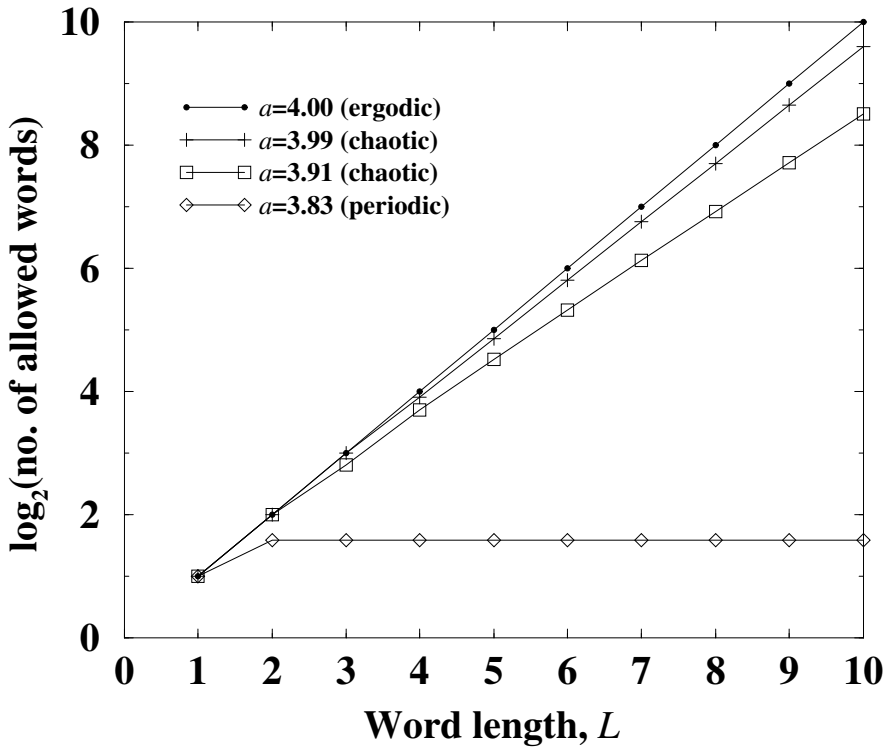

Figure 4: *The number of allowed words versus L for different values of a.*

3. Case (*ii*): the periodic potential wells

We have analysed the sequences produced by solving the periodic potential well equation with the following parameters: $K = 0.05$, $\omega = 1$, $\Omega = 1.1$ and $A = 1.12$. These parameters result in a chaotic solution in which the particle hops from well to well.

The forbidden words always appear in pairs, with a word and its inverse both being present, *e.g.* if 0100 is forbidden, then so too is 1011. The forbidden words for $L \leq 10$ are:

L	Forbidden words				
5	01001	01101	01110		
6	011001				
8	01111001	01111101			
9	010111101	011000001	011000010	011110101	011111001
10	0101000001				

and inverses. This symmetry exists because the periodic potential well system, and therefore the word probabilities, are symmetrical. For 2^{20} samples, it was found that

$P(0) = 0.4997, P(1) = 0.5003$
$P(00) = 0.4435, P(11) = 0.4441, P(01) = 0.0562, P(10) = 0.0562$

The forbidden word structure is seen to be more complicated than the logistic map example. Several new forbidden words can appear for each L, whereas apparently only one new word appears for the logistic map. The larger number of forbidden words also leads to a slower growth in the number of allowed words, as can be seen from figure 5. We conclude that this is the result of the higher dimensionality of the system that produces the sequence.

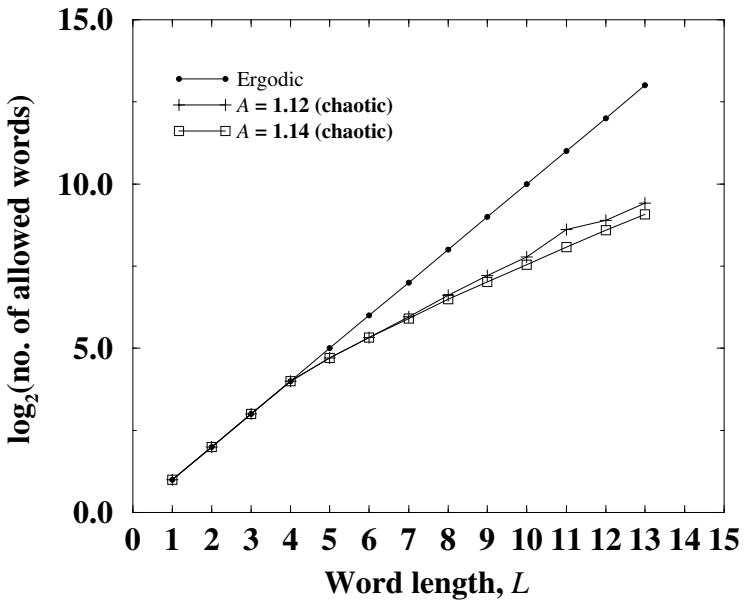

Figure 5: *The number of allowed words versus L for different values of* $A = 1.12$ *in the periodic potential well system. The ergodic line is for comparison only — no parameters have been found that give rise to ergodic behaviour in this system.*

4. Discussion

We have presented some results on binary sequences generated by chaotic systems when they behave ergodically, chaotically and periodically. Graphs of $\log_2$ (number of allowed words of length L) against L have been plotted. The slope m of such a graph must lie between 0 and 1. It cannot be negative since the number of allowed words must at least stay the same as L increases, so $m \geq 0$; and there are exactly 2^L binary sequences of length L, giving $m \leq 1$. Furthermore, $m = 0$ as $L \to \infty$ shows that S is periodic; and $m = 1$ as $L \to \infty$ shows that S is ergodic.

What can we say about values of m that lie in between these two extremes? We have suggested above that m can be interpreted as a measure of complexity of the system producing S. It is arguable that the larger the value of m, the more complex the system, which makes ergodic systems the most complex of all, and this position is defensible. The logistic map with $a = 4$ could in principle produce any sequence at all— *e.g.* the entire works of Shakespeare encoded in ASCII — given (a) the right initial condition x_0 (specified to enormous precision) and (b) an arbitrarily high precision computer to do the calculations. There is nothing remarkable about this: the x_0 required is just one possible way of encoding the works of Shakespeare. From the point of view of *modelling*, however, the ergodic case is easy: a generator of independent, identically distributed random numbers is all that is required.

By contrast, the difficult cases to model are those for which $0 < m < 1$ where there is dependence in S (s_i depends on s_{i-k}, $k > 0$), the probabilities are not equal and certain words provably have probability 0 (*i.e.* are forbidden). The dependence observed is of long range too — see figure 1. Furthermore, we see no reason to suppose that the number of forbidden words is finite, nor that they obey any simple pattern that allows them to be predicted in advance. Hence, we conjecture that an infinite amount of information would be needed to model such a system.

These observations imply that the more restricted the number of allowed words, the less complex is the behaviour of the system that produces S.

5. Further work

Many interesting if speculative questions are raised by this work, including:

- Can natural language (Pinker, 1999) be analysed in this way? There are various kinds of constraints at work in English, among them
  - Pronouncability constraints (*e.g.* 'jc' is a forbidden string);
  - Orthographical constraints (*e.g.* the rule that 'q' is always followed by 'u' means that 'qa' is a forbidden string);
  - Etymological constraints (no words happen to have evolved that contain the string 'cv')

  These and other constraints naturally lead to forbidden words in natural languages. It might be instructive to compare the complexity of languages in a way similar to that outlined above for chaotic symbol sequences. The comparison process would be different in some respects: for chaotic sequences, a forbidden word of length l_1 remains forbidden in words of length $l_2 > l_1$, which is not always true for languages. For instance 'sa' is not a word — so is a forbidden word of length two — but 'sat', which contains this string, *is* allowed.

  A further possibility along these lines but at a higher level, would be to look on words themselves as the symbols, rather than letters, and to investigate how these are aggregated into sentences. Again, there would be forbidden sequences of words — 'you am' — and some insight into complexity of language on the word-by-word level could be obtained from such analysis. Carrying this out would be likely to be hampered by the large number of 'symbols', compared to just two used in the analysis presented in this paper.

- Which words are forbidden is a characteristic of the underlying chaotic system. We conjecture that the word probabilities are not only robust, up to a point, in the presence of noise, but are also globally characteristic of the system and do not depend sensitively on the particular trajectory which is an expression of a particular initial state. Differences between the word probabilities in the case of a simulation without noise (but perhaps with rounding error) and a real system with imperfections and added noise, may possibly be taken as a measure of the amount of added noise in the system. There may be applications to the replication dynamics of DNA, where it is known that errors are introduced routinely in the copying process, perhaps regarded as added noise, and yet the overall DNA functionality is not compromised from generation to generation when combined with selection pressures.

- In a chaotic system, is the number of forbidden words infinite in general? Is there any way to predict which words will be forbidden?

- In the set of all words, how are the forbidden ones arranged? In particular, are they arranged in a self-similar manner?

  The set of allowed sequences can be coded as a number x where $0 \leq x \leq 1$, simply by putting a decimal point in front of the sequence and interpreting that as a binary decimal. The resulting set of numbers is likely to be a fractal with Hausdorff dimension lying between 0 and 1. (For instance, the set of real numbers between 0 and 1 whose binary expansions do not contain the forbidden word 000 has dimension $\ln 7 / \ln 8$ (Falconer, 1990).) In practice, taking into account several, or worse still, an infinite number of forbidden words makes the dimension calculation very difficult (Falconer, 1990).

References

K Falconer (1990), Fractal geometry: mathematical foundations and applications, John Wiley and sons, ISBN 0-471-92287-0.

S Pinker (1999), Words and Rules, Weidenfeld and Nicholson, ISBN 0-297-81647-0.

HG Schuster (1984), Deterministic chaos, Physik-Verlag, ISBN 3-87664-101-2.

S Wiggins (1988), Global bifurcations and chaos, Springer-Verlag, ISBN 0-387-96775-3.