APPENDIX

Statistical Science
1994, Vol. 9, No. 3, 429-438 (abridged)

Equidistant Letter Sequences in the Book of Genesis

Doron Witztum, Eliyahu Rips and Yoav Rosenberg

Abstract. It has been noted that when the Book of Genesis is written as two-dimensional arrays, equidistant letter sequences spelling words with related meanings often appear in close proximity. Quantitative tools for measuring this phenomenon are developed. Randomization analysis shows that the effect is significant at the level of 0.00002.

Key words and phrases: Genesis, equidistant letter sequences, cylindrical representations, statistical analysis.

1. INTRODUCTION

The phenomenon discussed in this paper was first discovered several decades ago by Rabbi Weissmandel [7]. He found some interesting patterns in the Hebrew Pentateuch (the Five Books of Moses), consisting of words or phrases expressed in the form of equidistant letter sequences (ELS's)--that is, by selecting sequences of equally spaced letters in the text.

As impressive as these seemed, there was no rigorous way of determining if these occurrences were not merely due to the enormous quantity of combinations of words and expressions that can be constructed by searching out arithmetic progressions in the text. The purpose of the research reported here is to study the phenomenon systematically. The goal is to clarify whether the phenomenon in question is a real one, that is, whether it can or cannot be explained purely on the basis of fortuitous combinations.

The approach we have taken in this research can be illustrated by the following example. Suppose we have a text written in a foreign language that we do not understand. We are asked whether the text is meaningful (in that foreign language) or meaningless. Of course, it is very difficult to decide between these possibilities, since we do not understand the language. Suppose now that we are equipped with a very partial dictionary, which enables us to recognise a small portion of the words in the text: "hammer" here and "chair" there, and maybe even "umbrella" elsewhere. Can we now decide between the two possibilities?

Not yet. But suppose now that, aided with the partial dictionary, we can recognise in the text a pair of conceptually related words, like "hammer" and "anvil." We check if there is a tendency of their appearances in the text to be in "close proximity." If the text is meaningless, we do not expect to see such a tendency, since there is no reason for it to occur. Next, we widen our check; we may identify some other pairs of conceptually related words: like "chair" and "table," or "rain" and "umbrella." Thus we have a sample of such pairs, and we check the tendency of each pair to appear in close proximity in the text. If the text is meaningless, there is no reason to expect such a tendency. However, a strong tendency of such pairs to appear in close proximity indicates that the text might be meaningful.

Note that even in an absolutely meaningful text we do not expect that, deterministically, every such pair will show such tendency. Note also, that we did not decode the foreign language of the text yet: we do not recognise its syntax and we cannot read the text.

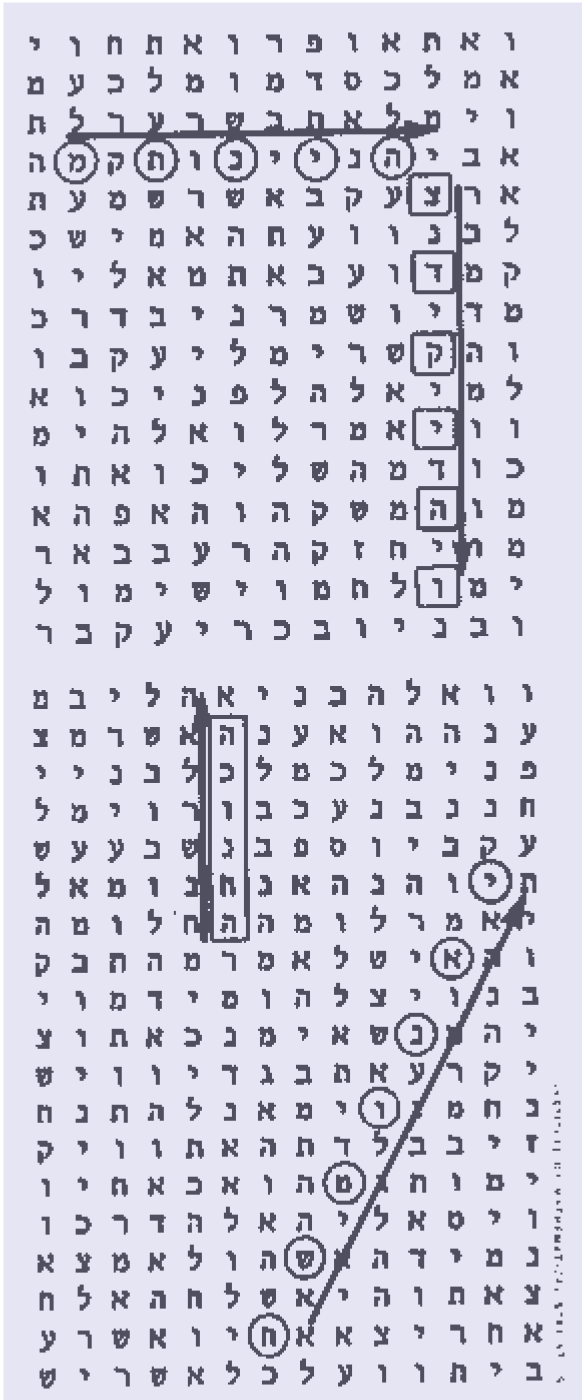
This is our approach in the research described in the paper. To test whether the ELS's in a given text may contain "hidden information," we write the text in the form of two-dimensional arrays, and define the distance between ELS's according to the ordinary two-dimensional Euclidean metric. Then we check whether ELS's representing conceptually related words tend to appear in "close proximity."

Suppose we are given a text, such as Genesis *(G)*. Define an equidistant letter sequence (ELS) as a sequence of letters in the text whose positions, not counting spaces, form an arithmetic progression; that is, the letters are found at the positions

$n, n+d, n+2d, \dots, n+(k-1)d$.

We call d the *skip*, n the *start* and k the *length* of the ELS. These three parameters uniquely identify the ELS, which is denoted (n,d,k).

Let us write the text as a two-dimensional array--that is, on a single large page--with rows of equal length, except perhaps for the last row. Usually, then, an ELS appears as a set of points on a straight line. The exceptional cases are those where the ELS "crosses" one of the vertical edges of the array and reappears on the opposite edge. To include these cases in our framework, we may think of the two vertical edges of the array as pasted together, with the end of the first line pasted to the beginning of the second, the end of the second to the beginning of the third and so on. We thus get a cylinder on which the text spirals down in one long line.



It has been noted that when Genesis is written in this way, ELS's spelling out words with related meanings often appear in close proximity. In Figure 1 we see the example of 'patish-ùéèô' (hammer) and 'sadan-ðãñ' (anvil); in Figure 2, 'Zidkiyahu-åäé÷ãö' (Zedekia) and 'Matanya-äéðúî' (Matanya), which was the original name of King Zedekia (Kings II, 24:17). In Figure 3 we see yet another example of 'hachanuka-äëåðçä' (the Chanuka) and 'chashmonaee -éàðåîùç' (Hasmonean), recalling that the Hasmoneans were the priestly family that led the revolt against the Syrians whose successful conclusion the Chanuka feast celebrates.

Indeed, ELS's for short words, like those for 'patish-ùéèô' (hammer) and 'sadan-ðãñ' (anvil), may be expected on general probability grounds to appear close to each other quite often, in any text. In Genesis, though, the phenomenon persists when one confines attention to the more "noteworthy" ELS's, that is, those in which the skip $|d|$ is *minimal* over the whole text or over large parts of it. Thus for 'patish-ùéèô' (hammer), there is no ELS with a smaller skip than that of Figure 1 in all of Genesis; for 'sadan-ðãñ' (anvil), there is none in a section of text comprising 71% of G; the other four words are minimal over the whole text of G. On the face of it, it is not clear whether or not this can be attributed to chance. Here we develop a method for testing the significance of the phenomenon according to accepted statistical principles. After making certain choices of words to compare and ways to measure proximity, we perform a randomization test and obtain a very small p-value, that is, we find the results highly statistically significant.

2. OUTLINE OF THE PROCEDURE

In this section we describe the test in outline. In the Appendix, sufficient details are provided to enable the reader to repeat the computations precisely, and so to verify their correctness. The authors will provide, upon request, at cost, diskettes containing the program used and the texts G, I, R, T, U, V and W (see Section 3).

We test the significance of the phenomenon on samples of pairs of related words (such as hammer-anvil and Zedekia-Matanya). To do this we must do the following:

(i) define the notion of "distance" between any two words, so as to lend meaning to the idea of words in "close proximity";

(ii) define statistics that express how close, "on the whole," the words making up the sample pairs are to each other (some kind of average over the whole sample);

(iii) choose a sample of pairs of related words on which to run the test;

(iv) determine whether the statistics defined in (ii) are "unusually small" for the chosen sample.

Task (i) has several components. First, we must define the notion of "distance" between two given ELS's in a given array; for this we use a convenient variant of the ordinary Euclidean distance. Second, there are many ways of writing a text as a two-dimensional array, depending on the row length; we must select one or more of these arrays and somehow amalgamate the results (of course, the selection and/or amalgamation must be carried out according to clearly stated, systematic rules). Third, a given word may occur many times as an ELS in a text; here again, a selection and amalgamation process is called for. Fourth, we must correct for factors such as word length and composition. All this is done in detail in Sections A.1 and A.2 of the Appendix.

We stress that our definition of distance is not unique. Although there are certain general principles (like minimizing the skip d) some of the details can be carried out in other ways. We feel that varying these details is unlikely to affect the results substantially. Be that as it may, we chose one particular definition, and have, throughout, used *only* it, that is, the function $c(w,w')$ described in Section A.2 of the Appendix had been defined before any sample was chosen, and it underwent no changes. [Similar remarks apply to choices made in carrying out task (ii).]

Next, we have task (ii), measuring the overall proximity of pairs of words in the sample as a whole. For this, we used two different statistics P_1 and P_2 , which are defined and motivated in the Appendix (Section A.5). Intuitively, each measures overall proximity in a different way. In each case, a small value of P_i indicates that the words in the sample pairs are, on the whole, close to each other. No other statistics were *ever* calculated for the first, second or indeed any sample.

In task (iii), identifying an appropriate sample of word pairs, we strove for uniformity and objectivity with regard to the choice of pairs and to the relation between their elements. Accordingly, our sample was built from a list of personalities (p) and the dates (Hebrew day and month) (p') of their death or birth. The personalities were taken from the *Encyclopedia of Great Men in Israel* [5].

At first, the criterion for inclusion of a personality in the sample was simply that his entry contain at least three columns of text and that a date of birth or death be specified. This yielded 34 personalities (the *first list*--Table 1). In order to avoid any conceivable appearance of having fitted the tests to the data, it was later decided to use a fresh sample, without changing anything else. This was done by considering all personalities whose entries contain between 1.5 and 3 columns of text in the *Encyclopedia*; it yielded 32 personalities (the *second list*--Table 2). The significance test was carried out on the second sample only.

Note that personality-date pairs (p,p') are not word pairs. The personalities each have several appellations, there are variations in spelling and there are different ways of designating dates. Thus each personality-date pair (p,p') corresponds to several word pairs (w,w'). The precise method used to generate a sample of word pairs from a list of personalities is explained in the Appendix (Section A.3).

The measures of proximity of word pairs (w,w') result in statistics P_1 and P_2 . As explained in the Appendix (Section A.5), we also used a variant of this method, which generates a smaller sample of word pairs from the same list of personalities. We denote the statistics P_1 and P_2 , when applied to this smaller sample, by P_3 and P_4 .

Finally, we come to task (iv), the significance test itself. It is so simple and straightforward that we describe it in full immediately.

The second list contains of 32 personalities. For each of the 32! permutations π of these personalities, we define the statistic P_1^π obtained by permuting the personalities in accordance with π, so that Personality i is matched with the dates of Personality $\pi(i)$. The 32! numbers P_1^π are ordered, with possible ties, according to the usual order of the real numbers. If the phenomenon under study were due to chance, it would be just as likely that P_1 occupies any one of the 32! places in this order as any other. Similarly for P_2, P_3 and P_4. This is our null hypothesis.

To calculate significance levels, we chose 999,999 random permutations π of the 32 personalities; the precise way in which this was done is explained in the Appendix (Section A.6). Each of these permutations π determines a statistic P_1^π; together with P_1, we have thus 1,000,000 numbers. Define the *rank order* of P_1 among these 1,000,000 numbers as the number of P_1^π not exceeding P_1; if P_1 is tied with other P_1^π, half of these others are considered to "exceed" P_1. Let ρ_1 be the rank order of P_1, divided by 1,000,000; under the null hypothesis, ρ_1 is the probability that P_1 would rank as low as it does. Define ρ_2, ρ_3 and ρ_4 similarly (using the same 999,999 permutations in each case).

After calculating the probabilities ρ_1 through ρ_4, we must make an overall decision to accept or reject the research hypothesis. In doing this, we should avoid selecting favorable evidence only. For example, suppose that $\rho_3 = 0.01$, the other ρ_i being higher. There is then the temptation to consider ρ_3 only, and so to reject the null hypothesis at the level of 0.01. But this would be a mistake; with enough sufficiently diverse statistics, it is quite likely that just by chance, some one of them will be low. The correct question is, "Under the null hypothesis, what is the probability that at least one of the four ρ_i would be less than or equal to 0.01?" Thus denoting the event "$\rho_i <= 0.01$" by E_i, we must find the probability not of E_3, but of "E_1 or E_2 or E_3 or E_4." If the E_i were mutually exclusive, this probability would be 0.04; overlaps only decrease the total probability, so that it is in any case less than or equal to 0.04. Thus we can reject the null hypothesis at the level of 0.04, but not 0.01.

More generally, for any given δ, the probability that at least one of the four numbers ρ_i is less than or equal to δ is at most 4 δ. This is known as the Bonferroni inequality. Thus the overall significance level (or p-value), using all four statistics, is $\rho_0 := 4 \min \rho_i$.

Up to Section 1 Up to Section 2 Down to Appendix

3. RESULTS AND CONCLUSIONS

In Table 3, we list the rank order of each of the four P_i among the 1,000,000 corresponding P_i^π. Thus the entry 4 for P_4 means that for precisely 3 out of the 999,999 random permutations π, the statistic P_4^π was smaller than P_4 (none was equal). It follows that $\min \rho_i = 0.000004$ so $\rho_0 = 4 \min \rho_i = 0.000016$. The same calculations, using the same 999,999 random permutations, were performed for control texts. Our first control text, R, was obtained by permuting the letters of G randomly (for details, see Section A.6 of the Appendix). After an earlier version of this paper was distributed, one of the readers, a prominent scientist, suggested to use as a control text Tolstoy's *War and Peace*. So we used text T consisting of the initial segment of the Hebrew translation of Tolstoy's *War and Peace* [6]--of the same length of G. Then we were asked by a referee to perform a control experiment on some early Hebrew text. He also suggested to use randomization on words in two forms: on the whole text and within each verse. In accordance, we checked texts I, U and W: text I is the Book of Isaiah [2]; W was obtained by permuting the words of G randomly; U was obtained from G by permuting randomly words within each verse. In addition, we produced also text V by permuting the verses of G randomly. (For details, see Section A.6 of the Appendix.) Table 3 gives the results of these calculations, too. In the case of I, $\min \rho_i$ is approximately 0.900; in the case of R it is 0.365; in the case of T it is 0.277; in the case of U it is 0.276; in the case of V it is 0.212; and in the case of W it is 0.516. So in five cases $\rho_0 = 4 \min \rho_i$ exceeds 1, and in the remaining case $\rho_0 =$ 0.847; that is, the result is totally nonsignificant, as one would expect for control texts.

We conclude that the proximity of ELS's with related meanings in the Book of Genesis is not due to chance.

TABLE 3

Rank order of P_i among one million P_i^π

	P_1	P_2	P_3	P_4
G	453	5	570	4
R	619,140	681,451	364,859	573,861
T	748,183	363,481	580,307	277,103
I	899,830	932,868	929,840	946,261
W	883,770	516,098	900,642	630,269
U	321,071	275,741	488,949	491,116
V	211,777	519,115	410,746	591,503

APPENDIX: DETAILS OF THE PROCEDURE

In this Appendix we describe the procedure in sufficient detail to enable the reader to repeat the computations precisely. Some motivation for the various definitions is also provided.

In Section A.1, a "raw" measure of distance between words is defined. Section A.2 explains how we normalize this raw measure to correct for factors like the length of a word and its composition (the relative frequency of the letters occurring in it). Section A.3 provides the list of personalities p with their dates p' and explains how the sample of word pairs (w, w') is constructed from this list. Section A.4 identifies the precise text of Genesis that we used. In Section A.5, we define and motivate the four summary statistics P_1, P_2, P_3 and P_4. Finally, Section A.6 provides the details of the randomization.

Sections A.1 and A.3 are relatively technical; to gain an understanding of the process, it is perhaps best to read the other parts first.

A.1 The Distance between Words

To define the "distance" between words, we must first define the distance between ELS's representing those words; before we can do that, we must define the distance between ELS's in a given array; and before we can do that, we must define the distance between individual letters in the array.

As indicated in Section 1, we think of an array as one long line that spirals down on a cylinder; its *row length h* is the number of vertical columns. To define the distance between two letters x and x', cut the cylinder along a vertical line between two columns. In the resulting plane each of x and x' has two integer coordinates, and we compute the distance between them as usual, using these coordinates. In general, there are two possible values for this distance, depending on the vertical line that was chosen for cutting the cylinder; if the two values are different, we use the smaller one.

Next, we define the distance between fixed ELS's e and e' in a fixed cylindrical array. Set

f := the distance between consecutive letters of e,

f' := the distance between consecutive letters of e',

l := the minimal distance between a letter of e and one of e',

and define $\delta(e, e') := f^2 + f'^2 + l^2$. We call $\delta(e, e')$ the *distance* between the ELS's e and e' in the given array; it is small if both fit into a relatively compact area. For example, in Figure 3 we have $f = 1, f' = \sqrt{5}$, $l = \sqrt{34}$ and $\delta = 40$.

Now there are many ways of writing Genesis as a cylindrical array, depending on the row length h. Denote by $\delta_h(e, e')$ the distance $\delta(e, e')$ in the array determined by h, and set $\mu_h(e, e') := 1/\delta_h(e, e')$; the larger $\mu_h(e, e')$ is, the more compact is the configuration consisting of e and e' in the array with row length h. Set $e = (n,d,k)$ (recall that d is the skip) and $e' = (n',d',k')$. Of particular interest are the row lengths $h = h_1, h_2, \ldots$, where hi is the integer nearest to $|d|/i$ (1/2 is rounded up). Thus when $h = h_1 = |d|$, then e appears as a column of adjacent letters (as in Figure 1); and when $h = h_2$, then e appears either as a column that skip alternate rows (as in Figure 2) or as a straight line of knight's moves (as in Figure 3). In general, the arrays in which e appears relatively compactly are those with row length h_i with i "not too large."

Define h_i' analogously to h_i. The above discussion indicates that if there is an array in which the configuration (e,e') is unusually compact, it is likely to be among those whose row length is one of the first 10 h_i or one of the first 10 h_i'. (Here and in the sequel 10 is an arbitrarily selected "moderate" number.) So setting

$$\sigma(e, e') := \sum_{i=1}^{10} \mu_{h_i}(e, e') + \sum_{i=1}^{10} \mu_{h_i'}(e, e'),$$

we conclude that $\sigma(e, e')$ is a reasonable measure of the maximal "compactness" of the configuration (e, e') in any array. Equivalently, it is an inverse measure of the minimum distance between e and e'.

Next, given a word w, we look for the most "noteworthy" occurrence or occurrences of w as an ELS in G. For this, we chose those ELS's e = (n,d,k) with $|d| >= 2$ that spell out w for which $|d|$ is minimal over all of G, or at least over large portions of it. Specifically, define the *domain of minimality* of e as the maximal segment T_e of G that includes e and does not include any other

ELS $\hat{e} = (\hat{n}, \hat{d}, \hat{k})$

for w with

$\hat{d} < |d|$.

If e' is an ELS for another word w', then $T_e \cap T_{e'}$ is called the *domain of simultaneous minimality* of e and e'; the length of this domain, relative to the whole of G, is the "weight" we assign to the pair (e,e'). Thus we define $\omega(e,e') := \lambda(e, e')/\lambda(G)$, where $\lambda(e, e')$ is the length of $T_e \cap T_{e'}$, and $\lambda(G)$ is the length of G. For any two words w and w', we set

$$\Omega (w, w') := \sum \omega (e, e') \sigma (e, e'),$$

where the sum is over all ELS's e and e' spelling out w and w', respectively. Very roughly, $\Omega (w, w')$ measures the maximum closeness of the more noteworthy appearances of w and w' as ELS's in Genesis--the closer they are, the larger is $\Omega (w, w')$.
When actually computing $\Omega (w, w')$, the sizes of the lists of ELS's for w and w' may be impractically large (especially for short words). It is clear from the definition of the domain of minimality that ELS's for w and w' with relatively large skips will contribute very little to the value of $\Omega (w, w')$ due to their small weight. Hence, in order to cut the amount of computation we restrict beforehand the range of the skip $|d| <= D(w)$ for w so that the expected number of ELS's for w will be 10. This expected number equals the product of the relative frequencies (within Genesis) of the letters constituting w multiplied by the total number of all equidistant letter sequences with $2 <= |d| <= D$. [The latter is given by the formula $(D-1)(2L-(k-1)(D+2))$, where L is the length of the text and k is the number of letters in w.] The same restriction applies also to w' with a corresponding bound $D(w')$. Abusing our notation somewhat, we continue to denote this modified function by $\Omega (w, w')$.

A.2 The Corrected Distance

In the previous section we defined a measure $\Omega(w,w')$ of proximity between two words w and w' -- an inverse measure of the distance between them. We are, however, interested less in the absolute distance between two words than in whether this distance is larger or smaller than "expected." In this section, we define a "relative distance" $c (w,w')$, which is small when w is "unusually close" to w', and is 1, or almost 1, when w is "unusually far" from w'.

The idea is to use perturbations of the arithmetic progressions that define the notion of an ELS. Specifically, start by fixing a triple (x,y,z) of integers in the range {-2,-1,0,1,2}; there are 125 such triples. Next, rather than looking for ordinary ELS's (n,d,k), look for "(x,y,z)-perturbed ELS's" $(n,d,k)^{(x,y,z)}$, obtained by taking the positions

$n, n + d,..., n + (k\text{-}4)d, n + (k\text{-}3)d + x, n + (k\text{-}2)d + x + y, n + (k\text{-}1)d + x + y + z,$

instead of the positions n, $n + d$, $n + 2d,..., n + (k\text{-}1)d$. Note that in a word of length k, k-2 intervals could be perturbed. However, we preferred to perturb only the three last ones, for technical programming reasons.

The *distance* between two (x,y,z)-perturbed ELS's $(n,d,k)^{(x,y,z)}$ and $(n',d',k')^{(x,y,z)}$ is defined as the distance between the ordinary (unperturbed) ELS's (n,d,k) and (n',d',k').

We may now calculate the "(x,y,z)-proximity" of two words w and w' in a manner exactly analogous to that used for calculating the "ordinary" proximity $\Omega(w,w')$. This yields 125 numbers $\Omega^{(x,y,z)}(w,w')$, of which $\Omega(w,w') = \Omega^{(0,0,0)}(w,w')$ is one. We are interested in only some of these 125 numbers; namely, those corresponding to triples (x,y,z) for which there actually exist some (x,y,z)-perturbed ELS's in Genesis for w, and some for w' [the other $\Omega^{(x,y,z)}(w,w')$ vanish]. Denote by $M(w,w')$ the set of all such triples, and by $m(w,w')$ the number of its elements.

Suppose (0,0,0) is in $M(w,w')$, that is, both w and w' actually appear as ordinary ELS's (i.e., with $x = y = z = 0$) in the text. Denote by $v(w,w')$ the number of triples (x,y,z) in $M(w,w')$ for which $\Omega^{(x,y,z)}(w,w') >= \Omega(w,w')$. If $m(w,w') >= 10$ (again, 10 is an arbitrarily selected "moderate" number),

$$c(w,w') := v(w,w') / m(w,w').$$

If (0,0,0) is not in $M(w,w')$, or if $m(w,w') < 10$ (in which case we consider the accuracy of the method as insufficient), we do not define $c(w,w')$.

In words, the corrected distance $c(w,w')$ is simply the rank order of the proximity $\Omega(w,w')$ among all the "perturbed proximities" $\Omega^{(x,y,z)}(w,w')$; we normalize it so that the maximum distance is 1. A large corrected distance means that ELS's representing w are far away from those representing w', on a scale determined by how far the *perturbed* ELS's for w are from those for w'.

A.3 The Sample of Word Pairs

The reader is referred to Section 2, task (iii), for a general description of the two samples. As mentioned there, the significance test was carried out only for the second list, set forth in Table 2. Note that the personalities each may have several appelations (names), and there are different ways of designating dates. The sample of word pairs (w, w') was constructed by taking each name of each personality and pairing it with each designation of that personality's date. Thus when the dates are permuted, the total number of word pairs in the sample may (and usually will) vary.

We have used the following rules with regard to Hebrew spelling:

1. For words in Hebrew, we always chose what is called the *grammatical orthography*--"ktiv dikduki." See the entry "ktiv" in Even-Shoshan's dictionary [1].

2. Names and designations taken from the Pentateuch are spelled as in the original.

3. Yiddish is written using Hebrew letters; thus, there was no need to transliterate Yiddish

names.

4. In transliterating foreign names into Hebrew, the letter "alef-à" is often used as a *mater lectionis*; for example, "Luzzatto" may be written "åèöåì" or "åèàöåì." In such cases we used both forms.

In designating dates, we used three fixed variations of the format of the Hebrew date. For example, for the 19th of Tishri, we used éøùú è'é, éøùú è'éá and éøùúá è'é. The 15th and 16th of any Hebrew month can be denoted as ä'é or å'è and å'é or æ'è, respectively. We used both alternatives.
The list of appellations for each personality was provided by Professor S. Z. Havlin, of the Department of Bibliography and Librarianship at Bar Ilan University, on the basis of a computer search of the "Responsa" database at that university.
Our method of rank ordering of ELS's based on (x, y, z)-perturbations requires that words have at least five letters to apply the perturbations. In addition, we found that for words with more than eight letters, the number of (x, y, z)-perturbed ELS's which actually exist for such words was too small to satisfy our criteria for applying the corrected distance. Thus the words in our list are restricted in length to the range 5-8. The resulting sample consists of 298 word pairs (see Table 2).

A.4 The Text

We used the standard, generally accepted text of Genesis known as the *Textus Receptus*. One widely available edition is that of the Koren Publishing Company in Jerusalem. The Koren text is precisely the same as that used by us.

A.5 The Overall Proximity Measures P_1, P_2, P_3 and P_4

Let N be the number of word pairs (w, w') in the sample for which the corrected distance $c(w, w')$ is defined (see Sections A.2 and A.3). Let k be the number of such word pairs (w, w') for which $c(w, w') <= 1/5$.
Define

$$p_1 := \sum_{j=k}^{N} \binom{N}{j} \left(\frac{1}{5}\right)^{j} \left(\frac{4}{5}\right)^{N-j}$$

To understand this definition, note that *if* the $c(w, w')$ were independent random variables that are uniformly distributed over [0,1], *then* P_1 would be the probability that at least k out of N of them are less than are equal to 0.2. However, we do *not* make or use any such assumptions about uniformity and independence. Thus P_1, though calibrated in probability terms, is simply an ordinal index that measures the number of word pairs in a given sample whose words are "pretty close" to each other [i.e., $c(w, w') <= 1/5$], taking into account the size of the whole sample. It enables us to compare the overall proximity of the word pairs in different samples; specifically, in the samples arising from the different permutations of the 32 personalities.
The statistic P_1 ignores all distances $c(w,w')$ greater than 0.2, and gives equal weight to all distances less than 0.2. For a measure that is sensitive to the actual size of the distances, we calculate the product $\Pi c(w,w')$ over all word pairs (w,w') in the sample. We then define

$$P_2 := F^N := \left(\Pi\, c(w, w')\right),$$

with N as above, and

$$F^N(X) := X \left(1 - \ln X + \frac{(-\ln X)^2}{2!} + \ldots + \frac{(-\ln X)^{N-1}}{(N-1)!} \right).$$

To understand this definition, note first that if $x_1, x_2, ..., x_N$ are independent random variables that are uiformly distributed over [0,1], then the distribution of their product $X := x_1 x_2 \ldots x_N$ is given by Prob($X <= X_0$) $= F^N(X_0)$; this follows from (3.5) in [3], since the -ln x_i are distributed exponentially, and $-\ln X = \Sigma i$ $(\ln x_i)$. The intuition for P_2 is then analogous to that for P_1: *If* the $c(w,w')$ were independent random variables that are uniformly distributed over [0,1], *then* P_2 would be the probability that the product Π $c(w,w')$ is as small as it is, or smaller. But as before, we do not use any such uniformity or independence assumptions. Like P_1, the statistic P_2 is calibrated in probability terms; but rather than thinking of it as a probability, one should think of it simply as an ordinal index that enables us to compare the proximity of the words in word pairs arising from different permutations of the personalities.

We also used two other statistics, P_3 and P_4. They are defined like P_1 and P_2, except that for each personality, all appellations starting with the title "Rabbi" are omitted. The reason for considering P_3 and P_4 is that appellations starting with "Rabbi" often use only the given names of the personality in question. Certain given names are popular and often used (like "John" in English or "Avraham" in Hebrew); thus several different personalities were called Rabbi Avraham. If the phenomenon we are investigating is real, then allowing such appellations might have led to misleadingly low values for $c(w,w')$ when π matches one "Rabbi Avraham" to the dates of another "Rabbi Avraham." This might have resulted in misleadingly low values P_1^π and P_2^π for the permuted samples, so in misleadingly low significance levels for P_1 and P_2 and so, conceivably, to an unjustified rejection of the research hypothesis. Note that this effect is "one-way"; it could not have led to unjustified acceptance of the research hypothesis, since under the null hypothesis the number of P_i^π exceeding P_i is in any case uniformly distributed. In fact, omitting appellations starting with "Rabbi" did not affect the results substantially (see Table 3); but we could not know this before perfoming the calculations.

An intuitive feel for the corrected distances (in the original, unpermuted samples) may be gained from Figure 4. Note that in both the first and second samples, the distribution for R looks quite random, whereas for G it is heavily concentrated near 0. It is this concentration that we quantify with the statistics P_i.

A.6 The Randomizations

The 999,999 random permutations of the 32 personalities were chosen in accordance with Algorithm P of Knuth [4], page 125. The pseudorandom generator required as input to this algorithm was that provided by Turbo-Pascal 5.0 of Borland Inter Inc. This, in turn, requires a seed consisting of 32 binary bits; that is, an integer with 32 digits when written to the base 2. To generate this seed, each of three prominent scientists was asked to provide such an integer, just before the calculation was carried out. The first of the three tossed a coin 32 times; the other two used the parities of the digits in widely separated blocks in the decimal expansion of π. The three resulting integers were added modulo 2^{32}. The resulting seed was 01001 10000 10011 11100 00101 00111 11.
The control text R was constructed by permuting the 78,064 letters of G with a single random permutation, generated as in the previous paragraph. In this case, the seed was picked arbitrarily to be the decimal integer 10 (i.e., the binary integer 1010). The control text W was constructed by permuting the words of G in exactly the

same way and with the same seed, while leaving the letters within each word unpermuted. The control text *V* was constructed by permuting the verses of *G* in the same way and with the same seed, while leaving the letters within each verse unpermuted.

The control text *U* was constructed by permuting the words within each verse of *G* in the same way and with the same seed, while leaving unpermuted the letters within each word, as well as the verses. More precisely, the Algorithm *P* of Knuth [4] that we used requires *n* - 1 random numbers to produce a random permutation of *n* items. The pseudorandom generator of Borland that we used produces, for each seed, a long string of random numbers. Using the binary seed 1010, we produced such a long string. The first six numbers in this string were used to produce a random permutation of the seven words constituting the first verse of Genesis. The *next* 13 numbers (i.e., the 7th through the 19th random numbers in the string produced by Borland) were used to produce a random permutation of the 14 words constituting the second verse of Genesis, and so on.

REFERENCES

[1] EVEN-SHOSHAN, A. (1989). *A New Dictionary of the Hebrew Language*. Kiriath Sefer, Jerusalem.

[2] FCAT (1986). The Book of Isaiah, file ISAIAH.MT. Facility for Computer Analysis of Texts (FCAT) and Tools for Septuagint Studies (CATSS), Univ. Pennsylvania, Philadelphia. (April 1986.)

[3] FELLER, W. (1966). *An Introduction to Probability Theory and Its Applications* **2**. Wiley, New York.

[4] KNUTH, D. E. (1969). *The Art of Computer Programming* **2**. Addison-Wesley, Reading, MA.

[5] MARGALIOTH, M., ed. (1961). *Encyclopedia of Great Men in Israel; a Bibliographical Dictionary of Jewish Sages and Scholars from the 9th to the End of the 18th Century* **1**-**4**. Joshua Chachik, Tel Aviv.

[6] TOLSTOY, L. N. (1953) *War and Peace*. Hebrew translation by L. Goldberg, Sifriat Poalim, Merhavia.

[7] WEISSMANDEL, H. M. D. (1958). *Torath Hemed*. Yeshivath Mt. Kisco, Mt. Kisco.

EQUIDISTANT LETTER SEQUENCES IN THE BOOK OF GENESIS: II. THE RELATION TO THE TEXT

Doron Witztum, Eliyahu Rips, and Yoav Rosenberg

ABSTRACT. It has been noted that when the Book of Genesis is written as two-dimensional arrays, equidistant letter sequences spelling words often appear in close proximity with portions of the text which have related meaning. Quantitative tools for measuring this phenomenon are developed. Randomization analysis is done for three samples. For one of them the effect is significant at the level of .000000004.

Key words and phrases. Genesis, Equidistant letter sequences, Strings of letters, Cylindrical representations, Statistical analysis.

---

1. Introduction.

There is an old Jewish tradition about a "hidden text" in the Hebrew Pentateuch (the Five Books of Moses), consisting of words or phrases expressed in the form of equidistant letter sequences (ELS's) -- that is by selecting sequences of equally spaced letters in the text. Since this tradition was passed orally, only few expressions that belong to the "hidden text" were preserved in writing (Rabbenu Bahya, 1492 and Cordovero, 1592); actually almost all the words and the syntax are unknown. Rabbi H.M.D. Weissmandel (Weissmandel, 1958) was the first to try to show the existence of such a "hidden text", by finding interesting patterns consisting of ELS's.

In a previous paper (Witztum *et al*, 1994), we developed a methodology for systematic and rigorous studies of the same nature; namely, for attempts to show objectively the existence of the"hidden text" in the Hebrew Pentaceuch. This methodology was applied to study the "hidden text" of the Book of Genesis.

The approach we have taken in our research can be illustrated by the following example. Suppose we have a text written in a foreign language that we do not understand. We are asked whether the text is meaningful (in that foreign language) or meaningless. Of course, it is very difficult to decide between these possibilities, since we do not understand the language. Suppose now that we are equipped with a very partial dictionary, which enables us to recognise a small portion of the words in the text: "hammer" here and "chair" there, and maybe even "umbrella" elsewhere. Can we now decide between the two possibilitities?

Not yet. But suppose now that, aided with the partial dictionary, we can recognise in the text a pair of conceptually related words, like "hammer" and "anvil". We check if there is a tendency of their appearances in the text to be in "close proximity". If the text is meaningless, we do not expect to see such a tendency, since there is no reason for it to occur. Next, we widen our check; we may identify some other pairs of conceptually related words: like "chair" and "table", or "rain" and "umbrella". Thus we have a sample of such pairs, and we check the tendency of each pair to appear in close proximity in the text. If the text is meaningless, there is no reason to expect such a tendency. However, a strong tendency of such pairs to appear in close proximity indicates that the text might be meaningful.

Note that even in an absolutely meaningful text we do not expect that, deterministically, every such pair will show such tendency. Note also, that we did not decode the foreign language of the text yet: we do not recognise its syntax and we cannot read the text.

In our research we consider the set of all ELS's spelling out words or phrases in the language of the text. The approach described in this example suggests the two following lines of investigation:

A) A study of the mutual location of ELS's spelling out conceptually related words or expressions.

B) A study of the mutual location of ELS's spelling out words or expressions with conceptually related portions of the text.

Suppose we are given a text, such as Genesis (G). Define an *ELS* (equidistant letter sequence) as a sequence of letters in the text whose positions, not counting spaces, form an arithmetic progression; that is, the letters are found at the positions

$n, n+d, n+2d,..., n+(k-1)d.$

We call d the *skip,* n the *start*, and k the *length* of the ELS. These three parameters uniquely identify the ELS, which is denoted (n, d, k).

Let us write the text as a two-dimensional array -- i.e., on a single large page -- with rows of equal length, except perhaps for the last row. Usually, then, an ELS appears as a set of points on a straight line. The exceptional cases are those where the ELS "crosses" one of the vertical edges of the array and reappears on the opposite edge. To include these cases in our framework, we may think of the two vertical edges of the array as pasted together, with the end of the first line pasted to the beginning of the second, the end of the second to the beginning of the third, and so on. We thus get a cylinder on which the text spirals down in one long line.

It has been noted that when Genesis is written in this way, and the distance between ELS's is defined according to the ordinary two-dimensional Euclidean metric -- ELS's spelling words with related meaning often appear in close proximity. It has also been noted that ELS's spelling words often appear in close proximity with portions of the text which have related meaning.

Thus, our research focuses on two phenomena:

Phenomenon A

The appearance of "noteworthy" ELS's spelling words with related meanings in close proximity, on two-dimensional arrays. (the "noteworthy" ELS's are those for which the skip $|d|$ is minimal on the whole text, or on large parts of it; for short we call them *minimal* ELS's).

Our paper (Witztum *et al*., 1994) deals with Phenomenon A. There we developed a method for testing the significance of the phenomenon according to accepted statistical principles. After making certain choices of words to compare and ways to measure proximity, we performed a randomization test and obtained a very small p-value, i.e. we found the results highly statisically significant.

Phenomenon B

The appearance of minimal ELS's spelling words in close proximity, on two-dimensional arrays, with conceptually related words or expressions appearing in the string of letters of the text: i.e. with skip ± 1.

Example

On Figure 1 we see a pair of words: the word פרטיים (private) appearing as an ELS (with skip 855) with the word שמות (names) appearing in the text (Gn 26:18).

Each ELS determines a series of tables with row lengths $h = h_1, h_2,...$, where h_i is the integer nearest to $|d|/i$ (1/2 is rounded up).

The rows in our table has 428 letters (only 21 of them are shown in Figure 1). The number 428 is the nearest integer to 855/2, so the word פרטיים (private) appears every second row.

The table is determined by this ELS of the word פרטיים (private), which is minimal in a section of the text comprising 76% of G. The word שמות (names) appears in the text of G seven times as SL's. Note that G contains 78,064 letters. More examples are given in Appendix A.3.1.

The measuring scheme for Phenomenon A (see Witztum *et al*., 1994) is applicable, with minor changes, to study Phenomenon B.

In this paper we make certain choices of words to compare and perform similar randomization tests. We obtain very small p-values; that is, we find that the results are highly statistically significant.

2. Outline of the Procedure.

In this section we describe the test in outline. In an appendix, sufficient details are provided to enable the reader to repeat the computations precisely, and so to verify their correctness.

We test the significance of the phenomenon on samples of pairs of related words. To do this we must do the following:

i. define the notion of "distance" between any two words, so as to lend meaning to the idea of words in "close proximity";
ii. define statistics that express how close, "on the whole," the words making up the sample pairs are to each other (some kind of average over the whole sample);
iii. choose a sample of pairs of related words on which to run the test; and
iv. determine whether the statistics defined in (ii) are "unusually small" for the chosen sample.

Notice that the procedure here described is identical to that used in (Witztum *et al*., 1994), except for minor changes in the details of task (i), due to fact that here we consider "distance" between an ELS and a word in the string of letters (SL) of the text, and not a "distance" between two ELS's (as in Witztum *et al*., 1994).

Task (i) has several components. First, we must define the notion of "distance" between an ELS and an SL of the text in a given array; for this we use a convenient variant of the ordinary Euclidean distance. Second, there are many ways of writing a text as a two-dimenional array, depending on the row length; we must select one or more of these arrays, and somehow amalgamate the results (of course, the selection and/or amalgamation must be carried out according to clearly stated, systematic rules). Third, a given word may occur many times as an ELS in a text; here again, a selection and amalgamation process is called for. Fourth, we must correct for factors such as word length and composition. All this is done in detail in Sections A.1 and A.2 of the Appendix.

Next, we have task (ii), measuring the overall proximity of pairs of words in the sample as a whole. For this, we used two different statistics, P_1 and P_2, which are defined and motivated in the Appendix (Section A.6). Intuitively, each measures overall proximity in a different way. In each case, a small value of P_i indicates that the words in the sample pairs are, on the whole, close to each other.

To accomplish task (iii) we composed three samples (**Sample B1**, **Sample B2** and **Sample B3**) of pairs of expressions (w, w'), where w's are words appearing as ELS's, and w' 's are words appearing as SL's (i.e. with $d' = \pm 1$).

Preliminary test was done for each sample, in order to check how the *subject* of the sample appears as ELS's and SL's in *Genesis*, and consequently to decide whether to test the sample itself. For details see Appendix, Section A.3.1.

Sample B1 is built on the basis of the Hebrew alphabet. For every letter *'x'* of the Hebrew alphabet we consider pairs (w,w'), where w' 's are found as SL's and have the meaning "a name beginning with *'x'* " or "names beginning with *'x'* ", while w's are names beginning with *'x'* taken from "A Treasury of Men`s Names" (which is included as an appendix in Even-Shoshan's famous Hebrew dictionary (Even-Shoshan, 1989). For a detailed definition of the **Sample B1** see Appendix (Section A.3.2).

Sample B2 is built exactly in the same way as Sample B1, except for the fact that the names are taken from "A Treasury of Women's Names" from the same dictionary.

Sample B3 is built on the basis of the list of the seventy descendents of Noah's sons: the Semites, the Hamites, and the Japhetites, found in *Genesis* Chapter 10. Jewish tradition teaches, that these seventy descendents became the Seventy Nations which constitute Humanity. This concept is well known, and is usually found in biblical Encyclopaedias under the title "The Table of Nations" (see for instance *Encyclopedia Biblica,* 1962). **Sample B3** consists of pairs (w, w') where w' 's are names from this list, and w`s are expressions from a fixed set of expressions describing basic aspects of nationality (such as name, country, language etc.) For details see Appendix, Section A.3.3 and Table 8.

Finally, we come to Task (iv), the significant test itself. We apply the same procedure for all three samples. For **Sample B3** we describe it here: for the other two samples the (similar) details are given in Appendix A.4.

The list of Seventy Nations consists of 68 *different* names (in two cases nations have the same name). For each of the 68! permutations π of these names, we define the statistic P_1^π obtained by permuting the names in accordance with π, so that Name i is matched with the set of expressions defined for Name $\pi(i)$. The 68! numbers P_1^π are ordered, with possible ties, according to the usual order of the real numbers. If the phenomenon under study were due to chance, it would be just as likely that P_1 occupies any one of the 68! places in this order as any other. Similarly for P_2. This is our null hypothesis.

To calculate significance levels, we chose 999,999 random permutations π of the 68 names; the precise way in which this was done is explained in the Appendix (Section A.7). Each of these permutations π determines a statistic P_1^π together with P_1, we have thus 1,000,000 numbers. Define the *rank order* of P_1, among these 1,000,000 numbers as the number of P_1^π not exceeding P_1; if P_1 is tied with other P_1^π, half of these others are considered to "exceed" P_1. Let ρ_1 be the rank order of P_1, divided by 1,000,000; under the null hypothesis, ρ_1 is the probability that P_1 would rank as low as it does. Define ρ_2 similarly (using the same 999,999 permutations in each case).

For **Sample B3** we performed an additional test with 999,999,999 random permutations. In this case only statistic P_2^π was calculated. The time needed for the computation of P_1^π for 999,999,999 random permutations is at present not within the reach of our possibilties.

After calculating the probabilities ρ_1 and ρ_2, we must make an overall decision, for each sample, to accept or reject the null hypothesis. Thus the overall significance level (or p-value) for each sample, using the two statistics. is $\rho_0 := 2 \min \rho_i$.

3. Results and Conclusions.

In Tables 1, 2 and 3 we present the results for the three samples. Table 1 shows the results for **Sample B1**. There we list the rank order of P_1 and P_2 among the 1,000,000 corresponding P_1^π and P_2^π. Thus the entry 1139 for P_2 means that for 1138 out of the 999,999 random permutations π, the statistic P_2^π was smaller than P_2. It follows that $\min \rho_i$ = .000442, so $\rho_0 = 2 \min \rho_i$ = .000884.

We conclude that for **Sample B1** the null hypothesis is rejected with significance level (p-value) .000884.

The same calculations, using the same 999,999 random permutations, were performed for a control text V (see Witztum *et al.,* 1994). The text V was obtained from G by permuting the verses of G randomly. (For details, see Appendix; Section A.7).

Table 1 gives the results of these calculations too. In the case of V, $\min \rho_i$ is approximately .169, being non-significant.

Table 2 shows the results for **Sample B2** for G. The results are non-significant. We saw no reason to perform any further tests for this sample.

Table 3 shows the results for **Sample B3**. In part A of it, the results for G, as well for the control text V for 999,999 random permutations are summarized. In the case of V, $\min \rho_i$ is approximately .559.

In part B, the results for G for 999,999,999 random permutations are given. Notice that only the rank order of P_2 was calculated. It turned out to be 4.

We conclude that for Sample B3 the null hypothesis is rejected with significance level (p-value) .000000004.

APPENDIX: Details of the Procedure.

In this Appendix we describe the procedure in sufficient detail to enable the reader to repeat the computations precisely. Some motivation for the various definitions is also provided.

In Section A.1 of this Appendix, a "raw" measure of distance between words is defined. Section A.2 explains how we normalize this raw measure to correct for factors like the length of a word and its composition (the relative frequency of the letters occurring in it). Section A.3 explains how the three samples B1, B2 and B3 are constructed. Section A.5 identifies the precise text, of Genesis that we used. In Section A.6, we define and motivate the statistics P_1 and P_2. The details of the task (iv) are described in Section A.4. Finally, Section A.7 provides the details of the randomization.

A.1 The Distance Between Words.

To define the "distance" between words, we must first define the distance between an ELS representing a word and a string of letters (SL) in the text, (i.e. with $d = \pm 1$) representing the other word. Before we can do that, we must define the distance between ELS and SL in a given array: and before we can do that, we must define the distance between individual letters in the array.

As indicated in Section 1, we think of an array as one long line that spirals down on a cylinder; its *row length h* is the number of vertical columns. To define the distance between two letters x and x', cut the cylinder along a vertical line between two columns. In the resulting plane each of x and x' have two integer coordinates. and we compute the distance between them as usual, using these coordinates. In general, there are two possible values for this distance, depending on the vertical line that was chosen for cutting the cylinder; if the two values are different, we use the smaller one.

Next, we define the distance between fixed ELS e and SL e' in a fixed cylindrical array. Set

f := the distance between consecutive letters of e,

f' := the distance between consecutive letters of $e' = 1$.

l := the minimal distance between a letter of e and one of e',

and define $\delta(e, e') := f^2 + f'^2 + l^2 + 1$. We call $\delta(e, e')$ the *distance* between the ELS e and the SL e' in the given array; it is small if both fit into a relatively compact area.

Now there are many ways of writing Genesis as a cylindrical array, depending on the row length h. Denote by $\delta_h(e, e')$ the distance $\delta(e, e')$ in the array determined by h, and set $\mu_h(e, e') := 1/\delta_h(e, e')$; the larger $\mu_h(e, e')$ is, the more compact is the configuration consisting of e and e' in the array with row length h. Set $e = (n, d, k)$ (recall that d is the skip). Of particular interest are the row lengths $h = h_1, h_2,....$ where h_i is the integer nearest to $|d| / i$ (1/2 is rounded up). Thus when $h = h_1 = |d|$, then e appears as a column of adjacent letters and when $h = h_2$, then e appears either as a column that skips alternate rows or as a straight line of knight's moves. In general, the arrays in which e appears relatively compactly are those with row length h_i with i "not too large."

The above discussion indicates that if there is an array in which the configuration (e,e') is unusually compact, it is likely to be among those whose row length is one of the first ten h_i. (Here and in the sequel 10 is an arbitrarily selected "moderate" number). So setting

$$\sigma(e, e') := \sum_{i=1}^{10} \mu_{h_i}(e, e'),$$

we conclude that $\sigma(e, e')$ is a reasonable measure of the maximal "compactness" of the configuration (e, e') in any array. Equivalently, it is an inverse measure of the minimum distance between e and e'.

Next, given a word w, we look for the most "noteworthy" occurrence or occurrences of w as an ELS in G. For this, we chose ELS's $e = (n,d,k)$ with $|d| \geq 2$ that spell out w for which $|d|$ is minimal over all of G, or at least over large portions of it. Specifically, define the *domain of minimality* of e as the maximal segment T_e of G that includes e and does not include any other ELS $\hat{e} = (\hat{n}, \hat{d}, \hat{k})$ for w with $|\hat{d}| < |d|$. The length of T_e, relative to the whole of G, is the "weight" we assign to e. Thus we define $\omega(e) := \lambda(T_e)/\lambda(G)$, where $\lambda(T_e)$ is the length of T_e, and $\lambda(G)$ is the length of G. For any two words w and w', we set

$$\Omega(w, w') := \sum \omega(e)\sigma(e, e'),$$

where the sum is over all ELS's e spelling out w and over all SL's e' spelling out w'. Roughly, $\Omega(w, w')$ measures the maximum closeness of the more noteworthy appearances of w as ELS's and w' as SL's in Genesis--the closer they are, the larger is $\Omega(w, w')$.

When actually computing $\Omega(w, w')$, the size of the list of ELS's for w may be impractically large (especially for short words). It is clear from the definition of the domain of minimality that ELS's for w with relatively large skips will contribute very little to the value of $\Omega(w,w')$ due to their small weight. Hence, in order to cut the amount of computation we restrict beforehand the range of the skip $|d| \leq D(w)$ for w so that the expected number of ELS's for w will be 10. This expected number equals the product of the relative frequencies (within Genesis) of the letters constituting w multiplied by the total number of all equidistant letter sequences with $2 \leq |d| \leq D$. (The latter is given by the formula $(D - 1)(2L - (k - 1)(D + 2))$, where L is the length of the text and k is the number of letters in w). Abusing our notation somewhat, we continue to denote this modified function by $\Omega(w,w')$.

A.2 The Corrected Distance.

In the previous section we defined a measure $\Omega(w, w')$ of proximity between two words w and w' -- an inverse measure of the distance between them. We are, however, interested less in the absolute distance between two words, than in whether this distance is larger or smaller than "expected". In this section, we define a "relative distance" $c(w, w')$, which is small when w is "unusually close" to w', and is 1, or almost 1, when w is "unusually far" from w'.

The idea is to use perturbations of the arithmetic progressions that define the notion of an ELS. Specifically, start by fixing a triple (x,y,z) of integers in the range $\{-r,...,0,...,r\}$; there are $(2r + 1)^3$ such triples. In Witztum *et al*. (1994) and also here we put $r = 2$. which gives us 125 triples. Next, rather than looking for ordinary ELS's (n,d,k), look for "(x,y,z)-perturbed ELS's" $(n,d,k)^{(x,y,z)}$ obtained by taking the positions

$$n, n + d,...,n + (k - 4)d, n + (k - 3)d + x, n + (k - 2)d + x + y, n + (k - 1)d + x + y + z,$$

instead of the positions $n, n + d, n +2d,...,n +(k - 1)d$. Note that in a word of length k, k-2 intervals could be perturbed. However, we preferred to perturb only the 3 last ones, for technical programming reasons.

The *distance* between the (x,y,z)-peturbed ELS $(n, d, k)^{(x,y,z)}$ and the SL $(n', \pm 1,k')$ is defined by the same formulae as in the non-perturbed case, where f is taken to be the distance between the first two letters of (x,y,z)-perturbed e.

We may now calculate the "(x,y,z)-proximity" of two words w and w' in a manner exactly analogous to that used for calculating the "ordinary" proximity $\Omega(w, w')$. This yields 125 numbers $\Omega^{(x,y,z)}(w, w')$, of which $\Omega(w,w')=\Omega^{(0,0,0)}(w,w')$ is one. We are interested in only some of these 125 numbers; namely, those corresponding to triples (x,y,z) for which there actually exist some (x,y,z)-perturbed ELS's in Genesis for w (the other $\Omega^{(x,y,z)}(w,w')$ vanish). Denote by $M(w, w')$ the set of all such triples, and by $m(w, w')$ the number of its elements.

Suppose (0,0,0) is in $M(w, w')$, i.e., w actually appears as ordinary ELS (i.e., with $x = y = z = 0$) in the text. Denote by $v(w,w')$ the number of triples (x,y,z) in $M(w,w')$ for which $\Omega^{(x,y,z)}(w,w') \geq \Omega(w,w')$. If $m(w,w') \geq 10$ (again, 10 is an arbitrarily selected "moderate" number),

$$c(w,w') := v(w,w')/m(w,w').$$

If (0, 0,0) is not in $M(w, w')$, or if $m(w, w') < 10$ (in which case we consider the accuracy of the method as insufficient), we do not define $c(w,w')$.

In words, the corrected distance $c(w,w')$ is simply the rank order of the proximity $\Omega(w,w')$ among all the "perturbed proximities" $\Omega^{(x,y,z)}(w,w')$; if $\Omega(w,w')$ is tied with other $\Omega^{(x,y,z)}(w,w')$, half of these others are considered to "exceed" $\Omega(w,w')$. We normalize it so that the maximum distance is 1. A large corrected distance means that ELS's representing w are far away from the SL's representing w', on a scale determined by how far the *perturbed* ELS's for w are from the SL's for w'.

A.3 The Samples of Word Pairs.

In this Section we describe the three samples that were tested in this research. Each of them consists of pairs of expressions (w,w'), where according to our procedure, we are looking for w to appear as ELS's and for w' to appear as SL's.

Our method of rank ordering of ELS's based on (x,y,z)-perturbations requires that words have at least 5 letters to apply the perturbations. In addition, we found that for words with more than 8 letters, the number of (x,y,z)-perturbed ELS's which actually exist for such words was too small to satisfy our criteria for applying the corrected distance. Thus the words in our list are restricted in length to the range 5-8, exactly as in Witztum *et al*. (1994). However. there is no restriction on the words or expressions appearing as SL's.

A.3.1 Preliminary Tests.
Sample B1 deals with men's names. Appendix B in Even-Shoshan's Hebrew dictionary (Even-Shoshan. 1989) is "A Treasury of Private Names". It is divided into two parts: "A Treasury of Men's Names" and "A Treasury of Women's Names". Originally, we intented to check a sample based on the (much bigger) Treasury of men's names. The Treasury contains names from the Hebrew Bible and from various periods of the Hebrew language. We decided to include in our sample only names taken from the Hebrew Bible (they are indicated as such in this Treasury). These are original Hebrew names, or names which are etymologically Semitic or Hebrew. (See the foreword to the Treasury).

We describe the *subject* of the sample by the following set of pairs of expressions:

1. פרטיים (private) - שמות (names),
2. שמיים (Semitic) - שמות (names),
3. עבריים (Hebrew) - שמות (names),
4. מקוריים (original) - שמות (names).

The Hebrew term for "The Hebrew Bible" is either מקרא (Mikra) or תנ"ך (Tanach). We used both:

5. מן ה מקרא (from the Hebrew Bible) - שמות (names),
6. מן ה תנ"ך (from the Hebrew Bible) - שמות (names).

Then we describe the fact that we deal with men's names, by the following pairs:

a. אנשים (of men) - שמות (names),
b. האנשים (of the men) - שמות (names),
c. לאנשים (for men) - שמות (names)
d. של אנשים (of men) - שמות (names),
e. של ה אנשים (of the men) - שמות (names).

The first pair is shown above in Fig. 1 (in the Introduction). There, a minimal ELS for the word פרטיים (private) appears in close proximity with an appearance of the word שמות (names) as a SL. We *mark this appearance* of שמות (names), and check how close, each of the other expressions in the pairs 2) to 6) and a) to e) - appearing as a minimal ELS - "hits" at it. For example, this appearance of שמות (names), is shown in Fig. 2: this table is determined by an ELS for מקוריים (original), which is minimal in a section of the text comprising 86% of *G*.

Tables 4A and 4B give the values of $c(w,w')$ for each pair from the above lists (where w' is שמות (names), and w is the corresponding expression). Recall that $c(w, w')$ is defined only when w appears as an ELS, and that w is restricted in length to the range 5-8.

The total for the 11 pairs is: $P_1 = 0.0000042$, $P_2 = 0.000397$ (for the definition of P_1 and P_2 see section A.6).

A randomization test that fits for this type of samples, where the same word is "paired" with a list of expressions, is the subject of our next paper.

Sample B2 deals with women's names. We describe this fact exactly as we had done with men's names; i.e. by the following pairs:

a. נשים(of women) - שמות (names),
b. ה נשים(of the women) - שמות (names),
c. ל נשים(for women)- שמות (names),
d. של נשים (of women) - שמות (names),
e. של ה נשים (of the women) - שמות (names).

Here too, the word שמות(names) is in the *above mentioned* appearance as SL.

Table 5 gives the values of $c(w,w')$ for each pair from the above list (where w' is שמות (names), and w, is the corresponding expression).

To do a preliminary test for Sample B3, we proceed in a similar way and check the following expressions with the same appearance of שמות(names) as above:

a. עמים(of nations) - שמות (names),
b. ה עמים(of the nations) - שמות (names),
c. ל עמים(for nations) - שמות (names):
d. של עמים (of nations) - שמות (names),
e. של ה עמים (of the nations) - שמות (names).

We checked more specifically the subject of Sample B3: the title שבעים ה עמים "The Seventy Nations" can be also written as ע' ה עמים (using the well-known numerical value of the Hebrew letters. ע'= 70). The alternative title in use is שבעים בני נח ("The Seventy Descendants of Noah"). Notice that in Hebrew בני נח means both the sons and the descendants of Noah.) This expression can be also written as ע' בני נח, where again ע' stands for 70. Thus we have the pairs:

f. שבעים ה עמים (the Seventy Nations) - שמות (names),
g. ע' ה עמים (The Seventy Nations) - שמות (names),
h. שבעים בני נח (The Seventy Descendants of Noah) - שמות (names),
i. ע' בני נח (The Seventy Descendants of Noah) - שמות (names).

Table 6 gives the values of $c(w, w')$ for each pair from the above list (where w' is שמות (names), and w is the corresponding expression). The total for the 9 pairs is: $P_1 = 0.0170$, $P_2 = 0.0154$.

Seeing these results as encouraging, we proceed to check even more closely the subject of Sample B3.

The expression בני נח (the descendants of Noah) exists in G as SL's five times. So we can look directly for meetings between this expression and the word שבעים(seventy) as ELS's:

1. שבעים(seventy) - בני נח (the descendants of Noah)

In Fig. 3 we see the best meeting of these expressions. This table is determined by an ELS for the word שבעים (seventy) with skip 28. We mark *this appearance* of בני נח (the descendants of Noah), and check how close, each of the other expressions in the following pairs - appearing as a minimal ELS- "hits" it.

2. ה עמים (the nations) - בני נח (the descendants of Noah),
3. ע' ה עמים (The Seventy Nations) - בני נח (the descendants of Noah),
4. שבעים ה עמים (The Seventy Nations) - בני נח (the descendants of Noah),
5. ע' בני נח (The Seventy Descendants of Noah) - בני נח (the descendants of Noah),
6. שבעים בני נח (The Seventy Descendants of Noah) - בני נח (the descendants of Noah),
7. מבני נח (some of the descendants of Noah) - בני נח (the descendants of Noah),
8. בני בניו (his descendants) - בני נח (the descendants of Noah).
9. ע' בני בניו (the seventy descendants of him) - בני נח (the descendants of Noah),
10. שבעים בני בניו (the seventy descendants of him) - בני נח (the descendants of Noah),
11. מבני בניו (some of his descendants) - בני נח (the descendants of Noah),
12. צאצאיו (his descendants) - בני נח (the descendants of Noah),
13. ע' צאצאיו (the seventy descendants of him) - בני נח (the descendants of Noah),
14. שבעים צאצאיו (the seventy descendants of him) - בני נח (the descendants of Noah),
15. מצאצאיו (some of his descendants) - בני נח (the descendants of Noah).

The expressions בני בניו and צאצאיו also mean descendants: and מבני נח (some of descendants of Noah), מבני בניו (some of his descendants) and מצאצאיו (some of his descendants) express the fact that not all of his descendants became nations.

In Fig. 4 we see the pairs 1), 2) and 5) appearing together. This table is determined by an ELS for ע' בני נח (The Seventy Descendants of Noah) with skip -100, each row containing 25 = 100/4 letters. The ELS for שבעים (seventy) is the same as shown in Fig. 3. The ELS's for ע' בני נח (The Seventy Descendants of Noah) and for ה עמים (the nations) are minimal over the whole text of *G*.

Table 7 gives the va1ues of $c(w,w')$ for each pair from the above list (where w' is בני נח (the descendants of Noah), and w is the corresponding expression). The total for the 15 pairs is: $P_1 = 0.0000779$, $P_2 = 0.0000377$.

A.3.2 Sample B1.

"A Treasury of Men's Names" is included as an appendix in Even-Shoshan's Hebrew dictionary (Even-Shoshan, 1989). From this appendix we chose all the names which are mentioned in the Hebrew Bible (Pentateuch, Prophets and Writings) as proper names - as indicated in the Treasury itself - and contain 5 to 8 letters.

This collection of names is designated as *S*. Let *HA* := Hebrew Alphabet (which consists of 22 letters).

For every $x \in HA$ we define

$W(x) := \{w \in S \mid w \text{ begins with the letter } x\}$.

We define 22 sets $A(x)$, $x \in HA$, where the elements of each set are the following three expressions:

i. 'x' ב שם (a name beginning with 'x')
ii. 'x' ב שמות (names beginning with 'x')
iii. 'x' ב שמת (names beginning with 'x')

The plural form of שם(=name) is written in Genesis either as שמות, or as שמת.

Then we define

$W'(x) := \{w' \in A(x) \mid w'$ appears in the text with $d' = \pm 1\}$.

Now we define

$HA' := \{x \in HA \mid W(x) \neq \emptyset$ and $W'(x) \neq \emptyset\}$.

and, for every $x,y \in HA'$

Pair $B1(x,y) := \{(w, w') \mid w \in W(x), w' \in W'(y)\}$.

For instance, ('א' ב שם, אליעחד) $\in$ *Pair* $B1$(א, א).

The sample is defined by:

$$\textit{Sample } B1 := \bigcup_{x \in HA'} \textit{Pair } B1(x,x).$$

For *Genesis* it contains 457 pairs of expressions.

Sample B2 is defined precisely in the same way as Sample B1, except that we used 'A Treasury of Women's Names" from the same dictionary. For *Genesis* it contains 38 pairs of expressions.

A.3.3 The Sample B3.

Chapter 10 of *Genesis* contains the names of the seventy descendents of Noah's sons: the Semites, the Hamites and the Japhetites (see Table 8). Jewish tradition teaches, that these seventy descendents became the Seventy Nations, which constitute Humanity. The source for this tradition is found in *Targum Yonathan* (Jonathan's Translation) to *Deuteronomy* 32:8 (see Kasher, 1983, p.294). This concept is well known, and is found in biblical Encyclopedias under the title "the Table of Nations" (see for instance, *Encyclopedia Biblica*, 1962).

Jewish tradition further tells us (Hagra, 1905), that a nation has the following four characteristics:

1. its name,
2. its country,
3. its language.
4. its script.

Sample B3 is constructed on the basis of the "Table of Nations" using these four categories. Let us denote

TN:= the Table of Nations

For every $x \in TN$ we define below a set *List* $B3(x)$ of expressions.

For every $x,y \in TN$ we define

Pair $B3(x,y) := \{(w,y) \mid w \in$ *List* $B3(x)$ and w has from 5 to 8 letters $\}$.

The resulting sample is:

$$\textit{Sample } B3 := \bigcup_{x \in TN} \textit{Pair } B3\,(x,x).$$

Now let us explain how *List B3*(x) is constructed. Take, for example, one of the names in the list-- כנען (Canaan - item 18 in Table 8):

1. To express the fact that כנען is a name of a nation, we use the (Biblical) expressions עם כנען (the *Nation* of Canaan) and כנענים (the *Canaanites*).
2. For the *Country* of Canaan there is the Biblical expression: ארץ כנען.
3. For the *Language* of Canaan there is the Biblical expression: שפת כנען.
4. For the *Script* of Canaan we do not have a Biblical expression like כתב כנען, but it is a Biblical formation.

In the course of history names of nations and countries changed. In certain cases, the Jewish tradition, as expressed by *Targum Yonathan* to *Genesis* , gives the new names for the nations and countries under consideration. We add these new names to the corresponding categories (1) and (2). Thus, for example, the new name for Javan (item 4 in Table 8) is מקדוניא (Macedonia), whereas for Canaan there are no new names.

The same procedure is applied to each name from the Table of Nations. For every $x \in TN$ List B3 (x) consists of the following expressions:

1. 'עם x' (the nation of x), 'ים-x' (the people of x), NewNationName(x);
2. 'ארץ x' (the country of x), NewCountryName(x);
3. 'שפת x' (the language of x);
4. 'כתב x' (the script of x).

In Table 8 for every name $x \in TN$ the corresponding List B3(x) is given.

Remark 1. The expressions 'עם x', 'ארץ x' and 'שפת x' were chosen for categories (1), (2) and (3) because at least for some $x \in TN$ they appear as Biblical expressions (see the above example of Canaan). For category (4) we used 'כתב x'; although it does not appear for any $x \in TN$, it is the only Biblical formation suitable for this category.

Remark 2. Concerning the expressions 'ים-x ', we are aware of the fact that only valid lingustic formations should be used; therefore we use these expressions if, at least, their singular form 'י-x' appears in the Hebrew Bible. For instance, for גמר (item 1) we do not take גמרים, since the expression גמרי does not appear in the Hebrew Bible; so we cannot know whether it has any meaning at all. The check was done using Even-Shoshan's *New Concordance of the Bible* (Even-Shoshan, 1981).

Remark 3. In Talmudic Literature we find various new names for nations and countries, and we cannot decide in favor of one against the other. We preferred to use a single source. The Aramaic translation of *Genesis* that gives the largest number of such new names is *Targum Yonathan*. We used *Targum Yonathan* as printed in *Torah Shelemah* (Kasher, 1929) (with *Torah Shelemah*'s corrections according to the Ginzburger manuscript and others).

Targum Yonathan identifies the new name for the country for items 1 to 15, 17, and 19 to 23 in Table 8; and the new name for the nation for items 24 to 33, and 41 to 44.

Targum Yonathan is a translation into Aramaic. We need the identification of the names, but we do not intend to check whether the Aramaic appears non-randomly as ELS's. Thus, where the Hebrew formation exists, we used it alone.

Remark 4. Here, as in Witztum *et al.* (1994), we used the grammatical orthography - *ktiv dikduki.*

A.4 The Permutations.

Here we define the permuted samples for B1, B2 and B3.

A.4.1 Samples B1 and B2.

Let π be a permutation on the set HA'. Then we define the permuted sample

$$Sample\ B1(\pi) := \bigcup_{x \in HA'} Pair\ B1(x, \pi(x)).$$

The set HA' consists of 12 elements, thus there are 12! different permutations. To calculate significance levels, we chose 999,999 random permutations π as described in Section A.7.

For Sample B2 we proceed similarly.

A.4.2 Sample B3.

For a permutation π on the set TN we define the permuted sample

$$Sample\ B3(\pi) := \bigcup_{x \in TN} Pair\ B3(x, \pi(x)).$$

The set TN consists of 68 different elements (the names שבא and חוילה appear twice), thus there are 68! different permutations. To calculate significance levels, we first chose 999,999 random permutations π as described in Section A.6. Then we did a new measurement which 999,999,999 random permutations, only for statistic P_2.

A.5 The Text.

We used the standard, generally accepted text of Genesis known as the "Textus Receptus." One widely available edition is that of the Koren Publishing Company in Jerusalem. The Koren text is precisely the same as that used by us.

A.6 The Overall Proximity Measures P_1 and P_2.

Let N be the number of word pairs (w, w') in the sample for which the corrected distance $c(w,w')$ is defined (see Section A.2 and A.3 above). Let k be the number of such word pairs (w,w') for which $c(w, w') \leq 1/5$. Define

$$P_1 := \sum_{j=k}^{N} \binom{N}{j} \left(\frac{1}{5}\right)^j \left(\frac{4}{5}\right)^{N-j}$$

To understand this definition, note that *if* the $c(w, w')$ were independent random variables that are uniformly distributed over [0,1], *then* P_1 would be the probability that at least k out of N of them are ≤ 0.2. However, we do *not* make or use any such assumptions about uniformity and independence. Thus P_1, though calibrated in probability terms, is simply an ordinal index that measures the number of word pairs in a given sample whose words are "pretty close" to each other (i.e. $c(w, w') \leq 1/5$), taking into account the size of the whole sample. It enables us to compare the overall proximity of the word pairs in different samples; specifically, in the samples arising from the different permutations π.

The statistic P_1 ignores all distances $c(w,w')$ greater than 0.2, and gives equal weight to all distances less than 0.2. For a measure that is sensitive to the actual size of the distances, we calculate the product $\prod c(w,w')$ over all word pairs (w,w') in the sample. We then define

$$P_2 := F^N (\prod c(w, w')),$$

with N as above, and

$$F^N(X) := X(1 - \ln X + \frac{(-\ln X)^2}{2!} + \cdots + \frac{(-\ln X)^{N-1}}{(N-1)!}).$$

To understand this definition, note first that if $x_1, x_2, ..., x_N$ are independent random variables that are uniformly distributed over [O, 1], then the distribution of their product $X := x_1, x_2 ... x_N$ is given by $\mathrm{Prob}(X \leq X_0) = F^N(X_0)$; this follows from (3.5) in Feller (1966), since the $-\ln x_i$ are distributed exponentially, and $-\ln X = \sum_i (-\ln x_i)$. The intuition for P_2 is then analogous to that for P_1: If the $c(w, w')$ were independent random variables that are uniformly distributed over [0,1], *then* P_2 would be the probability that the product $\prod c(w,w')$ is as small as it is, or smaller. But as before, we do not use any such uniformity, or independence assumptions. Like P_1, the statistic P_2 is calibrated in probability terms; but rather than thinking of it as a probability, one should think of it simply as an ordinal index that enables us to compare the proximity of the words in word pairs arising from different permutations.

A.7 The Randomizations.

The 999,999 random permutations for the three tests were chosen in accordance with Algorithm P of Knuth (Knuth: 1969, page 125). The pseudorandom generator required as input to this algorithm was that provided by Turbo-Pascal 5.0 of Borland Inter Inc. We used the same seed as in Witztum *et al.* (1994): 01001 10000 10011 11100 00101 00111 11.

The control text V was constructed by permuting the verses of G with a single random permutation, generated like in the previous paragraph. In this case, the seed was picked arbitrarily to be the decimal integer 10 (i.e., the binary integer 1010).

Acknowledgment.

We express our gratitude to Yaakov Rosenberg who prepared the software for the permutation test.

We thank Mr. Bernard Goldstein of London and Mr. Yaakov Aharonov of Bnei Brak for their help with computer facilities.

We express our sincere gratitude to Mishmereth Stam (Bnei Brak) for the text of the Book of Genesis on a disc.

We wish to express our thanks to Dr. Shalom Srebrenik for helpful discussions and valuable suggestions.

We thank Yaakov Orbach for help on linguistic matters.

REFERENCES

1. Cordovero, M. (1592) *Pardes Rimmonim*, Cracow.
2. *Encyclopledia Biblica* (1962), Jerusalem: Bialik Institute.
3. Even-Shoshan, A. (1981) *A New Concordance of the Bible*, Jerusalem: Kiriath Sefer.
4. Even-Shoshan, A. (1989) *A New Dictionary of the Hebrew Language*, Jerusalem: Kiriath Sefer.
5. Feller, W. (1966) *An Introduction to Probability Theory and Its Applications*, 2, New York: Wiley.
6. Hagra, (1905) *A Commentary to the Book of Job*, Jerusalem.
7. Kasher, M. M. (1929) *Torah Shelemah. Talmud-Midrashic Encyclopedia on the Pentateuch*, 1, Jerusalem: Beth Torah Shelemah.
8. Kasher, M. M. (1953) *Torah Shelemah. Talmud-Midrashic Encyclopedia on the Pentateuch*, 35, Jerusalem: Beth Torah Shelemah.
9. Knuth, D. E. (1969) *The Art of Computer Programming*, 2, Reading: Addison-Wesley.
10. Rabbenu Bahya, (1492) *A Commentary on the Pentateuch*, Naples.
11. Weissmandel, H. M. D. (1958) *Torath Hemed*. Mt. Kisco: Yeshivath Mt. Kisco.
12. Witztum, D., Rips. E. and Rosenberg, Y. (1994) Equidistant Letter Sequences in the Book of Genesis. *Stat. Science*, 9, No 3, 429 - 438.