Cambridge Semitic Languages and Cultures

Studies in Semitic Vocalisation and Reading Traditions

Edited by Aaron D. Hornkohl and Geoffrey Khan

UNIVERSITY OF CAMBRIDGE
Faculty of Asian and Middle Eastern Studies

https://www.openbookpublishers.com

Updated digital material and resources associated with this volume are available at https://doi.org/10.11647/OBP.0207#resources

Semitic Languages and Cultures 3.

ISSN (print): 2632-6906
ISSN (digital): 2632-6914

ISBN Paperback: 978-1-78374-935-5
ISBN Hardback: 978-1-78374-936-2
ISBN Digital (PDF): 978-1-78374-937-9
DOI: 10.11647/OBP.0207

Cover image: Detail from a bilingual Latin-Punic inscription at the theatre at Lepcis Magna, IRT 321 (accessed from https://it.wikipedia.org/wiki/File:Inscription_Theatre_Leptis_Magna_Libya.JPG). Leaf of a Syriac prayer book with Western vocalisation signs (source: Wikimedia Commons). Leaf of an Abbasid-era Qurʾān (vv. 64.11–12) with red, yellow, and green vocalisation dots (source: Wikimedia Commons). Genizah fragment of the Hebrew Bible (Gen. 11–12, Cambridge University Library T-S A1.56; courtesy of the Syndics of Cambridge University Library). Genizah fragment of a Karaite transcription of the Hebrew Bible in Arabic script (Num. 14.22–24, 40–42, Cambridge University Library T-S Ar. 52.242; courtesy of the Syndics of Cambridge University Library). Greek transcription of the Hebrew for Ps. 22.2a in Matt. 27.46 as found in Codex Bezae (fol. 99v; courtesy of the Syndics of Cambridge University Library).

Cover design: Anna Gatti

AN EXPLORATORY TYPOLOGY OF NEAR-MODEL AND NON-STANDARD TIBERIAN TORAH MANUSCRIPTS FROM THE CAIRO GENIZAH

Estara Arrant

1.0. Introduction[1]

The present study is a codicological and linguistic classification of 296 Torah codices in the Genizah collections of Cambridge University Library that have nearly all of the characteristics of 'model' codices[2] and that have standard and non-standard Tiberian vocalisation patterns. Such a study is warranted due to multiple gaps in modern scholarship on the codicology and vocalisation of the Hebrew Bible.

In previous scholarship in the field, attention has been focused on the most codicologically-sophisticated manuscripts.

[1] I wish to thank Prof. Geoffrey Khan for his support and comments; Nick Posegay for proofreading; Dr David Wright and Prof. Andrew Lang for their guidance and support with the statistical analysis; and Prof. Judith Olszowy-Schlanger for her assistance with the palaeography.

[2] These have been termed in scholarship 'complete' Bibles (Yeivin 1980, 11–12) or 'great' Bibles (Sirat 2002, 42–43).

https://doi.org/10.11647/OBP.0207.12

There has not been sufficient differentiation and study of Bibles that are sophisticated, but lack the full range of the features associated with exemplar manuscripts, such as Codex Leningradensis.[3] In previous scholarship, descriptions of 'model' codices generalised specific feature groupings that, in fact, appear to be distinct from each other, hiding important differentiation in manuscript features. For example, Yeivin states:

> The majority of older texts and Geniza fragments are beautifully written and "complete" (that is, masoretic notes and vowel and accent signs were systematically added). They were written on parchment, with great care taken over the forms of the letters and over corrections, and they contain the Mm, Mp, and vowel and accent signs. They were written with two or three columns to a page.[4]

In this article I introduce a new category of Torah codex: the 'near-model' codex, and I show how the different feature patterns in this type of codex fall into statistically-verifiable subtypes. Near-model codices have nearly all, but not the complete range, of the codicological and textual features that exemplar Tiberian Bibles have. Because none of these exemplar codices have fewer than three columns, I question Yeivin's grouping two-column manuscripts with the most complete, model Bibles, and I consider two-column codices with masoretic notes, vocalisation, and cantillation to be near-model. Moreover, there are many three-

[3] By exemplar, I mean specifically specimens such as Codex Leningradensis, the Aleppo Codex and the Cairo Codex of the Prophets.

[4] Yeivin (1980, 11).

column manuscripts that fall just shy of the ‘complete’ criteria that Yeivin lists above. These I also consider near-model and show to be statistically distinct from their two-column peers.

Within all of the Torah manuscripts that have Tiberian vocalisation there is a substantial group of manuscripts that use Tiberian vowels in non-standard ways. There have been some studies of this type of Tiberian vocalisation, which is referred to by a variety of terms, the most common being ‘Palestino-Tiberian’ vocalisation.[5] In such studies, however, there has not been sufficient attention on the diversity of non-standard vocalisation patterns that exist in Genizah manuscripts. In this article I show that there were many non-standard Tiberian (hereafter, NST)[6] patterns, and I delineate an exploratory typology of these patterns in Genizah Torah manuscripts using statistical methods.

[5] The best literature reviews of this subject are found in Fassberg (1991, 55); Saenz-Badillos (2008, 92–94); Blapp (2017, 8–32); Khan (2017, 265–266). This kind of vocalisation is generally characterised in scholarship by an ‘extended’ use of *dagesh* and *rafe*, the vowel interchanges of *pataḥ/qameṣ* and *segol/ṣere*, and the non-standard placement of *shewa* and *ḥaṭef* vowels.

[6] Blapp (2017) was the first to introduce the term ‘non-standard Tiberian’ (or NST) outside of the Davis-Outhwaite catalogues. I follow Blapp here in using this term to delineate any pattern of deviation from the standard Tiberian (ST) of Codex Leningradensis that uses Tiberian vowel signs.

Another gap in scholarship on the Hebrew Bible that this study addresses is the lack of communication between codicological and textual studies on manuscripts.[7] In preliminary case-studies of the corpus I observed that not only do there appear to be sub-types of NST, but that various codicological features present in near-model codices also appear to be arranged into definite subtypal patterns. Moreover, it seemed that NST subtypes tended to correlate with these codicological subtypes. The aim of this study is to map NST diversity onto near-model Torah codicology in order to demonstrate (statistically) that the correspondence is not completely random.

1.1. Terminology, Structure, and Hypotheses

The key descriptors of codices that I am using in this paper are as follows:

- 'Model Codex': these codices look identical in style to exemplar Tiberian Bible codices such as Codex Leningradensis. They have the following combination of features: (1) a parchment base; (2) three columns; (3) a standard Tiberian (hereafter, ST) text; (4) full Masoretic notes—both Masorah Parva and Masorah Magna.

[7] Yeivin (1980, 11–12) mentions codicology briefly in his exploration of the development of the Tiberian Masorah and Diez-Macho (1971, 91–92) attempts a codicological typology of paper Bibles. These attempts to synthesise codicology and textual features are, however, limited in scope.

- ‘Near-Model’ Codex: these codices nearly attain the status of ‘model’, as defined above, except that the full four-part pattern is not present. For example, an otherwise model manuscript may lack full Masoretic notes, or may only have two columns instead of three. Manuscripts with NST automatically are considered ‘near-model’ for purposes of this study, but there are a substantial number of NST Torah codices that have all of the other features of a model codex.[8]

This fuller study of 296 fragments is built upon observations from preliminary case studies on 150 of these Genizah fragments. These specific observations have determined the structure of the study. Because none of the exemplar Bibles have two columns, it seemed appropriate to label two-column parchment Torah copies with full Masorah and vocalisation as near-model. It is not assumed, however, that these are homogeneous with three-column near-model Bibles present in the corpus, and so the study tests them separately to see if there is a statistically-verifiable difference.

Another critical factor indicated by preliminary observations regards Masoretic notes. For near-model Bibles, two-column parchment manuscripts without Masorah tend to vary widely and contain many poorly-made specimens. However, three-column

[8] Many of them are visually indistinguishable in style from exemplar manuscripts, and are set apart only by deviations in their vocalisation patterns. This seems to suggest that NST was part and parcel of sophisticated Bible codex production in the main Genizah period (ninth–twelfth centuries CE).

parchment manuscripts without Masoretic notes still retained a high degree of careful execution. It seems, therefore, that greater column numbers can be associated with a higher level of codicological sophistication, but this is not the case with the lack of Masoretic notes. Lack of Masoretic notes is not a sophisticating factor for three-column Torahs. It is, however, a major de-sophisticating factor for two-column Torahs.[9]

The present research is guided by two hypotheses that are tested through statistical, codicological, and linguistic analysis:

1. Near-model Torah parchment manuscripts with two or three columns in the Genizah have distinguishable patterns in their codicological features that indicate the presence of sub-groups in the manuscript corpus. Moreover, column number is a major factor in distinguishing these sub-groups, because nearly-model manuscripts with two columns are codicologically distinct from nearly-model manuscripts with three columns.
2. There are statistically distinguishable patterns in the NST vocalisation of these manuscripts, indicating sub-groups of NST vocalisation. These patterns can be linguistically validated. Moreover, these patterns tend to correlate with the codicological patterns of hypothesis 1.

The findings can be summarised as follows: first, a tentative, yet statistically-sound, typology of near-model manuscripts

[9] There is not space here to analyse the large population of two-column parchment codices without Masoretic notes; they are addressed in my PhD thesis.

can be established and subtypes within this typology can be identified. Second, NST is not a monolithic phenomenon, but contains significant subtypes. These subtypes reflect regional patterns of scribal activity comprising various streams of diversity in pronunciation traditions and in the application of Tiberian vowel signs to represent the pronunciation. Finally, subtypes of NST map onto codicological features in a broad sense. This indicates that there is a linkage between the codicology of a manuscript and the features of the written text that it contains.

1.2. The Evidence Threshold

As a general rule, predictive statistical tests are considered significant if they have a probability value (p-value) of at least 0.1. This indicates that there is less than a 10 percent probability that the particular statistical relationship tested for happened by chance. However, p-values are not meant in this study to be used as a definitive marker of typology: a p-value which approaches significance, but which fails the full test, is still treated as meaningful and placed on a spectrum alongside the significant results.[10]

[10] The current attitude of researchers towards p-values is that they should be interpreted on a continuum indicating weakness or strength in the results, not treated as categorical, black-and-white measures of the subject being studied (Amrhein, Greenland, and McShane, 2019). This is the approach that I embrace in the present research.

2.0. METHODOLOGY

2.1. Sampling Strategy

The data in this study consist of fragments of two- or three-column parchment codices of the Torah with complete dimensions from the extant Taylor-Schechter and Lewis-Gibson Genizah collections in the Cambridge University Library. Wherever possible, the data were collected via first-hand assessment of the manuscripts, with the support of the metadata and photographs from the Davis-Outhwaite catalogues, the Cambridge University Digital Library's Lewis-Gibson entries, and the Friedberg Genizah Project. In order to limit the study to a reasonable size, the corpus is split into two groups based on number of columns, with different criteria for inclusion in each group:

Three-column group criteria:

- A parchment base.
- Any combination of Masoretic notes: no notes, full Masoretic notes (Masorah Parva and Masorah Magna), or partial Masoretic notes (either Masorah Parva or Masorah Magna).
- Either unvocalised or have NST vocalisation. Also included are fragments with ST vocalisation which lack full Masoretic notes.

I found 142 three-column manuscripts in Cambridge that meet these criteria.

Two-column group criteria:

- A parchment base.

- Either full or partial Masoretic notes. Two-column parchment manuscripts without any Masorah are excluded because they vary so widely in their features (see Section 1.1).
- Any vocalisation type: none, ST, or NST.

I found 154 Torah fragments meeting these specifications in the Genizah collections in Cambridge.

In total, 296 two- and three-column fully dimensioned fragments meet the aforementioned conditions for the study. This is an estimated 98–99 percent of manuscripts with these codicological features in Cambridge (as always, it is possible that some manuscripts may have been overlooked, so I do not assume complete comprehensiveness). The research is therefore representative for the Genizah collections in Cambridge.

2.2. Palaeography

A cautious approach was taken regarding palaeographic assessment. Each of the manuscripts in the corpus which had NST vocalisation was assigned a general palaeographic identification, with a focus on determining the provenance rather than on pinpointing an exact date. The assessments involved establishing the palaeographic type of script on the basis of comparative samples and estimating a date spanning two centuries.[11] Below are the categories used as general palaeographic descriptors for region:

[11] It is fully expected that further research may (and should) correct and clarify some of the palaeographic assertions made in this study. The palaeographic estimations were based on comparative sources and used the methods developed in the following scholarly resources: Birnbaum

- ‘Oriental’: manuscripts with a ‘Northeastern’ or ‘Southwestern’[12] Oriental script style.
- ‘Palestinian-Byzantine’: manuscripts with a script style that is characteristic of manuscripts produced in a region ranging from the Levant to Asia Minor.
- ‘Italian-Byzantine’: manuscripts with a script style that is characteristic of manuscripts produced in a region ranging from Italy to Asia Minor.
- ‘Sephardi’: manuscripts with a clear Sephardi style of script.

The regional labels I attach to specific scripts should be seen as approximations rather than fixed assessments. The mobility of scribes and the variability of script styles in the Genizah often makes the exact pinpointing of regions and dates problematic. For purposes of this typology, the regional labels should be taken as wide estimations rather than exact diagnoses.

(1971); Beit-Arie, Engel and Yardeni (1987); David (1990); and Yardeni (2002). Judith Olszowy-Schlanger also assisted in the assessment of a number of the manuscripts and provided me with methodological insight and feedback.

[12] Olszowy-Schlanger (2015) introduces these terms and describes the differences between Southwestern Oriental and Northeastern Oriental scripts. It is important to note that palaeographic typological features appear on a spectrum and that overlap between regions is likely. Most notably, Olszowy-Schlanger explains here that the ‘Northeastern Oriental’ Hebrew script spread from Mesopotamia to the rest of the Islamic world rapidly, and so many Egyptian manuscripts are written in what we call a ‘Northeastern’ script style.

2.3. Statistical Procedures

The statistical approach taken in this study was non-experimental and relied mainly (but not exclusively) on non-parametric statistical tests (meaning that no statistical prediction/probability was involved). Data were stored in an SQL database which I created especially for the research. In collecting linguistic data, only one page (single or conjoined) was read per manuscript in order to avoid assigning multiple-page manuscripts greater weight than single leaves (multiple pages of a manuscript generate more linguistic data and this could bias the statistics against single-leaf manuscripts).

The general descriptive statistics (basic distributions of features) are reported first. Then three kinds of clustering algorithms are performed on the data (k-means, k-modes, and mean-shift clustering), because their different mechanisms elucidate different aspects of the data. The computer ran each algorithm up to ten times: the data are clustered and re-clustered by the computer until the numerical distance between each group is optimal.[13]

Codicological and linguistic features were assessed separately. The results of the codicological clustering are given in section 4, and the results of the linguistic clustering are given in section 5. In the conclusion of the study, the results of the codicological and linguistic clusters are compared: the major finding is that manuscripts that cluster together in the codicology also tend to cluster together in the linguistic groups.

[13] See section 4.2 for a more in-depth explanation of clustering algorithms and relevant literature.

2.4. Textual and Linguistic Analysis

The textual data of the manuscripts were compared with photographs of Codex Leningradensis[14] and the BHS. Due to the size of the corpus, I did not find it helpful to generate a ratio comparing the number of occurrences of an NST feature against the size of the manuscript or passage involved.[15] Any deviation from Leningradensis/BHS was noted. I did not, however, record *rafe*, due to the fact that it varies greatly even across standard Tiberian manuscripts.[16] Cantillation was likewise not assessed. After the clustering was performed and the patterns established, their linguistic characteristics were assessed in-depth, and the patterns and resulting examples are shown in Section 5.

3.0. Comprehensive Descriptive Statistical Analysis: Codicology and Linguistic Features

The following report on the feature distributions of codicology concerns all 296 manuscript fragments which are the subject of this study. The report on linguistic feature distributions concerns the 55 NST manuscript fragments which were found in the corpus of the whole 296.

[14] National Library of Russia, I Firkovitch Evr. I B 19a.

[15] Blapp (2017) uses such a ratio very successfully, because his corpus of manuscripts is small. I have found that with a large corpus, such a ratio provides only diminishing returns.

[16] Thanks to Ben Outhwaite for his advice regarding this decision.

3.1. Descriptive Statistics: Codicology[17]

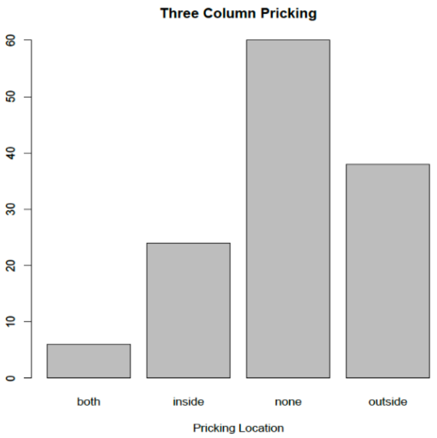
3.1.1. Format (Ratio of Width x Length)

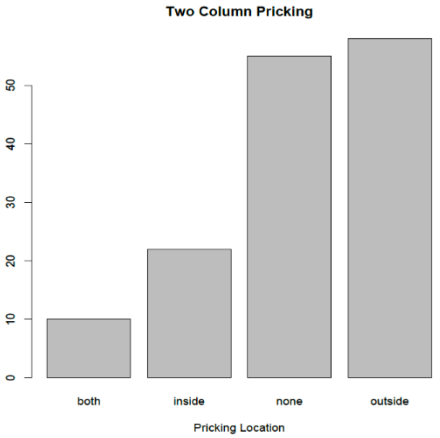
The two groups (two-column and three-column) have roughly equal distributions of formatting proportions: with ‘portrait’ format (ms length > ms width by more than 1cm) being the most common, and ‘square’ (width and length within 1 cm of each other) the second-most common. ‘Landscape’ (ms width > ms length by more than 1 cm) is the rarest.

3.1.2. Pricking (Holes in the Margins to Aid in Ruling a Page)

The majority of both groups has no visible pricking. The two-column group has significantly more manuscripts with pricking in the outside margin (58; 37.6 percent) than the three-column group (38; 26.7 percent):

[17] The following manuscript features are not reported here due to their homogeneity between the two manuscript groups: ruling (99 percent were ruled); regular parchment shape (~93 percent had regularly-shaped, high quality parchment); *petuḥa* and *setuma*: 99 percent had regular line breaks; Masorah (see section 1.1); graphical line-fillers to keep the margins even (the majority favoured a couple of line-fillers per page); correction extent (the majority of manuscripts had minimal corrections).





3.1.3. Margins

Manuscripts were visually assessed for their margin width in relation to the text and not measured numerically. ‘Regular’ margins = the margin width is average all around the text and not overly large or small. ‘All-wide’ margins = all margins are disproportionately wide in relation to the space the text takes on the page. There were other more unusual variations in the relation of margin width to the text, such as ‘bottom-wide’, where the bottom margin was disproportionately wide while the other margins were regular. Both groups favoured ‘regular’ margins. Differential results: two-column group: more ‘all-wide’ manuscripts. (45 manuscripts total had this feature = 29 percent) than the three-column group (26; 18.3 percent). As a group, the two-column manuscripts tended to have more variation in margin width than the three-column group, which was more homogeneous.

3.1.4. Illumination and Decoration

Extra-textual decoration was rare for both groups. Differential results:

- Two-column group: much variation: *parashot* decorations (23.3 percent; micrography 2.59 percent; 1 manuscript with extensive decoration; 1 manuscript with professional illumination).
- Three-column group: minimal variation: only small decoration surrounding *parashot* markers were found (30 manuscripts; 21.1 percent).

3.1.5. Script Type, Level of Sophistication, and Script Size

All manuscripts were assessed on the type of script (square or semi-cursive), the sophistication (scribal, average, or unprofessional), and size (small, average, medium, large) of the letters of the handwriting in proportion to the dimensions of the page. Differential results:

- Script type: 100 percent of manuscripts used a square script.

- Sophistication: 100 percent of three-column manuscripts had a professional script;[18] 5, or 3.24 percent, of the two-column manuscripts had an 'average' script which was either professional but overwritten (and less legible) or which was written in a less sophisticated hand.
- Script size: an 'average' size script (not overly large or small in proportion to the page) predominated in both groups. 'Small' was a significant minority in both (two-column: 57; 37 percent; three-column: 50; 35.2 percent). Outlier: T-S A3.15: a three-column fragment with a 'large' script.[19]

3.1.6. *Parashot/Sedarim*

Both groups favour no marking of a *parasha* (probably because the passages on the fragments did not begin a *parasha*). Differential results:

- Three-column preferred *parasha* markers over *sedarim* markers (17; 11.9 percent marked the *seder*);
- Two-column had a greater number with *sedarim* markers (35; 22.7 percent).
- A small minority of both groups marked both *parashot* and *sedarim.*

[18] T-S AS 1.249 has been crudely re-written on the verso.

[19] This manuscript was categorised as post-twelfth c. Oriental. Thanks to Judith Olszowy-Schlanger for her assistance.

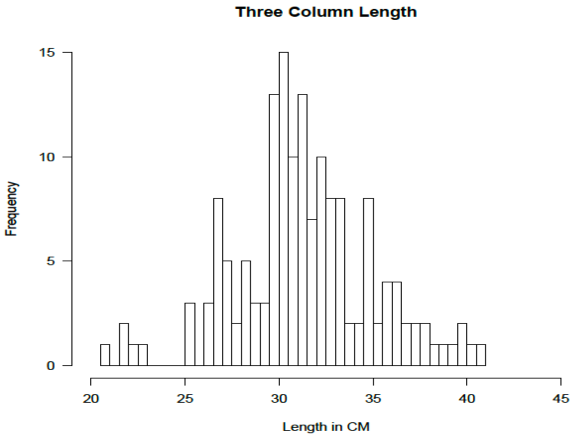
3.1.7. Vocalisation

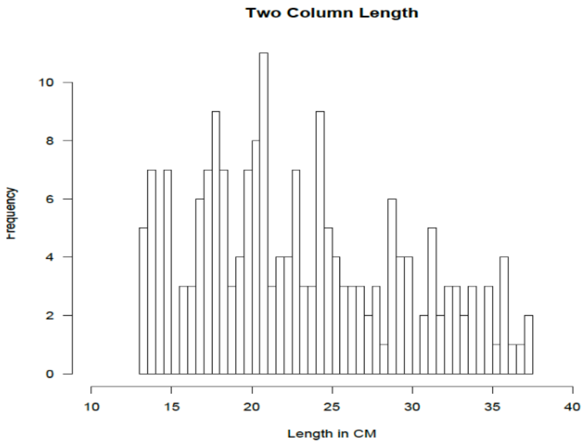
Both groups had a majority of manuscripts with ST vocalisation. Differential results: three-column manuscripts had proportionally more NST manuscripts (33; 23.2 percent) than the two-column manuscripts (22; 14.2 percent). This proportion no doubt would change if two-column manuscripts without Masorah were included.

3.1.8. Dimensions

The distribution of leaf length and width differ for the two groups:

Length:





The three-column group has a distribution that somewhat resembled a normal[20] distribution:

- Range: 20.6–40.9 cm
- Mean:[21] 31.3 cm
- Standard Deviation (a rating of variance in the lengths of manuscripts): 3.70.
- Quartiles: median: 31.1 cm, interquartile range (measure of dispersion): 29.6–33.1 cm
- Test of normality (Shapiro-Wilk test): p-value = 0.05

[20] 'Normal' here means that the shape of the distribution bars peaks at the median and tapers down symmetrically on both sides. This means that most three-column manuscripts have a typical length of approximately 31 cm, and those that differ from that size become rarer the more their length deviates from this value.

[21] This is the average length of a three-column parchment leaf.

The three-column group is quite uniform, and the average length of 31.3 cm is representative—meaning that the general three-column parchment 'near-model' Genizah Torah codex is likely to have a page length of around 31 cm. This is because the distribution is essentially normal and the standard deviation is low.[22] An outlier group of small three-column codices clearly occurs between 21 and 24 cm. The interquartile range is small, indicating homogeneity (not much variation in the majority of manuscripts). The Shapiro-Wilk result indicates that the distribution is for all intents and purposes normal.[23]

The two-column group varies considerably and does not resemble a bell curve.

- Range: 13.0–37.3 cm
- Mean: 23.2 cm
- Standard Deviation: 6.33 cm
- Quartiles: median: 22.2 cm, interquartile range: 18.1–27.9 cm
- Shapiro-Wilk: p-value = 0.00002

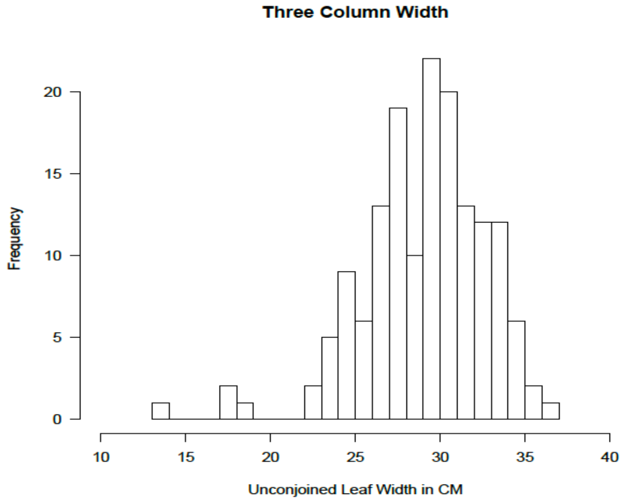
The standard deviation is double that of the three-column group, and so the average length of 23.2 cm is less representative

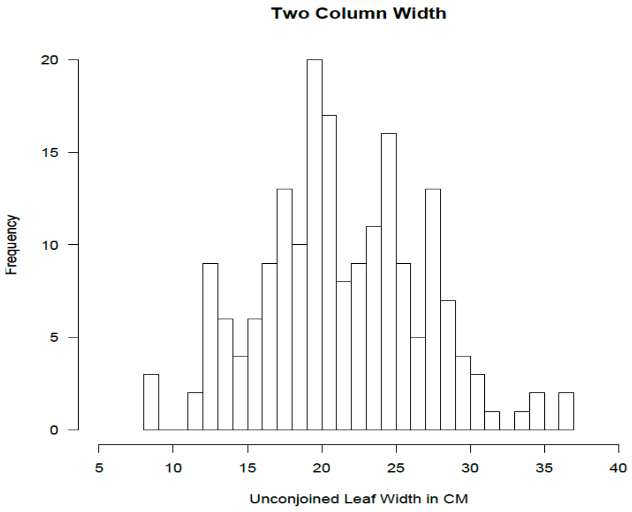
[22] A high standard deviation would indicate that many manuscripts differ from the average dimensions of the entire group. For three-column manuscripts, the low standard deviation means that many are close in size to the average.

[23] If $p > 0.05$ on a Shapiro-Wilk test result, the data are considered normally distributed and predictions can be more confidently made about the average and non-average features of the manuscript population.

of the whole group. The interquartile range is triple that of the three-column group, meaning more manuscripts vary in their length from the average. The extremely low result of the Shapiro-Wilk test indicates that the data are far from normally distributed. These results indicate that there are smaller sub-groups of similarly-sized manuscripts within this heterogenous data set.

Width:





The difference in distribution of widths between groups is noteworthy.

Three-column:

- Range: 13.8–36.7
- Mean: 29.0
- Standard deviation: 3.63
- Quartiles: median: 29.5 cm, interquartile range: 27.0–31.5 cm
- Shapiro-Wilk: p-value = 0.00007

Two-column:

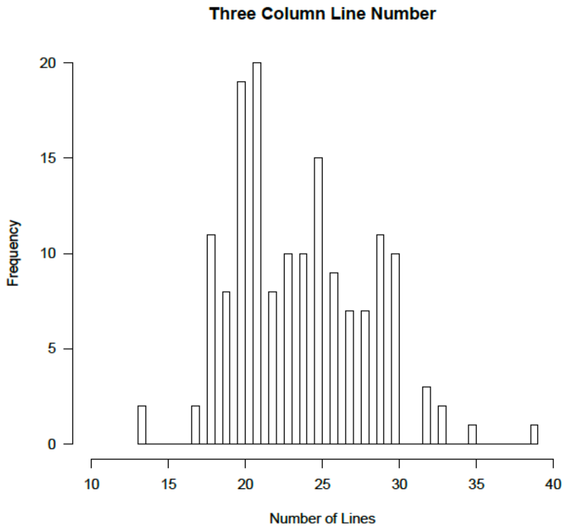
- Range: 8.85–36.9
- Mean: 21.3
- Standard deviation: 5.45
- Quartiles: median: 20.6 cm, interquartile range: 17.5–24.8 cm

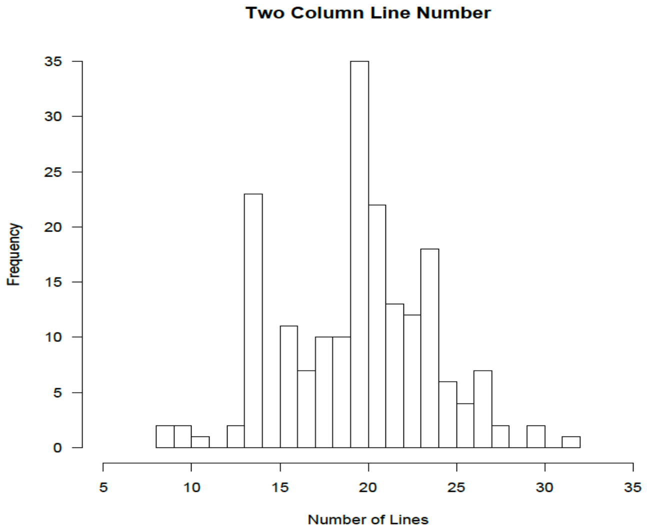
- Shapiro-Wilk: p-value = 0.4456

The average width of a manuscript in the three-column group is 29 cm, and the small standard deviation indicates that 29 cm is likely the true average width for the entire group. The median, or middle, width (29.5 cm) is close to the mean, or average width (29.0 cm), which further confirms that the average width is representative for the group. The Shapiro-Wilk result, however, indicates that the data are far from normally distributed, no doubt because of the outlying group of small manuscripts (between 13–19 cm).

Though the two-column manuscript group has a higher standard deviation, and the mean and median are farther apart, it is safe to say that the average width of 21.3 cm is generally representative of the group. The Shapiro-Wilk test for this group is positive ($p > 0.05$), indicating that the data are likely distributed normally.

3.1.9. Line Number





Three-column:

- Range: 13–39 lines
- Mean: 23.7

- Standard deviation: 4.40
- Quartiles: median: 23 lines, interquartile range: 20–27 lines
- Shapiro-Wilk: p-value = 0.001

Two-column:

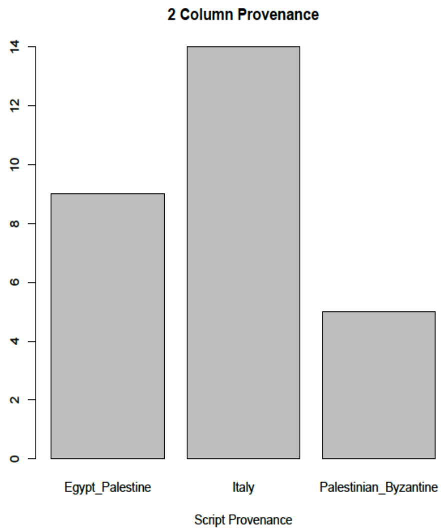
- Range: 8–32 lines
- Mean: 20.0 lines
- Standard deviation: 4.21
- Quartiles: median: 20, interquartile range: 17.2–23 lines
- Shapiro-Wilk: p-value = 0.004

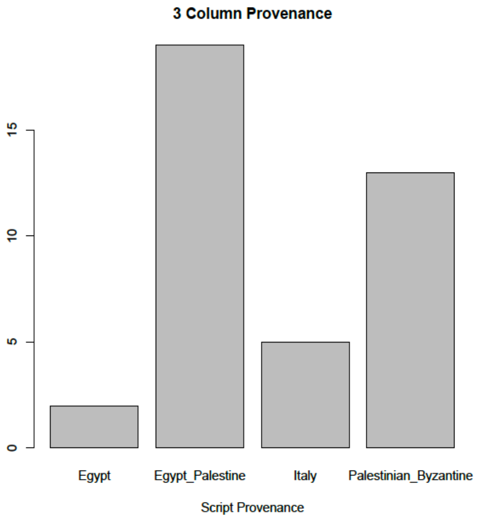
The average line number for both groups is generally representative. The Shapiro-Wilk tests, however, indicate that neither group is normally distributed regarding line number ($p < 0.05$ in both sets), and this indicates the possibility of sub-groups of similar manuscripts within this heterogeneous corpus.

3.1.10. Palaeography, Provenance, and Date

While manuscripts were assigned a typological category based on their palaeography, only the NST manuscripts were carefully assessed for their provenance and date. The data shown below reflects only manuscripts with NST vocalisation (55 manuscripts total).

Differential results: There are many more Italian-Byzantine NST manuscripts in the two-column group (9; 40.9 percent). The three-column group has significantly fewer Italian-Byzantine specimens (4; 12.1 percent). Oriental manuscripts (both Northeastern and Southwestern) predominate in the three-column group (29; 87.8 percent) and are large minorities in the two-column group (13; 59 percent). In the charts below, ‘Egyptian-Palestinian’ indicates scripts with a ‘Northeastern’ Oriental script style (which had spread to the Levant and to Egypt: see footnote 12).





3.1.11. Discussion of Descriptive Codicological Statistics

The descriptive statistical findings indicate three levels of codicological feature distribution, viz. common, less common, and infrequent features (but not necessarily all in the same manuscript in all three levels of occurrence).

Common features in both groups include a portrait format, no evident pricking holes, regular/even margins, minimal decoration, Masoretic line breaks, a square and professional script that is balanced in size and with an 'Oriental' (either Northeastern or Southwestern) palaeography, an ST vocalisation, 23–33 cm long x 20–30 cm wide, and 20–23 lines.

Less common features include square manuscripts, wider margins, a greater amount of decoration, a small and professional script that is Byzantine or Italian, NST vocalisation, more variation in size and number of lines. It is likely that there are multiple sub-groups of Bible types indicated by these data that can be uncovered through correlational statistics and clustering.

Finally, infrequent features include a landscape format, pricking on both margins, narrow or unbalanced margins, very late Oriental or Italian scripts, complex illumination, no line breaks, no vocalisation, and extremes in size and number of lines.

The most important finding of these descriptive statistics is that they clarify the differences and similarities between Torahs with two and three columns. The two groups of manuscripts had at least one significant difference in the distribution of features for each feature presented above. For example, there are many more Italian-Byzantine near-model Bibles with two columns, while more Oriental near-model Bibles tend towards three columns (§3.1.10). Ultimately, the data show that the two- and three-column manuscripts are related on many points, but distinct in a significant number of ways.

The most noteworthy trend regards dimensions. Two-column Bibles are more heterogenous in terms of dimensions and line number, which indicates that multiple sub-groups may be more clearly defined in the corpus. Three-column manuscripts, on the other hand, are much more homogeneous, which means that while sub-groups exist, they may be less distinct.

Ultimately, while two- and three-column 'near-model' Torah codices can be grouped together in terms of average shared features, it is clear that we should not conflate them based on their commonalities; they are better characterised as close sisters within the same family.

3.2. Descriptive Statistics: Linguistic Features

Within the corpus, the three-column group contains 33 manuscripts with NST vocalisation, and the two-column group contains 22 manuscripts with NST vocalisation (55 total NST manuscripts). By comparing these manuscripts with Codex Leningradensis (hereafter, L), I identified 103 distinct types of variation in all of the manuscripts. Of the total of 103 types of variation, 76 are relevant to the present study.[24]

[24] Features such as *plene* and defective spellings, *qere* in place of *ketiv*, and textual differences were not incorporated into the statistics presented here. *Rafe* was also not a factor in the statistics due to the unpredictability of its usage. As Blapp (2017) points out, all the exemplar

The two-column group had fewer distinct vocalisation or diacritical features (60) than the three-column group (92). The general distributional trends of these features are presented below.

3.2.1. Feature Frequency Distributions

There are three kinds of distributions of NST features in the corpus of manuscripts:

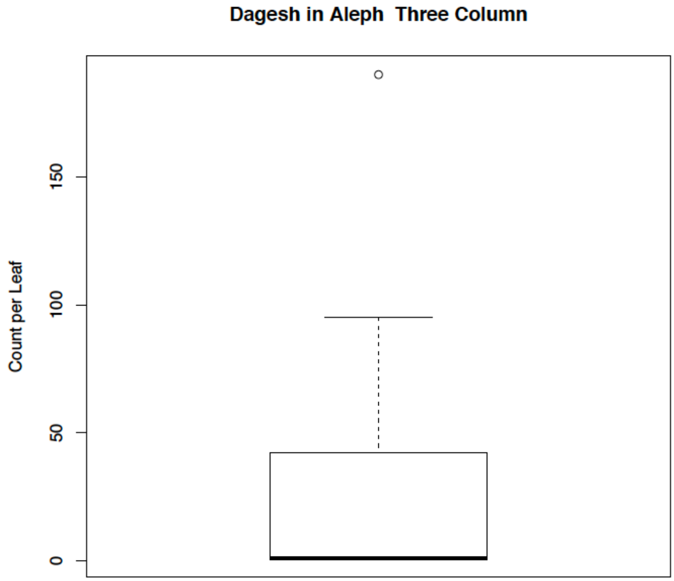
A. Infrequent occurrences: There are a significant number of features in both groups that occur once or at most twice in a manuscript. Either the feature is the only deviation from L present in the manuscript, or the feature is the result of a larger pattern of more complex phonological changes in the pronunciation of the vowels in the text.

B. Even distributions: some features occur evenly through a spread of multiple manuscripts. For example, the feature '*dagesh* in an *ʾalef*' occurs at regularly increasing intervals between one and fifty times in two-column manuscripts. These kinds of distributions are rare, making up at most 10 percent of the data. They indicate that the feature is generally common for that group.

C. Uneven distributions: These are distributions in which a particular feature occurs infrequently in many manuscripts,

codices use *rafe* in a different way, and "this observation suggests that *rafe* has not been standardised, which makes it necessary to study *rafe* in each manuscript" (223).

alongside extreme outliers where the same feature occurs more than two-hundred times in a single manuscript:



This boxplot shows us that for three-column manuscripts, the majority of the data are concentrated in manuscripts that have *dagesh* in *ʾalef* fifty or fewer times. Then, at the very top of the plot, we see one manuscript which has the feature over two-hundred times. While this distribution pattern occurs in both groups, it is more typical in the three-column group. Many three-column manuscripts have large quantities of one NST feature (alongside more moderate counts of other NST features), while the two-column group's manuscripts typically have a more balanced distribution of NST features.

3.2.2. Systematic Understanding of Feature Types

It is clear that not every NST feature is equal in its frequency of occurrence in the corpus, or in its role in the larger pattern(s) of features within a given manuscript. Some features predominate and seem to set the trend for less-common features. The features that occur the most frequently across the corpus, and that seem to set the trend for patterns observed, are listed below alongside the highest attested count of occurrence in a manuscript.

- Missing *dagesh* (209 times)
- *Dagesh* in *ʾalef* (190 times)
- Unexpected *dagesh* (116 times)[25]
- *Pataḥ* for *qameṣ* (90 times)
- *Pataḥ* for *ḥaṭef pataḥ* (54 times)
- *Pataḥ* for *shewa* (40 times)
- Word-Final *shewa* (37 times)
- *Ṣere* for *segol* (35 times)
- *Pataḥ* for *segol* (32 times)
- *Shewa* for *ḥaṭef pataḥ* (30 times)
- Unexpected *shewa* (25 times)[26]
- *Segol* for *ḥaṭef segol* (23 times)
- Missing *shewa* (20 times)
- *Shewa* for *pataḥ* (12 times)
- *Segol* for *ṣere* (12 times)

[25] This category simply describes an instance where a manuscript has *dagesh* and L does not; differentiated types of unexpected dagesh were analysed after the statistical clustering and are described in section 5.

[26] See the above footnote; the same applies for ‘unexpected *shewa*’.

- Missing *mappiq* (10 times)

The above list indicates the NST features that predominate in the corpus and that seem to play the most critical roles in the patterns of NST vocalisation. There are, however, many other deviations from L that occur at lower frequencies, but that are still important for shaping differences in sub-groups of vocalisation.

3.3. Discussion

These data complement findings stated in previous scholarship on NST vocalisation. Blapp is indeed correct when he states "we have to be aware that the degree of non-standardness of all the manuscripts [in his thesis] varies".[27] This applies also to the present corpus. Blapp noted, furthermore, that some manuscripts in his corpus, for example, T-S A13.18, contain very few NST features.[28] Likewise, in the present study, there are specific groups of features that occur once or twice in an otherwise fully ST manuscript.

Most notably, the predominating features in Blapp's study were the following interchanges:

- *Qameṣ* with *pataḥ*
- *Ṣere* with *segol*
- *Ḥireq* with *shewa*
- *Ḥolem* with *qameṣ*
- *Ḥaṭef* vowels with *shewa*
- *Shewa* for furtive *pataḥ*

[27] Blapp (2017, 199).

[28] Ibid.

He noted, in addition, extensive non-standard use of *dagesh*. Apart from the interchanges of *ḥolem/qameṣ* and *ḥireq/shewa*, all of these features predominate to a high degree in my larger corpus of 55 manuscripts.

4.0. Patterns of Codicology and Text: Cluster Analyses of Codicological and Linguistic Data

4.1. Methodology Review

The statistical methodology was chosen with the aim of exploring meaningful patterns within the dataset and was therefore non-experimental. The main focus was upon finding patterns using appropriate clustering algorithms and then verifying their linguistic and codicological meaningfulness. The general methodology took three steps:

1. Three clustering algorithms, k-means, k-modes, and mean-shift (defined in section 4.2), were run on the data in order to establish the initial boundaries of large patterns in codicological and linguistic data. The clustering algorithms assessed all of the manuscripts and grouped them based on which features (codicological and linguistic, respectively) certain manuscripts share, and how often those features occur per manuscript in the group. The results of the algorithms are lists of manuscripts that share features.
2. These patterns were analysed in order to identify the most critical factors and to refine the clustering process by identifying and removing distracting variables.

3. Where applicable, traditional tests of significance (ANOVA, Chi-Squared, etc.) were run to clarify the strength of correlations between specific codicological or linguistic features that were unearthed by the clustering results.

4.2 Cluster Analyses

Statistical clustering is a branch of unsupervised machine learning that is targeted towards data mining and towards establishing the shape of patterns in large-scale data.[29] It is, therefore, an appropriate strategy for identifying patterns in Torah manuscripts in the Genizah.[30] Different clustering algorithms group the data together based on similarities, which, when compared in person by the researcher, allow for cross-validation and a more complete picture of patterns within the dataset.

K-means is the most commonly used algorithm, because it works with the mean (average) of numeric data of a manuscript

[29] An explanation of the statistical processes used in this research can be found in the following introductory volume: James, Witten, Hastie, Tibshirani (2015). More technical papers are cited in the footnotes below.

[30] In one instance, the computer found separate leaves of the same manuscript and placed them together in the same cluster. This was confirmed by Zina Cohen, who kindly performed her microscopic reflectography method on some of the manuscripts in this corpus (Cohen, Olszowy-Schlanger, Hahn and Rabin 2017). The results of the reflectography showed that the manuscripts shared the same kind of ink. Many thanks to Zina Cohen for conducting the reflectography analysis for the present research.

feature (for example, codicology: number of lines; linguistics: counts of a vowel interchange per manuscript), in order to determine clusters based on how similar each manuscript is to the others (using Euclidean distance measures).[31] K-means requires the researcher to anticipate the number of clusters in the dataset in advance. As this was not known, I ran calibration by increasing the number of clusters until the Euclidean distance between clusters stopped dropping dramatically between tests (meaning the features of all the manuscripts in a given cluster were relatively homogeneous).[32]

K-modes, on the other hand, works with the mode, not the mean, to establish clusters in both numerical (quantitative) and non-numerical (qualitative) data.[33] Since it works with the central point of a group of data, which is less affected by outliers

[31] For codicology, I used this algorithm on the dimensions and line number counts; for linguistics, each feature within a manuscript was counted on the basis of its occurrences per word, and thus could be analysed by this algorithm. A respected paper on k-means clustering is MacQueen and James (1967).

[32] Euclidean distance here means a rating of variance between manuscript features in a cluster; the more clusters in a dataset, the smaller the distance between manuscripts within one cluster (i.e., the more codicologically or linguistically similar the manuscripts in a particular cluster are). The cited work in footnote 31 deals more with Euclidean distance. A paper on optimising the number of clusters using the method as described above (known as the 'elbow method') is Kodinariya and Makwana (2013).

[33] Quantitative data are only numeric (number of lines = 15, 16, 17); qualitative data are non-numeric (script size = small, medium, large,

(e.g., very rare features, or manuscripts with very high counts of an NST feature), k-modes is appropriate when manuscripts have extremely large or small amounts of features, because it is less affected by the outliers and produces more reliable clusters.

Mean shift clustering is another numerical algorithm that was performed to act as a supplement to k-means/k-modes. Mean shift clustering does not require the researcher to anticipate how many clusters may be in the data in advance; it finds the number of clusters automatically. It can, however, be thrown off by large or small outliers in the data.[34] Nonetheless, because of its ability to find clusters without advance prediction, it was used to help validate the number of clusters found by k-means and k-modes. With all three clustering algorithms performed together, I was able to arrive at the optimal number of clusters in the manuscript data and therefore all of the sub-groups of the manuscripts are statistically reliable and visually apparent and distinct.

It is tempting to test every single codicological or linguistic variable, no matter how infrequently it appears in the data. The present study found, however, that this does not produce useful results, because clustering algorithms are sensitive to outliers and can be distracted by numerous variables. This can result in the creation of false groups, separating similar manuscripts and grouping together dissimilar manuscripts. For example, when the

average). A resource for k-modes clustering is Chaturvedi, Green and Caroll (2001, 35–55).

[34] A paper on mean shift clustering: Cheng (1995, 790–99).

computer considered too many outlying variables, two manuscripts which shared many codicological features would be artificially separated on the basis of an inconsequential difference.

On the whole, it is better to test on fewer, more critical features, rather than many. Controlling the number of variables produces the best results and can sometimes find the most critical features in the typology. Whilst this method may be susceptible to bias, I was careful to avoid bias by investigating outliers and outlier clusters separately. It, therefore, does not increase the risk of missing out on rare features, because manuscripts which lack the more common, tested features are placed by the computer in an 'outlier' group. This allows the researcher to further investigate and find the rare features that set them apart.

Therefore, avoiding the inclusion of rare features and reducing the number of different factors for the computer to analyse results in clearer groups. Most notably, features that are not included in the clustering, if they truly are part of a pattern, will self-organise around the features that are tested, and the researcher will catch important details.

4.3. Codicological Cluster Analysis and Results

After the cluster analyses, the next step was to identify the major factors that distinguished the clusters. As some features were identified as biasing the clustering results, they were removed and the clustering was re-performed. The critical features that were included in the final round of codicological clustering were: format, pricking location, margin width, illumination, script size, presence or absence of Masorah, *parashot* or *sedarim* markers, extent of line fillers, dimensions, and number of lines.

The most crucial variables for establishing meaningful clusters were dimensions and line number. These features established themselves as independent variables: when performing clustering on **only** dimensions and number of lines, *every other codicological feature self-organised into the pattern without being tested*. For example, I did not include palaeography in the clustering, yet the groups established by differences in dimensions and line number also each had their own unique palaeographic tendencies.

This is a find of crucial importance. It appears that typological variation in codicology can be solidly established solely on the basis of dimensions and number of lines of manuscripts in a dataset. Manuscripts with similar sizes and numbers of lines are likely to share the same palaeography (and other codicological features). This may indicate that regional scribal practices are distinguishable mainly on the basis of size and line number.

4.3.1. General Characteristics of the Codicological Clusters

The clustering of all 296 manuscripts (including ST and NST manuscripts) resulted in thirty distinct subtypes across both the two- and three-column groups. While there is not space to give the details for each group, there are distinct, general trends that are meaningful for assessing the correlation between linguistic and codicological features. The following typology is organised by dimensions, and then by observations of the general level of sophistication of each subgroup. Individual features are tested with significance tests where necessary to determine the strength of feature correlations within the subgroup.

4.4. Codicological Manuscript Sub-Groups[35]

The following subtypes are selected representatives of the full thirty subtypes found across the 296 manuscripts that were clustered.

Small Italian-Byzantine Codex[36] (Two-column)

This was the smallest and most homogeneous group in the typology.

- 13.1–14.7 cm in length x 11.4–13.1 cm in width.
- 20–21 lines
- Italian or late Byzantine script style
- Portrait format (two are square)
- The square manuscripts have wide-bottom margins and a small script
- Unpricked, average script size
- All mark the Palestinian triennial reading cycle
- Full Masorah (one has only Masorah Parva)

[35] The manuscripts within these subgroups were either Sephardi (late: fourteenth c.), Italian-Byzantine, or Palestinian-Byzantine (i.e., Southwestern Oriental to Byzantine) in their palaeography.

[36] Members: T-S Misc 3.49 (Southwestern Oriental script type); T-S Misc 9.8; T-S NS 24.36; T-S NS 9.31; T-S NS 8.8; T-S NS 14.35; T-S NS 173.92; T-S AS 64.206; Or 1080.A1.2.

- Generally sophisticated in formatting (rare use of line-fillers to keep an even margin)
- 50 percent had NST vocalisation, and all of these had all-wide margins

Large Monumental Levantine Codex[37] (Three-column)

- 35–38.2 cm long x 32–35 cm wide
- 25–30 lines
- Portrait (one square manuscript included)
- Pricking on the outside, or absent
- Wide margins (bottom widest)
- Sparse decoration
- Small-average script size
- Full Masorah favoured
- *Parashot* marked most often.
- NST predominates, and the majority have full Masorah (Fisher's Exact = 0.0238, $\chi^2 = 0.0611$).

[37] The manuscripts within this group are either Northwestern Oriental or Southwestern Oriental in their palaeography and are likely to come from the Levant: T-S NS 77.3; T-S NS 77.2 (join with T-S NS 77.3); T-S NS 12.22; T-S NS 248.2; T-S NS 248.3 (join with T-S NS 248.2); T-S A 4.30; T-S A2.1; T-S NS 20.14; T-S NS 12.2.

Monumental Bare Wide-Ranging (Oriental to Italian;[38] Three-column)

These manuscripts are smaller than those of the aforementioned groups, and all lack Masoretic notes.

- 27–30.5 cm long x 24.4–29.5 cm wide
- 32–39 lines
- Mainly square format
- Majority not pricked
- Wide margins predominate
- Decoration only on one manuscript
- All scripts are small
- Reading cycles generally unmarked, but, where occurring, mark the *sedarim*
- Majority have NST vocalisation

Small Oriental Codex[39] (Two-column)

This group is a relatively homogeneous group of manuscripts, which, like the small Italian-Byzantine manuscripts above, are

[38] The majority of the manuscripts in this group are Egyptian (late) or Southwestern Oriental–Italian-Byzantine. Members: T-S A 2.30 (Egyptian, post-eleventh c.); T-S NS 51.22 (Southwestern Oriental or Italian-Byzantine; T-S NS 282.69 (may be a join with T-S NS 51.22) T-S AS 64.242 (Southwestern Oriental or Italian-Byzantine); T-S AS 66.52 (Egypt, post-eleventh c.).

[39] Members: Or 1080.A4.10 (Northwestern Oriental, probably Egypt); T-S AS 28.259 (Southwestern Oriental); T-S Misc 9.80 (Egyptian, post-eleventh c.); T-S Misc 1.46 (Egyptian, post-eleventh c.); T-S A 1.2 (probably Southwestern Oriental); T-S NS 19.16 (probably Egyptian).

small. It can be seen as a counterpart to the Small Italian-Byzantine Codex.

- 14.6–17 cm long x 12.5–14.6 cm wide
- 19–25 lines which are set together and very compact
- Portrait format (with one square manuscript)
- Pricking on the inside margin (except for a square manuscript which pricks the outside, $\chi^2 = 0.0820$, Fishers' Exact = unsignificant)
- Decoration is rare, and associated with manuscripts marking the *parasha* (manuscripts marking the *seder* do not have decoration)
- No manuscripts have full Masorah
- Margins are average, except for the one NST manuscript, which has narrow vertical margins
- Inverse relation between the size of the script and the number of lines; manuscripts with a 'small' script size could have more than 20 lines, but manuscripts with an 'average' script size did not have more than 20 lines

Oriental Bare Square Group[40] (Three-column)

This is the only three-column group to have manuscripts with an Oriental (Egypt-Palestine) script style and to include scripts from no other regions.

- 30.6–32.8 cm long x 31.5–36.7 cm wide
- 28–32 lines
- Square-landscape format
- Inside pricking

[40] Members: T-S NS 19.29; T-S NS 56.5; T-S NS 65.34; T-S NS 67.6.

- Narrow-regular margins
- Small script size
- Masorah is rare (hence the 'bare' label)
- *Sedarim* marked twice, the *parasha* marked once
- No NST vocalisation occurs in the group

Large Monumental Egyptian[41] (Two-column)

These manuscripts are very homogeneous as a group, and they have one feature which connects them to the Small Italian group in the section above: the tendency to mark the Palestinian Triennial *Seder.*

- 31.4–37.2 cm long x 25.2–31 cm wide
- 23–25 lines
- All have portrait format
- Outside pricking (except for the NST manuscripts, Fisher's Exact = 0.09524).
- The majority have wide bottom margin
- Sparse decoration
- Average script
- Most of the manuscripts have full Masorah
- Mainly mark the *sedarim*
- The only manuscript with a small script size is also the only manuscript to mark both the *parashot* and the *sedarim*

[41] Members: T-S A 4.4; T-S A5.3; T-S A 4.8; T-S A 4.9; T-S NS 68.22; T-S NS 74.43; T-S A 2.5; T-S NS 78.31.

Average Monumental Oriental[42] (Two-column)

This group is the most informal of all the groups represented in the two-column corpus. This is due mainly to the fact that most of them are either re-written in a very clumsy hand, or the hand is not very sophisticated. Regardless, these manuscripts still contain sophisticated codicological features.

- 18.3–23 cm long x 15.1–18.13 cm wide
- 16–18 lines
- Portrait format
- Regular-wide margins
- Decoration occurs in only one manuscript
- Average-medium script (on account of overwriting or lack of sophistication)
- Most have full Masorah
- Most do not mark any reading cycle
- Palaeography difficult to identify due to overwriting, but they appear mainly Oriental

Square Egyptian-Palestinian[43] (Two-column)

This is a group of Oriental manuscripts which all have a square format and most of them typically have full Masorah. They are

[42] Members: T-S NS 12.4; T-S NS 17.30; T-S NS 51.31; T-S NS 57.22; T-S NS 73.4; T-S NS 161.270; T-S NS 279.74; T-S NS 282.59; Or 1080.A4.16.

[43] Members: Or 1080.A3.9; Or 1080.A1.18; T-S NS 65.32; T-S NS 24.38; T-S NS 23.25; T-S NS 22.22; T-S NS 20.25; T-S NS 57.20; Lewis-Gibson Bible 3.25; T-S NS 72.4; T-S NS 77.25; T-S NS 78.44; T-S NS 20.20; T-S NS 65.29; T-S NS 73.13; T-S NS 8.42; T-S Misc 2.74.

typically smaller than the Monumental group, but still have many sophisticated features.

- 19.1–24.3 cm long x 19.3–24.5 cm wide
- 14–17 lines
- Tend to have outside pricking
- Margins typically all wide, or bottom-wide
- Sparse decoration
- Wide range of script size
- Full Masorah
- Half mark the *sedarim*, half mark the *parashot*

Monumental Bare Oriental (Egyptian-Palestinian)[44] (Three-column)

- 25.1–30.9 cm long x 22.6–28.6 cm wide
- 20–24 lines
- Divided between portrait and square format
- Inside, outside, and no pricking present
- Majority do not mark reading cycles; those that do are square
- Wide-regular margins predominate
- Small-average script
- Masorah is rare, and only Masorah Parva present
- Majority are ST; NST manuscripts have a small script ($\chi^2 = 0.0764$, Fisher's Exact = 0.0833)

[44] Members: T-S A 5.8; T-S NS 18.5; T-S NS 65.26; T-S NS 65.39; T-S NS 65.46; T-S NS 76.48; T-S NS 180.54; T-S NS 319.101; T-S A 2.45; T-S NS 7.24; T-S NS 23.14; T-S NS 66.12; T-S NS 75.12; T-S NS 75.25; T-S NS 77.25; T-S NS 77.5; T-S AS 8.123; Lewis-Gibson Bible 2.37.

Monumental Oriental[45] (Three-column)

- 27.1–35 cm long x 27.9–33.9 cm wide
- 17–22 lines
- Majority portrait
- Pricking mainly on outside or not visible
- Wide-bottom or all-wide margins predominate
- Sparse decoration
- Average script size
- Full Masorah is uncommon (those with full Masorah have NST: $\chi^2 = 0.0154$, Fisher's Exact = 0.0119).

Oriental-Byzantine Landscape[46] (Two-column)

This is the smallest group identified by the algorithms, containing only a few manuscripts. These manuscripts, however, are distinct from any other group in that they have a landscape format (width longer than the length). No correlational statistics could be run to test the strength of their features since they all are so alike.

- 14.8–19.1 cm long x 18.8–26.2 cm wide

[45] Members: T-S A 2.42; T-S A 2.41; T-S A 2.29; T-S A 1.25; T-S Misc 1.122; T-S NS 8.6; T-S NS 24.31; T-S NS 72.18; T-S NS 73.31; T-S NS 75.20; T-S NS 76.24; Lewis-Gibson Bible 1.56; T-S AS 27.75; T-S NS 21.40; T-S A 2.51; T-S A 4.20; T-S NS 24.25; T-S NS 23.1; T-S A 4.28; T-S A 5.12; T-S NS 13.37; T-S NS 21.29; T-S AS 1.249; Lewis Gibson Bible 3.42; T-S A 1.23; T-S NS 19.23; T-S NS 23.6; T-S A 3.14; T-S A 3.23; T-S A 1.11.

[46] Members: T-S A1.56; Lewis-Gibson Bible 1.12; Lewis-Gibson Bible 1.14; T-S A41.18; T-S NS 65.24; Lewis-Gibson Bible 1.12 and Lewis-Gibson Bible 1.14 are joins.

- 8–18 lines
- Favour pricking on the outside margin
- Regular to all-wide margins
- Medium-average script size
- All have full Masorah

Large Monumental Egypt-Palestine Codex[47] (Three-column)

- 32.8–36.3 cm long x 28.2–31.8 cm wide
- 29–30 lines
- All have portrait format
- All have outside pricking
- All have wide-bottom margins
- Decoration is sparse
- Half have an average script size, half have a small script size
- Only one manuscript has full Masorah
- NST vocalisation predominates

4.5. Discussion of Clustering Results

Though only a few of the thirty total groups found in the research are presented here, the results indicate two main findings.

Firstly, the most important variables for codicological clustering tend to be dimensions and number of lines.

Secondly, the codicological groups exist on a spectrum: on one side are the groups containing mainly (or only) Italian-Byzantine manuscripts; in the middle are groups containing wide-ranging manuscripts, from Sephardi to Italian-Palestinian to

[47] Members: T-S NS 77.1; T-S NS 78.34; T-S NS 173.81; T-S AS 67.131.

Egyptian; at the other end are groups containing mainly (or only) Egyptian manuscripts. This indicates that some codicological formats were perhaps regional, while others were more widespread. Most importantly, the manuscripts are also visually similar to the others within their respective groups.

5.0. A Linguistic Typology of Non-Standard Tiberian Vocalisation: The Presentation of The Clustering Results

The linguistic findings presented below were clustered using the three clustering algorithms discussed above. Then the clusters were assessed by a thorough linguistic analysis. The results of the clustering generally fit into the schema that appears below, which was developed independently from the statistical analysis, through rigorous linguistic analysis of the data.[48] Due to limited space, I have chosen to prioritise the presentation of the linguistic results of the clustering analysis over the specific statistical details behind the results.

The findings are organised first by presenting the manuscripts of the main groups established by the clustering and linguistic analysis. Then, manuscripts which are connected to the main groups, but which are outliers in some way, are presented separately and the reason for their uniqueness is described. Furthermore, the two-column group had a small subgroup of individual outliers which did not connect clearly with any main group; these are summarised in footnote 49.

[48] Thanks to Geoffrey Khan for his assistance in developing this schema.

In the schema below, there are two hierarchies of vowel interchange. Patterns X and Y are notational, while the numbered patterns 1 and 2 (and the subtypes) may reflect phonetic changes induced by language contact.

Phonological Background	***Vowel Interchange Patterns***	
Notational interchanges of the *shewa* sign for other signs with the same sound	**Pattern X:** *Shewa-pataḥ* interchange (reflecting traditions where *shewa* was pronounced [a])	**Pattern Y:** *Shewa-ḥireq-ṣere* interchange (reflecting traditions where *shewa* was pronounced as a high vowel)
Reflecting a 'Palestinian' pronunciation with five vowels (one /a/ and one /e/) and/or phonetic Aramaic language contact	**Pattern 1:** *Pataḥ-qameṣ* and *Ṣere-segol* interchange	**Pattern 1a:** *Ṣere-segol* interchange *Pataḥ* and *qameṣ* **do not** interchange
Different patterns reflecting a reduced vowel inventory to three vowels, indicative of phonetic Arabic language contact	**Pattern 2**: *Pataḥ-segol* interchange	
	Pattern 2a: *Pataḥ-segol-qameṣ* interchange	
	Pattern 2b: *Pataḥ-segol-ṣere* interchange	
	Pattern 2c: *Ṣere-ḥireq* interchange; *Pataḥ* and *segol* **do not** interchange	
	Pattern 2d: *Pataḥ-segol-ṣere-qameṣ* interchange	
	Pattern 2e: *Pataḥ-segol-ṣere-qameṣ-ḥireq* interchange	

5.1. Two-column manuscripts: NST Linguistic Typology

The results below describe the language features of selected manuscripts within all of the clustering groups found (alongside their corresponding schema patterns). Not all manuscripts within the groups are presented here. The full lists of manuscripts are in the corresponding footnotes for each group. Note that specific vowel interchanges are reported with the vowel that appears in the manuscript first, and the vowel which appears in L second, after a hyphen. For example, a *pataḥ* for *segol* interchange is written: *pataḥ-segol*.

There were a few main groups established by the clustering: (1) the Byzantine trio: Italian-Byzantine manuscripts which all had a specific pattern of NST use of diacritics; (2) the Orthoepic group, which contained manuscripts that used NST features to reinforce an ST pronunciation; (3) Lexically-Specific NST manuscripts: those which had only NST features on specific words; (4) a group of manuscripts exhibiting a three-way interchange between *ṣere*, *segol*, and *pataḥ*.[49]

[49] There also were four manuscripts which were found by the computer to be unique individual outliers unconnected to these four main groups. These are: T-S NS 248.5, which has the Byzantine trio with a more extensive profile of vowel interchange than expected, viz. Schema 2a; Or 1080.A1.2, which has partial features of the Byzantine trio with a different profile of vowel interchange, viz. Schema 2; T-S AS 65.125, which has sign interchange, and fits the closest to the 2e schema, but lacks any interchange involving *qameṣ*; T-S NS 17.30, which both has sign interchanges and appears to fit schema 2e, although it is very damaged and the readings are tentative.

5.1.1. The Byzantine Trio of Features (Schema Patterns X, Y, 1, 1a)[50]

The following collection of two-column manuscripts contains a clear pattern which I have called the ‘Byzantine trio of features’. This pattern was found solely by the computer clustering. The Byzantine trio is as follows:

- *Dagesh/Mappiq*[51] occurs in consonantal *ʾalef*, contrasting with *rafe* on *mater lectionis ʾalef* and on historical spellings of *ʾalef* that have no consonantal pronunciation. Its function is to differentiate consonantal and non-consonantal *ʾalef*s, thereby ensuring that consonantal pronunciation is preserved. *Mappiq* is typically also extended from word-final *heh* to word-initial and word-medial *heh* and has the same function of marking the *heh* as consonantal.

[50] Other members: T-S NS 248.16; T-S NS 248.9 (no word-final *shewa* occurs because the passage does not have a word-final *ʿayin* or *ḥet*); T-S NS 248.17;

[51] There is controversy around whether this dot should be identified as *mappiq* or *dagesh*. It can be seen to function as *mappiq* in that it differentiates consonantal from non-consonantal *ʾalef*. It also, however, ensures the pronunciation of consonantal *ʾalef*. The Karaite grammarian Ibn Nūḥ treated this feature as gemination of *ʾalef*, and Karaite Arabic transcriptions of the Bible place a *shadda* (the Arabic gemination sign) on consonantal *ʾalef* (Khan 2020, §I.1.1) This allows for the possibility that the scribes using this sign considered it a *dagesh* rather than *mappiq*.

- Extended use of *dagesh* to certain ‘weak’ consonants after a vowelless consonant: mainly *lamed*, *mem*, and *nun*, but occasionally on sibilants such as *sin*, *shin*, and *samekh*, and the emphatics *ṭet*, *ṣade*, and *qof*. In some manuscripts in the group, these consonants without the *dagesh* take *rafe*.
- The presence of a silent *shewa* on word-final *ʿayin* and *ḥet*. This has the function of ensuring a word-final guttural is pronounced by explicitly marking that the consonant closes the syllable.

While these features can independently appear in manuscripts from other groups, they occur together in this trio only in manuscripts with Italian/Byzantine or distinct Palestinian scripts. The most noteworthy manuscripts with the trio are as follows:

T-S NS 21.6 places *dagesh/mappiq* in consonantal *ʾalef* consistently. It places *rafe* over the *ʾalef*s in יִשְׂרָאֵֿ֣ל ‘Israel’, and in וַֽיֹּאֿמְרוּ׃ ‘and he said’.[52] It puts *dagesh* in ‘weak’ consonants after a vowelless consonant: mainly in *lamed*, *mem*, and *nun*, but also three times in *samekh* (אֶֽל־סִּיחֹ֗ון ‘to Siḥon’ Num. 21.21, etc.), once in *ṣade* (בְּאַרְצֶּ֔ךָ ‘in your land’ Num. 21.22), and once in *qof* (וְנִשְׁקָּ֖פָה ‘and overlooking’ Num. 21.20). It puts word-final *shewa* on *ʿayin* and *ḥet* to close a syllable (וַיִּשְׁמַּ֔עְ ‘and he heard’ Num. 21.3).

[52] The pronunciation of this word in this scribe’s tradition apparently elided the glottal stop and combined the two vowels together in a diphthong: [yisrael] instead of [yisraʾel].

T-S NS 248.11 is in keeping with the patterns of the manuscripts above. It also places *rafe* on *mater lectionis* *ʾalef* (לַֽעֲזָאֿזֵֿל ‘to ʿAzazel’ Lev. 16.26). It has extended use of *dagesh* on ‘weak’ consonants after vowelless consonants and places *rafe* on consonants without *dagesh* (including *yod* and *ṣade*: וְיָֿצָֿא ‘and he will come out’ Lev. 16.24).

Or 1080.A4.18 regularly places *dagesh* in consonantal *ʾalef* (though it is sometimes omitted). It also places *dagesh* on word-internal and word-initial *heh* with a vocalisation sign (for example, יִהְּיוּ ‘they shall be’ Num. 28.19, instead of יִהְיוּ). *Rafe* occurs on *mater lectionis* *ʾalef* consistently. Similarly, ‘extended’ *dagesh* on weak consonants after vowelless consonants occurs. Word-final *shewa* occurs twice on *ḥet* to indicate the closing of a syllable; it also occurs twice to replace furtive *pataḥ* with *shewa* (for example, נִיֽחֹחְ ‘pleasant’ Num. 28.24, instead of נִיֽחֹחַ).

The general patterns of vowel interchanges within this group are all consistently similar and minimal (interchanges do not occur more than a few times per manuscript). The manuscripts generally fit into the schema patterns X, Y, 1, and 1a. This possibly indicates an underlying ‘Palestinian’ vowel system with one /a/ and one /e/ vowel. Noteworthy examples:

All but two manuscripts[53] in the group interchange *ṣere*-*segol* at least once (T-S NS 21.6: נִטֵּה֙ for נִטֶּה֙ ‘spread out’

[53] T-S NS 248.16 and T-S NS 248.17 do not have a *segol-ṣere* interchange. They do, however, have a slight profile of raised vowels. For example, T-S NS 248.16 has *ḥireq* for vocalic *shewa* once: לִגְלְעָ֔ד for לְגִלְעָ֔ד ‘to Gilʿad’

Num. 21.22; Or 1080.A4.18: לֵהָבָ֫ה for לֶהָבָ֫ה ‘flame’ Num. 21.28).

All but one (T-S NS 248.9) have *pataḥ/qameṣ* interchange (T-S NS 248.16: גַ֞ד for גָ֞ד ‘Gad’ [Num. 26.15]; Or 1080.A4.18: וַיִּירָ֫שׁ for וַיִּירַ֫שׁ ‘and he seized’ Num. 21.24, and בַשַּׁ֑מֶן for בַשָּׁ֑מֶן ‘of oil’ Num. 28.28).

There is a slight tendency to interchange *ḥireq* with *shewa* and *ḥireq* with *ṣere* (Or 1080.A4.18: וְנֵשְׁעַ֑ן for וְנִשְׁעַ֑ן ‘leaning’ Num. 21.15).

5.1.2. Byzantine Trio Outlier: T-S Misc 2.75 (Schema Patterns X, Y, 1a)

This manuscript was separated by the clustering algorithm from the aforementioned manuscripts because of its extremely high count of *dagesh* in *ʾalef* (66 times) and unexpected *dagesh* in ‘weak’ consonants (95 times). The manuscript, however, contains the full ‘Byzantine trio of features’ as well as two additional vowel interchanges. These are: *shewa* for *qameṣ* (חְפַ֖צְתִּי for חָפַ֖צְתִּי, ‘I [do not] want’ Deut. 25.8) and *ṣere* for *segol* (אֵ֣בֶן for אֶ֣בֶן ‘stone’ Deut. 25.13).

Num. 26.29, T-S NS 248.17 has (clearly) a *ḥireq* for a *pataḥ*: מִלְאַ֨ךְ for מַלְאַ֨ךְ ‘angel of’ (Num. 22.35). Thus they fit within schema patterns X and Y.

5.1.3. Orthoepy: NST use of Tiberian Vowel Graphemes for Orthoepic Purposes (No Schema Pattern)

These manuscripts use the non-standard placement of Tiberian *dagesh* and *mappiq* as orthoepic measures to ensure that weak consonants are correctly pronounced. Apart from a few sporadic vowel interchanges, the vocalisation of the manuscripts is otherwise ST and the pronunciation is ST with some orthoepic enhancements in the form of geminated weak consonants. The vowel interchanges are, for the most part, sign interchanges that do not represent a phonetic deviation from ST pronunciation.

- T-S A3.8: all /bgdkft/ letters in this manuscript without *dagesh* have *rafe*. Quiescent *ʾalef* takes *rafe* (for example, וַיֹּאֿמֶר ‘and he said’ Lev. 10.3, etc.), but consonantal *ʾalef* does not have *dagesh*. Three times the scribe reinforces ‘weak’ consonants (sibilants and sonorants) after a vowel with *dagesh* (קֳדָשִּֽׁים for קֳדָשִֽׁים ‘holies’ Lev. 10.12; בְּמָּקֹ֣ום for בְּמָקֹ֣ום ‘in [the] place’ Lev. 10.13; רָֽאשֵּׁיכֶ֣ם ‘your heads’ Lev. 10.6). *Mappiq* is marked in non-final consonantal *heh* (הִּ֔וא for הִ֔וא ‘she’ Lev. 11.6). The only vowel interchange is *ḥaṭef qameṣ* for *qameṣ* once (הֳעֵדָ֔ה for הָעֵדָ֔ה ‘the community’ Lev. 10.17).
- T-S AS 66.179: this is an Italian-Byzantine manuscript that exhibits extended use of *dagesh* in only a few instances: once in *lamed*, and twice in *ʿayin* (הַצְּעִּירָ֑ה ‘the younger’ Gen. 19.31), and once in *qof* (נַשְׁקֶּ֫נּוּ ‘let us drink’ Gen. 19.32). *Dagesh* also occurs on a ‘weak’ consonant at the end of the word after a vowel (אֲנִ֥י־אֵ֣לּ ‘I am God’ Gen.

17.1) and also in word-final *mater lectionis yod* in רֳאִ֑י 'seeing' Gen. 16.13). The first *heh* of the Tetragrammaton takes *mappiq* in two cases. Word-final *shewa* occurs twice in *ʿayin*.[54] This manuscript has sporadic sign interchanges: once *pataḥ* is substituted for *ḥaṭef pataḥ* (חַשֵׁכָ֥ה for חֲשֵׁכָ֥ה 'darkness' Gen. 15.12), and twice *pataḥ* is used in place of *ṣere* (תִקָֽבַ֖ר for תִקָֽבֵ֖ר 'you will be buried' Gen. 15.15 and קָֽדַ֖שׁ for קָֽדֵ֖שׁ 'holy' Gen. 16.14). Despite the minor vowel interchange, the holistic picture indicates a basic ST pronunciation with orthoepic features.

5.1.4. Orthoepic Group Outlier: T-S AS 64.206 (No Schema Pattern)

This Italian-Byzantine manuscript has features inherently connected to the orthoepic group. Its features, however, are not sporadic, but rather systematic. The comprehensive details of this manuscript are published elsewhere.[55]

[54] The manuscript does not have *dagesh* in *ʾalef*, and so it does not belong in the 'Byzantine Trio' group.

[55] I give a comprehensive overview of this manuscript in my Genizah Fragment of the Month article, April 2019: http://www.lib.cam.ac.uk/collections/departments/taylor-schechter-genizah-research-unit/fragment-month/fotm-2019/fragment-2

5.1.5. Orthoepic Group Outlier: T-S NS 248.23 (Schema Pattern 1—minimal)

This Italian-Byzantine manuscript is associated with the orthoepic group because it has one orthoepic NST feature: the placement of *dagesh* in every consonantal *ʾalef* (and *rafe* placed over every quiescent *ʾalef*). It is unique because the NST features are otherwise minimal. Examples: אֱּלֹהֶ֯יךָ ‘your god’ (six times), כַּאֲּשֶׁר ‘as’ (twice), אֶּת object marker (ten times). An example of *rafe* over quiescent *ʾalef*: הֻוּאֿ ‘him’ (verso, col. 1, line 8). Three times *dagesh* is placed in *lamed* in word-initial position after a vowelless consonant to strengthen it (לְֹּא ‘no’). There are only two vowel interchanges: one instance of *ḥireq* for *ḥolem* (לִשְׁמִ֝ר for לִשְׁמֹ֝ר ‘to keep’ Deut. 13.19) and one of *qameṣ* for *pataḥ* (מַעְשָׂ֝ר for מַעְשַׂ֝ר ‘tithe’ Deut. 14.28). It is ‘orthoepic’ in nature and placed with this particular group because it marks consonantal versus quiescent *ʾalef*.

5.1.6. ST Codices with Lexically-Specific NST features (No Schema Pattern)

This group is the most standard of the two-column manuscripts. It consists of those manuscripts which contain a few one-off NST features that do not form a particular pattern, alongside one NST feature that occurs in a lexically-specific pattern on only one word throughout. This feature is the placement of *shewa* for *ḥaṭef segol* on the word אֱלֹהְ֯ים ‘God, gods’. This probably does not represent a difference in pronunciation, particularly as all the other vowels are all represented with ST orthography. These manuscripts are both Oriental (Egypt-Palestine) in their palaeography. The manuscripts in this group are:

- T-S NS 68.22 (אְלֹהֶ֗יךָ ‘your god’ three times,)[56]
- T-S NS 78.47 (אְלֹהִ֔ים ‘God, gods’ three times).

5.1.7. Three-Way Interchange: *ṣere-segol-pataḥ* (Pattern 2b, X)[57]

These manuscripts all present this three-way interchange and lack interchanges with *qameṣ*. They also have *ḥaṭef* vowel sign interchanges which are not phonetic. but only notational. They exhibit Palestinian and Byzantine palaeography. The most noteworthy member of this group is:

T-S AS 67.133: Vowel interchanges: *pataḥ-segol* (once: יֵרָ֨אַה for יֵרָ֨אֶה ‘he will appear’ Deut. 16.16); *segol-pataḥ* (once: וּבִשֶׁלְתָּ֨ for וּבִשַּׁלְתָּ֨ ‘you cook’ Deut. 16.7), *segol-ṣere* (מַעֲשֶׂה for מַעֲשֵׂה, ‘deed’ Deut. 14.29); *ṣere-segol* (יִהְיֵה֙ for יִהְיֶה֙ ‘he will be’ three times); *ṣere-pataḥ* (בֵּבָּקָ֖ר for בַּבָּקָ֖ר ‘for cattle’ Deut. 14.25). Sign interchanges: *pataḥ-shewa* (גַּבֻלְךָ֖ for גְּבֻלְךָ֖ ‘your border’ Deut. 16.4) and vice versa (בְּשָּׁנָ֖ה for בַּשָּׁנָ֖ה ‘in the year’ Deut. 14.28) *segol ḥaṭef-segol* (lexically-specific: אֶלֹהֶ֫יךָ ‘your God’ 23 times); *pataḥ ḥaṭef-pataḥ* (consistent); *ḥaṭef pataḥ-shewa* (בְּאָ֨סְפֲּךָ֔ for בְּאָ֨סְפְּךָ֔, ‘when you gather’ Deut. 16.13); *ḥaṭef pataḥ-pataḥ* (once).

[56] This manuscript also has an unexpected *mappiq* in מַעֲשֵׂ֖הּ ‘work of’ Deut. 28.12.

[57] Other members: T-S Misc 1.46 (very Oriental script); T-S A4.3.

5.1.8. The outlier: Lewis Gibson Bible 1.75 (Schema Pattern 2, X)

This manuscript is connected to the above three-way interchange group in that it has *pataḥ* and *segol* interchanges, but is an outlier because it lacks any interchange with *ṣere*, making it unique. Like the previous group, it lacks *qameṣ* interchange and has a high level of non-phonetic sign interchange. Vowel interchanges: *pataḥ-segol* (חַלְקַת for חֶלְקַת ‘portion’ Gen. 27.16); *segol-pataḥ* (once: נֶפְתָּלִי for נַפְתָּלִי ‘Naftali’ Gen. 30.8); *shewa-segol* (once: תְּבָרְכְךָ for תְּבָרֶכְךָ ‘my soul may bless you’ Gen. 27.25); *pataḥ-ḥaṭef segol* (הַוֵה for הֱוֵה ‘be’ Gen. 27.29). Sign interchanges: *pataḥ-ḥaṭef pataḥ* (25 times); *pataḥ-shewa* and *shewa-pataḥ* (once each); *segol-ḥaṭef segol* (lexically specific: אֱלֹהֶיךָ ‘your God’ five times).

5.2. Three-column Manuscripts: Non-Standard Linguistic Typology

The main difference between the two-column manuscript data and the three-column data is that manuscripts in the two-column corpus tend to have small, discrete counts of features with a moderate number of vowel interchange. The three-column corpus has a few manuscripts with extremely high counts of one or two types of vowel interchange. It also has manuscripts with complex patterns of vowel interchange, while the two-column corpus tends to have simpler interchange patterns. Because of these outliers and complexity, I relied only on the k-modes algorithm, as it is less affected by high or low feature counts.

The main groups found were: (1) the Minimal Application group: one group of one manuscript with very minimal, lexically specific NST; (2) The Orthoepic group: manuscripts which mainly used NST features to reinforce ST pronunciation (alongside some vowel interchange possibly indicative of a Palestinian Hebrew substrate); (3) the two-way interchange group fitting with Schema 2; (4) the three-way interchange group fitting Schema 2b; (5) the three-way interchange group fitting Schema 2c; (6) the five-way interchange group fitting Schema 2e; (7) the largest outlier, which fit Schema 2d.

5.2.1. Minimal Application of NST

Unlike the two-column group, there is only one manuscript in the three-column group that has a minimal application of NST: T-S NS 76.32 (Italian-Byzantine). It only has the lexically-specific application of *shewa* for the *ḥaṭef segol* in אְלֹ֒הִים for אֱלֹ֒הִים 'God, gods' eight times).

5.2.2. Orthoepic Features with Interference from a Palestinian Substrate[58]

The manuscripts in this group tend to have some orthoepic use of *dagesh*, alongside vowel interchanges reflecting a Palestinian type of pronunciation, as well as sign interchanges involving *shewa* and *ḥaṭef* vowels.

Noteworthy manuscripts in this group include:

[58] Other members: T-S NS 248.20; T-S NS 248.12; T-S NS 248.2 (regularly places *dagesh* in word-final *ʾalef*; T-S NS 75.8 (occasionally places *dagesh* in *qof* and *ʿayin* (for example, עִּמּֽוֹ for עִמּֽוֹ 'with him' Gen. 32.7, קָּטֹ֙נְתִּי for קָטֹ֙נְתִּי 'I am unworthy' Gen. 32.11); T-S A2.30; Or 1080.A3.21 (Patterns X, 1); T-S NS 283.23; T-S A5.12.

T-S NS 248.18 (Schema Patterns Y, 1):

Dagesh occurs in ‘weak’ consonants after vowelless consonants and in consonantal *ʾalef*. *Pataḥ* for *qameṣ* occurs twice (for example, מִסְפַּ֣ר for מִסְפָּ֣ר ‘number’ Num. 9.20). *Ḥireq* for *shewa* occurs once (יִהִיֶ֖ה for יִהְיֶ֖ה ‘he will be’ Num. 9.21). It is not, however, a perfect fit with Pattern 1: it lacks a *segol-ṣere* interchange.

T-S NS 78.34 (Schema Patterns X, 1a):

This manuscript would belong to group 1a according the schema presented above. It is a fragment with a Palestinian-Byzantine script that has occasional use of *dagesh* to fortify weak consonants (but does not have *dagesh* in *ʾalef*). It exhibits the vowel interchange *segol* for *ṣere* (twice) and the sign interchange *shewa* for *ḥaṭef pataḥ* (twice).

T-S AS 67.131 (Schema Patterns X, 1a):

Pataḥ for *ḥaṭef pataḥ* (seventeen times), *pataḥ* for *shewa* (בַּעַֽרְבֹ֣ת for בְּעַֽרְבֹ֣ת ‘in the steppes [of Moab]’ Num. 26.3, reflecting the pronunciation of vocalic *shewa*; and וָמֲעַלָה for וָמֲעְלָה ‘and higher’ Num. 26.4, where ST has a silent *shewa*). *Shewa* for *ṣere* occurs once. *Ṣere* and *segol* interchange in both directions occurs three times.

Lewis-Gibson Bible 3.34 (Schema Patterns X, Y, 1a)

Occasional patterns of *dagesh/rafe* on *ʾalef* occur. Vowel interchange: *ṣere-segol*, regularly (אֶלְעָלֶ֑א for אֶלְעָלֵ֑א ‘Elʿale’ Num. 32.37; וַיֹּ֖ורֵשׁ for וַיֹּ֖ורֶשׁ ‘and he disposessed’ Num. 32.39). The following can be identified as sign interchanges: *pataḥ* with *ḥaṭef pataḥ*; *shewa* with *ḥireq* (סְיחֹ֜ן for סִיחֹ֜ן ‘Siḥon’ Num. 32.33).

T-S NS 77.1 (Schema Group 1):

On one occasion it shows use of *dagesh* in a weak letter after a vowelless consonant, i.e., in consonantal *ʾalef*, and multiple times on word-final consonantal *waw* (עֵשַׂ֣וּ for עֵשָׂ֣ו ‘Esau’ three times). Vowel interchange: *ṣere-segol* (35 times) and *segol-ṣere* (twelve times); *pataḥ-qameṣ* (ninety times); *qameṣ-pataḥ* (twice). There are also sign interchanges involving *ḥaṭef* vowels.

5.2.3. Two-Way Interchange: Schema Group 2[59]

The manuscripts here all have a very simple pattern of vowel interchange that fits into Schema Group 2, have very few orthoepic features, and often fail to put *dagesh* where expected. All of the manuscripts in this group have an Oriental (Egypt-Palestine) palaeography. Noteworthy members:

T-S A1.25:

This is an Oriental manuscript that interchanges *pataḥ* for *segol* (three times) and interchanges *segol* for *pataḥ* (once). The *naqdan* also places *shewa* with quiescent *ʾalef* (for example, לֵאְמֹ֣ר ‘saying’ twice). Once the *qere* is written rather than the *ketiv* (גוִֹים for גֹיִים, ‘nations’ Gen. 25.23).

T-S A2.1:

Pataḥ-ṣere occurs once (הַעַ֤דֹתָה for הַעֵ֤דֹתָה ‘you warned’ Exod. 19.23), but this is not consistent in the text and so does not form

[59] Other members: T-S NS 20.14; T-S NS 78.41.

a pattern. Instead, *pataḥ-segol* (including *ḥaṭef* vowels) interchanges much more regularly (six times). There is also the sign interchange *pataḥ-ḥaṭef pataḥ*.

5.2.4. Three-Way Interchange: Schema Group 2b[60]

These Oriental manuscripts are similar to the group above in that they have slight orthoepic features and many instances of missing *dagesh*, but they differ in that *qameṣ* is included in their vocalic interchange pattern.

T-S NS 24.16:

This has some orthoepic features, such as *dagesh* in weak letters after a vowel (e.g. *mem*: עִזִּ֣ים 'goats' Num. 29.25, *lamed*, *ʿayin*); also *dagesh/mappiq* in consonantal *yod* (כַּאֲרִ֖יּ 'as lions' Num. 24.9) and consonantal *ʾalef* (אֵּילִ֣ם 'rams' Num. 29.13). Normal use of *dagesh lene* and *forte* is mainly missing (absent 131 times). Vowel interchanges: *pataḥ-qameṣ*; *qameṣ-pataḥ*; *pataḥ-segol*. The following can be identified as sign interchanges: *shewa-ḥaṭef pataḥ*; *shewa-ḥireq* (שְׁנָ֣יְם for שְׁנָ֣יִם 'two' Num. 29.26).

T-S NS 18.5:

This Egyptian manuscript[61] has sporadic orthoepic features involving *dagesh* alongside an extensive pattern of vowel interchange. Vowel interchanges: *Qameṣ-segol* (וַיֹּ֣אמָר for וַיֹּ֣אמֶר 'and he said' Num. 14.41; תִּנֶּ֣גְפ֔וּ for תִּנָּ֣גְפ֔וּ 'you stumble' Num. 14.42).

[60] Other members: T-S NS 23.31; T-S AS 8.123; T-S NS 284.85

[61] I arrived at this conclusion upon consultation with Judith Olszowy-Schlanger.

Segol-pataḥ (אֶתֶּ֑ם for אַתֶּ֑ם ‘you’ [pl.] Num. 14.41) and *pataḥ-ḥireq* (לָאַ֙יַל֙ for לָאַ֙יִל֙ ‘for a ram’ Num. 15.6). There are also *ḥaṭef* vowel sign interchanges. Finally, the scribe places *shewa* on *ʾalef* to close a syllable (חָטָ֥אְנוּ for חָטָ֥אנוּ ‘we have sinned’ Num. 14.40).

5.2.5. Three-Way Interchange Outlier (Schema Pattern 2a, X):

Lewis Gibson Bible 3.12:

This manuscript is an outlier which is connected to the Group 2 interchange manuscripts in that it exhibits pattern 2a, but is separate because it places *shewa* at the end of the word to close the syllable 37 times on many letters: *lamed*, *taw*, *mem*, *resh*, *heh*, *dalet*, *ʿayin* (notable examples: נָשִׂ֖יאְ for נָשִׂ֖יא ‘chief’ (Num. 7.42); פַּ֣רְ for פַּ֣ר ‘bull’; שֶׁ֥קֶלְ for שֶׁ֥קֶל ‘shekel’; syllable-initial *ʿayin* בֶּן־דְּעְוּאֵ֖ל for בֶּן־דְּעוּאֵ֖ל ‘son of Deuel’). Vowel interchange: *Qameṣ-pataḥ* once each (וּלְזֶ֣בָח for וּלְזֶ֣בַח ‘and for a sacrifice’ Num. 7:59 and הְשָׁלָמִים֮ for הַשְּׁלָמִים֮ ‘the peace offerings’ Num. 7.58). *Pataḥ-qameṣ* twice. *Pataḥ-segol* once (בֶּן־בָּקָ֔ר for בַּן־בָּקָ֔ר ‘of the herd’ Num. 7.51. Frequent sign interchange involving *ḥaṭef* signs, *pataḥ* and *shewa.*

5.2.6. Three-Way Interchange (Schema Pattern 2c)

T-S AS 66.52:

Egypt, post-eleventh c. Vowel interchanges: *segol-ṣere*, *shewa*, and *pataḥ* (one each); *qameṣ-shewa* (וּמָכָ֣רוּ for וּמְכָר֣וֹ ‘or selling him’ Deut. 24.7), and *ḥaṭef* vowel sign interchanges.

T-S A3.15:

Egypt, post-twelfth c. This manuscript has sporadic orthoepic features: *dagesh* in *ʾalef* and *mappiq* in non-final consonantal *heh* (once each). Vowel interchanges: *Ḥireq-segol*, *ḥireq-ṣere*, *pataḥ-segol* (five times), *segol-pataḥ* (once), *pataḥ-ṣere* (once: הַמִּזְֽבַּ[חַ] for הַמִּזְֽבֵּחַ 'the altar' Lev. 4.30).

5.2.7. Five-Way Interchange (Schema Pattern 2e)[62]

T-S A5.7:

An Egyptian manuscript. Dagesh in *ʿayin* occurs twice (צֹֽעַּר 'Zoar' Deut. 34.3; בְּעֵּינֶ֫יךָ 'in your eyes' (Deut. 34.4). Vowel interchanges: *ḥireq-pataḥ* (נִפְתָּלִ֔י for נַפְתָּלִ֔י 'Naftali' Deut. 34.2), *ḥireq-segol* twice (אִתְּנֶ֑נָּה for אֶתְּנֶ֑נָּה 'I will give it' Deut. 34.4); *pataḥ-qameṣ* (יַדָ֑ע for יָדָ֑ע 'he [does not] know' Deut. 33.9; also *qameṣ ḥatuf* קַדְקֹד for קָדְקֹד 'scalp' Deut. 33.20); *qameṣ-pataḥ* (וּבְגָאַ[בָ]תֹו for וּבְגַאֲוָתֹו 'and in his majesty' Deut. 33.26[63]); *pataḥ-segol*; *segol-ṣere*; *ṣere-segol* occurs twice (בַּרְזֵ֑ל for בַּרְזֶ֑ל 'iron' Deut. 33.25).

[62] Other members: T-S NS 67.20; Lewis-Gibson Bible 1.56.

[63] The *bet* was placed above the word as a substitute for the consonantal *waw* (see the verso, col. 3, line 18). This indicates that fricative *bet* had the same phonetic realisation as consonantal *waw*. This phenomenon is also seen in a Genizah manuscript of the Torah written by an unprofessional writer, i.e., a child or layman (determined by the unsophisticated nature of the handwriting): T-S A21.125, where the manuscript has הַֽחַבִילָ֔ה for הַֽחֲוִילָ֔ה 'Ḥavilah' (Gen. 2.11).

T-S NS 282.69:

This Italian-Byzantine manuscript has a few orthoepic features: *dagesh* occurs once in *mem* after a vowel and once in *ʿayin* after a vowel (בָּעֶּ֫רֶבּ for בָּעֶ֫רֶב ‘in the evening’ Deut. 16.6). *Dagesh* in consonantal *ʾalef* occurs once (וְכָאַּ֖יָּל ‘and as the deer’ Deut. 15.22). Vowel interchanges: *ṣere-segol* (פֵּ֣סַח for פֶּ֣סַח ‘Passover’ twice,); *shewa-segol* occurs once, as well as for *ḥaṭef* vowels; *qameṣ*-silent *shewa* occurs once (לְבָבָ֔ךְ for לְבָבְ֔ךְ ‘your heart’ Deut. 15.7, thereby adding a syllable to the word). *Qameṣ-pataḥ*, and *qameṣ-ḥolem* once (עֳנִ֑י for עֹ֑נִי ‘affliction’ Deut. 16.3). *Ḥireq-pataḥ* (אִתָּ֔ה for אַתָּ֔ה ‘you’ Deut. 16.11). *Shewa* occurs on the first *heh* of the Tetragrammaton.

5.2.8. The Major Outlier T-S NS 72.1 (Schema Pattern 2d)

This Egyptian manuscript (twelfth c.) was consistently placed alone in the clustering. It has the highest concentration of NST features of all the manuscripts. In twelve columns of text (with 30 lines per column), 454 words had NST features. The manuscript has seventeen different vowel interchanges (of varying distributions), but the main features are *pataḥ-ṣere-segol-qameṣ* all interchanging as allophones of /a/:

- *Pataḥ-shewa* (בַּעֲ֣ד for בְּעֲ֣ד ‘through’ Gen. 26.8)
- *Pataḥ-ṣere* (once: וַיַּ֥לֶךְ for וַיֵּ֥לֶךְ ‘and he went’ Gen. 26.17)
- *Pataḥ-ḥireq* (once פְּלַשְׁתִּ֑ים for פְּלִשְׁתִּ֑ים ‘Philistines’ Gen. 26.8)
- *Pataḥ-segol* (וַיִּישַׂם for וַיִּישֶׂם ‘and was set’ Gen. 24.33 *ketiv*)
- *Segol-ṣere* (זַרְעֶ֔ךְ for זַרְעֵ֔ךְ ‘your offspring’ Gen. 24.60)
- *Ṣere-segol* (עֵבֶד for עֶבֶד ‘servant’ multiple times)

- *Segol-qameṣ* (once, בָּאֱרֶץ for בָּאָרֶץ ‘in the land’ Gen. 26.22)
- *Qameṣ-segol* (once, אָל for אֶל ‘to’ Gen. 26.18)
- *Qameṣ-ḥaṭef qameṣ* (once, אָהָלֹו for אָהֳלֹו ‘his tent’ Gen. 26.25)
- *Qameṣ-pataḥ* (אָברָהָם for אַבְרָהָם ‘Abraham’ Gen. 24.59)
- *Qameṣ-shewa* (וְיִצָּחָק for וְיִצְחָק ‘and Isaac’ Gen. 24.62)
- *Ṣere-ḥireq* (once, כֵּי for כִּי ‘for’ Gen. 26.16)

5.3. Concluding Discussion: Linguistic Typology

The above typology for two- and three-column NST near-model Torah codex fragments from the Genizah collections in Cambridge University Library is virtually comprehensive. All of the subtypes established by the clustering, which assessed every near-model NST fragment with full dimensions in Cambridge which I found (a total of 55 fragments), are reported above, with descriptions of selected examples. A general schema of vocalic interchange patterns was constructed independently of the statistics, and it was generally found that the clustering complemented this general schema. The results indicate that certain patterns of vowel interchange may be indicative of a few separate phenomena:

- A striving to reproduce the pronunciation of ST, but doing so by using Tiberian vowel graphemes in a non-standard way (orthoepy).
- Lexically-specific NST features that occur in otherwise ST manuscripts, which are probably learned spellings particular to the scribe or to the community that produced the text.

- Sign interchange (specifically, *shewa* and *ḥaṭef* vowels, or vocalic *shewa* and *pataḥ*), which is only notational and does not represent a phonetic shift in vowels.
- Vocalic interchange patterns of varying degrees of complexity, often occurring alongside the non-standard use of diacritics such as *dagesh* or silent *shewa*, and which are likely to reflect pronunciations influenced by Aramaic or Arabic.

The most crucial finding uncovered by the clustering algorithms was that the feature frequencies differ between the two- and three-column manuscripts. This affected not only which clustering algorithm was most appropriate for the specific group, but the typology. Two-column manuscripts had the following general features:

- They exhibited on average a moderate amount of vocalic interchange, and the outlier manuscripts could usually be clearly tied to a specific group (or more than one specific group).
- Many of the manuscripts were either from the Southwestern Oriental (Palestinian-Byzantine) or Italian-Byzantine group.
- The pronunciation behind the vocalic interchange seemed to be associated with influence due to Aramaic language contact, as seen in the schema patterns.
- Orthoepic features that reinforced ST pronunciation in a non-standard way are associated with the two-column group.

The three-column group had the following different general features:

- Within this group were manuscripts with extreme counts of NST features, or extremely complex patterns of vocalic interchange, including the manuscript with the most NST features (T-S NS 72.1).
- The extremity of the outlying features indicated that only the k-modes algorithm was appropriate to assess the group statistically, because other clustering algorithms would be biased by the outliers.
- Patterns with extended use of *dagesh* were associated with the three-column group.
- The majority of the manuscripts in this group were clearly Oriental (Egypt and Palestine, especially twelfth c. Egypt). Moreover, the various patterns of vowel interchange seemed to be associated with the levelling of vowel phonemes, reflecting convergence with the Arabic vowel system.

The results indicate that two- and three-column manuscripts are distinct in their patterns of NST features. There are clear regional and language contact differences, which can be seen when comprehensive data are taken into account. Moreover, clustering, validated by rigorous linguistic assessment, is useful for the analysis of large amounts of NST features, especially when the researcher is careful not to perform the clustering on a large number of features at once. The coherency of the clustering results and the linguistic validation by means of the schemas supports the hypothesis that there are statistically and linguistically valid subtypes of NST vocalisation.

6.0. Conclusions: The Correlation between Codicology and Linguistic features

At the beginning of the study it was hypothesised that both the codicological and linguistic features of the near-model manuscripts in the Cambridge Genizah collections have clear subtypes that can be validated through statistical analysis, and this has been shown to be the case. It was, however, also hypothesised that linguistic patterns would generally correlate with codicological subtypes. This concluding section presents the data in support of the latter hypothesis and brings the study to a close with some final assessments concerning how to carry the analysis forward in future research.

6.1. The Correlation between Codicological and Linguistic Subtypes

In general, the linguistic patterns found above were distinct not only regarding differences between two- and three-column manuscripts, but also regarding the fact that manuscripts with similar linguistic patterns tended to group together in either the same codicological subgroup, or in related codicological subgroups:

6.1.1. Two-column Manuscripts

- Byzantine Trio pattern (Section 5.1.1). These manuscripts all came from various groups that exhibit a broad palaeographical relationship, which included Sephardi, Italian-Byzantine, and Palestinian-Byzantine manuscripts.
- Orthoepic, Nearly Standard (5.1.3, including outliers: 5.1.4 and 5.1.5). These manuscripts came from groups with the most diverse palaeographic regions, from groups with Sephardi manuscripts, to Oriental (Egypt-Palestine), and Byzantine groups. This may indicate that every region produced some nearly-standard, orthoepic manuscripts.
- Lexically-specific (5.1.6): These two manuscripts came from Monumental Oriental (Egypt-Palestine) groups: specifically, T-S NS 68.22 came from the 'Large Monumental Egyptian' group described in section 4, and T-S NS came from another Egyptian group which was not described as an example group in this study.
- Three-way Interchange: *ṣere-segol-pataḥ* (5.1.7). All of the manuscripts in this group came from Arabic-speaking regions, as their palaeography indicates areas ranging from Egypt to Palestine-Byzantine (Southwestern Oriental) areas. Specifically, T-S Misc 1.46 is a late Egyptian manuscript from the subgroup Small Oriental Codex.
- Two-column Outliers (5.1.8 and 5.1.9): as these manuscripts were all outliers, they all came from different regional groups.

6.1.2. Three-column Manuscripts

- Minimal Application of NST (5.2.1). This manuscript with a Southwestern Oriental (Palestinian) script type came from a group with other Italian-Byzantine and Palestinian-Byzantine manuscripts.
- Orthoepic Features (5.2.2). All of these manuscripts were from Monumental groups, mainly from Egypt. Those represented in the sample codicological subgroups above are: T-S NS 78.34, T-S AS 67.131, T-S NS 77.1 (Large Monumental Egypt-Palestine Group); T-S NS 248.2 (Monumental Levantine Codex Group); T-S A2.30 (a late Egyptian manuscript from the Monumental Bare Wide-Ranging [Oriental to Italian] Group); T-S A5.12 (the Monumental Oriental Group).
- Two-way Interchange: Schema Group 2 (5.2.3). All of the manuscripts in this group came from either the Monumental Oriental Group (T-S A1.25) or the Large Monumental Levantine Codex Group (T-S A2.1, T-S NS 20.14), or other closely-related Egyptian Monumental groups not exhibited above.
- Three-way Interchange: Schema Group 2b (5.2.4 and 5.2.5). The manuscripts all came from Arabic-speaking regions. They involve the Egypt-Palestine Monumental groups, one of which is represented in the examples above: Monumental Bare Oriental Codex (T-S NS 18.5, T-S AS 8.123). The outlier Lewis Gibson Bible 3.12 comes from the Square Monumental Egyptian-Palestinian codex (not reported in section 4).

- Three-way Interchange, Schema Pattern 2c (5.2.6). Also from Arabic-speaking regions. Both manuscripts, are in different, but related, Oriental groups and are both late Egyptian in their palaeography (T-S AS 66.52: post-eleventh c.) and T-S A3.15 (post-twelfth c.).
- Five-way Interchange: Schema Pattern 2e (5.2.7). The manuscripts from this group belong to wide-ranging regions, mainly from Arabic-speaking areas. T-S A5.7 and T-S NS 67.20 belong to the Square Monumental Oriental Group (not reported as example); the Lewis-Gibson Bible 1.56 is an Egyptian-Palestinian (Northwestern Oriental) manuscript from the Monumental Oriental Group; T-S NS 282.69 is in the Monumental Bare Wide-Ranging Group.
- Finally, the major outlier, T-S NS 72.1, is in the same group as T-S A5.7, T-S NS 67.20, and Lewis Gibson Bible 3.12, which are all Egypt-Palestinian in their palaeography.

The results of these general correlations show that, while linguistic features do co-occur in patterns alongside codicological subtypes, these co-occurrences are in wider regional swaths of similarity. It is also to be noted that the specific date of the scripts was not a major factor in this study. Apart from a few late manuscripts that grouped together, further analysis may refine these correlational findings by clarifying the palaeographic date of the manuscripts. It can safely be said, however, that subtypes of NST can be regionally defined and generally correlate with regional patterns of codicology.

6.2. Final Conclusions

The analysis in this paper is, to date, the most comprehensive assessment of a large number of manuscripts on many grounds: both codicological and linguistic. It has introduced a new methodology that allows the researcher to analyse effectively thousands of individual data points and 296 manuscript fragments. The results clarify our understanding of near-model and NST vocalisation phenomena in the Genizah.

Firstly, it can be affirmed that near-model manuscripts exist as a conceptual category of codex type within the Genizah, and that, when considered as parts of larger groups, those with two columns are distinct, both codicologically and linguistically, from those with three columns. These kinds of manuscripts represent the threshold of the standard, exquisite Bibles, which have been the focus of scholarship, and show that rich diversity lies just below the surface of what has been analysed in the past.

Secondly, it has been demonstrated that codicology can be regionally defined and that styles of book-making practices and scribal habits differed slightly (and in a statistically verifiable way) from region to region in the Genizah. Most importantly, dimensions and line number are the most reliable measures for distinguishing differences in codicological styles across regions.

Thirdly, NST can be considered a hypernym for what is in fact an internally diverse phenomenon with distinct subtypes. These subtypes can represent many things, ranging from an adherence to the pronunciation of the ST text (but non-adherence in notation), to a completely different phonological profile, which is most likely due to language contact and regional pronunciations of Biblical Hebrew in Egypt, the Levant, Asia Minor, and Italy.

Finally, this study has shown that language and codicological features complement each other and, when studied together, can aid the researcher in understanding the larger picture of the background of the manuscript. Since codicological styles varied by region, and since NST language features also varied by region, codicology and language can indeed be used to help clarify each other. This demonstrates that medieval Hebrew manuscripts are holistic entities, which, in order to be studied properly, must have both their physicality and their language features taken into account

This study is a first, exploratory step in using the methodology that I have developed here. The methodology should be applied to other groups of manuscripts in order to refine it properly, to find pitfalls, and to calibrate it for further improvements of analysis. It has great potential to allow scholars to look at the wider picture of a corpus of manuscripts without sacrificing detail. Furthermore, statistical clustering puts the researcher above the data and allows for the prioritisation of the most critical data and details.

Avenues for future research include applying this same analysis to other groups of non-standard Hebrew Bible codices (which is the topic of my current PhD research[64]), as well as re-

[64] Working title: "A Codicological and Linguistic Typology of Non-standard Torah Codices from the Cairo Genizah.

fining the typology presented above by means of further investigation into specific aspects. These include patterns of Masorah, cantillation, or, especially, the extreme outliers identified in this paper. In any case, it is hoped that the present study has not only opened conceptual doors to further bolster our study of medieval Jewish manuscripts, but has also introduced a new methodology and set of tools by which to do so.

7.0. References

Amrhein, Valentin, Sander Greenland, and Blake McShane. 2019. ‘Scientists Rise Up against Statistical Significance’. *Nature* 567: 305–7. https://www.nature.com/articles/d41586-019-00857-9

Arrant, Estara, J. 2019. ‘Standard Tiberian Pronunciation in a Non-Standard Form: T-S AS 64.206’. Cambridge University Library: Genizah Research Unit *Fragment of the Month:* April 2019. https://www.lib.cam.ac.uk/collections/departments/taylor-schechter-genizah-research-unit/fragment-month/fotm-2019/fragment-2

Beit-Arié, Malachi, Edna Engel, Ada Yardeni, and Comité De Paléographie Hébraïque. 1987. *Asupot Ketavim ʿIvriyim Mi-Yeme-ha-Benayim.* 3 vols. Jerusalem: The Israel Academy of Sciences and Humanities.

Birnbaum, S.A. 1971. *The Hebrew Scripts.* 2 vols. Leiden: Brill.

Blapp, Samuel. 2017. ‘The Non-Standard Tiberian Hebrew Language Tradition according to Bible Manuscripts from the Cairo Genizah’. PhD dissertation, University of Cambridge.

Chaturvedi, Anil, Paul E. Green, and J. Douglas Caroll. 2001. 'K-modes clustering'. *Journal of Classification* 18 (1): 35–55.

Cheng, Yizong. 1995. 'Mean Shift, Mode Seeking, and Clustering'. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 17 (8): 790–99.

Cohen, Zina, Judith Olszowy-Schlanger, Oliver Hahn, and Ira Rabin. 2017. 'Composition Analysis of Writing Materials in Geniza Fragments'. In *Jewish Manuscript Cultures: New Perspectives,* edited by Irina Wandrey. Studies in Manuscript Cultures 13. Berlin: De Gruyter.

David, Ismar. 1990. *The Hebrew Letter: Calligraphic Variations.* Northvale, NJ: Jason Aronson Press.

Davis, M. C., Henry Knopf, and Ben Outhwaite. 1978. *Hebrew Bible Manuscripts in the Cambridge Genizah Collections.* Cambridge University Library Genizah Series 1–4. Cambridge: Cambridge University Press.

Díez Macho, Alejandro. 1971. *Manuscritos Hebreos y Arameos de la Biblia: Contribución al Estudio de las Diversas Tradiciones del Texto del Antiguo Testamento*. Studia Ephemeridis Augustinianum 5. Rome: Institutum Patristicum Augustinianum.

Fassberg, Steven, E. 1991. *A Grammar of the Palestinian Targum Fragments from the Cairo Genizah.* Harvard Semitic Studies 38. Leiden: Brill.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2015. *An Introduction to Statistical Learning: With Applications in R*. Springer Texts in Statistics 103. New York: Springer Science & Business Media.

Khan, Geoffrey. 2017. 'The Background of the So-Called 'Extended Tiberian' Vocalisation of Hebrew'. *Journal of Near Eastern Studies* 76 (2): 265–73.

———. 2020. *The Tiberian Pronunciation Tradition of Biblical Hebrew: Including a Critical Edition and English Translation of the Sections on Consonants and Vowels in the Masoretic Treatise Hidāyat al-Qāriʾ ‘Guide for the Reader.’* 2 vols. Cambridge Semitic Languages and Cultures 1. Cambridge: University of Cambridge and Open Book Publishers.

Kodinariya, Trupti M. and Prashant R. Makwana. 2013. ‘Review on Determining Number of Cluster in K-Means Clustering’. *International Journal of Advance Research in Computer Science and Management Studies* 1 (6): 90–94.

MacQueen, James. 1967. ‘Some Methods for Classification and Analysis of Multivariate Observations’. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, 281–97. Berkeley: University of California Press.

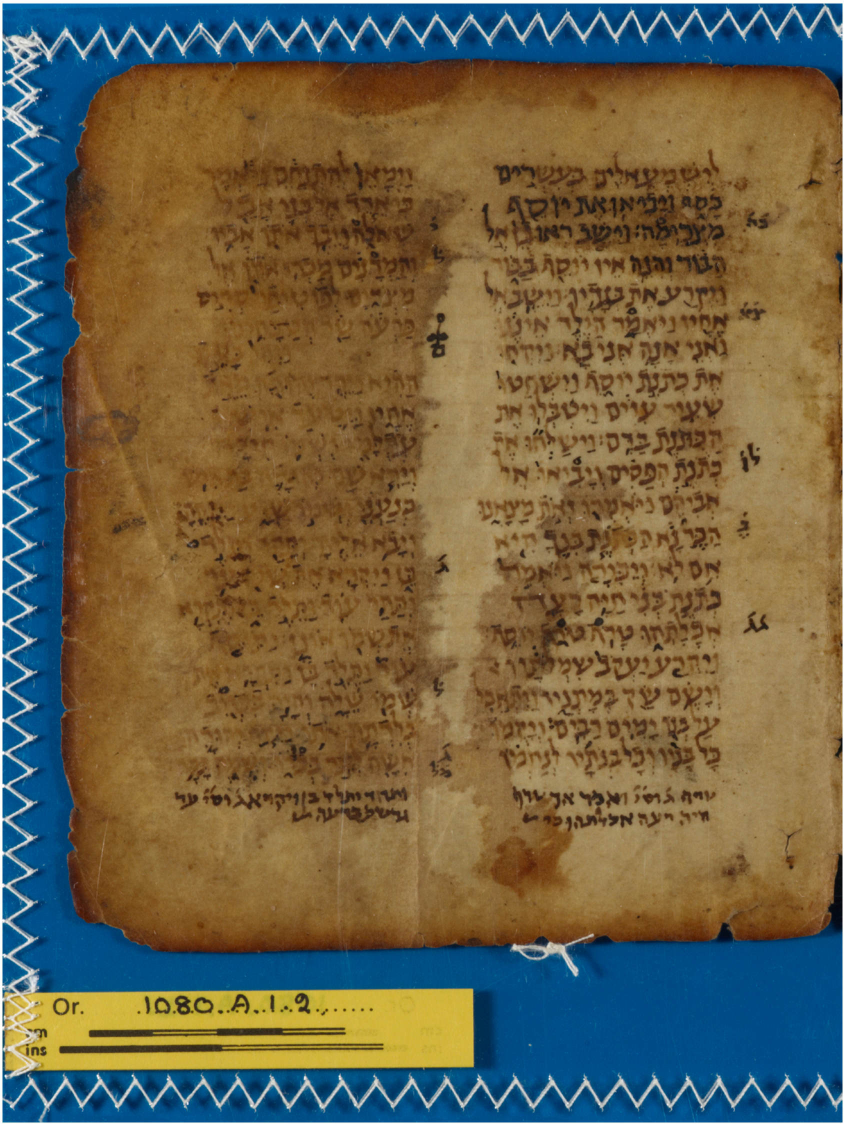
R Core Team. 2018. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Sáenz-Badillos, Angel. 2008. *A History of the Hebrew Language*. Cambridge: Cambridge University Press.

Sirat, Colette. 2002. *Hebrew Manuscripts of the Middle Ages*. Translated by Nicholas R. de Lange. Cambridge: Cambridge University Press.

Yardeni, Ada. 2002. *The Book of Hebrew Script: Palaeography, Script Styles, Calligraphy and Design*. London: New Castle.

Yeivin, Israel and E. J. Revell. 1980. *Introduction to the Tiberian Masorah.* Masoretic Studies 5. Missoula, MT: Scholars Press.

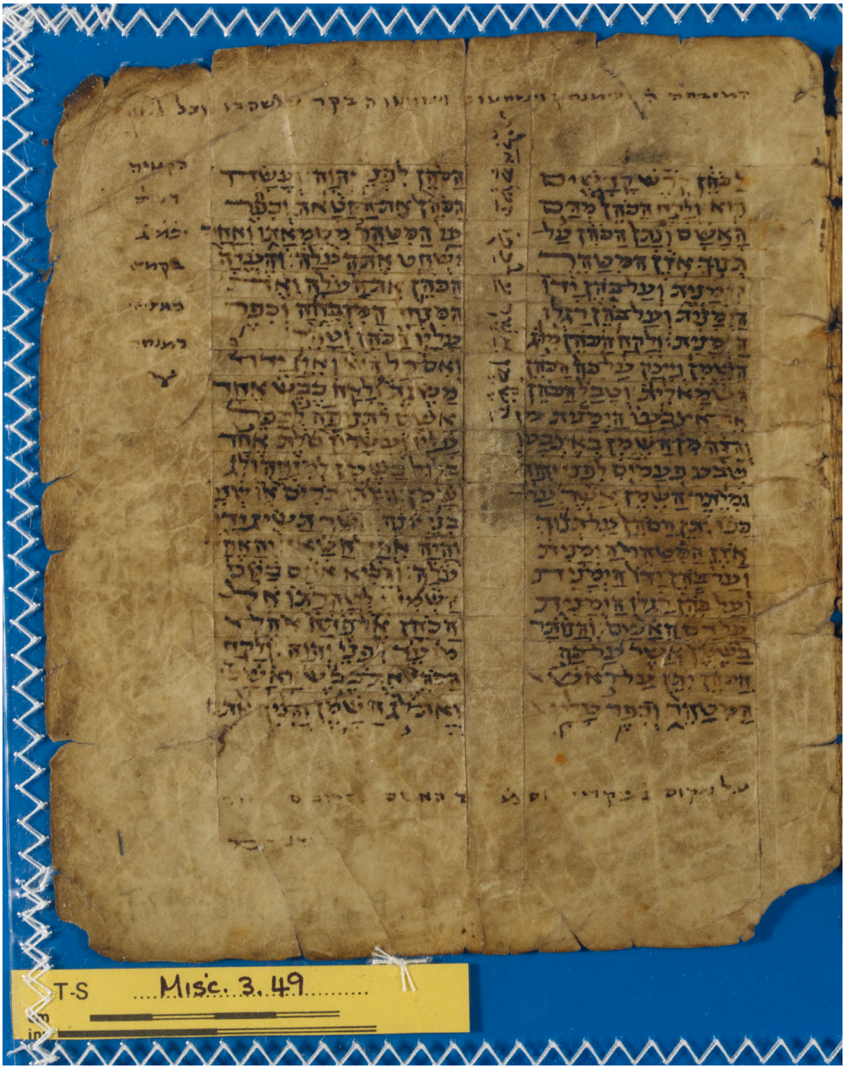


וימאן להתנחם ויאמר
כי ארד אל בני אבל
שאלה ויבך אתו אביו
והמדנים מכרו אתו אל
מצרים לפוטיפר סריס
פרעה שר הטבחים
[illegible]

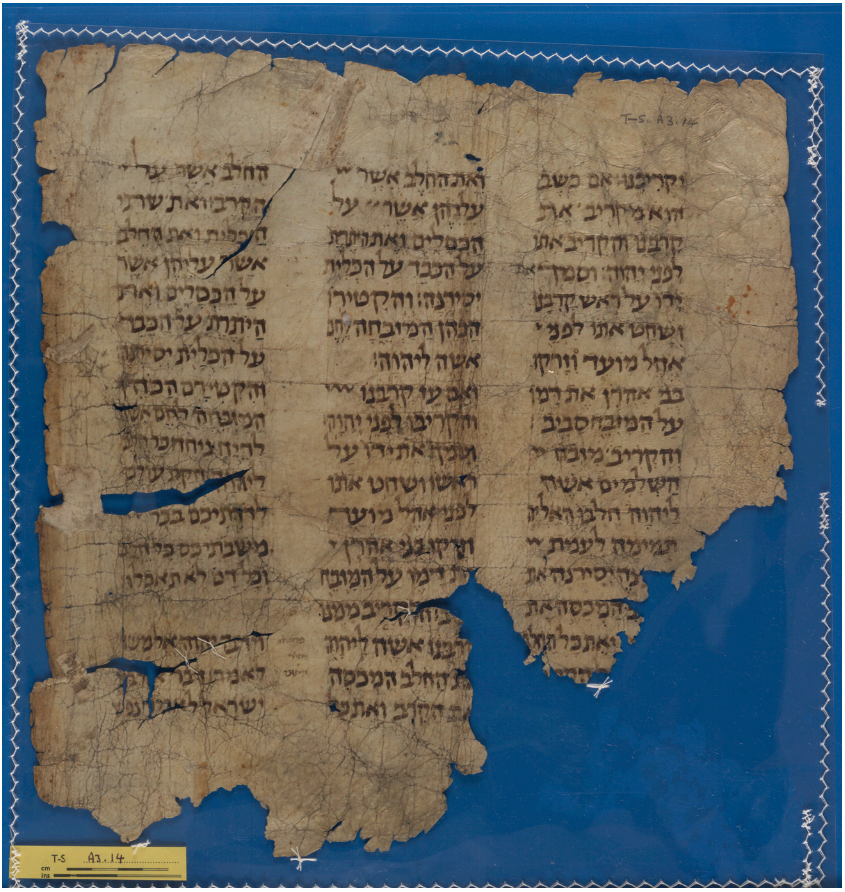
לישמעאלים בעשרים
כסף ויביאו את יוסף
מצרימה: וישב ראובן אל
הבור והנה אין יוסף בבור
ויקרע את בגדיו: וישב אל
אחיו ויאמר הילד איננו
ואני אנה אני בא: ויקחו
את כתנת יוסף וישחטו
שעיר עזים ויטבלו את
הכתנת בדם: וישלחו את
כתנת הפסים ויביאו אל
אביהם ויאמרו זאת מצאנו
הכר נא הכתנת בנך היא
אם לא: ויכירה ויאמר
כתנת בני חיה רעה
אכלתהו טרף טרף יוסף
ויקרע יעקב שמלתיו וי
וישם שק במתניו ויתאבל
על בנו ימים רבים: ויקמו
כל בניו וכל בנתיו לנחמו

Cambridge University Library Or 1080.A.1.2.

Small Italian-Byzantine Codex



Cambridge University Library T-S Misc 3.49
Small Italian-Byzantine Codex

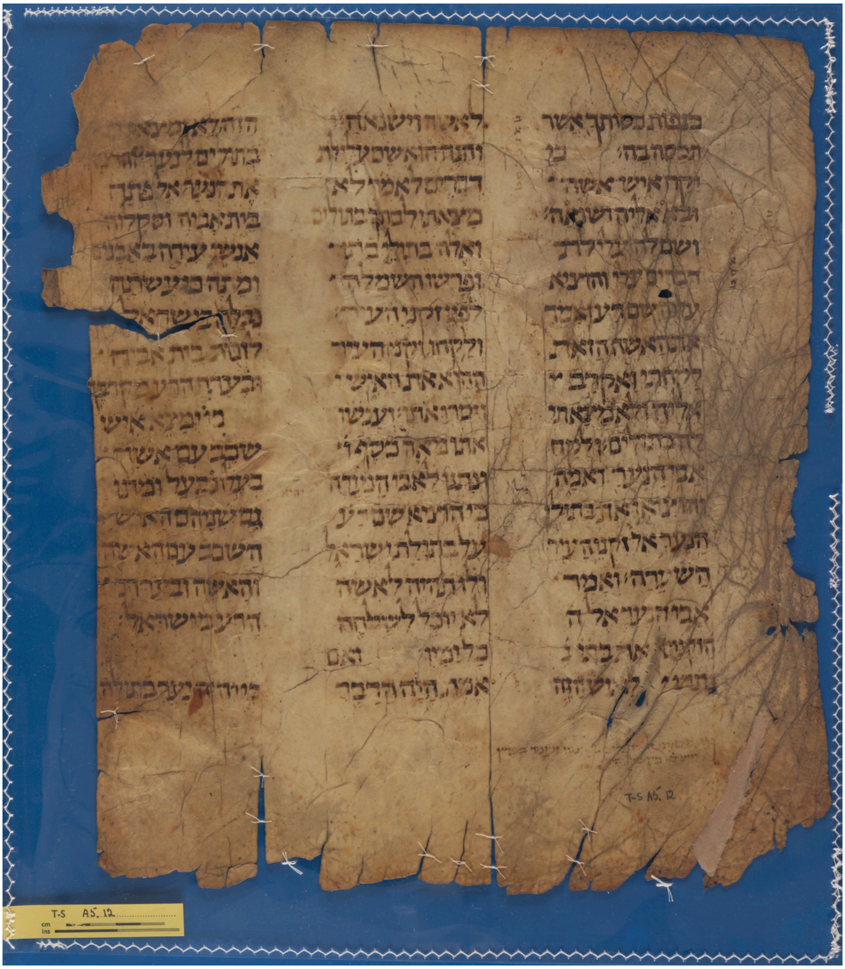


יקריבנו אם כשב
הוא מקריב את
קרבנו והקריב אתו
לפני יהוה וסמך את
ידו על ראש קרבנו
ושחט אתו לפני
אהל מועד וזרקו
בני אהרן את דמו
על המזבח סביב
והקריב מזבח
השלמים אשה
ליהוה חלבו האליה
תמימה לעמת
העצה יסירנה ואת
החלב המכסה את
ואת כל החל

ואת החלב אשר
עלהן אשר על
הכסלים ואת היתרת
על הכבד על הכלית
יסירנה והקטירו
הכהן המזבחה לחם
אשה ליהוה
ואם עז קרבנו
והקריבו לפני יהוה
וסמך את ידו על
ראשו ושחט אתו
לפני אהל מועד
וזרקו בני אהרן
את דמו על המזבח
והקריב ממנו
קרבנו אשה ליהוה
החלב המכסה
הקרב ואת

החלב אשר על
הקרב ואת שתי
הכלית ואת החלב
אשר עליהן אשר
על הכסלים ואת
היתרת על הכבד
על הכלית יסירנה
והקטירם הכהן
המזבחה לחם אשה
לריח ניחח כל חלב
ליהוה חקת עולם
לדרתיכם בכל
משבתיכם כל חלב
וכל דם לא תאכלו
וידבר יהוה אל משה
לאמר דבר אל בני
ישראל לאמר נפש

Cambridge University Library T-S A3.14
Monumental Oriental Codex



Cambridge University Library T-S A5.12
Monumental Oriental Codex