Patterns of Co-Linear Equidistant Letter Sequences and Verses

Nachum Bombach
Economist
Certified Public Accountant
Ramat Gan, Israel
nachum@bombachschiff.co.il

Harold Gans
Senior Cryptologic Mathematician (retired)
U.S. Department of Defense
Fort George G. Meade, Maryland, U.S.A.
haroldgans@aol.com

Abstract

It has been shown ([4], [3]) that equidistant letter sequence (ELS) pairs in the book of Genesis (G) form more compact geometric patterns on the surface of a cylinder than is expected at random. This phenomenon has been demonstrated in G for specific lists of biographical data. We extend these results and show that:

1. *The compactness phenomenon holds for triplets of elements where two elements are words, taken from a lexicon derived from all the words in the Pentateuch, which form co-linear ELSs, and the third element is a verse in the text that contains two words with the same meanings as the first two elements respectively.*
2. *This phenomenon manifests itself in the entire Hebrew Pentateuch.*

The p-level obtained for this experiment is 6×10^{-8}.

1. Introduction

Witztum et al.([4]) and Gans et al.([3]), provide strong statistical evidence for the non-random coincidence of equidistant letter sequence (ELS) pairs in the Hebrew text of the book of Genesis (G). An ELS (n, d, k), in a text T, is defined as a sequence of letters in T found at positions $n, n+d, n+2d, \ldots, n+(k-1)d$. d is called the "skip distance". This phenomenon was demonstrated in G for pairs of ELSs spelling words taken from specific lists of biographical data. We extend these results in the following ways:

1. We consider a more complex pattern, comprising two co-linear ELSs and a verse in the text.
2. The linguistic relationships between the three elements of the pattern are more general than those considered in [4] and [3]. In particular, the relationships are semantical as opposed to historical.
3. The words spelled by the ELSs are taken from a lexicon consisting of all words in the Pentateuch (P) as opposed to being limited to historical data. One important consequence of this choice, first suggested by R. Haralick, is that the list is obviously *a priori*.
4. The phenomenon manifests itself in P as opposed to only in G.

2. The experiment

For two ELSs, $e = (n, d, k)$ and $e' = (n', d', k')$, we say that e and e' are co-linear if $d = d'$ and either $n' = n + kd$ or $n = n' + k'd'$. If e spells a word W and e' spells a word W', we call the pair (W, W') a "PLS" if e and e' are co-linear. We start with a lexicon consisting of all words in P that are at least 5 letters long. We now take all possible pairs of words from this lexicon and test each pair to see if it is a PLS, using skip distances, d, $2 \leq d \leq 1000$. $6,060$ PLSs were found. We now call a PLS a "phrase" if it has potential meaning as a Hebrew phrase. For example, in English, the phrase "green chair" would satisfy this condition while "green tall" would not. We now define a PLS to be (semantically) related to a verse in P if the following conditions are satisfied:

1. The PLS must have one word appearing in the verse.
2. The other word of the PLS shares the same root (in Hebrew, " שורש ") as a word in the verse. For example, the words may be verbs in different tenses; or one may be singular and the other plural. In Hebrew, one may be masculine and the other feminine, etc.
3. The PLS is a "phrase".
4. The meanings of the words in the phrase are the same as the meanings of their corresponding words in the verse. For example, in English consider the phrase "stern man" and the sentence "The man stood at the

0-7695-2521-0/06/$20.00 (c) 2006 IEEE

stern of the ship". This phrase and sentence would not satisfy this condition because the word "stern" has different meanings in the phrase and in the sentence.

There are two obvious alternatives to conditions 1 and 2 above, viz:

- require that both words satisfy condition 1, or
- require that both words satisfy condition 2.

The former yields a data size of 22 which, after conditions 3 and 4 are applied, leaves a data size of only 9. The latter choice yields a data size of $12,694$ making application of conditions 3 and 4, which must be done manually, impractical. Conditions 1 and 2 yield a workable data size of 796 PLSs; the total number of PLS - verse pairs is $1,698$. Application of conditions 3 and 4 result in 74 phrases that are related to at least one verse; the total number of related phrase - verse pairs is 113. This data set and the details about the linguistic decision process, can be found at [1].

The patterns formed by triplets of co-linear ELSs and related verses are tested for significant compactness using an adaptation of the techniques used in [4]. The compactness measure is described in section 3. In section 4 we develop a statistical technique for computing the p-level of the compactness measures obtained. The results of the experiment, and our conclusions, are presented in section 5.

3. The compactness measure

We define the compactness measure for each set of triplets (each co-linear pair of ELSs with its related verses) following the methodology of [4] with appropriate changes designed to accommodate the more complex patterns involved. Let $e = (n, d, k)$ and $e' = (n', d', k')$ be co-linear ELSs in P, and $V = \{v_1, \ldots, v_N\}$ be a set of verses in P that are related to e and e' (designated $Rel(e, e', V)$). $\delta_h(e, e', v_i)$ is defined by writing P as a single helix of letters spiraling down a cylinder with h vertical columns of letters and setting $\delta_h(e, e', v_i) = f^2 + g^2 + 1$, where f is the usual Euclidean distance (in rows and columns of letters) between two consecutive letters of e on the surface of the cylinder, and g is the minimal Euclidean distance between a letter of e or a letter of e' to a letter of v_i on the surface of the cylinder. Then $\mu_h(e, e', v_i) = 1/\delta_h(e, e', v_i)$ is directly related to the compactness of the configuration formed by e, e' and v_i on the cylinder for the given h. In general, setting $h = h(j)$ = the nearest integer to $|d|/j$ tends to make f small for small j, so we let $h(j)$ = the nearest integer to $|d|/j$ and define

$$\sigma(e, e', v_i) = \sum_{n=1}^{10} \mu_{h(j)}(e, e', v_i) \tag{1}$$

Note that $\sigma(e, e', v_i)$ tends to be large provided that there is a relatively compact configuration of e, e' and v_i on the surface of a cylinder with $h(j)$ columns for at least one of $h(j), j = 1, \ldots, 10$. We now set

$$\Omega(e, e', V) = \sum_{n=1}^{N} \sigma(e, e', v_i) \tag{2}$$

$\Omega(e, e', V)$ is the aggregate compactness measure of the co-linear pair of ELSs (e, e') and the set, V, of related verses. Note that for our data each co-linear pair (e, e') is unique.

4. The significance level of the compactness measure.

For each $\Omega(e, e', V)$, $e = (n, d, k)$, $e' = (n', d', k')$, let r be a pseudo-random variable uniformly distributed on $[0, 1]$, and $\lambda(P)$ be the number of letters in P. We then define $\eta(r, e, e') = [1 + r(\lambda(P) - |d|(k + k' - 1) - 1)]$ if $d > 0$ and $\eta(r, e, e') = [1 - d(k + k' - 1) + r(\lambda(P) - |d|(k + k' - 1) - 1)]$ if $d < 0$ so that $\eta(r, e, e')$ is uniformly distributed over all possible starting points of a $k + k'$ long phrase with skip d in P. We now define $\delta_h^r(e, e', v_i)$ in the same way as $\delta_h(e, e', v_i)$ except that we use $\eta(r, e, e')$ as the starting position of the phrase in the calculation of g; f is unaffected. We then define $\sigma^r(e, e', v_i)$ and $\Omega^r(e, e', V)$ in the same way that $\sigma(e, e', v_i)$ and $\Omega(e, e', V)$ are defined, but using $\delta_h^r(e, e', v_i)$ instead of $\delta_h(e, e', v_i)$. Thus, $\Omega^r(e, e', V)$ is the aggregate compactness measure of the randomly placed phrase (e, e') and the set V of related verses. This calculation is performed for random variables $r = r(i), i = 1, \ldots, 9999$. Then the p-level associated with any set $Rel(e, e', V)$ is $P(e, e', V) = (card(\{r(i) | \Omega^{r(i)}(e, e', V) \geq \Omega(e, e', V)\}) + 1)/10^4$. We calculate two overall statistics for the experiment by combining the 74 p-levels obtained in a way paralleling the techniques used in [4]. We define

$$P_1 = \sum_{i=k}^{74} \binom{74}{i} (0.2)^i (0.8)^{74-i} \tag{3}$$

where $k = card\{P(e, e', V) | P(e, e', V) \leq 0.2\}$. If the $P(e, e', V)$ are independent then P_1 is the binomial probability that at least k of the $P(e, e', V)$ would be less than or equal to 0.2. We define $P_2 = F^{74}(\prod(P(e, e', V)))$ where $F^N(X) = X(1 - \ln X + (-\ln X)^2/2! + \cdots + (-\ln X)^{N-1}/(N-1)!)$. If the $P(e, e', V)$ are independent then $F^N(X)$ is the probability that $\prod(P(e, e', V)) \leq X$ (reference [2] formula (3.5)).

99,999,999 permutations were done in which the phrases were pseudo-randomly permuted relative to the sets of related verses, so that each phrase was matched to a set

of verses that were not related to it. Each permutation π_i determines statistics ${P_1}^{\pi_i}$ and ${P_2}^{\pi_i}$. Then $P_1 = (card\{\pi_i | {P_1}^{\pi_i} \leq P_1\} + 1)/10^8$ is the probability under the null hypothesis that P_1 would rank as low as it does among the ${P_1}^{\pi_i}$. The same was done for calculating P_2, except that its number of permutations was 999,999,999.

5. Results and conclusions

The process of deciding if a phrase is related to a verse is a human one, and therefore necessarily subjective. It is, however, a relatively simple and in most cases an unambiguous task for anyone fluent in the language. This decision process was made by the author before any compactness measures were calculated. To test the reliability and stability of the results obtained, the linguistic decision process was replicated by two linguistic editors: Mrs Riva Rothman (Hebrew Language BA from Bar Ilan University), and Mrs Noah Eitam (Tanach and Computer Science BA from Jerusalem college (Michlalla)) – henceforth: the "consultants". They were not privy to the author's results. The first row in Table 1 shows the results of the experiment with the data prepared by the author. The second row shows the independent results obtained by the consultants. The third row shows the results obtained on the intersection of the author's and consultants' data sets. The fourth row shows the results obtained on the union of the author's and consultants' data sets.

Table 1. N is the total number of related phrase-verse pairs, n is the number of phrases related to at least one verse, P_1 and P_2 are the p-levels obtained for the two statistics.

	N	n	P_1	P_2
Author (A)	113	74	3.0×10^{-8}	3.3×10^{-8}
Consultants (C)	138	84	1.0×10^{-6}	1.2×10^{-7}
(A) and (C)	108	73	1.5×10^{-6}	5.0×10^{-8}
(A) or (C)	143	85	2.7×10^{-7}	1.0×10^{-7}

We conclude that:

1. The compactness of patterns formed on the surface of a cylinder by co-linear ELSs and semantically related verses is smaller than can be attributed to chance. Specifically, the Bonferroni inequality yields a p-level of 6×10^{-8} against the null hypothesis of random distribution of the compactness measures.

2. The p-levels obtained are stable relative to the linguistic decision process.

Acknowledgment

We acknowledge the contribution of Yuri Pikover of Los Angeles, whose generous sponsorship of this research made this paper possible.

References

[1] N. Bombach. Web site for attachments to "Patterns of co-linear equidistant letter sequences and verses". http://www.torahcodes.org/patterns/attachments.htm.

[2] W. Feller. *An Introduction to Probability Theory and its Applications*, volume 2. Wiley and Sons, New York, London, Sydney, 1966.

[3] H. Gans, Z. Inbal, and N. Bombach. Patterns of equidistant letter sequence pairs in Genesis. In *Proceedings of the 18th International Conference on Pattern Recognition*, August 2006.

[4] D. Witztum, E. Rips, and Y. Rosenberg. Equidistant letter sequences in the book of Genesis. *Statistical Science*, 9(3):429–438, August 1994.