Stylometry and the Septuagint
Applying Anthony Kenny’s Stylometric Study to the LXX

Rick Brannan
Logos Bible Software

BibleTech:2009
Bible Technologies Conference
March 27, 2009

ABSTRACT


While Kenny’s primary application of his method was in the area of authorship studies, this paper is more interested in the general style of the LXX, and not at all interested in authorship theories or assigning a ‘hand’ to different passages. For better or worse, this paper treats the LXX as a corpus, and has little interest in its relationship with the underlying Hebrew text.

Once the analysis has been detailed, some points of interest (known only when the analysis is complete as the nature of the study is exploratory) will be further explored. Areas in which the work could be further developed will also be reviewed.

1 Author email: rick@pastorelepistles.com
INTRODUCTION


However, there is a small problem. Actually, the problem is not so small, it is rather large. Kenny spends 124 pages to detail portions of his analysis of the New Testament. The New Testament (NA27 text) consists of 27 “books”, which consist of a total 260 chapters, which consist of 138,019 words. This is a lot of words. Kenny ends up with over 1000 words of the corpus represented per printed page. (138,019w / 124pg > 1000 w/pg)

By comparison, the corpus of the LXX analyzed for this paper consists of 59 books and around 450,000 words. To allay your fears, I have no intention of reading a 450 page paper here today. But this illustrates my conundrum: I have three to four times the amount of data to hand that Kenny had, and I only have 40 minutes in which to talk with you about it. Actually, it’s more like 38 minutes now, and that doesn’t account for questions.

With this in mind, my presentation today will be less about using stylometric data to explore and work through particular problems of style, and more about simply trying to get a grip on what sorts of information can be gathered and evaluated from above the word level. We will focus on two higher levels of discourse, the discourse (equivalent with the “book”) and the chapter. One is authentic (the book); the other (the chapter) is somewhat contrived but serves well as an intermediate level by which to examine how particular stylometric features are distributed throughout a given discourse. Finally, an example of Kenny’s method will be applied to a particular feature in the Pentateuch.

DIVING INTO THE DATA, FROM THE TOP DOWN

The first step in Kenny’s analysis is to count. Count everything. Count anything that might be worthy of counting. Well, that’s not completely true; Kenny’s first step is to develop a list of whatever might be worthy of counting. This is called a “feature list”. Kenny’s feature list consists of 99 items, including features like:

- Feature 1: Number of occurrences of καί
- Feature 27: Number of occurrences of εν
- Feature 36: Total number of articles
- Feature 49: Number of nouns in the accusative

You get the idea. Kenny’s method involves asking a “yes or no” question of each word token in the corpus. If the answer to the question is “yes”, then it is counted as an instance of that feature. If the answer is “no”, then it is not counted. Is the word καί? Yes. So count it as an instance of that feature.

---

2 Kenny, 123-124.
In practice this means that whatever can be counted for a particular word based on its morphology or specified lexical form (for a select group of important or frequent words) should be counted. Kenny came up with 99 items based on the categories present in the Friberg morphology. In reviewing the Logos Bible Software morphology for the LXX, I came up with a list of 109 features. This paper, however, is only concerned with a subset of that list (having to do with specific morphological criteria) that has been visualized for the corpus under analysis. The subset includes:

- Part-of-speech (book proportion)
- Aggregated noun/adjective/pronoun/article/participle case counts (by chapter)
- Aggregated noun/adjective/pronoun/article/participle number counts (by chapter)
- Aggregated noun/adjective/pronoun/article/participle gender counts (by chapter)
- Verb tense (by chapter)
- Verb voice (by chapter)
- Verb mood (by chapter)

Because there is so much data to evaluate, the process has been to start examining a few items (listed above) across the whole of the corpus to determine if anything interesting may be going on.

**Looking at Book-Level Part-of-Speech Distribution throughout the Whole Corpus**

Starting from the top down makes book-level part-of-speech distribution a good place to start. But reporting simple numbers as counts doesn’t allow for easy visual comparison of data. Instead, I developed a colored bar where each color represents percentage of each part-of-speech found in the book. With these colored bars stacked one on top of the other, differences are more readily apparent. Also, the width of each book’s bar is calculated based on the number of words in the book. So larger books have thicker bars. Because the disparity in book size is so great (e.g., compare Obadiah to Jeremiah, word-count wise) a minimum width was established.

---

3 See appendix.
This is not to be confused with the old Atari “Breakout” game:

While the bar-style table is somewhat reminiscent of the old Atari 2600 Breakout game, the visualization also gives a very rudimentary idea of what sort of part-of-speech distribution to expect in a book. It also allows one to compare the distribution by book length (width of the bars). One could also examine distribution considering overall genre (narrative vs. poetry) or overall historical grouping (Pentateuch vs. historical books).

**Looking at Chapter-Level Distributions**

Book-level proportions give a fairly blunt and, let’s face it, not too useful look at what is going on inside of a book of the LXX. Getting an idea of how the distribution ebbs and flows within a discourse (book) can give a better idea to how the author uses particular criteria throughout the discourse. So I’ve done a similar analysis for each book, with color bars representing chapters.

**Nouns and other Substantives**

**Case Distributions**

Greek, like any language, has grammatical “rules” that are regularly broken. Some of these regularly-broken rules involve the role of case within a clause. Substantives in the nominative case are typically, but not always subjects (or, like articles and adjectives, directly modifying subjects). Substantives in the accusative case are typically direct objects, and substantives using the dative case are typically indirect objects. These sorts of “rules” are valid more often than not, but there are other usages of each case.

That said, tracking the distribution of case use from chapter to chapter across a discourse can be useful. Consider the case distribution of the book of Joshua:

4 Source: http://www.videogamecritic.net/images/2600/breakaway_iv.png
It is easy to see how some chapters have more instances of one case than the other. For example, genitive usage seems rather consistent across the whole of the book, but accusatives seem to have more variation. Also, what is it about chapters 5 and 7 that require the vocative case? Why does chapter 11 have so many accusatives, proportionally? Are these differences stylistically relevant?

Before an visualization such as this was applied to the book, one may have had an intuitive notion of such frequencies and infrequencies. But constructing the picture provides a consistent basis by which to examine a slice of data (here morphological case) across a large portion of data (here the book of Joshua).

The balance of chapter distribution visualizations are all variations on this same theme.

**Number Distribution**

Here is an example of distribution of number within Joshua:

Singular is the most common, but some chapters have far more plural items than singular. This sort of information could be valuable when doing a participant analysis within a discourse.

**Gender Distribution**

Gender in this context is a grammatical quality that has little to do with sex. Here again are distributions within Joshua:
Even with a quality such as morphological gender, some differences are evident. Masculine appears to be most common, but in some chapters the feminine is more common than the masculine. This could indicate particular subject matter, and be worthy of further examination.

**Verbs**

Verbs are similarly analyzed, with proportions of tense, voice and mood plotted through all chapters. The sample book used below is Leviticus.

**Tense Distribution**

Tense in Leviticus is interesting because of the wide fluctuations in tense between chapters. In most modern Koine/Hellenistic Greek Grammars the Aorist tense is thought of as the default tense in narrative;\(^5\) when the discourse diverges from the aorist, the use is marked as regards the discourse. That is, the author uses a different tense to achieve a different task.

Tense in Leviticus makes an interesting case study because of the concentration of future tense verbs in most chapters.

The abundance of aorist verbs with the juxtaposed paucity of future tense verbs stand outs in chapters 8-10. Is this usage of the future tense typical? Is there any significance to it?

**Voice Distribution**

Similar disparities are visible when examining per-chapter proportion of voice.

![Voice in Leviticus (proportion, by chapter)](chart)

Is there significance to the abundance of active voice verbs in chapters 1 and 3 (and chapters 8-9, interestingly enough)? And what about the seemingly large number of passives in chapter 6?

**Mood Distribution**

As well, an overview of distribution of mood can bring other questions to light.

![Mood in Leviticus (proportion, by chapter)](chart)

Indicative is, understandably, the most common as it is the default mood in narrative. But what of the high relative proportions of imperatives in chapters 9, 20 and 24? What about participle usage? And what of the range of subjunctives?

**Some Further Thoughts on Chapter-Level Distributions**

While visualizations such as these provide a wealth of information, and while they can open our eyes to some grammatical trends in the books in question, they don’t really tell us much about style. They give us no real standard of comparison. The unit analyzed (chapter) is variable; some are large and others are small. While we may know that in one book, 20% of the verbs used are in the future tense; and that within that book, there are chapters that have more than 20% and less than 20% of verbs in future tense, this does

---

not help us in knowing what is expected and thus “normal”, and what is outside of the statistical bound of normalcy. Looking at chapters gives us a better idea than looking at the whole book, but what is really needed is a smaller unit of analysis, consistent between books, and an approach to quantify what is recorded stylistically. Anthony Kenny gives us this approach.

Looking at Distributions in Consistently-sized Chunks: Future Tense in Leviticus

Anthony Kenny summarizes where we’ve been (book and chapter distributions) and points the way that we should go:

If we are to use statistical methods to study and compare the style of different portions of the New Testament it is obviously not sufficient simply to record raw data. It is not enough to know that, say, St John's Gospel contains 271 instances of οτι, or that the Epistle to the Ephesians contains a sentence which is 141 words long. Clearly, we need to set these numbers in their correct context and measure them in some way against the totality of the data available for the relevant text. 7

Kenny’s method of setting “numbers in their correct context and measuring them in some way against the totality of the data available for the relevant text” can be summarized as follows:

1. Break the text into chunks of 50 consecutive words, starting from the first word in the book.
2. Count occurrences of each feature in each chunk.
3. Analyze the result. For each feature:
   b. Use the binomial distribution to generate a theoretical expected distribution of the feature in the text.
   c. Compare the expected (theoretical) distribution of the feature to the actual distribution using a chi square test.
   d. Determine if the results indicate a significant or insignificant difference from the expected (theoretical) distribution.

An Example of Kenny’s Method Applied: Nouns in Leviticus

An example application is necessary to understand the method before using it to dig into a particular feature, such as future tense use in Leviticus. We will use nouns in Leviticus as an example.

Counting Feature Instances

To utilize Kenny’s method, the first step is to break the book of Leviticus into 50 word chunks, making 381 complete chunks, plus 32 words left over, from 19082 words total.

7 Kenny, 17.
8 See Appendix for list of features counted in this study.
Within each 50-word chunk, feature occurrences (here nouns) must be counted. For Kenny, this is feature 46, the number of nouns. This maps to feature 37 on the feature list in the Appendix. Counting total instances in each chunk and then listing by totals gives us:

<table>
<thead>
<tr>
<th>Instances in Group</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>17</td>
</tr>
<tr>
<td>8</td>
<td>24</td>
</tr>
<tr>
<td>9</td>
<td>30</td>
</tr>
<tr>
<td>10</td>
<td>44</td>
</tr>
<tr>
<td>11</td>
<td>38</td>
</tr>
<tr>
<td>12</td>
<td>61</td>
</tr>
<tr>
<td>13</td>
<td>40</td>
</tr>
<tr>
<td>14</td>
<td>46</td>
</tr>
<tr>
<td>15</td>
<td>27</td>
</tr>
<tr>
<td>16</td>
<td>24</td>
</tr>
<tr>
<td>17</td>
<td>12</td>
</tr>
<tr>
<td>18</td>
<td>3</td>
</tr>
<tr>
<td>19</td>
<td>5</td>
</tr>
<tr>
<td>20</td>
<td>3</td>
</tr>
<tr>
<td>21</td>
<td>0</td>
</tr>
<tr>
<td>22</td>
<td>1</td>
</tr>
</tbody>
</table>

In other words, there is one instance of a fifty-word chunk with only three nouns; two chunks with five nouns, etc. One chunk has 22 nouns.

---

9 Kenny, 124.
This table shows a few things. First, even though just over 24% of the words in the book are nouns, this doesn’t mean that there will be 12 nouns in every fifty-word chunk. Instead, there are some chunks that have more, and some chunks that have less. The actual range is from three nouns to 22 nouns. On this phenomenon, Kenny comments (regarding nouns in Philippians):

> If 31% of the whole Epistle is nouns, then the average number of nouns in a fifty-word section must be 15.5. But sections vary wildly from this, with as many as twenty-six in the first section, and as few as seven in the antepenultimate one. Does not this show the follow of trying to pin down the free creative spirit of a writer within the soulless confines of a statistical distribution?¹⁰

**Calculating Expected Feature Instances**

The key, according to Kenny, is not to apply a straight proportion (in our case, ~24%) across the board as the expectation. Instead, one must use the binomial distribution as a measure to approximate the expected distribution, and then compare reality with the expectation.

The binomial distribution allows, given a proportion, prediction of how many chunks will contain a certain number of feature instances. The process is two steps. First is to use the binomial distribution¹¹ to generate a probability or likelihood of a fifty-word chunk having \( n \) instances. Second is to apply that likelihood to the current set of chunks using simple multiplication. Here is the above chart with a new column added, reflecting the likelihood of each grouping:

<table>
<thead>
<tr>
<th>Instances in Group</th>
<th>Likelihood</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1.04E-06</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1.65E-05</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0.000128</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0.00065</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>0.002422</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0.007071</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>0.016825</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>0.033554</td>
<td>17</td>
</tr>
<tr>
<td>8</td>
<td>0.057221</td>
<td>24</td>
</tr>
<tr>
<td>9</td>
<td>0.084722</td>
<td>30</td>
</tr>
<tr>
<td>10</td>
<td>0.110208</td>
<td>44</td>
</tr>
<tr>
<td>11</td>
<td>0.127148</td>
<td>38</td>
</tr>
<tr>
<td>12</td>
<td>0.131107</td>
<td>61</td>
</tr>
</tbody>
</table>

---

¹⁰ Kenny, 21.

¹¹ All calculation done using Microsoft Excel 2007 unless otherwise cited.
Thus, using the binomial distribution, there is approximately a 0.06% likelihood that a fifty-word chunk will have three nouns.

The next step is to calculate the expected number of chunk instances based on the likelihood. This is done with simple multiplication, multiplying the likelihood by the number of fifty-word groups. The completed table is below:

<table>
<thead>
<tr>
<th>Instances in Group</th>
<th>Likelihood</th>
<th>Expected</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1.04E-06</td>
<td>0.000397</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1.65E-05</td>
<td>0.006291</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0.000128</td>
<td>0.048898</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0.00065</td>
<td>0.248224</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>0.002422</td>
<td>0.925365</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0.007071</td>
<td>2.701048</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>0.016825</td>
<td>6.427249</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>0.033554</td>
<td>12.81773</td>
<td>17</td>
</tr>
<tr>
<td>8</td>
<td>0.057221</td>
<td>21.85855</td>
<td>24</td>
</tr>
<tr>
<td>9</td>
<td>0.084722</td>
<td>32.36381</td>
<td>30</td>
</tr>
<tr>
<td>10</td>
<td>0.110208</td>
<td>42.09931</td>
<td>44</td>
</tr>
<tr>
<td>11</td>
<td>0.127148</td>
<td>48.57064</td>
<td>38</td>
</tr>
<tr>
<td>12</td>
<td>0.131107</td>
<td>50.08282</td>
<td>61</td>
</tr>
<tr>
<td>13</td>
<td>0.12159</td>
<td>46.44731</td>
<td>40</td>
</tr>
</tbody>
</table>
Comparing Expected Instances to Actual Instances

Now, we actually have a basis by which to compare what occurs in the text with what is expected in the text. And there is some discrepancy, as the below graph\textsuperscript{12} shows:

![Graph comparing expected to actual instances](image)

The horizontal axis is count per 50 word group.\textsuperscript{13} The vertical axis is count of groups. The above graph shows a few places where expectations are not in line with what is witnessed in the text. But are the differences significant?

---

\textsuperscript{12} Graph created using MSExcel 2007.

\textsuperscript{13} Plus 1, thanks to MSExcel. The automated chart function did not (easily) allow me to set the values for this baseline and assumed a starting value of one instead of zero.
**Determining Significance**

Comparing an expected set of data to data that actually occurs is a common statistical practice. Following Kenny’s footsteps, the chi-squared test will be used.\(^\text{14}\) Kenny explains this as follows, in his explanation of part of speech distribution in Philippians:

> It will be seen that the actual results do not differ greatly from the expected results, and that the actual distribution is reasonably close to the expected one. But there are undoubtedly differences between the two distributions, and we need to know whether to attach importance to them. In order to discover this, we have to apply another statistical test, the chi-squared test. This is a very generally useful test to determine whether the differences between a set of predicted and observed results are statistically significant, or whether they are simply the kind of differences to be expected between one hand and another fairly dealt from the same pack of cards. One common use of the chi-squared test is, as here, to test whether a set of observed results is a good fit to a theoretical distribution such as the binomial.\(^\text{15}\)

Microsoft Excel provides a function, CHITEST, which will complete the necessary analysis for a set of expected and actual values. It returns a value for something called “the two-tailed P value”.\(^\text{16}\) With this measure (a decimal between zero and one), the lower the number, the more independent the datasets are.

However, before proceeding, the numbers need to be massaged a little. For items where expected or actual values are less than five, some consolidation needs to take place.

<table>
<thead>
<tr>
<th>Instances in Group</th>
<th>Expected</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-6</td>
<td>10.35747</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>12.81773</td>
<td>17</td>
</tr>
<tr>
<td>8</td>
<td>21.85855</td>
<td>24</td>
</tr>
<tr>
<td>9</td>
<td>32.36381</td>
<td>30</td>
</tr>
<tr>
<td>10</td>
<td>42.09931</td>
<td>44</td>
</tr>
<tr>
<td>11</td>
<td>48.57064</td>
<td>38</td>
</tr>
<tr>
<td>12</td>
<td>50.08282</td>
<td>61</td>
</tr>
<tr>
<td>13</td>
<td>46.44731</td>
<td>40</td>
</tr>
<tr>
<td>14</td>
<td>38.94626</td>
<td>46</td>
</tr>
<tr>
<td>15</td>
<td>29.65572</td>
<td>27</td>
</tr>
<tr>
<td>16</td>
<td>20.58203</td>
<td>24</td>
</tr>
<tr>
<td>17</td>
<td>13.0602</td>
<td>12</td>
</tr>
<tr>
<td>18</td>
<td>7.596659</td>
<td>3</td>
</tr>
</tbody>
</table>

\(^{14}\) Kenny, 22.

\(^{15}\) Kenny, 22.

The results of the CHITEST function (the “two-tailed P value”) for this set of data is 0.386429612. However, Excel doesn’t tell us interpret the number. So I ran the same set of numbers (with decimal values in expected values rounded to nearest integer) through a different statistical calculator, which also provides some explanation of the result:

**P value and statistical significance:**

Chi squared equals 13.963 with 13 degrees of freedom.
The two-tailed P value is less than 0.3764
By conventional criteria, this difference is considered to be not statistically significant.

The P value answers this question: If the theory that generated the expected values were correct, what is the probability of observing such a large discrepancy (or larger) between observed and expected values? A small P value is evidence that the data are not sampled from the distribution you expected.

So, what does all of this mean?

Recall that we used the binomial distribution to estimate the expected values. The most general statement that can be made is that the binomial distribution of nouns and the actual distribution of nouns are compatible; it is capable of accurately modeling the distribution of nouns in Leviticus. In other words, the distribution of nouns appears “normal”. As Kenny would put it, “… they are simply the kind of differences to be expected between one hand and another fairly dealt from the same pack of cards.”

The purpose of all of this is to establish some sort of a measure by which to compare features. If the binomial distribution can be that measure, then one can determine where there are statistically significant deviations from that measure; this in turn may give insight as to when uses a given feature outside the bounds of the normal measure.

Is the binomial distribution an adequate model to use to determine this measure? Kenny writes regarding the distribution of words and parts of speech in Philippians:

In no case [of part-of-speech distribution in Philippians] is the value of chi-square significant, and we can conclude that all the parts of speech are binomially distributed in the Epistle to the Philippians.

One may wonder whether individual words, like parts of speech, are also binomially distributed. A partial is already in the tables [referenced on the same page]: for the definite article, while forming a class on its own, as a determiner, is also of course a single vocabulary item, the commonest word, indeed, in the Epistle. …

However, as [the probability of feature occurrence] drops, and equally if [the size of group, at 50 in examples] drops so that the text is divided into shorter and shorter sections, the binomial distribution

---

17 GraphPad QuickCalcs: chi squared calculator. Online: [http://graphpad.com/quickcalcs/chisquared1.cfm](http://graphpad.com/quickcalcs/chisquared1.cfm). Accessed March 18, 2009. Further results formatted in this manner are results from the same chi squared calculator. The difference in two-tailed P value calculation between Excel and GraphPad’s chi squared calculator is due to GraphPad’s restriction that the column totals of compared datasets match exactly. Excel uses the decimal values; GraphPad uses integers, at least for this particular calculator.

18 Kenny, 24.
cannot be expected to provide such a good fit to the data, and this for reasons both of general statistical theory and of the nature of the literary subject matter.\textsuperscript{19}

Thus, the appropriateness of using the binomial distribution to produce an expected distribution of a particular feature is contingent upon a large enough sample, and a large enough occurrence of the feature within the sample. Kenny reports that for Philippians, the binomial distribution can serve as a method to approximate part-of-speech distribution. He relies on similar analysis for the rest of the New Testament, so we may deduce it accurately models the rest of the New Testament corpus as well.

Using Kenny’s Method on Future Tense Verbs in Leviticus

Because the distribution of the future tense in Leviticus seems extraordinary on the surface, it should be analyzed further, and will serve as our test case. As a reminder, here is the graph of distribution of verb tense by chapter in Leviticus:

The red portion represents the future tense; the light blue portion represents the aorist tense.

On occurrences per 50 word chunk, the expected and actual numbers are below:

\begin{table}[h]
\begin{tabular}{|c|c|c|}
\hline
\textbf{ Instances in Group} & \textbf{ Expected} & \textbf{ Actual} \\
\hline
0 & 12.00017 & 38 \\
1 & 42.99734 & 38 \\
2 & 75.49044 & 52 \\
3 & 86.55581 & 68 \\
4 & 72.88168 & 75 \\
5 & 48.04968 & 59 \\
6 & 25.82475 & 35 \\
7 & 11.63255 & 12 \\
\hline
\end{tabular}
\end{table}

\textsuperscript{19} Kenny, 23-24.
There are large discrepancies between “expected” and “actual”; the largest are in bold. Instead of finding 12 chunks with no instances, there are 38. Instead of finding 75 chunks with two instances, there are 52. And instead of finding 86 chunks with three instances, there are 68. These are significant differences.

Below is the graph comparing the expected distribution (using the binomial distribution) to the actual distribution of future tense verbs in Leviticus:

It looks similar, but the basic shape is shifted over (to the right, so higher concentrations of feature occurrences) and down (so fewer groups of those higher-concentrated items). The differences described above are visible now. Essentially, instead of the evenness of the binomial distribution, the actual distribution has significantly more groups with zero feature instances, vastly fewer lower-count instances (1-3 instances per group, the left side of the curve) and more higher-count instances (5-6 instances per group, on the right side of the curve).

Put in more plain language, this means that future tense verbs are concentrated in pockets more highly than the binomial distribution predicts, and this seems to agree with what was seen in the chapter-level tense distribution graphic.

But is this a significant difference? The two-tailed P value, according to Excel, is 7.16801E-13, which is a very small number (much less than 0.0001, one ten-thousandth). Here is a different presentation of the numbers:

<table>
<thead>
<tr>
<th>Chunk Size</th>
<th>Expected</th>
<th>Actual</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>8</td>
<td>38</td>
</tr>
<tr>
<td>1</td>
<td>9</td>
<td>52</td>
</tr>
<tr>
<td>2</td>
<td>86</td>
<td>68</td>
</tr>
</tbody>
</table>

P value and statistical significance:
Chi squared equals 73.975 with 8 degrees of freedom.
The two-tailed P value is less than 0.0001
By conventional criteria, this difference is considered to be extremely statistically significant.20

While GraphPad reports that the differences are “extremely statistically significant”, this does need some explanation. This means that if the binomial distribution is an accurate method to describe the distribution of the future tense, the dataset that we see in Leviticus varies enough to say that it cannot be accounted for by the binomial distribution. That is, they are not “fairly dealt from the same pack of cards”. There must be some other explanation.

The “other explanation” is not determined by the statistical output. It could be that the assumption of the binomial distribution as appropriate is wrong. It could be that, as mentioned above,21 the probability of the feature we are measuring is too small. It could be that the translational nature of the text plus genre and subject matter combine in such a way as to require a more frequent use of the future tense. Whatever the case, the binomial distribution does not accurately model the distribution witnessed in Leviticus.

**Checking Future Tense Distributions in Other Books**

One way to tell if the binomial distribution is an appropriate measure is to test the same feature in other books and examine the results. Here are the balance of the books in the Pentateuch.

**Genesis**

![Graph showing future tense distributions in Genesis](image)

**P value and statistical significance:**

Chi squared equals 288.730 with 4 degrees of freedom.
The two-tailed P value is less than 0.0001
By conventional criteria, this difference is considered to be extremely statistically significant.

---

20 GraphPad QuickCalcs. Chi square calculator.

21 When we quoted Kenny, pp. 23-24.
Excel’s CHITEST function calculates the two-tailed P value at 4.41601E-63, much smaller than the number for Leviticus. Genesis has many more groups with no instances (400 vs >300 expected) with a slight bump at groups containing four future verbs.

**Exodus**

![Graph showing expected vs actual counts](image)

**P value and statistical significance:**

Chi squared equals 545.545 with 6 degrees of freedom.
The two-tailed P value is less than 0.0001
By conventional criteria, this difference is considered to be extremely statistically significant.

Excel’s CHITEST function calculates the two-tailed P value at 1.7465E-113. This number is much smaller than even Genesis; it has 100 more decimal places than the number for Leviticus. It is infinitesimally small, again with many more groups containing zero future tense verbs, but with more groups containing higher counts (five, six, seven and eight count groups). As with the others, this seems to show that when the future tense occurs, it occurs in groups, not sporadically.

**Numbers**
P value and statistical significance:

Chi squared equals 428.098 with 5 degrees of freedom.
The two-tailed P value is less than 0.0001
By conventional criteria, this difference is considered to be extremely statistically significant.

Excel’s CHITEST function calculates the two-tailed P value at 9.3867E-91. Another incredibly small number, though larger than the number for Exodus. But the pattern is the same: More zero-count groups, fewer low-count groups, and more higher-count groups than the binomial distribution models.

Deuteronomy

P value and statistical significance:

Chi squared equals 147.652 with 6 degrees of freedom.
The two-tailed P value is less than 0.0001
By conventional criteria, this difference is considered to be extremely statistically significant.
Excel’s CHITEST function calculates the two-tailed P value at 3.90394E-29. This as well is a low number, but it is closest to the two-tailed P value for Leviticus. And the same sorts of differences are evident: more zero-count groups, fewer low-count groups, and here slightly more higher-count groups.

Conclusions

The first conclusion appears obvious: verbs in the future tense do not follow the binomial distribution within the Pentateuch. Interestingly enough, the actual distribution looks most similar to the expected distribution in Leviticus, the book containing the feature distribution that prompted this examination (it’s just shifted a bit) and also in Genesis. The other three books compared (Exodus, Numbers, Deuteronomy) are least like the expected distribution and return incredibly, infinitesimally low numbers for the two-tailed P value. Leviticus actually has the highest two-tailed P value of the lot.

But it may not actually be so obvious. It may have more to do with the probability of occurrence of verbs in the future tense. Recall Kenny’s final comment in his section on distribution of words and parts of speech:

However, as [probability of feature] drops … the binomial distribution cannot be expected to provide such a good fit to the data, and this for reasons both of general statistical theory and of the nature of the literary subject matter.22

Perhaps the probability of future tense verbs occurring is too low to use Kenny’s method. But the probability of a word being a future tense verb in Leviticus is around 7%, which is well above values in the worked example that Kenny includes in his book.23 Kenny gives no insight on determining when a feature probability is too low, so we can only guess at this point.

My own guess is that not that occurrences are too low, it is that the binomial distribution does not accurately model the expected distribution of future tense verbs. The distribution of future tense verbs in Leviticus can be further explained by examining the subject matter and content of chapters 1-7, 8-10, and 11-27, along with considering the necessities of translating Hebrew into Greek:

Future tense verbs in Greek are generally going to be translations of the imperfect aspect (without waw consecutive), whereas the imperfect waw-consecutive is usually translated as an aorist. There are very few imperfects without a waw-consecutive in Lev 8-10, thus the relative scarcity of future tenses in that section.

The reason for the radical shift in tenses/aspects used in Lev 8-10 is that the passages before chapter 8 are law code material (you shall kill animals in such and such a way) and then 8-10 consists of narrative material (telling the story of the first consecration of the the priests—Aaron and his sons) and then in chapter 11, we’re back to law codes (clean and unclean animals).24

If that is the case, what good is all of this?

Whether accurate or not, using the binomial distribution as a standard of measure allows us to see when things are not evenly distributed. It allows us to see where concentrations of a feature occur, and where lack of a feature occurs. These ebbs and flows of a given feature can provide insight to grammatical tools

---

22 Kenny, 24.


that a given author uses in a document. They can help us ask better questions about grammatical phenomenon that actually occur in the text.

In this case, the “pockets” or concentrations of future tense verbs in Leviticus can be explained by appealing to genre and translational issues (as was done above). Perhaps even more interesting to examine, in light of the distributional analysis, would be the isolated instances of the future tense.

Examining stylometric features in this way can bring these bumps and wiggles to light, and help the person examining the text to ask informed, focused questions on how an author communicates information to the reader/hearer. Stylometry can be less about “who wrote it?” and more about “what was written?” And that has surely got to be a good thing, valuable for understanding and exegesis.
BIBLIOGRAPHY


APPENDIX: LIST OF FEATURES

Features shared with Kenny

1. Number of occurrences of και
2. Number of occurrences of ειναι (so, ειμι)
3. Number of Conjunctions (C)
4. Number of occurrences of αλλα
5. Number of occurrences of δε
6. Number of occurrences of γαρ
7. Number of occurrences of ουν
8. Number of occurrences of εαν
9. Number of occurrences of ει
10. Number of occurrences of ου
11. Number of occurrences of μη
12. Number of occurrences of ινα
13. Number of occurrences of όπως or ωστε
14. Number of occurrences of στι
15. Number of occurrences of ως
16. Number of Particles (T)
17. Number of Prepositions (P)
18. Number of occurrences of εν
19. Number of occurrences of εις
20. Number of occurrences of εκ
21. Number of occurrences of απο
22. Number of occurrences of επι
23. Number of occurrences of προς
24. Number of occurrences of δια
25. Number of occurrences of κατα
26. Number of occurrences of μετα
27. Number of Articles (D)
28. Number of occurrences of nominative articles
29. Number of occurrences of dative articles
30. Number of occurrences of genitive articles
31. Number of occurrences of accusative articles
32. Number of occurrences of masculine articles
33. Number of occurrences of feminine articles
34. Number of occurrences of neuter articles
35. Number of occurrences of singular articles
36. Number of occurrences of plural articles
37. Number of Nouns (N)
38. Number of occurrences of nominative nouns
39. Number of occurrences of dative nouns
40. Number of occurrences of genitive nouns
41. Number of occurrences of accusative nouns
42. Number of occurrences of masculine nouns
43. Number of occurrences of feminine nouns
44. Number of occurrences of neuter nouns
45. Number of occurrences of singular nouns
46. Number of occurrences of plural nouns
47. Number of occurrences of θεος
48. Number of occurrences of third person pronouns
49. Number of occurrences of third person pronouns in the nominative
50. Number of occurrences of third person pronouns in the dative
51. Number of occurrences of third person pronouns in the genitive
52. Number of occurrences of third person pronouns in the accusative
53. Number of occurrences of third person pronouns in the masculine
54. Number of occurrences of third person pronouns in the feminine
55. Number of occurrences of third person pronouns in the neuter
56. Number of occurrences of third person pronouns in the singular
57. Number of occurrences of third person pronouns in the plural
58. Number of occurrences of αὐτός
59. Number of Adjectives (J)
60. Number of occurrences of nominative adjectives
61. Number of occurrences of dative adjectives
62. Number of occurrences of genitive adjectives
63. Number of occurrences of accusative adjectives
64. Number of occurrences of πᾶς
65. Number of Adverbs (B)
66. Number of Verbs (V)
67. Number of occurrences of verbs in the subjunctive mood
68. Number of occurrences of verbs in the optative mood
69. Number of occurrences of verbs in the imperative mood
70. Number of occurrences of verbs in the infinitive mood
71. Number of occurrences of verbs in the third-person singular indicative
72. Number of occurrences of verbs in the present tense
73. Number of occurrences of verbs in the future tense
74. Number of occurrences of verbs in the imperfect tense
75. Number of occurrences of verbs in the aorist tense
76. Number of occurrences of verbs in the perfect tense
77. Number of occurrences of verbs in the pluperfect tense
78. Number of occurrences of verbs in the active voice
79. Number of occurrences of verbs in the middle voice
80. Number of occurrences of verbs in the passive voice
81. Number of occurrences of λέγειν (so, λέγω)

**Because of differences with Kenny in classification of voice**

82. Number of occurrences of third-person singular indicative verbs in the either-middle-or-passive voice

**Because of differences with Kenny in classification of Participles**

83. Number of occurrences of verbs in the participle mood
84. Number of occurrences of participles in the nominative
85. Number of occurrences of participles in the dative
86. Number of occurrences of participles in the genitive
87. Number of occurrences of participles in the accusative
88. Number of occurrences of participles in the masculine
89. Number of occurrences of participles in the feminine
90. Number of occurrences of participles in the neuter
91. Number of occurrences of participles in the singular
92. Number of occurrences of participles in the plural

**Proper Nouns**

93. Number of occurrences of Proper Nouns (test: when lemma is upper-cased)

**Interjections**

94. Number of occurrences of Interjections (I)

**Vocative nouns/adjectives/articles**

95. Number of occurrences of vocative articles
96. Number of occurrences of vocative nouns
97. Number of occurrences of vocative adjectives

**Other Pronoun info**

98. Number of occurrences of Pronouns (R)
99. Number of occurrences of Relative Pronouns
100. Number of occurrences of Reciprocal Pronouns
101. Number of occurrences of Demonstrative Pronouns
102. Number of occurrences of Correlative Pronouns
103. Number of occurrences of Interrogative Pronouns
104. Number of occurrences of Indefinite Pronouns
105. Number of occurrences of Reflexive Pronouns
106. Number of occurrences of Possessive Pronouns
107. Number of occurrences of Personal Pronouns (or is this caught with αὐτοῦ?)

**Comparative/Superlative adjectives**

108. Number of occurrences of comparative adjectives
109. Number of occurrences of superlative adjectives