Discrete Dynamics in Nature and Society, Vol. 4, pp. 293–295
Reprints available directly from the publisher
Photocopying permitted by license only

A Comparative Dynamical Analysis of Hebrew Texts

ABRAHAM BOYARSKY* and PAWEŁ GÓRA

Department of Mathematics and Statistics, Concordia University,
7141 Sherbrooke Street West, Montreal, Quebec H4B 1R6, Canada

(Received 10 May 1999)

Three Hebrew texts, one of them the Hebrew bible, are investigated using dynamical analysis. First, the average mutual information for each of the texts is determined. The first minimum occurs at $T = 3$ in all three cases, suggesting that 3-letter words are sufficient to study the dynamical properties of the texts. Using 3-letter words as the state space for each of the text, we construct a Markov chain model and compute the relative measure–theoretic entropy for each of the texts and use this tool as a means of comparing the information content of the three texts.

Keywords: Text analysis, Markov chains, Entropy, Dynamical systems, Mutual information

1 INTRODUCTION

The Hebrew language consists of 22 letters and a space that separates words. Five of the letters have also final forms, which are used at the end of the words. Some vowels in the middle of a word are omitted in the old language while they are put in the modern texts. We removed these vowels from all the texts in order to obtain texts that can be more uniformly compared. Vowels which we ignored inside a word in order to produce more uniform word data in all three texts are: alef, vav, yud, ayin.

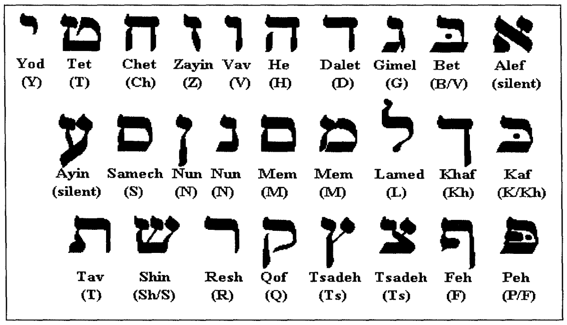

FIGURE 1

Let X denote the space of 28 basic letters (Fig. 1). To study the statistical dynamics of a text we consider a transformation T that takes the first 3 letters of a word to the first 3 letters of the next word. If a word has only 2 letters, the space following the word is appended to the word itself. If a word has only one letter we append two spaces after it to create a triple representing this word.

* Corresponding author. E-mail: boyar@vax2.concordia.ca, pgora@vax2.concordia.ca.

In this way we obtain a data set consisting of a sequence, each element of which is a 3-letter sequence. Depending on the text under consideration there are approximately 7000–8000 such 3-letter admissible sequences (maximum is $27 \times 28 \times 28$).

2 STATE SPACE

We use 3-letter sequences because that size presents a respectable indication of word structure without creating an overly large space. Four-letter sequences would create a space with hundreds of thousands of elements, making it difficult to analyze even with the most powerful computers. Since the average word size is between 3 and 4, 3-letter words give a good indication of the word itself.

We suggest a number of measures for "richness" of text. One such measure is simply based on an average attainable set of 3-sequences. Let xyz represent a valid 3-letter sequence in a Hebrew text. We compute all the 3-letter sequences that are attainable from xyz in the given text. We repeat this for all possible 3-letter sequences and then compute the average number of attainable sequences over all admissible 3-letter sequences. A rich text will have a relatively large such average.

The following argument supports our choice of 3-letter representation of words. For all three texts, we calculated the average mutual information $\mathrm{AMI}(T)$ ([1]) between the first letter $x_1^{(i)}$ and the (T)th letter $x_T^{(i)}$ of the word, $T = 2, \ldots, 6$.

$$\mathrm{AMI}(T) = \sum_i P(x_1^{(i)}, x_T^{(i)}) \log\left(\frac{P(x_1^{(i)}, x_T^{(i)})}{P(x_1^{(i)})P(x_T^{(i)})}\right),$$

where $P(x_j^{(i)})$ and $P(x_j^{(i)}, x_k^{(i)})$ are the probabilities (frequencies) of individual letters and pairs of letters in the text, and i changes from 1 to the number of words in the text (approximately 8000). For all texts the first minimum of $\mathrm{AMI}(T)$ occurs at $T = 3$. We interpret this fact as follows: the dependence of the letter x_T on the letter x_1 of the word decreases for $T = 2, 3$ and then starts increasing, attaining the minimum at $T = 3$. This shows that the fourth letter of a word is more dependent on the first letter than the third letter of the word. Hence the fourth letter provides less information about the text than does the third letter.

3 ENTROPY FOR A MARKOV CHAIN

We model a text by a Markov chain on 3-letter sequences. (Textual analysis is what originally motivated Markov to introduce Markov chains in 1911 [2] and was a source of examples in Shannon's work on communication theory [3]. See also [4,5].)

Motivated by the results of Section 1, we use only the first 3 letters of each word (adding extra spaces for 1- and 2-letter words). Each word is considered as a state and the transition probabilities are calculated from the flow of the text. We describe this in detail now.

Let N denotes the number of words in the text. Our texts contain around 80 000 words each. In general, we assume that the number of words is "large". For any admissible triple of letters uvw let N_{uvw} denote the number of words represented by uvw in the text but without including the last word, i.e., considering only words $1, 2, \ldots, N-1$. Similarly, for any admissible triple of letters uvw let N'_{uvw} denote the number of words represented by uvw in the text but without including the first word, i.e., considering only words $2, \ldots, N-1, N$. The numbers N_{uvw} and N'_{uvw} are equal for all except at most two admissible triples, and for these two they differ at most by 1. We can artificially add one word at the end to the text to make them always equal. In the sequel we assume that they are equal, as we are interested only in the ratios N_{uvw}/N.

To calculate transition probabilities, for any uvw we count the number of consecutive pairs represented by uvw and xyz, $N_{uvw,xyz}$ and we set

$$P_{uvw,xyz} = \frac{N_{uvw,xyz}}{N_{uvw}} \quad \text{for any admissible } uvw \text{ and } xyz. \qquad (1)$$

We will show that the numbers $P_{uvw} = N_{uvw}/N$ are stationary probabilities for this Markov chain, if we assume $N_{uvw}/N = N'_{uvw}/N$, for all uvw.

PROPOSITION *If $N_{uvw}/N = N'_{uvw}/N$, for all uvw, then the numbers $P_{uvw} = N_{uvw}/N$ are stationary probabilities for the Markov chain with transition probabilities given by* (1).

Proof We have to show that for any xyz

$$P_{xyz} = \sum_{uvw} P_{uvw} P_{uvw,xyz}. \qquad (2)$$

Obviously, we have $N'_{xyz} = \sum_{uvw} N_{uvw,xyz}$, for any xyz. If $N_{uvw}/N = N'_{uvw}/N$, for all uvw, we can rewrite this as

$$N_{xyz}/N = \sum_{uvw} (N_{uvw}/N) N_{uvw,xyz}/N_{uvw},$$

which is equivalent to (2).

Now the entropy of the Markov chain is given by the formula

$$\text{entropy} = \sum_{uvw} \sum_{xyz} P_{uvw} P_{uvw,xyz} \log(P_{uvw,xyz}),$$

where the logarithm is to the base 2.

4 RESULTS

In this section we present the numerical results obtained from analysis of the texts. The average number of the possible different following words and standard deviation of these numbers:

Text	Average	Standard deviation
Hebrew bible	106.798	138.84
Hebrew text 1	97.621	142.51
Hebrew text 2	76.016	115.66

In the following table we present the entropy and the maximal possible entropy (as defined in [3]) for the three texts. The maximal possible entropy is the logarithm (to base 2) of the number of states in the Markov chain. The numbers of different admissible triples are different in the three texts.

Text	Entropy	Number of triples	Max. entropy
Hebrew bible	3.56	3586	11.808
Hebrew text 1	3.69	4545	12.150
Hebrew text 2	3.80	4692	12.196

The text used in analysis were: Koren text of Hebrew bible and two books obtained in electronic form from Bar Ilan University:

Hebrew text 1: “Hitganvut Yehidim” (Heart Murmur) by Joshua Kenaz, published by Am Oved Tel-Aviv, ISBN 965-13-0413-8.

Hebrew text 2: “An Autobiography” by N. Lorekh, now a historian, formerly a military officer and diplomat.

Acknowledgment

We are grateful to Prof. Y. Choueka of Bar Ilan University for providing us with digitized versions of two modern Hebrew texts. The Koren text of the Hebrew bible was obtained from “http://cs.anu.edu.au/people/bdm/dilugim/torah.html”.

References

[1] Abarbanel, H.D.I., *Analysis of Observed Chaotic Data*, Springer, New York, 1996.

[2] Iosifescu, M., *Finite Markov Processes and Their Applications*, J. Wiley & Sons, New York, 1980.

[3] Shannon, C.E. and Weaver, W., *The Mathematical Theory of Communication*, Indiana University Press, Urbana, 1949.

[4] Gani, J., Stochastic models for type counts in a literary text, *Perspectives in Probability and Statistics* (Gani, J., Ed.), Academic Press, London, 1975, pp. 313–323.

[5] Brainerd, B., On the relation between types and tokens in literary text, *J. Appl. Prob.* **9** (1972), 507–518.