The Shannon Guessing Game Simulation and Long Torah Code Phrases

Preprint · August 2023

DOI: 10.13140/RG.2.2.33595.18723

2 authors:

Eliyahu Rips
Hebrew University of Jerusalem

Art Avraham Levitt
Torah Codes Research

The Shannon Guessing Game Simulation and Long Torah Code Phrases

Abstract

Torah Code research is concerned with a particular type of letter sequence, formed by extracting equally spaced letters from the Torah (the first five books of the Hebrew Bible). This is called an ELS (equidistant letter sequence). Numerous examples of long meaningful ELS phrases have been discovered. In this and a companion paper [3], we develop tools to study their significance, in particular using the redundancy of a meaningful text, that is, the correlation between its letters or words.

The quantitative study of redundancy was initiated by Claude Shannon who introduced the notion of entropy (the measure of uncertainty) applied to natural languages ([13],[4]). To estimate it, Shannon invented the Shannon Guessing Game (SGG), which measures the success rate of human subjects to guess some of the characters of a meaningful text, given a context of other characters of the text.

We use an analog of this idea to assess the significance of long Torah Code phrases. Namely, we replace part of the letters in these sequences by "wildcards". We find that they are restorable from the context, with very few alternatives (see Figure 1, where the locations of the wildcards are indicated in the column, in gray). The large incidence of guessable letters (well above one third of them) demonstrates a significant pattern.

1. Description of the SGG Method

1.1. Overview

To form an ELS by letter extraction, we ignore all punctuation and inter-word spaces. For example, the ELS phrase (or *string*), "tin tops" can be found starting with the first "t" in the word "punctuation" in the preceding sentence, and using a skip distance of +4 (that is, counting forward every 4 letters from the starting position).

Starting with a given keyword ("*anchor*") we find its appearance as an ELS and attempt to extend it in both directions to form a longer ELS string. In numerous cases, this extension reads as a meaningful phrase or phrases. The phrase in Figure 1 was discovered by Dr. Leib Schwartzman, starting with the anchor "bin Laden", which has a skip distance of 6598. The figure shows the underlying Torah text, of which the ELS is a part, with each row separated from the one above it by 6598 letters, so that the encoded ELS appears vertically in a column. Each word of the ELS (also called the phrase or string) is highlighted in the column, and translated in English nearby.

To estimate the rarity of meaningful phrases, we can compare them to a large number of random strings. This was done in [9] for the bin Laden ELS, yielding a high significance. This method compared the phrase to over 7 million strings, from over 600,000 control texts. The data was rated and cross-checked by 85 reviewers under double blind protocol.

Another method for estimating the rarity is to draw on the ideas of Claude Shannon. His language entropy gives an average estimate of the share of meaningful texts among random letter strings of length L. Namely, if the Shannon entropy of the language is H bits per character, then there are roughly 2^{HL} meaningful texts of length L. The entropy of a randomly extracted string (with letter frequencies reflecting their occurrence in the language) is equal to the first order entropy of the language, namely

$$H_1 = -\sum_{i=1}^{k} p_i \log_2 p_i$$

where p_i denotes the frequency of letter i in the language, which has an alphabet of k letters. Consequently, there are roughly $2^{H_1 L}$ random strings of length L (for a more precise statement, see ([4]), *Theorem 1.3.1*). Hence, the rate of meaningful texts of length L among all such random strings of length L is $2^{L(H-H_1)}$.

Applying entropy to a Torah ELS string, the frequency of its letters are estimated from the Torah's (Hebrew) letter frequencies, since the string is a sampling of Torah letters. The first order entropy of the Hebrew language, calculated with the frequencies of the letters as they appear in the Torah, is $H_1 = 4.07$, without considering inter-word spaces. According to the data available to us, the entropy of the Hebrew language is estimated by 1.8 ([12],[5]) . Making a simple proportional adjustment for deletion of spaces gives the

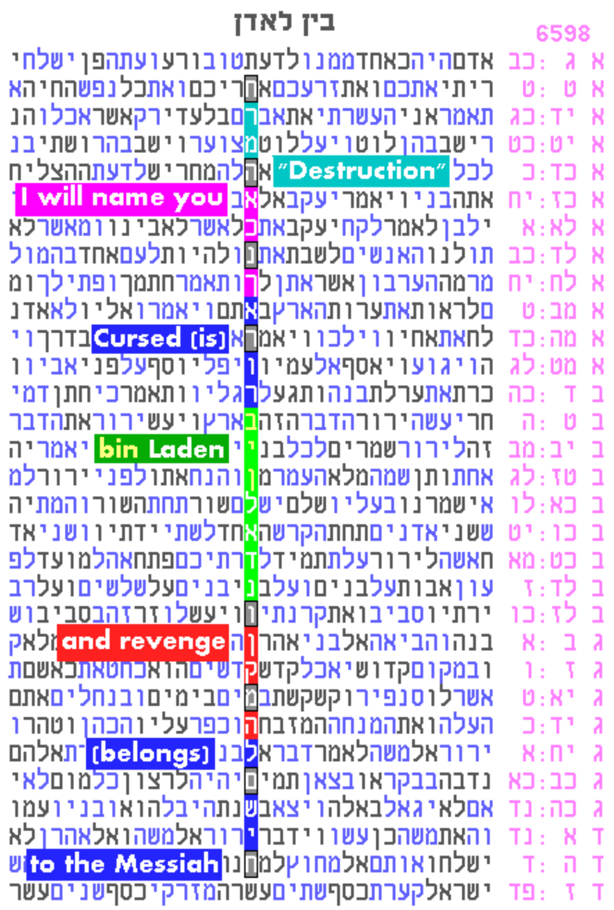

Figure 1. A Torah Code phrase. The bin Laden "*anchor*" is extended by 22 letters

value $H = 2.27$ based on the Torah's average word length of 3.81. Accordingly, under the null hypothesis of no Torah Codes, the portion of random Torah ELS strings of length 22 (as in Figure 1) which are meaningful, is estimated to be $2^{22(2.27-4.07)}$ = 1.25E-12. However, the Shannon entropy comes from experiments measuring the redundancy of the language and it relates not to a single expression but to the language as a source of meaningful messages. While the above estimate has an indicative power, it is necessary to subject an individual ELS string to a direct checking of its redundancy, because that redundancy may depend on individual factors, especially the average word length. We do this checking with our simulation of the SGG.

1.2. The SGG simulation for letters

Claude Shannon measured his concept of language entropy in the English language by means of the Shannon Guessing Game (SGG). We invite the reader to sample an online version of the game, which is highly instructive [2].

In the original SGG, one guesses each consecutive letter in order. Our simulation is done by guessing only a certain subset of the letters, having full access to the others, and we do not insist that the letters be guessed in a specific order. For this purpose we construct a *template* from our original ELS string. We use English only to illustrate the concept, but the actual method is applied to the code, always in Hebrew. Suppose that our string is:

CURSED IS BIN LADEN AND REVENGE BELONGS TO THE MESSIAH

As we have shown above, the first order entropy of Hebrew is almost twice as much as its Shannon entropy, so theoretically we could try to guess every second letter. However, the resources required for this exceed practicality, and we therefore guess every third letter. This gives us only a lower bound on the redundancy, and accordingly an upper bound on the p-value.

Here is the template, with every third letter (excluding the anchor) designated as a "wildcard":

*UR*ED *S BIN LADEN *ND *EV*NG* BE*ON*S T* TH* ME*SI*H

From this template, we try to guess letters for the wildcards, filling in all possibilities that form a meaningful text that is also relevant to bin Laden. We consider and reject a large number of meaningless candidates such as:

CURLED AS BIN LADEN AND SEVEN GO BELONGS TO THE MESSIAH

This candidate has the additional feature of requiring more words than the original, and is rejected for that reason as well. We may in fact find that the original and the following minor variation are the only possibilities:

CURSED *AS* BIN LADEN AND REVENGE BELONGS TO THE MESSIAH

But even this candidate is not as meaningful as the original. Typically, we find zero or very few alternatives, in line with the Shannon-based estimates.

2. Application of the SGG Method

2.1. The probability space

The original string, S_1, from the Torah (text T) follows:

חרמה אכנך ארור בין לאדן ונקמה למשיח

Translation: I will dub you "Destruction". Cursed is **bin Laden** and revenge belongs to the Messiah.

We create our template by placing wildcards into S_1, using the same spacing as in our English example:

*רמ*אכ*כא*ור בין לאדן *נק*הל*שי*

We consider the probability space of strings with wildcards replaced by independent random variables $X_i (1 \leq i \leq 8)$ that take as values Hebrew letters with probabilities equal to the frequencies of these letters in T and with all the other letters fixed as above:

X_1רמX_2אכX_3כאX_4ורבין לאדןX_5נקX_6הלX_7ש'X_8

We designate by $\boldsymbol{E}$ the favorable event of a phrase appearing that is as meaningful and relevant to bin Laden as string S_1. We calculate the probability of $\boldsymbol{E}$ based on how often we observe it, as in the following section.

2.2. The SGG outcome

Now is the decisive step of our procedure. We list on our review site (`http://www.torahcodes.net/sgg.html`) all possible substitutions for the X_i that form valid words. We find that there are no other word boundaries (positions of spaces) that can be used to form a member of $\boldsymbol{E}$, and that 6 of the 8 X_i have only one choice for inclusion in $\boldsymbol{E}$, taking on the values of S_1. For X_2 we can have the letter *heh* as in S_1, which forms the word for "Destruction", or the letter *yod*, which forms the word for "My Destruction"; and we can have *reish* or *aleph* for X_4, the former being the word for "cursed" and the latter being the word for "I will curse". In total, then, there are only 4 competing strings (including the original) that can be formed from this template, by choosing one of the two possibilities for X_2 and one of the two for X_4. Thus, we obtain a very simple description of $\boldsymbol{E}$.

2.3. Probability measurement

2.3.1. Initial calculation. In view of the above description of $\boldsymbol{E}$, $P(\boldsymbol{E})$ = 5.07E-10, namely, the product of the letter frequencies for all meaningful combinations: we use the product of the frequencies of the original letters occupying positions X_1, X_3, X_5, X_6, X_7 and X_8; multiplied by the sum of the frequencies of *heh* and *yod* for X_2; and by the sum of the frequencies of *aleph* and *reish* for X_4.

2.3.2. Adjustments Typically, many ELSs exist for a given anchor in T, but our first choice is always the minimal skip, followed by larger skips, ordered by the magnitude of the skip.

The anchor used for S_1 is the 5th minimal occurrence of this ELS in T, which is an adjustment factor of 39.6, using a weighted Bonferroni inequality detailed in the Appendix. The anchor can be anywhere within or adjoining the 22 non-anchor letters, not just in the position of our template. This is a factor of 23. These factors reduce the result to 4.6E-7. In view of the exponential decay of meaningful phrases with rate $H - H_1$ per unit of length (see 1.1), we can account for all possible competing phrases of greater length, with a factor smaller than 2. This yields a final p-level = 9.2E-7, about 1 in 1.1 million, the upper bound of the probability that a string comparable to S_1 appeared in T merely by chance.

2.3.3. Comparison to the previous result. Our p-level is over 20 times more significant than the one obtained in the previous study, when they are compared with like measures, yet both are highly significant. The previous study rated S_1's apparent intelligibility (compared to thousands of control strings), whereas the current study rates its relevance and meaning, in the context of the anchor. As the SGG concept would predict, a text that goes beyond mere intelligibility and contains focused meaning on a topic, is in fact easier to guess, and is therefore more significant and rarely seen in a random string.

3. Reinforcing Observations

Figure 2. The "*mirror*" encoding

When we present the table of Figure 1 with rows of half their original width, we observe additional occurrences of

Messiah and *takes revenge* (see Figure 2,discovered by Igor Picetsky). The original message is repeated in a modified form. And this example is by no means unique. It seems as though we are encountering a *mirror* encoding.

Likewise, we see that the words preceding the bin Laden anchor are "echoed" by horizontal occurrences near them (see Figure 2 on our review site). On an intuitive level, this kind of repetition is a strong indication of intentional encoding. Moreover, these patterns suggest a possibility of modifying our SGG. To aid our guessing of the content of the wildcards we can use hints from the surrounding encodings as well as the plain text. We leave a more detailed study of this to a further investigation.

4. A Further Example: The lasting power of Esau's tears

4.1. Background

We obtained our second example phrase, S_2, starting with the anchor, Titus:

אורב עשו בטיטוס ישא בכיה אחת ימשך כה

Translation: Esau ambushes via **Titus**; he will raise one cry; it will continue like this (see Figure 3, read from bottom to top).

Titus was the Roman emperor who led the war against Judea, conquering Jerusalem, destroying the Holy Temple and exiling the Jewish people [6]. This exile has continued for almost 2000 years, the major portion of Jewish history.

4.2. Analysis

4.2.1. The template. We construct the template
X_9כהX_8ימשךX_7אחתX_6בכיהX_5ישאX_4בטיטוסX_3עשוX_2אורבX_1
and review all possible substitutions in the same manner as before. We find that the set of favorable events, $\boldsymbol{E}$, as described in section 2.1, consists of 24 phrases, adding 23 to the original one, as follows. For the first word, we have the choice of the word אורב, *ambush*, or צורר, *enemy*; the words אשא, נשא and תשא can be used in addition to the original ישא, all of these being conjugations of the same verb, *to raise*; finally, the word כה can optionally be replaced by כך or כן, all of which are synonymous in this context. These variations can be combined to form $24 = 2 \times 3 \times 4$ competing phrases. Accounting for all these possibilities, as in section 2.3.1, we obtain $P(\boldsymbol{E})$ = 4.35E-10.

4.2.2. Adjustment and final result. Our weighted Bonferroni adjustment is 26.6 since Titus in S_2 is the 4th minimal skip in T. We also have a factor of 25 for alternative anchor locations, and the factor of 2 for competing phrases of greater length.

Figure 3. The Titus "*anchor*" is extended by 24 letters

After these adjustments, the final p-value is 5.8E-7, about 1 in 1.7 million, the upper bound for the probability that a string comparable to S_2 appeared in T merely by chance.

4.3. Beyond the numbers

In this section we explain the historical context, which helps to reveal the meaning and significance of the Titus phrase, S_2. The appearance of Esau in the Torah phrase generated from the anchor Titus is especially interesting because the Talmud (Gittin 56b) indicates that Titus descended from Esau. And famous Kabbalist Rabbi Menachem Azaria of Fano [11] even says that Titus was a reincarnation of Esau.

According to Jewish tradition based on the vision in the book of Daniel (chapter 7), the Jewish people will be subjected to the dominion of Four Kingdoms (Babylonian, Persian, Greek, and currently Roman, also known as the Exile of Edom or Esau). As an early Jewish source states [1], chapter 35 (see also [10], p. 350), the dominion of the first three kingdoms is for a known amount of time, unlike the fourth one, the Exile of Esau, which is continuing for a very long time without having a known end. This is precisely reflected by the last component of our phrase, "it will continue like this".

According to the Torah, the origin of Esau's enmity to Jacob (Yaakov) was the bitter rivalry over their father Isaac's blessings. When Esau realized that the blessings were given to his brother Yaakov and not to him, "Esau raised his voice and cried." (Genesis 27:38). The middle part of our ELS phrase is actually a paraphrase of this quote.

We can see the tight link between all parts of the code from the Zohar, which states, "these [Esau's] tears lowered Israel into exile. As soon as [the influence of] these tears will cease, through the weeping of Israel, they will come out of the exile." (Zohar Shemot 12b)

4.4. The connection to Genesis 27:38

We note that our ELS string S_2 contains four words that are derived from the verse in section 4.3, Genesis 27:38, namely: *Esau; will raise; one; cry*. To estimate this closeness quantitatively, we produce a population of random letter strings to determine how many of them contain words derived from this verse to a comparable extent. In a population of 280 million random strings, we find only 103 which contain four or more words derived from the verse, but none of them make grammatical sense. The details are documented on our review site. This observation further reinforces the significance of S_1. We have encountered other instances of ELS strings paraphrasing (parts of) verses, and this very interesting phenomenon is worth a separate study.

5. Discussion

While the overwhelming majority of decisions required for the determination of the set $\boldsymbol{E}$ are straightforward, we feel that they should be augmented by an independent review.

A word of comparison with the WRR paper [15] is in order. Our study differs from [15] in (at least) two important matters: (1) we search the entire Torah, not only the book of Genesis, and (2) we order the appearances of a keyword by the magnitudes of the observed skips.

Both Titus and bin Laden are on the short list of major historical and contemporary figures. We know that some other major figures, such as President Bush and Arafat, form interesting Torah Code phrases as well. But even without having a complete picture, the exceptionally high significance for the two current examples provide strong evidence for Torah Codes. This is in addition to other current studies discussed in [8].

A. Appendix

We apply the weighted Bonferroni inequality described below to situations where we have an ordered series of trials.

Consider testing (null) hypothesis $H_1, H_2, \ldots$ with corresponding p-values $P_1, P_2, \ldots$. Let $w_1, w_2, \ldots$ be a sequence of weights with w their sum. If a specific hypothesis H_i is rejected when $P_i \leq \alpha \cdot w_i/w$, then the weighted Bonferroni inequality

$$\Pr\{\cup(P_i \leq \alpha \cdot w_i/w)\} \leq \alpha$$

ensures that the probability of rejecting at least one hypothesis when all are true is no greater than α.

We cannot choose $w_i = 1/i$ because the harmonic series is diverging. Therefore we set $w_i = 1/[(i+1)\cdot(\ln(i+1))^2]$ to obtain a convergent series.

Unlike the usual form of the weighted Bonferroni inequality ([14], [7]) this form assigns to the null hypothesis H_i a weight which is dependent only on i and not on the total number of null hypotheses. Therefore, it can be used in a situation where the number of hypotheses is not pre-specified, but their order is established. We meet this situation when investigating the ELS strings generated by consecutive appearances of a given anchor.

Acknowledgments

The authors wish to thank Yuri Pikover and Rabbi Earl David for their generous support, and Harold Gans and Professor Robert Haralick for their valuable advice.

References

[1] *Pirkei Rabbi Eliezer*. Eshkol Publishing House, Jerusalem. [in Hebrew].

[2] Shannon's experiment to calculate the entropy of English (online simulation). `http://www.math.ucsd.edu/~crypto/java/ENTROPY/`.

[3] Component analysis of Torah Code phrases. ICPR06, submission, 2006.

[4] R. Ash. *Information Theory*. Dover, New York, 1965.

[5] Y. Choueka and S. Yeshurun. Statistical aspects of modern Hebrew prose. *Proc. 5th National Conference of Israeli Association for Information Processing*, pages 295–319, 1969. [in Hebrew].

[6] J. Flavius. *The Jewish War*.

[7] Y. Hochberg. A sharper Bonferroni procedure for multiple tests of significance. *Biometrica*, 75:800–802, 1988.

[8] A. Levitt. Torah Codes primer and survey of the latest research. http://www.torahcodes.net.

[9] A. Levitt, N. Bombach, H. Gans, R. M. Haralick, L. Schwartzman, and C. Stal. Long phrases in Torah Codes. In *2nd Annual Speech and Language Conference*, U. of Belgrade, 2004. http://www.torahcodes.net/papers.html.

[10] R. (Nachmanides). *Commentary on the Torah*. Shilo Publishing House, New York, 1971. [translated by C. B. Chavel].

[11] Rabbi M. Azaria of Fano. *Reincarnation of Souls*. Yareed Hasfarim, Jerusalem, 1998. [in Hebrew].

[12] Z. Sela. Oral communication.

[13] C. E. Shannon. Prediction and entropy of printed English. *Bell System Technical Journal*, 30(1):50–64, 1951.

[14] R. J. Simes. An improved Bonferroni procedure for multiple tests of significance. *Biometrica*, 73:751–754, 1986.

[15] D. Witztum, E. Rips, and Y. Rosenberg. Equidistant letter sequences in the book of Genesis. *Statistical Science*, 9(3):429–438, August 1994. http://www.torahcodes.co.il/wrr1/wrr1.htm.