Towards a Digital Critical Edition of the Mishnah*

Hayim Lapin

This essay describes a project to develop a "born-digital" critical edition of the Mishnah.[1] The Mishnah is a compendium of rabbinic laws, compiled in about 200 CE, running about 188,000 words.[2] It is an important text for several reasons. As a heritage text of Judaism, it is regularly studied, either alone or as part of the Talmud, by hundreds of thousands of people. As a historical document, by scholarly consensus it is the first rabbinic document to reach something like its present form, so that it serves as a benchmark for the emergence of rabbis as religious and intellectual movement in Palestine. As a document of ethnically specific, and often utopian, laws by a group of Jews from Palestine under Roman rule, it is also an important document of Romanization or its absence in the late second century.

The Digital Mishnah will provide a dynamic edition of the Mishnah that takes advantage of the digital medium to provide multiple and customizable presentations of the text, as well as analytical tools that will allow the user to study variability between witnesses as well as other features. A great

* I am pleased to acknowledge the assistance of Travis R. Brown, Tim Finney, Shamma Y. Friedman, Leor Jacobi, Trevor Muñoz, Desmond Schmidt, Daniel Stoekl Ben Ezra, and, as always, Maxine Grossman.

[1] Abbreviations of Mishnah witnesses referred to below:

Kauf	Library of the Hungarian Academy of Sciences, Kaufmann A50 ParmA De Rossi 138–39
Leid	Leiden Scaliger Or. 4720, 1289
Camb	Cambridge Add. 470.1
G1	Cairo Genizah T-S E1.99 f. 1r–2v (Cambridge)
Maim	Maimonides Autograph MS
Nap	Mishnah, *Editio Princeps*, Naples 1492
Paris	Bibliothèque nationale de France, MS Hébreu 328–29, 230b–31b Mun Bayerische Staatsbibliothek Handschrift Cod.Hebr.95 (1342)
Hamb	Cod. Hamburg. 165.

A working demo, with limited capabilities, is available at http://dev.digitalmishnah.org. Github serves as the repository for the files of the digital Mishnah project: http://github.com/umd-mith/mishnah.

[2] Academy for the Hebrew Language, "The Historical Dictionary Project" [Hebrew], http://hebrew-treasures.huji.ac.il/.

proportion of Peter Schäfer's work has been dedicated to text editing, and to elucidating the knowledge that comes of – or is destabilized by – rigorous editorial work. It seems only fitting to offer this preliminary description as a tribute.

Is There a Text in This Corpus?

Despite its importance, no scholarly edition of the Mishnah has been produced. The reasons for this are complicated, and certainly have to do with the inherent difficulty of the task. The gap between compilation and the first manuscripts is at least seven or eight centuries, with clear evidence for scholarly intervention in the text in Talmudic discussion. There are numerous still unpublished fragments, and many citations and paraphrases in medieval commentaries and early modern commentaries. The cultural and religious significance of the text, when compounded with the technical difficulties, probably has served as an inhibitor as well.

Moreover, the field of rabbinic textual scholarship, like all fields of the humanities, has been shaped by the postmodern turn in recent decades. In many cases, this has meant a retreat from that most "scientizing" of classical fields, textual criticism, toward "readings" that are attentive to the textual evidence in varying degrees. This is especially true of North American scholarship. Among those primarily interested in textual developments, even among the least explicitly postmodern, we have seen a tendency toward multiplicity in the presentation of the data. Critical editions of individual tractates in the last several decades have presented two texts of the Mishnah, representing two families, in parallel columns. Inspired, in part, by the significant work of Shamma Friedman, Israeli Talmudic scholarship "*ʿal derek ha-meḥqar*" presents the Talmudic text by witness in an alignment table. Peter Schäfer and his students and coworkers and followers have opted instead for a synoptic format, in which each witness is presented in a separate column.[3]

[3] D. Rosenthal, "Mishna Aboda Zara: A Critical Edition" [Hebrew] (PhD diss., Hebrew University, 1980); A. Goldberg, *Commentary to the Mishnah, Shabbat* [Hebrew] (Jerusalem: Jewish Theological Seminary of America, 1976); and Goldberg, *The Mishna Treatise Eruvin* [Hebrew] (Jerusalem: Jewish Theological Seminary of America, 1986) all present two texts. For the approaches of Schäfer and Friedman see, e.g., P. Schäfer, ed., *Synopse zur Hekhalot-Literatur* (Tübingen: Mohr Siebeck, 1981); P. Schäfer and H.-J. Becker, eds., *Synopse zum Talmud Yerushalmi*, 4 vols. (Tübingen: Mohr Siebeck, 1991–2001); and S. Friedman, *Talmud arukh, BT Bava meziʿa VI* [Hebrew] (Jerusalem: Jewish Theological Seminary of America, 1990). Both formats are called by their practitioners "synopsis"; I use alignment table for the one and synopsis for the other to distinguish between them.

In the case of Schäfer, the choice of format quite explicitly accompanies a theory of the nature of the text to be discussed.[4] From the starting point of the manuscripts of the Hekhalot corpus, Schäfer argued that what prior scholars had identified as "texts" lacked redactional identity. There were more or less discrete "microforms," themselves composed of smaller units, that could circulate independently, out of which the larger "texts," or "macroforms," were composed, but those larger macroforms often had no redactional coherence. In some cases, the very notion that there was a particular identifiable text was a modern scholarly construct.

A dominant stream of classical stemmatics is based on the principle of the "common" or "disjunctive" error developed by Maas.[5] Where witnesses A and B each include an error or a change that is not likely to have been created independently (e.g., an explanatory gloss identically worded) that others do not, they are to be placed on the same branch of the stemma, and the other witnesses on another. If B includes all such errors attested in A, but includes additional errors not present in A, then B is a descendant of A. The presence of common errors in both A and B alongside errors in each that are not attested in the other leads to the hypothesis of a common ancestor. From this it is possible to develop a descent tree, often associated with the name of Karl Lachmann.

It does not take a particularly deep inquiry to recognize that the method worked out by Maas is ultimately deeply dependent upon the critic's judgment about what errors could or could not be autonomous, and how to distinguish what is an "error" from what the author would have written. Maas, in fact, offers a brief exhortation to the study of style, "even if [the critic] realizes that one man's lifetime is not long enough to allow a real mastery in this field to reach maturity."[6] It is, I surmise, this kind of striving to mastery and the application of judgment that Schäfer characterizes, critically, as the "ideal of the true scholar who – sitting in his *Studierstube*, and by virtue of

[4] P. Schäfer, "Handschriften zur Hekhalot-Literatur," in Schäfer, *Hekhalot Studien* (Tübingen: Mohr Siebeck, 1988; originally published in *FJB* 11 [1983]: 111–93), 154–233, and a number of other essays republished in that collection. Schäfer, "Research into Rabbinic Literature: An Attempt to Define the *Status Quaestionis*," in *Rabbinic Texts and the History of Late-Roman Palestine*, ed. M. Goodman and P.S. Alexander (Proceedings of the British Academy 165; London: Oxford University Press, 2011; originally published in *JJS* 37 [1986]: 139–52), 51–65, extends this approach to Rabbinic literature. The latter should be read together with C.J. Milikowsky, "The *Status Questionis* of Research in Rabbinic Literature," in *Rabbinic Texts and the History of Late-Roman Palestine* (originally published in *JJS* 39 [1989]: 201–11), 67–78; Schäfer and Milikowsky, "Current Views," 78–88.

[5] P. Maas, *Textual Criticism*, trans. B. Flower (Oxford: Clarendon, 1958).

[6] Maas, *Textual Criticism*, 10. Cf. W.W. Greg, *The Calculus of Variants: An Essay on Textual Criticism* (Oxford: Clarendon, 1927), for whom qualitative judgment does not disappear, but for whom the collection of patterns of variation and the consideration of the formal logical determinants of relationship play a decisive role.

the critical faculty of his superb human mind – indeed constructed a better text than any of those versions preserved in his manuscripts."[7]

On this, I am generally in agreement with Schäfer. However, it should be admitted that this is not entirely an apt characterization of textual criticism as it is practiced today, which is rather open about the hypothetical character of the outcome and uses methods that explicitly invoke judgment and probability, even if the techniques are increasingly computational.[8] Still, even conceding that as a general principal that some prior, antecedent text is reconstructable, it is an open question what such a reconstruction would teach us about the Mishnah in particular.

Beginning from the manuscripts, as Schäfer did, the Mishnah as a macroform does not devolve into a loose conglomeration of microforms. There are several whole manuscripts of the text, and other sizable pieces, and numerous fragments from the Cairo Genizah that situate the text as an item to be copied, to be dated perhaps to between the tenth and twelfth centuries. Maimonides, contemporary with these manuscripts, wrote the first extant commentary to the entire Mishnah, but a geonic commentary to Toharot survives, and there are references to, and some quotations and fragments from, other geonic commentaries, although apparently not earlier than Seʿadya in the tenth century.[9]

Working backwards from the manuscripts and commentaries to the Talmuds, by the time of the redaction of the Babylonian and Palestinian Talmuds (itself problematic), at least five orders of the Mishnah existed in sufficiently stable form to serve as the organizing frame for the Palestinian and Babylonian Talmuds. It seems unlikely that the Mishnah was called into being through the redaction of the Babylonian Talmud,[10] although, depending upon one's understanding of the transmission of Palestinian tradition to Babylonia, one could make that case for the Yerushalmi. This would bring a substantial proportion of the Mishnah back to the turn of the fifth century, leaving the historical development of overarching structure (orders, and their organization) uncertain.[11] Unless we flatten each of the Talmuds to a single layer, it seems certain that both incorporated earlier traditions that know "our" Mishnah, as distinct from other "tannaitic" material quoted, and have attributed its redaction to Judah the Patriarch, a figure whose chronology fits with the internal chronology of the Mishnah itself.[12] I have argued elsewhere that sifting through the amoraic traditions in Yerushalmi tractate *Sheviʿit* suggests that the Mishnah assumed a particular significance in the traditions of Palestinian amoraim; that rise in significance can perhaps be dated to the later third century.[13]

During this earlier period, when there are no manuscripts to attest directly to the transmission, it is quite possible that the Mishnah circulated in a state analogous to that diagnosed for Schäfer for medieval Hekhalot texts. Certainly, discussions in the Talmuds imply differences in transition at that stage and attest to interventions into the text of the Mishnah in such forms as interpolation or recasting that have made their way into the manuscript tradition.[14] In addition, some scholars have argued that the entire corpus of rabbinic literature, and especially law, was transmitted orally, and others, furthermore, have cast the Mishnah and Tosefta as two crystallizations of oral performance.[15] Whatever the medium (I am agnostic about the first of these positions and dubious about the second), underlying the relationship of the Tosefta to the Mishnah may lie a significant period of textual instability.[16]

The individual tractates of the Mishnah are generally topically coherent, and one of Neusner's important contributions was to demonstrate that a few formal syntactical structures overwhelmingly accounted for the Mish-

[7] In Schäfer's portion of Schäfer and Milikowsky, "Current Views," 81.

[8] G. Mink, "Problems of a Highly Contaminated Tradition: The New Testament," in *Studies in Sematology*, ed. P.T. van Reenen et al., 2 vols. (Amsterdam: John Benjamins, 2004), 2:26.

[9] R. Brody, *The Geonim of Babylonia and the Shaping of Medieval Jewish Culture* (New Haven: Yale University Press, 1998), 267–79.

[10] But see the views of M.S. Zuckermandel, *Tosefta, Mischna und Boraitha in ihrem Verhältnis zu einander*, 2 vols. (Frankfurt am Main: J. Kauffmann, 1908–1909), summarizing work begun in the 1840s.

[11] J.N. Epstein, *Introduction to the Mishnaic Text* [Hebrew], 2 vols. (3d ed.; Jerusalem: Magnes, 2000), 980–1006; later literature in H.L. Strack and G. Stemberger, *Introduction to the Talmud and Midrash*, trans. and ed. M.N.A. Bockmuehl (2d ed.; Edinburgh: T & T Clark, 1996), 118–24.

[12] J.N. Epstein, *Introduction to Tannaitic Literature: Mishna, Tosephta and Halakhic Midrashim* [Hebrew] (Jerusalem: Magnes, 1957), 200.

[13] H. Lapin, "Institutionalization, Amoraim, and Yerushalmi *Shebiʿit*," in *The Talmud Yerushalmi and Graeco-Roman Culture*, ed. P. Schäfer and C. Hezser (Tübingen: Mohr Siebeck, 2002), 3:161–84.

[14] Epstein, *Nusaḥ*.

[15] Orality: Y. Sussman, "Torah she-be-ʿal peh: peshuṭtah ke-mashmaʿah," in *Meḥqere talmud*, ed. Sussman and D. Rosenthal (Jerusalem: Magnes, 2005), 5:209–384; performance: M.S. Jaffee, *Torah in the Mouth: Writing and Oral Tradition in Palestinian Judaism, 200 BCE–400 CE* (Oxford: Oxford University Press, 2001), 100–25; E.S. Alexander, *Transmitting Mishnah: The Shaping Influence of Oral Tradition* (Cambridge: Cambridge University Press, 2006). At this stage, Jaffee assumed written transmission as well.

[16] See, e.g., T. Pesaḥ. 10 juxtaposed to M. Pesaḥ. 10, in which T assumes a different order of events in the Passover meal, but also appears to presuppose the Mishnah. It is possible that the two current competing positions – that the Tosefta is largely dependent upon the Mishnah (which I generally share), and that the Tosefta incorporates source material for the Mishnah (J. Hauptman, *Rereading the Mishnah: A New Approach to Ancient Jewish Texts* [Tübingen: Mohr Siebeck, 2005]; S.Y. Friedman, *Tosefta Atikta, Synoptic Parallels of Mishna and Tosefta Analyzed, with a Methodological Introduction* [Hebrew] [Ramat Gan: Bar-Ilan University, 2003]) – too narrowly characterize the early transmission of these texts.

nah's text.[17] However, within tractates there are digressions linked by form, keywords, or associative logic that are usually explained in terms of the use of sources or the original oral medium that might instead be investigated as examples of fluidity.

However, open ended fluidity does not appear to be characteristic of the *manuscript* tradition, although a full edition of all the manuscript materials might change that impression. There is certainly some raggedness – for instance, supplementary sections, notably at the end of tractates – but there is also a text.[18] Still, Schäfer's broader point is well taken. Sussman has argued that by the time we have a manuscript tradition, there were already two streams of transmission of the Mishnah, that of the more static "Palestinian" tradition, whose origin he places in the Byzantine world, and the more dynamic tradition that dominates elsewhere, in which the Mishnah circulated with the Babylonian Talmud and was not conceived of as a discrete text.[19] If it is correct that the practice of writing the Mishnah only began in the eighth century, the very fact of writing will have changed the nature of the text and of the ways in which its students approached it.[20] In many respects, the text of the Mishnah also must be assumed to be a highly "contaminated" tradition, not only because scribes may have been aware of more than one source manuscript or incorporated scribal corrections to their source document reflecting another textual stream, but more significantly because the Mishnah itself and the Talmud were known to those who received a classical education (and scribes were among these, presumably). Scribes may then have consciously or unconsciously incorporated changes large and small.

All this suggests that stemmatic analysis, allowing that it could be carried out for the Mishnah with some measure of confidence, may help to locate the Mishnah not in its late second- or early third-century context, but rather in the context of late ancient and early medieval patterns of study, text production, and transmission. Nor is the reconstruction of a notional archetypal text the only valuable goal of the study of manuscripts. Every copying of a manuscript represents at once a span in the work- and life-history of the scribe or scribes, a moment in the history of the text copied, and a point of reference in the making and disseminating of the book and of written textual knowledge.

Goals of the Digital Mishnah Project

Precisely because there are differences about the reconstructablility of authoritative texts and optimal presentation of textual editions, lack of clarity about what time, place, and social setting any such reconstructed text would correspond to, and the existence of a range of significant research questions not covered by such approaches, the Digital Mishnah project does not propose a single representation of the text of the Mishnah. Instead, the project is conceived of as providing a set of digital tools for scholars interested in the textual history of the Mishnah. The project is now in a pilot stage, using *Babaʾ Meṣiʿaʾ* Chapter 2 as the sample text. The following data and functionalities are anticipated:

(1) *Transcriptions of all manuscript witnesses*, including Genizah fragments, as well as the Naples edition of 1492. At present, vocalization is generally excluded, but punctuation is retained. Transcriptions will be encoded in XML following the Text Encoding Initiative's recommendations. The encoding includes markup that allows the viewer to view the text in its manuscript format, as well as the extracted text alone, and to select to view the text with or without the corrections of later hands. Additional scribal features such as decorated text and ligatures are described. Metadata for each manuscript includes a characterization of the hand, writing surface, approximate date when available, as well as format details including size of the page, number of text columns, and script.

(2) *Comparison and collation.* Once the texts are transcribed, for any given passage (the basic subdivision is the standard printed Mishnah passage) it is fairly straightforward to give a full parallel-column synopsis of either all the witnesses to a selected passage, or to chosen witnesses in a specified order. Thus, in theory, a user could select all Genizah fragments, or all Ashkenazic hands. In addition, the Digital Mishnah will provide collation and / or alignment tools such as the automated construction of an alignment table. At present we are experimenting with CollateX, a collaborative project under development in Europe.[21] Again, it is possible to select all or specified witnesses.

(3) *Analysis tools.* The Digital Mishah project provides visualization and grouping tools. These will include a heat map representing areas of the text that show the most variation, as well as the ability to calculate proximities of witnesses (how much they share in common or differ), and based on those proximities to cluster texts in groups or families. I have also experimented with applying phylogenetic software to those distances to model transmission stemmata.

[17] J. Neusner, *A History of the Mishnaic Law of Purities*, 22 vols. (SJLA 6; Leiden: Brill, 1974–1977).

[18] Academy for the Hebrew Language, "Historical Dictionary," lists two additional supplements totaling another thousand words. As the bibliography accompanying this suggests, the identification of supplements is based on Epstein, *Nusaḥ*, 946–79. In addition, to these, at 976–78, Epstein notes other more or less extensive sections not included in these listings.

[19] Y. Sussman, "Kitbe yad u-mesorot nusaḥ shel ha-mishnah," in *Proceedings of the Seventh World Congress for Jewish Studies*, ed. I. Gutmann (Jerusalem: Israel Academy of Sciences, 1980), 3:240–44.

[20] T. Fishman, *Becoming the People of the Talmud: Oral Torah as Written Tradition in Medieval Jewish Cultures* (Philadelphia: University of Pennsylvania Press, 2011).

[21] Interedition, "CollateX," collatex.sourceforge.net.

Conceived of as a tool rather than an edition in the traditional sense, the Digital Mishnah will provides an interface for doing close textual work on the witnesses, without committing the user to one particular theory of text presentation and study. Unlike print editions, the Digital Mishnah is dynamic, in the sense that new texts can be incorporated and errors corrected, but also in that users have the option of which texts to view and how to present them. While print editions incorporating synopses and alignment tables have become standard, they are static, inefficient (incorporating very little text per page or requiring a large and non-standard format), and frequently extraordinarily expensive. In practice, their editors have also had to filter out some of the information about the text in its manuscript setting to incorporate it into the selected format. The proposed digital format avoids these drawbacks and compromises.

Text Preparation

The texts for the project are encoded using the TEI guidelines, version P5, a specific application of XML designed for and widely used in digital humanities projects, as customized for this project.[22] Every document has two parts, a header and a body. The header provides metadata that allows both human and machine readers to use the document. Some metadata provides information about the particular document, for instance, in our case, an identifier for the manuscript, the repository where it is held, or dimensions. Other metadata links the particular document to others in the corpus, for instance, noting that a particular fragment joins with another that is separately encoded,[23] to geographical or prosopographical information, or to information outside the project, such as digital images of encoded documents.

At present, the Digital Mishnah project has four types of document each with different types of content in the body. Overwhelmingly, most documents are witness documents that contain the transcription together with the specialized XML markup. The markup notes the structure of the text (mishnah, chapter, tractate, order), the layout of the text (folio, page, column, line), manuscript features (gaps, damaged text, insertion or deletions), scribal features (ligatures, supplementary characters at the ends of lines, abbreviations and their expansions), items of interest (personal and geographical names, non-Hebrew words, or biblical verses), and can include

[22] TEI Consortium, "TEI P5: Guidelines for Electronic Text Encoding and Interchange," http://www.tei-c.org/Guidelines/P5/.

[23] E.g., CUL T-S NS 329.286 and T-S AS 78.69; other fragments appear to be part of this manuscript. For Baba Qamma, CUL T-S E1.99, T-S F6.3, and Bodl. MS Heb, c.21 8a–11b combine to make up a single manuscript that covers nearly the entire tractate.

notes. A special linked document called a schema is used to check that required elements are included and that all elements of the markup are used in the specified way.

The other three types of documents serve specific purposes. There is one reference document (suitably named ref.xml) that makes explicit the structure of the whole Mishnah, identifies and links to all the witness files, and provides unique identifiers for each section of the Mishnah. For instance, the identifier for *Baba' Meṣiʿa* 2:10 is ref.4.2.2.10, i.e., order 4 [*Neziqin*], tractate 2 [*Baba' Meṣiʿa*], chapter 2, mishnah 10. The text used for this reference document is based on the Vilna edition of the Mishnah as printed in the *Yakhin u-Boʿaz* edition.[24] Another type of documents is for collecting readings in secondary sources (the Palestinian and Babylonian Talmuds and citations in commentaries). These are similar in many respects to the witness files, but since they may collect multiple readings, do not present the whole text of the Mishnah, and may draw on modern printed editions, they have special rules. Finally, there are additional documents that contain prosopographic or geographical data.

If we focus on the witness files, the principal difference between a conventional print edition and a markup edition is that the markup uses coded elements, or tags, where print uses typographical conventions. (Figure 1 and Figure 2 provide a brief coding sample and the text it encodes.) This makes it simple to change display conventions, but more importantly to easily access and process specific data. For example, if we are providing a word-by-word collation of a text, we want to make sure that the expression that an abbreviation, especially one including more than one word (e.g., אע"פ), in one witness is properly aligned with the full expression (e.g., אף על פי) in another. We can do this by instructing the collator to use the expanded text that is encoded with the abbreviation, as the collation item. If we are interested in scribal features of the text, we can instruct a processor to omit the expansion, but to capture instead the abbreviation itself and even the format the scribe uses to mark the abbreviation.

Display and Comparison

Once there is a body of texts, they can be processed singly or in groups. For individual witness documents users are able to view a text in its manuscript context and format, ideally opposite an image of the witness. Options can

[24] Ed. Vilna, *Mishnayyot*. The choice is pragmatic in the absence of a critical text. This version is likely to be the longest (i. e., to include later additions), and it is readily available in print and electronic formats. This is also the version of the Mishnah that most users encounter on a regular basis.

include a "diplomatic" transcription or an edited restored transcription. Figure 3 shows formatted output of the recto of a small Genizah fragment from Cambridge (T-S AS 78.69).[25] For comparison, work to the present has focused on synopsis and collation. At a later stage, I envision allowing the user to complete the lacunae with text from other witnesses.

Synopsis is fairly easily done, and involves selecting the passages and witnesses to be aligned. A synoptic alignment of all the witnesses encoded as of January 2012 is available through the repository for the project. Collation is more complicated. Currently the Digital Mishnah has the ability to present an alignment table or the text of a chosen witness with apparatus. As noted, we are experimenting with CollateX, from which output can be constructed. At present, output is fairly error prone. CollateX generates a table that is on occasion as much as fifteen percent longer (or, occasionally, shorter) than the number of number of columns required for a correct alignment. In order for collation to take place the text needs to be divided into collatable items, or "tokens," and improvement will come from a tokenization process that passes both the actual text and a normalized version (as with abbreviations) to the processor. In one test sample, using ten witnesses, and standardizing the text by expanding all abbreviations and removing all *waws* and *yods*, provided a nearly perfect collation. This has the effect of separating out differences that are largely orthographic from differences in wording or grammatical form, with some kinds of variation falling in between. (For instance, the use of initial *waw* as a conjunction, potentially a substantive variation, is passed over when ignoring *waw*; while, the representation of the *mu* in the Greek word *emporia* as *mem* or *nun*, as in one case in the sample text, is an orthographic difference not captured by ignoring the Hebrew *matres lectionis*.) To capture those omitted differences, each reading that is identical when standardized will then need to be processed separately using the original spelling. There are, in addition, other software projects that we are exploring, including Juxta and nMerge.[26] An additional possible functionality (inspired by the previous two software projects) is the ability to view two (or perhaps more) text fields in parallel with the differences highlighted.

[25] A full version in html can be found at: http://umd-mith.github.com/mishnah/ samples/g4-formatted.html. The site includes formatted version of BM Chapter 2 in Kauf as well: http://umd-mith.github.com/mishnah/samples/BM_Ch2_Kauf-formatted.html.

[26] Juxta: "Juxta: Collation Software for Scholars," NINES, www.juxtasoftware.org. nMerge, developed by Desmond Schmidt, and now part of HRIT: "HRIT: Humanities Research and Infrastructure Tools," Center for Textual Studies and Digital Humanities, Loyola University, https://sites.google.com/a/ctsdh.luc.edu/hrit-intranet/.

Analysis

One advantage of having the text available as a digital string of characters is that it is possible to make use of statistical techniques of data comparison and representation. It will be useful, for instance, to have a heat map that represents a large expanse of text (on the level of a tractate or order) that shows areas of greater textual variation through color differences, and allows the user to identify and select highly variant (or invariant) passages for examination. In addition, I have been doing preliminary research and experimentation on clustering techniques using SPSS, a standard statistical package, with the goal of creating web-based implementations of techniques that are useful for showing the grouping of witnesses into families.[27] Useful procedures would need to be written in a language like Java for application in a website. Here I summarize preliminary results.

Of the possible techniques, the easiest to apply without a specialized algorithm is the distance matrix method, and it is this that I have focused on. An example of a distance matrix is the table provided on some maps that shows the driving distance between cities. It is possible to generate maps of points based on those distances. More than one map is possible and techniques to generate those calculated maps aim to find the distributions that best account for all the distances. The example of a table of driving distances is useful for raising the problem of dimensionality. We know (or most of us do) that the surface of the earth is roughly a sphere, and the locations at varying elevations relative to sea level on that sphere, while the representation of the distances on a paper map or a computer screen is in two dimensions. Intuitively, it makes sense for some data sets – a table of intercontinental flying distances, for instance – that the third dimension will more greatly distort the layout of the two dimensional map. In fact, for every N data points, it might be necessary to imagine a space in N–1 dimensions in order to account for all the differences between the data points.

Again, intuitively, we can understand that manuscripts that differ from each other by one character are closer to each other than manuscripts that differ by many characters, and that the fact that manuscript A and B differ from each other by one measure does not mean that they each differ from C to the same degree. In theory, it is possible to perform a character-by-character comparison along the lines of comparisons of DNA or protein sequences. Because CollateX generates output based on words, it was convenient to construct distance measure on the basis of words, not individual characters.[28] Distance was calculated using the squared Euclidean distance on binary data.

[27] SPSS Ver. 19.0 for Macintosh, IBM, Armonk, NY.

[28] In addition, the observation below that proper clustering requires considerable nor-

The first attempt began with a collation for the entire sample chapter using eight witnesses that had complete text for the chapter, and transferring the alignment table to Microsoft Excel where it could be corrected by hand. (The text collated represented the best reconstruction of what the original scribe wrote.) Each row in the table constitutes a witness. Distinct readings each occupy a column. Witnesses that share a reading in common each have text in the same column; those that lack that reading have nothing in that column. This has the effect of creating a sort of macrotext that includes all possible readings.[29] Each table of variations has a corresponding table in which, for each reading, the value of 1 was assigned if the reading is present in a given witness, and 0 if it is absent (henœ "binary"). It is worth noting that in contrast to the text-critical approach of Maas which seeks unique or disjunctive errors and demands of the critic to evaluate whether *errors* could have occurred independently, the statistical approach is interested in the consequences of *differences*, without attempting to assess the independence of their origin in each and every case.

I then selected all the loci (columns in an Excel spread sheet) that had variant readings and exported these to SPSS. Because of the large number of possible variations, I set some criteria to exclude variants that I supposed would not be significant. There were, in all, about 230 loci. SPSS created a distance matrix using squared Euclidean distance (a measure that emphasizes distance),[30] and generated a two dimensional map using a procedure called multi-dimensional scaling. These results did not confirm expectations. The Mishnah in the Leiden MS of the Yerushalmi and Parma A agreed appeared close to each other as, but Parma A was, on this calculation, closer to Paris and Maim than either Kauf or Camb, which are frequently grouped together with it by scholars. In fact, Kauf has been argued on classical text-critical grounds to be an antecedent of Parm A.[31]

Since the preceding excluded quite a bit of variation (and, on review of the data, contained errors as well as inconsistencies in including or excluding variations), and thinking that a more manageable number of variant loci more intensively handled might give different results, I focused on the first two *halakhot* of Chapter 2, included both variant and invariant loci, and incorporated a Genizah fragment (G1) and another Babylonian Talmud manuscript (Hamb) to give a wider range of the types of witnesses. Here, the presence of *waw* or *yod* representing vowels was ignored; and abbreviations expanded and not counted as variants unless the abbreviation itself suggested that the intended word was different than that appearing in one of the other witnesses.[32] Where a preposition could be written attached or unattached to the following noun (most frequently *shel*), the variant was counted once: שלנחתום against של נחתום counts as one variation; however, שלנחתום and של נחתומין count as two, one for the preposition, and once for the noun. There were, in all, 155 units of variation (columns in the alignment table). These results were a bit more consistent with expectations (Figure 4). The "Palestinian" manuscripts appeared in or adjacent to the lower left quadrant of the plot, near the origin. Visually, however, there was not clear clustering in groups that would imply strongly related families of manuscripts. However, two clustering procedures suggest that it would be appropriate to treat Kauf, Parm, Leid, Camb, and G1 as constituting one cluster, and Nap, Paris, Mun, and Hamb constituting a second (Table 1; they differ on Maim).

My impression, moreover, was that scribes were not entirely consistent with respect to some orthographic features (e.g., *plene* spelling, already excluded), and I wondered whether separating out orthographic from substantive features would not give different plots.[33] I went through the same collation of Mishnah 2:1–2 and extracting variants of these distinct types. Orthographic differences, (35 loci with variation, about 23 percent) included conjunctive use of *waw*, interchange of א and ע, final ם or ן in the masculine plural ending, or the six different spellings of *emporia*. Substantive difference (46 loci, about 30 percent) included different words, of course, but such variation as different gender or number endings. Figure 5 shows the multi-dimensional scaling plots for orthographic and substantive differences superimposed on one another.

In the case of substantive differences it seems clear that the "Palestinian" tradition does appear closely grouped, together with Maim, a close relation

malization of the text or the exclusion of some variation from the clustering comparison suggests that character-by-character comparison of non-normalized text will generate distorted data.

[29] The method is borrowed from T.J. Finney, "Potential Computer Applications in New Testament Textual Research," http://www.tfinney.net/ Potential/index.html, who has been very generous with his time. Analogous, although very different in implementation is D. Schmidt and R. Colomb, "A Data Structure for Representing Multi-version Texts Online," *International Journal Human-Computer Studies* 67 (2009): 497–514.

[30] The Euclidean distance sums the square of the difference between each reading, which, since these are binary, can only be 1s or 0s, and takes the square root of the sum. The squared Euclidean distance does not take the square root of the sums.

[31] Leid and ParmA: I.Z. Feintuch, "The Mishna of the MS Leiden of the Palestinian Talmud," [Hebrew], *Tarbiz* 45 (1976): 178–212; Kauf as antecedent to ParmA as well as Camb: Rosenthal, "Aboda Zara," 107–12.

[32] To take an example from elsewhere in the chapter, at M. BM 2:7, Mun reads דברי', which implies דברים (plural), while the other witnesses all have דבר (singular). However, since the adjacent abbreviations give evidence of the plural form, these were not marked as diverging from the singular forms.

[33] For the distinction I am indebted to Timothy Finney, personal communication. The terminology goes back to W.W. Greg, "The Rationale of Copy-Text," *Studies in Bibliography* 3 (1950): 19–36, who distinguished between substantive and accidental variation. The terminology is not without its problems.

in time and tradition, while the others are set off at some distance. This would suggest that variations in writing are driven by different factors than those that account for variation in substantive readings. At the same time, the focus on what were classed as "substantive" features has accentuated the distance of some witnesses relative to the others. Part of that must be due to plusses and minuses in the witnesses. Thus, for instance, in M. BM 2:5, Nap, which stands at the greatest distance from all the others, omits an eight word phrase present in all the others, presumably due to something akin to *homeoarchon*. This accounts for seventeen percent (eight out of forty-six loci) of all the substantive variation in Nap. Yet if it is a scribal omission, it presumably represents one alteration. Indeed, removing that omission from consideration and recalculating results in a rather tidy grouping of the Palestinian tradition in the center, Mun and Hamb (manuscripts of the Babylonian Talmud) to one side, and Nap and Paris (Mishnah with Maimonides' commentary in Hebrew translation) to the other (Figure 6). G1, interestingly, remains something of an outlier.

Subsequently, I re-tabulated the collation of all of Chapter 2 based on the criteria used for the shorter sample, and applied MDS on that larger set of variations (Figure 7). The results were reassuringly comparable. There is a core group of Kauf, Parm, and Leid, around which the others are arrayed as outliers; Mun and especially Nap are again quite distant from the center and each other. A somewhat larger grouping would include Maim and Camb, but also Paris. At the same time, the differences were significant, and the configuration does not map closely onto that of Figure 5.

One conclusion to draw from the discussion thus far, based primarily on a very small sample, is that certain types of variation do indeed tend to represent common tendencies among the witnesses. In other words, the witnesses do indeed cluster. However, it also appears that, if we are thinking of a spatial representation of groups of witnesses, different kinds of variation literally pull the witnesses in different directions. If the Digital Mishnah can provide a tool for examining the consequences of different types of variation in the Mishnah, and a reproducible model for how to do so with other texts, it will have made a contribution to the field of textual studies. However, the narrower question at hand is whether it is possible to provide tools that will offer users a necessarily rough, but useful, guide to the grouping of witnesses into classes or families, for it appears that the broadly orthographic differences as well as some substantive differences constitute "noise" that needs to be filtered out. For an online application, therefore, I would propose three strategies. It is possible to implement all of them.

The first strategy involves manual input. If we imagine the output in the form of an alignment table, with witness occupying a line, and each column representing a reading, it is possible to allow the user to tag each column. The advantage is the introduction of human judgment about the classification of differences. Based on the tags, one can not only perform grouping or clustering operations, but also statistically measure the impact of each of the categories for generating the distances between the witnesses. The disadvantages are human error (as demonstrated above), and subjective judgment.

The second, an automated strategy, has already been addressed indirectly. To the extent the project uses collation output from CollateX or a program like it, and the text must be heavily normalized for proper alignment, the output already roughly provides a sorting of the data into orthographic variation (which "disappears" when the text is normalized) and substantive (which does not disappear). It should be fairly simple to then treat these two classes of data differently. In addition, one could successively re-compare each column based on the reintroduction of each of the characters omitted or changed in normalization. Similarly, one could provide special handling for places where the unit of variation is longer than one word (i. e., omissions or additions). Again, it is possible to measure statistically the impact of these orthographic differences.

The third proposed strategy may again be automated, and uses principle component analysis.[34] Recall that for N witnesses, it may require N-1 dimensions to fully describe the distances between each of the witnesses. Principle component analysis in a sense rotates the space in which the variation takes place so that variation can be described as taking place along newly defined axes. Using the set of pairwise distances (the distance matrix) already constructed for multidimensional scaling, principle component analysis in SPSS found that three dimensions accounted for something over 45, 26, and 22 percent respectively, totaling 94 percent taken together. Component 1 captures considerable grouping. This is evident from the scatterplots based on components 1 and 2, which account for 72 percent of the variation (Figure 8), and of dimensions 1 and 3, accounting for 68 percent (Figure 9). The latter in particular could easily be analyzed in two conventional clusters, "Palestinian" (including, in this case, Maim) and "Babylonian."

Repeating the procedure on the whole chapter produced promising results, better than repetition of MDS with the expanded sample (Figure 10). Again, principal component analysis resolved the variation into three dimensions accounting for about 94 percent of the variation (41, 34, 19 percent respectively), with component 1 strongly contributing to clustering the "Palestinian" manuscripts, and the combination of components 1 and 3 (60

[34] I am grateful to Travis Brown for his suggestion of PCA in this context.

percent of variation) neatly clustering into two groups: Kauf, ParmA, Leid, Camb; and Maim, Paris, Nap, Mun.[35]

The Digital Mishnah project could thus present several views of the data based on PCA, or even a rotatable three-dimensional scatter plot that would allow user to get the best view into the data, along with clear information about how much variation a particular view accounts for.[36] Further research is needed to see if there is a statistically respectable way to provide automated clustering of the data as well, that will also be immediately useful information to the non-statistician user. If the current sample is roughly representative, it might be sufficient to present the results of clustering for between two and four clusters, using one standard method. The Ward linkage method generates what appear to be useful results, but further research is needed to see if this is an artificial result of tendencies in the method itself. Table 1 presents sample results for 2:1–2 and for all of chapter 2 using both the Ward linkage and K-means methods, with the overlaps in clustering between methods highlighted.[37] Of course, until unresolved issues such as the handling of plusses, minuses, or transpositions in the text are dealt with, the results of clustering is at best provisional.

It should be noted, finally, that while the grouping that is attested in Figure 8 and Figure 9 presumably captures what I called substantive differences above, the algorithm that does the splitting into principal components does not "know" this. It only configures the space in which the variation takes place into orthogonal dimensions and asks how much of the variation can be explained as a vector along that dimension. In addition, if principal component analysis is done "on the fly" with user-specified samples, the principal components will be different for each sample, as we have seen. To understand what features of the texts are responsible for which aspects of the resulting multidimensional distribution, we would need a more explicit and nuanced reading and coding of the text.

Thus far, experimentation has shown a fair amount of success in identifying at least one set of witnesses that has group affinity. I do not have a great deal of confidence in the results thus far for the grouping of the outlying witnesses, although a larger and more systematically treated dataset may alter that perception. For that reason, over and above any questions about the applicability of stemmatic approach to the evidence of the Mishnah, I am reluctant at this time to apply phylogenetic software. Based on a distance method (that of Fitch and Margoliash), for the short sample used in this paper (M. BM 2:1–2) we get an unrooted tree that is more or less believable, although not in its details (Figure 11, left).[38] Leid, for instance, is out of place. An unrooted tree gives the calculated genetic relationships between witnesses, introducing intermediate nodes to account for the attested distances. If we wanted suppose a single antecedent of all the witnesses, that is to root the tree, we would need to apply some procedure (e.g., rooting the tree between the two most distant witnesses, or calculating the statistically most likely tree) or manually select a root node based on external criteria. The default assumptions of the Phylip drawgram program generate a rooted tree that is remarkably problematic (Figure 11, right). The method used in the field of genetics of including an outgroup (a witness known to be more distant than all of the ones under consideration) to locate the common ancestor does not apply in this case.

Conclusions

We are still lacking a critical edition of the Mishnah. The Digital Mishnah project aims to fill this gap, but in a form that takes advantage of the dynamic capabilities of the computers and the internet. Rather than present a static text in a single format, the end project will allow users to choose the texts they wish to compare and the format in which they wish the data output. In describing the project, I have also reported on progress to date in collecting, collating, and presenting the text, and possible ways to provide analytical tools.

[35] In terms of "loading" on the components, Kauf, ParmA, and Leid are loaded on regression component 1; Camb and Maim on component 2 (with weaker loading for Nap and complex loading for Maim); and Paris on factor 3.

[36] See for instance Finney, "Computer Applications," at Example 6.

[37] The method described in introductory texts is to use a hierarchical procedure like Ward linkage to decide on the number of clusters and a procedure like K-means to describe the actual clusters. For the samples examined here a range between two and four seems appropriate, but more work is necessary. Because Ward linkage is agglomerative, witnesses clustered at an earlier stage, when there are more clusters, will also be grouped together when there are fewer. By contrast, methods like K-means require specification in advance of the number of clusters, but the clustering at a given number of clusters (say, four) is not directly dependent upon clustering at a higher or lower number (see, e.g., the placement of Maim in column two and Camb and Maim in column four).

[38] The tree was generated using the drawtree program of the Phylip package (J. Felsenstein, "Phylip," University of Washington, http://evolution.genetics.washington. edu/phylip/) through the interface of "Wageningen Bioinformatics Webportal: Phylogenetic Tree Plot," Laboratory of Bioinformatics, WUR, http://www. bioinformatics.nl/tools/plottree.html, with data formatting through V. Makarenkov, "T-Rex Online," Université du Québec à Montréal, http://www.trex.uqam.ca. For distance-based methods in genetics see W. M. Fitch and E. Margoliash, "Construction of Phylogenetic Trees," *Science* 155 (1967): 279–84; and the survey of different methods J. Felsenstein, "Phylogenies from Molecular Sequences: Inference and Reliability," *Annual Review of Genetics* 22 (1988): 521–65.

```
                                   <ab xml:id="Kauf.4.2.2.5">
             <c type="abbr"ך<milestone unit="MSMishnah" n="6"/><label>
 בכלל כל<lb n="13"/> אף השימלה היתה</c></label> 'rend="overmark">
<lb אילו ולמה יצאת להקיש אליה<lb n="14"/> אלא מה השימלה מיוחדת שיש בה
        </del><add '<del>ת <seg n="scribCorr">סימנין ויש להn="15"/>
   ואף</seg> ך</add>'</del><add hand="2h">ו<del>בע</add>וhand="2h">
    <add ת <seg n="scribCorr">שיש בו סימנין ויש לו<lb n="16"/> כל דבר
</seg> <seg ך</add>'</del><add hand="2h">ו<del>בע</add>וhand="2h">
               ואףn="scribCorr"><del hand="2h" rend="overline" seq="2">
  <del תובע <seg n="scribCorr">דבר שיש בו</del><lb n="17"/><del>כל
   חייב להכריז</seg></del></seg>ך</add>'</del><add seq="1">וseq="1">
                                                        </ab>
```

Figure 1: Coding Extract M. BM 2:5, Kauf

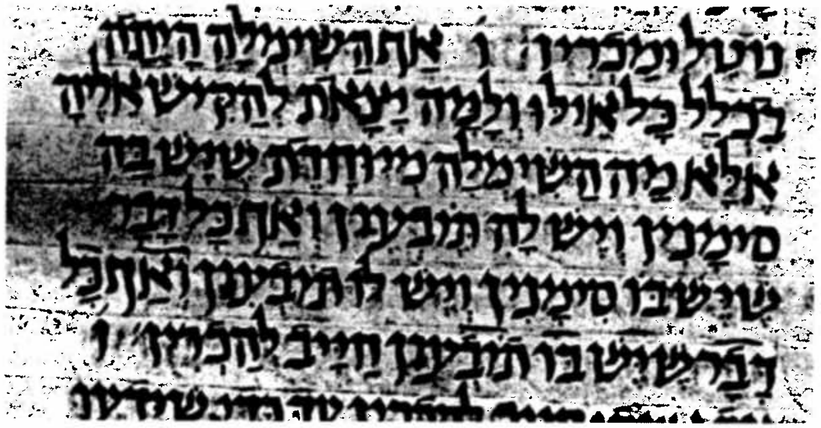
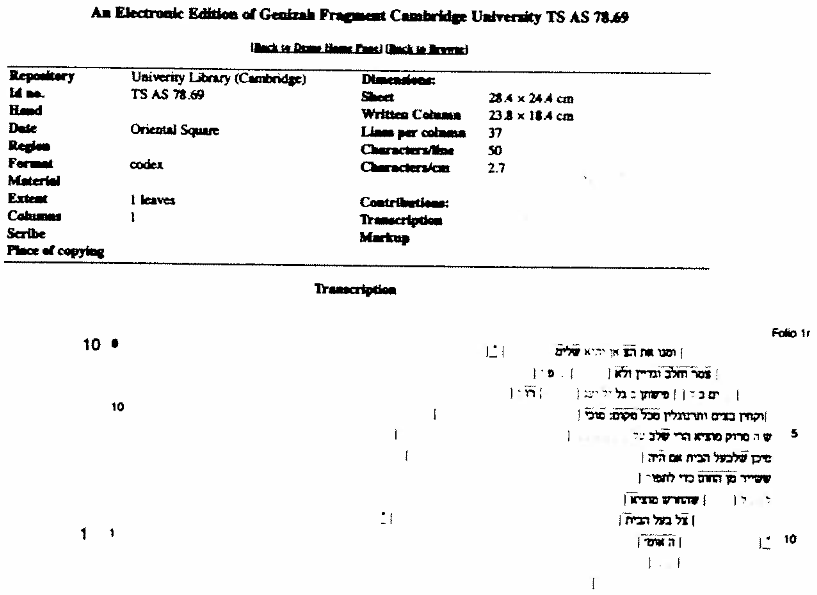


Figure 2: M. BM 2:5 in Kaufmann A50, 134r



An Electronic Edition of Genizah Fragment Cambridge University TS AS 78.69

[Back to Demo Home Page] [Back to Browse]

Repository	Univerity Library (Cambridge)	**Dimensions:**	
Id no.	TS AS 78.69	**Sheet**	28.4 × 24.4 cm
Hand		**Written Column**	23.8 × 18.4 cm
Date	Oriental Square	**Lines per column**	37
Region		**Characters/line**	50
Format	codex	**Characters/cm**	2.7
Material			
Extent	1 leaves	**Contributions:**	
Columns	1	**Transcription**	
Scribe		**Markup**	
Place of copying			

Transcription

Folio 1r

10 ● | ומנו את הב אן ...א שליש |

| צור חלב ובדין ולא | | . פ |

10 | ים כ | | פישתן ב גל ... | | ה |

| וקחין בבים ותרנגלין מכל מקום מוכי |

5 ש ה פרוק מציא הרי שלב ..

מקן שלבעל הבית אם היה |

ששייר מן החרס כדי לתפור |

| שהחרש מציא |

| צל בעל הבית |

1 1 | ה אומ | 10

Figure 3: Sample Formatted Text (CUL TS AS78.69 1r, Genizah Fragment)

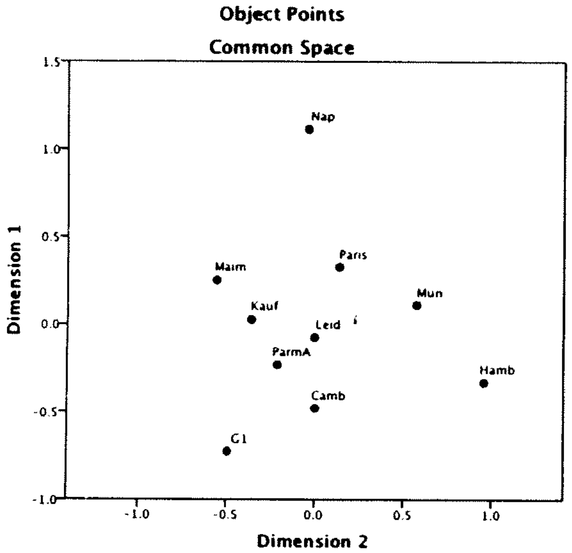


Figure 4: Multi-Dimensional Scaling Plot of M. BM. 2:1–2, Whole Sample

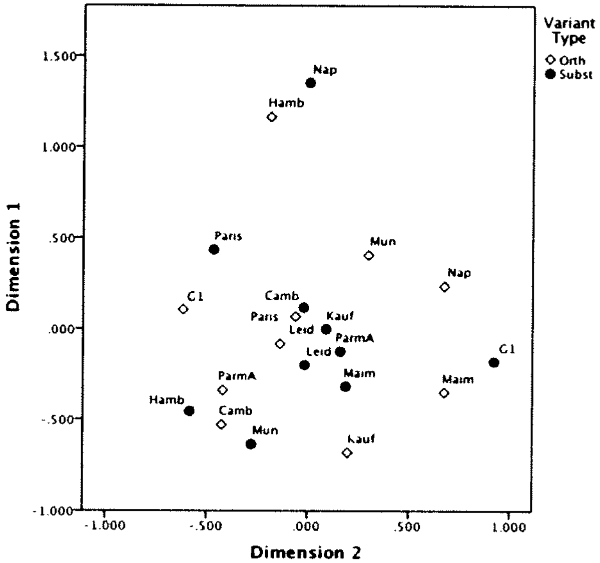


Figure 5: Multi-Dimensional Scaling Plot of M. BM. 2:1–2, Orthographic and Substantive Variation Plotted Separately

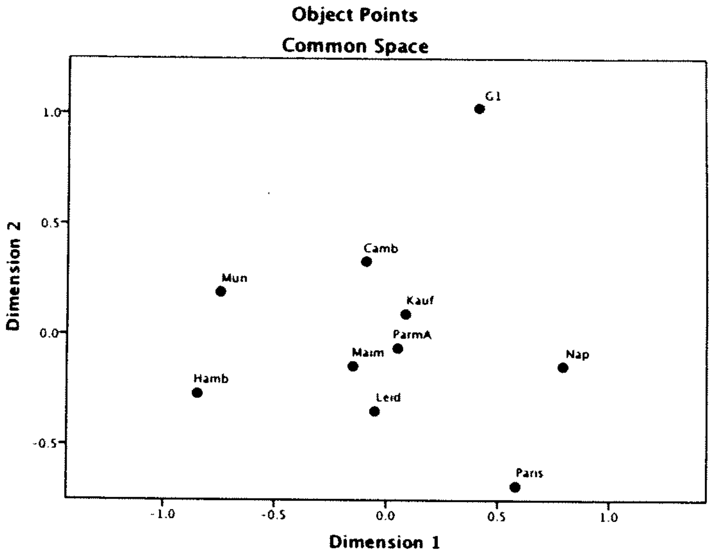


Figure 6: Multi-Dimensional Scaling Plot of M. BM. 2:1–2, Substantive Variation, Adjusting Nap

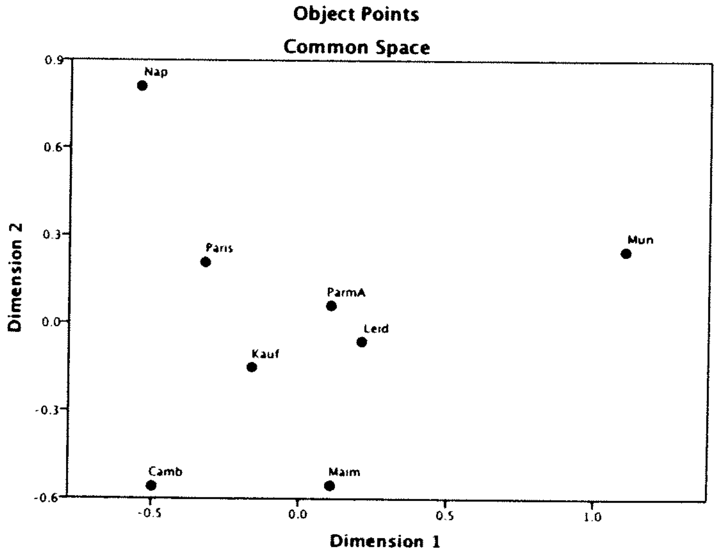


Figure 7: Multi-Dimensional Scaling Plot of M. BM. Chapter 2, Substantive Variation

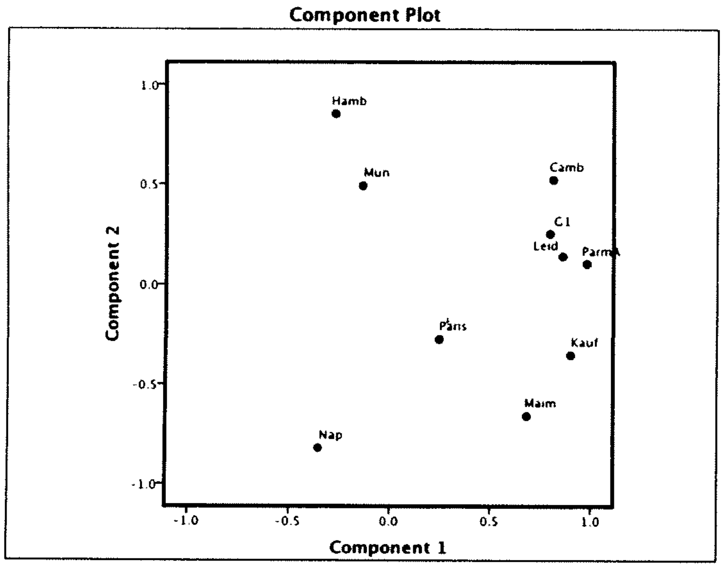


Figure 8: Scatterplot of Witnesses after Principal Component Analysis (a)

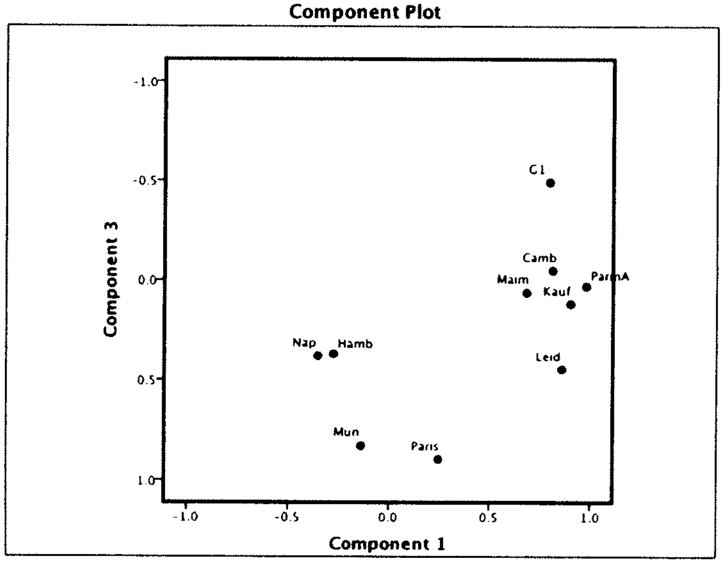


Figure 9: Scatterplot of Witnesses after Principal Component Analysis (b)

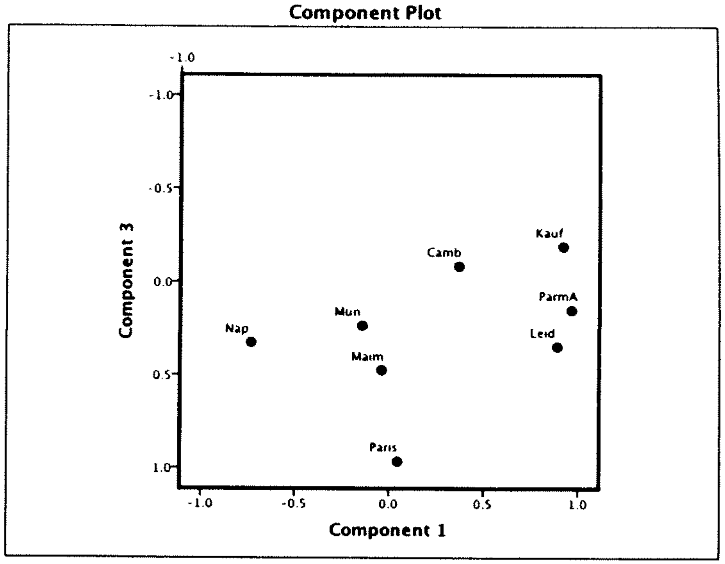


Figure 10: Scatterplot of Witnesses after Principal Component Analysis, Chapter 2, Entire

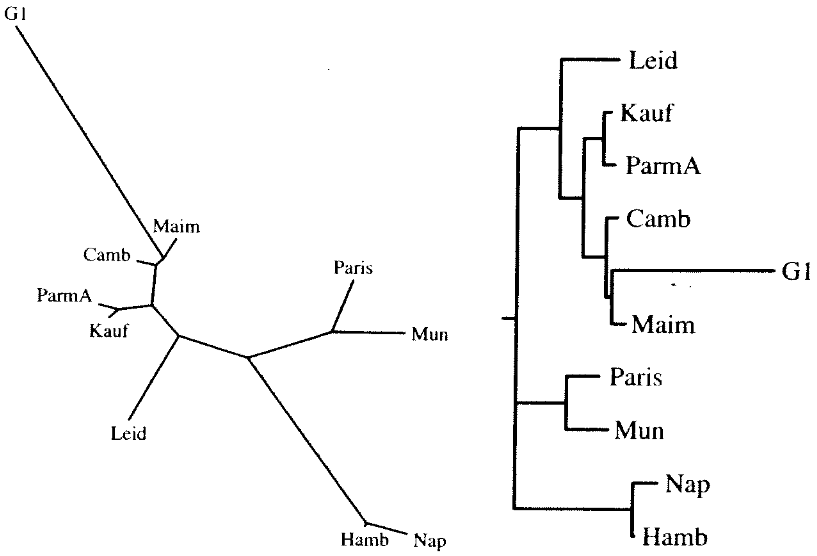


Figure 11: Unrooted (left) and Rooted (tree) Tree, BM Chapter 2, After PCA, Regression Factors 1 and 3 (Phylip)

Number of Clusters	Clustering*			
	BM 2:1–2		BM 2, Entire	
	Ward Linkage	K-Means	Ward Linkage	K-Means
2 a	**Kauf, ParmA, Leid, Camb, G1**	**Kauf, ParmA, Leid, Camb, G1,** Maim	**Kauf, ParmA, Leid**	**Kauf, ParmA, Leid,** Camb, Paris, Mun
b	Maim, **Paris, Nap, Mun, Hamb**	**Paris, Nap, Mun, Hamb**	Camb, **Maim, Nap,** Paris, Mun	**Maim, Nap**
3 a	**Kauf, ParmA, Leid, Camb, G1**	**Kauf, ParmA, Leid, Camb, G1,** Maim, Paris	**Kauf, ParmA, Leid**	**Kauf, ParmA, Leid,** Camb, Paris
b	Maim, Paris, **Mun, Hamb**	**Mun, Hamb**	Camb, **Maim, Nap,** Paris	**Maim, Nap**
c	Nap	Nap	Mun	Mun
4 a	**Kauf, ParmA, Leid,** Camb, G1	**Kauf, ParmA, Leid,** Maim, Paris	**Kauf, ParmA, Leid**	**Kauf, ParmA, Leid,** Paris
b	Maim, Paris, Mun	Camb, G1	**Camb, Maim,** Paris	**Camb, Maim**
c	Hamb	Mun, Hamb	Nap	Nap
d	Nap	Paris	Mun	Mun

* Bold highlights common clusters between Ward linkage and K-means. Underscoring indicates common groupings that appear in different contexts according to the two methods.

Table 1: Clustering Results, Ward Linkage and K-Means, for 2–4 Clusters (SPSS). Compare Figure 4

Mekhilta de-R. Yishmael

Some Aspects of its Redaction

Günter Stemberger

From the very beginnings of modern research into the halakhic midrashim in the late nineteenth century it was an accepted fact that these midrashim belonged to two different schools, those of R. Yishmael and of R. Aqiva. It was also soon recognized that the halakhic parts should be distinguished from the aggadic sections of these midrashim and that the attribution of a midrash to a school refers first of all to the halakhic parts; in more recent research the attribution to schools has become somewhat problematic, but the main lines of the argument are still almost generally accepted.

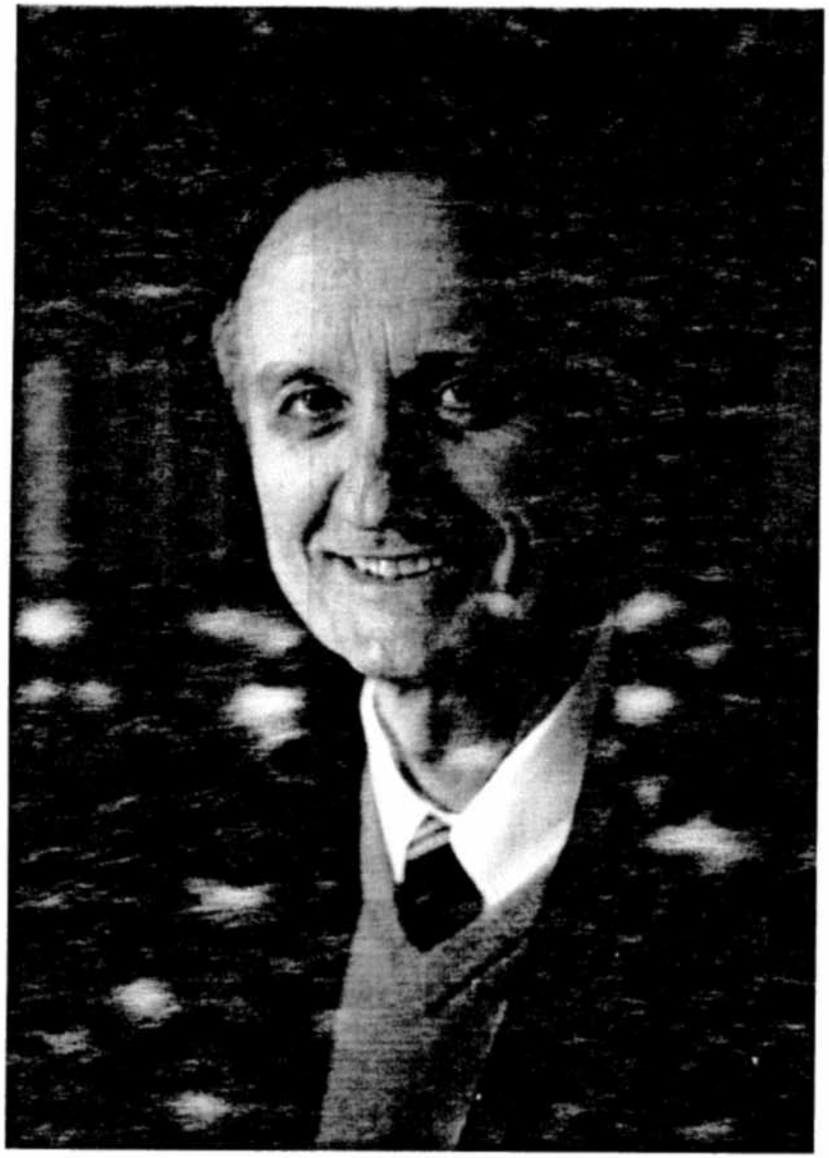
But even the individual midrashim in themselves are not literary units, but are composed of different traditions. As to the Mekhilta de-R. Yismael, already D. Hoffmann pointed to the uneven distribution of rabbis within the midrash;[1] other scholars, as, e.g., Jacob Epstein[2] and Jacob Neusner,[3] followed these observations, refined them and used them for the pre-history of the midrash, but normally without calling into question the traditional dating of the redaction of the Mekhilta in the decades after the redaction of the Mishnah. Epstein did not live to draw historical conclusions from his multiple observations on the lack of unity in the Mekhilta (his Mevo'ot were redacted after his death from his lecture notes by his student Ezra Melammed). J. Neusner in his later writings abandoned the source-critical approach to the Mekhilta (as to other rabbinic writings) and preferred to study it exclusively on the level of its final text, defining the Mekhilta as a Scriptural Encyclopedia,[4] since – contrary to the stunning thematic and logical coherence he thinks to have discovered in the other halakhic mi-

[1] D. Hoffmann, *Zur Einleitung in die halachischen Midraschim* (Berlin: M. Driesner, 1887), 38–40.

[2] J. N. Epstein, *Mevo'ot le-Sifrut ha-Tanna'im: Mishnah, Tosefta u-midreshe-halakhah* (Jerusalem: Magnes, 1957), 570–87.

[3] J. Neusner, *A History of the Jews in Babylonia*, vol. 1: *The Parthian Period* (SPB 9; 2d ed.; Leiden: Brill, 1969), 192–96.

[4] J. Neusner, *Mekhilta According to Rabbi Ishmael: An Introduction to Judaism's First Scriptural Encyclopaedia* (Atlanta: Scholars Press, 1988).



Envisioning Judaism

Studies in Honor of Peter Schäfer
on the Occasion of his Seventieth Birthday

Edited by

Raʿanan S. Boustan, Klaus Herrmann,
Reimund Leicht, Annette Yoshiko Reed,
and Giuseppe Veltri

with the collaboration of

Alex Ramos

Volume 1

Mohr Siebeck