Miqra as Oral Torah, Written Torah, and Digital Torah

By: SETH (AVI) KADISH and BEN DENCKLA

Abstract

The Hebrew Bible's twenty-four books were transmitted over the ages as *two* parallel traditions: that of the scribes and that of the readers. The oral tradition of the readers (*miqra*) was later committed to writing. The details of this story are well-known to scholars, yet most studies and surveys focus on narrow aspects of a single tradition, while others discuss "the masoretic text" (MT) as a single document. We argue that this lack of perspective has kept modern textual scholarship from understanding the MT's centrality within critical editions of the Hebrew Bible.

Our encoding of the MT as the digital dataset called *Miqra According to the Masorah* (MAM) leads to further conclusions: (1) Only through an effective dataset can the mature product of the masoretes be fully expressed. (2) The masoretic project is best understood by viewing it in retrospect as a kind of dataset. (3) Several significant oral elements that were given sporadic expression in the Tiberian masorah can be marked consistently in a modern dataset; such activity may be viewed as a modest extension of elements already present within the masoretic project, and as faithful to its spirit. Finally, we suggest that the best way to preserve and disseminate the masoretic project in digital form is through an open license.

Seth (Avi) Kadish teaches medieval Jewish thought, history, and biblical exegesis at the Oranim Academic College of Education in Kiryat Tivon, Israel. A graduate and *musmakh* of Yeshiva University, he subsequently completed his Ph.D. in Jewish Thought at the University of Haifa under the supervision of Prof. Menachem Kellner. His research areas include medieval Jewish philosophy, dogma, and ethics, as well as Jewish prayer. MAM (*Miqra According to the Masorah*), his hobby, is a labor of love.

Ben Denckla is a programmer whose specialties include programs that format pointed Hebrew texts of the Bible and Jewish liturgy. He has written such programs for publications by URJ, CCAR, and JPS. He received his A.B. at Harvard and his M.S. from MIT. He helps maintain the MAM sources as well as editions derived from MAM.

1. *Miqra* as Oral Torah[1]

Nechama Leibowitz, of blessed memory, once visited a class at an elementary school in Israel. She asked the children to explain the difference between the words *eikh* (אֵיךְ) and *keiẓad* (כֵּיצַד), both of which mean "how" in modern spoken Hebrew. The answer she expected was that *eikh* is found in the Tanakh (biblical Hebrew), while *keiẓad* is found in the Mishnah (rabbinic Hebrew). But that was not the answer she received. Instead, the children told her, "*We* say *eikh*, but *you* say *keiẓad*!" In other words, *eikh* is used in common speech, but *keiẓad* is a kind of literary Hebrew for sophisticated adults.[2]

A similar thing might be said about the Bible as a topic of study in Israel. In schools it is called "Tanakh" (תָּנָ"ךְ) and that is what most people normally say. But *Miqra* (מִקְרָא) is the term for an academic Bible department and is ubiquitous in high-level writing.

The acronym "Tanakh" has roots in ancient times: It reflects a three-part classification of sacred writings (*Torah*, *Nevi'im,* and *Ketuvim*) which Ḥazal knew and accepted. Each of these three titles was a convenient way to refer to part of the collection, and their combination could refer to the collection as a whole. For instance: "We find in the Torah, in the *Nevi'im*, and in the *Ketuvim* that the mate for a man comes only from the Holy One, Blessed be He" (*Bereshit Rabba* 68:3). But neither this combination nor its acronym ("Tanakh") served Ḥazal as a convenient title for the Bible as a whole.

Miqra is a classical term that goes back to the Bible itself, and variations of it were frequently used in several different ways by Ḥazal. The

[1] The story of *miqra* in its oral and written stages enhances our understanding of what "Written Torah" and "Oral Torah" mean, as well as our self-understanding as the People of Israel. It also provides critical background for meaningful discussion of the Hebrew Bible as a digital text. Although the details of the story are well-known to scholars of the masorah, we are unaware of any clear, balanced summary of the topic in its broadest outlines that stands on its own and is accessible to the public or even to scholars. We therefore devote this section, and much of the next one, to telling the story (and will return to it briefly at the end of the paper). We hope that this not only provides needed background and context to our discussion below of *miqra* as digital Torah, but will also contribute to public knowledge.

[2] Avi heard this story from Nechama Leibowitz at Yeshiva University's Gruss Kollel in Jerusalem, during the year 5751 (1990–1991), when she was already in her eighties.

three-letter root ק.ר.א, from which it is derived, has two completely different definitions as a verb in modern Hebrew: The first is to call out loud, usually to someone else (e.g., "Come here!"), and the second is to read a written text (as in quietly reading a book). But it is no accident that in ancient Hebrew, these two seemingly separate meanings were expressed by the very same word.[3] That is because in ancient times there was no cultural concept of silent reading as an independent activity. To "read" was to take a written text and call it out loud, usually in public.[4]

In biblical times, public readings of the instruction of God ("Torah") were rare. Instead, to seek out God's instruction most often meant to consult a prophet or a priest, or to find it in a dream or through an oracle. Moses himself was consulted in this way for judgment: "And Moses said to his father-in-law, 'For the people come to me to seek out God [***lidrosh*** *elo-him*]'" (Exodus 14:15). But later, during the Second Temple Era, to "call out" the Torah and other biblical books became central to Judean life and was done on a regular basis. Ezra and Nehemiah set the precedent for this new reality at a public event which took place under their guidance: "And they called out the book, the Torah of God, with explanation; and they gave the sense, and caused them to understand the *miqra* ['the calling out' or 'the reading']" (Nehemiah 8:8). From this point on, God's instruction would largely be sought out in His Torah, as did Ezra: "For Ezra had readied his heart to seek out the Torah of the Lord [***lidrosh*** *et* ***torat*** *Ado-nai*] and to do, and to teach in Israel statute and law" (Ezra 7:10). To "seek out" divine truth in the text of the Torah eventually came to be called *midrash*.

In the verse above about the public reading of the Torah, it can be clearly seen that two things are required to perform "the *miqra*": a written

[3] This is clear from a glance at any biblical dictionary. E.g., in BDB (*Brown-Driver-Briggs Hebrew and English Lexicon*) "to read aloud" to the public, for other listeners, is the sense in a significant minority of the places in which the verb is used. To privately read to oneself is less frequent, and even in those cases it was likely done out loud. But in most cases the verb simply means to call out loud to others, without any written text at all.

[4] This may be the plain sense of Rabbi Aqiva's statement (*Mishnah Sanhedrin* 10:1), that "one who reads [*ha-qore*] the external books" has no portion in the World to Come. It is not necessarily "reading" in the modern sense that is being condemned here, i.e., to mentally interpret and comprehend a written text. Rather, it is likely that Rabbi Aqiva denounces "reading" in the ancient sense (*miqra*): to "call out" a text publicly, in the synagogue, which is not part of *Torah, Nevi'im* or *Ketuvim*. Similarly, Menachem Kellner reports in private correspondence that Ḥakham Prof. José Faur once told him that Rabbi Aqiva's words mean "one who reads external books [out loud] with cantillation."

text ("the book"), and trained people capable of "calling it" out loud, clearly and accurately. And so it has been for thousands of years. To this very day, the reading of the Torah in any synagogue depends upon two different kinds of expertise: (1) There is a need for scribes (*soferim*), who are trained to copy the written letter-text from one scroll to another. (2) There is a need for readers (*qore'im*), who are trained to vocalize the text out loud according to a precise tradition that includes pronunciation, stress, and musical elements which express how adjacent words relate to one another.[5]

Ezra exemplified both roles at once. In general, however, the people who wrote Torah scrolls were not the same people who read the Torah in a synagogue. These were two distinct areas of expertise. Scribes need not be expert readers; as they copy the text, they may not know how to pronounce every word that they write, especially since ancient Hebrew writing lacks vowels beyond a limited use of *matres lectionis* (vowel letters). Similarly, readers need not be expert scribes. Although they have learned to vocalize the text aloud in a highly nuanced way, they may lack the skill to produce a written scroll and may not even know how to spell every word they call out (especially since the same vocalized word may be spelled with or without a *mater lectionis* in different places). The scribes and the readers may be thought of as two different professions or guilds. This means the Hebrew Bible has been transmitted for thousands of years in *two* parallel channels, via the tradition of the scribes and the tradition of the readers.

In this sense we may think of the Hebrew Bible in ancient times as "Oral Torah" no less than "Written Torah." It was indeed written; yet the most nuanced, sophisticated, and meaningful component of its transmission was entirely oral. It is therefore no accident that rabbinic dictums about the education of children speak of *miqra* alongside forms of oral learning. An important example is this: "A five-year-old for *miqra*, a ten-year-old for *mishnah*, a thirteen-year-old for *mizvot*, a fifteen-year-old for *talmud*" (*Avot* 5:21).[6] In this passage, *miqra*, *mishnah*, and *talmud* are not so

[5] Throughout this paper, "vocalization" refers to the full tradition of "reading out loud," which encompasses all of these elements.

[6] This statement is not actually part of the Mishnah, but rather a *baraita* appended to the end of *Avot* chapter 5 (much like chapter 6 of *Avot* which is entirely a *baraita*). The *baraita* in context is actually a continuation of the words of Shemu'el ha-Qatan in *Avot* 4:19, which means that the statement cited here is his, even though in the liturgical version of *Pirqei Avot* it immediately follows a passage by Judah ben Tema; see *Tosafot Yom Tov*, *Melekhet Shelomoh*, *Magen Avot*, and *Midrash Shemuel.*

much the titles of three different *texts* as they are three different kinds of *oral activity* which are appropriate to different ages: After he learns the letters, a five-year-old boy listens to his father or teacher call out the vocalization of a verse as they follow it together in a written scroll. Then the child repeats it. They repeat this until the child can vocalize that verse perfectly (and then they move on to the next verse). This oral activity is *miqra*.[7] For Ḥazal, reading the Torah was a skill to be acquired from a very young age, not a task to be deferred until just before *bar mizvah* at age 13! And that is indeed the basic *halakhah*: A minor who has mastered the art of *miqra* may be called up to the Torah and read for the whole community. Some synagogues do so in practice to this very day, especially in Israel.[8]

[7] Ḥazal used the term *miqra* in several related ways. It can mean, as it does here, the oral activity of "calling out" verses or doing a public reading. It can also refer to the oral tradition of a word *as opposed to* its written tradition, for instance: "the vocalization [*miqra*] is authoritative" (*yesh em la-miqra*; *Pesaḥim* 86b). Finally, for Ḥazal, *miqra* most often means the smallest, basic textual unit that is called out, namely a single verse. For instance: "A verse [*miqra*] never departs from its plain meaning" (*Shabbat* 63a), or in Aramaic "as the verse says" (*da'amar qera*). In general, when they cite verses, Ḥazal alternate between calling attention to the oral and written aspects. Common phrases that refer to the oral aspect include "as it is said" (*she-ne'emar*) and the aforementioned "as the verse says" (*da'amar qera*). Equally common phrases that refer to the written aspect include "as it is written" (*kemo she-katuv* or *dikhtiv*) and "the written [verse]" (*ha-katuv*).

[8] "All [people] count toward the quorum of seven [readers], even a minor and even a woman. But the Sages said: A woman should not read the Torah, out of respect for the congregation" (*Megillah* 23a). It seems appropriate that in our times, when there are voices that muster halakhic argumentation towards making a place for women in the reading of the Torah—something for which there was no sanction in the premodern world—that no less concern be shown for children, for whom there is ample and solid halakhic precedent to be called up for an *aliyah* and read it on their own from the Torah, as well as great educational merit in having them do so. This is desirable on an educational level for two reasons: On the one hand, it is a pedagogical travesty to press young adolescents to learn how to read the Torah from scratch in preparation for *bar mizvah*; it is simply the wrong age for that. But on the other hand, not only is the very same activity highly appropriate for much younger children, but a smaller child's mastery of *miqra* and personal experience of "calling out" the Torah to the community can have a lifelong, positive impact on commitment to the study and observance of Torah and to the People of Israel. Avi relates from personal experience at an Israeli synagogue in which he prayed regularly for nearly twenty years, that the experience of reading from the Torah in public had a tremendous, positive influence on children who grew up to be learned and committed Jews. For halakhic analysis of the practice and its educational value, see Rabbi Ovadiah Yosef, Responsa *Yeḥavveh Da'at* 4:23 (the conclusion may be found in *Yalkut*

Mishnah (from age 10) means the memorization of *halakhot* via oral repetition, out loud, but unlike *miqra* there is no written text at all to serve as a guide. It is appropriate to begin such activity at an older age than for *miqra*. Furthermore, the halakhic content of this activity is particularly appropriate in the years leading up to age 13, for that is when the boy becomes responsible to keep the *mizvot* and must therefore know *halakhot*. *Talmud* (from age 15) means mature oral analysis and argumentation related to the *halakhot* and their application.

Yet there is no expectation that every child will participate in the more sophisticated oral activities, namely *mishnah* and *talmud*. A midrash puts it this way (*Kohelet Rabba* 7): "In the way of the world, a thousand come into *miqra*; a hundred of them go out to *mishnah*; ten of them go out to *talmud*; one of them goes out to *hora'ah* (oral halakhic instruction), as it is said, 'I found one man in a thousand'" (Ecclesiastes 7:28). This means that the mostly-oral activity of *miqra* remains the main form of Torah study for most Jews throughout their lives.[9]

Many Jewishly-educated people today assume, when we speak of "Written Torah" and "Oral Torah," that the Written Torah really *was* written, but the Oral Torah was *not* truly oral. This conception is expressed when people say that Rabbi Yehudah ha-Nasi "wrote" the Mishnah, perhaps imagining a manuscript he produced and kept on a shelf. Yet reality during the times of Ḥazal—the collective masters of written and oral alike—seems to have been close to the opposite: On the one hand, even though the "Written Torah" was written, the greater part of its transmission was in fact *oral* (i.e., the activity of *miqra*). On the other hand, the Oral Torah really *was* oral: Rabbi Yehudah ha-Nasi did not *write* the Mishnah, he *spoke* it. He caused it to be memorized by his students through oral repetition (which is exactly what the word *mishnah* means). Rabbinic culture was a highly oral culture, and the only written texts frequently and commonly mentioned by Ḥazal are those that are "called out" in public, namely the books of the Bible.[10]

Yosef 282:8). In Israel, where Zionist families, communities, and schools are typically mixed (in terms of Sephardic and Ashkenazic backgrounds), custom need not be a barrier to this basic and important kind of *ḥinnukh*.

[9] Even for the general populace, *miqra* was nevertheless enlivened and supplemented through Targum and Midrash in ancient synagogues. It is in this way that most people learned ideas from the Oral Torah, both *halakhah* and *aggadah*. Since the middle ages, Rashi's commentary has been a popular, effective and beloved way to meet this very same need.

[10] On the orality of the Oral Torah, the most comprehensive study is Yaacov Sussmann, *Oral Law Taken Literally: The Power of the Tip of a Yod* (Jerusalem: Magnes,

This situation changed drastically during the geonic era, which is when it seems that most of the central oral traditions of Israel were reduced to writing.[11] These include not only the great compilations of the Oral Torah (e.g. Mishnah and Talmud, Midrash and Targum, blessings and prayers), but also the prominent oral aspect of the Written Torah, namely *miqra*: Symbols were invented during this same period to represent the oral nuances of the "calling out" of the Torah.

Yet the process of reduction to writing could not be the same for written and oral Torah. This is because a much-later written version of an early oral composition (like the Mishnah) strove to do no more than accurately represent that single oral tradition. The Mishnah stood on its own, whether as an oral text or a written one. But the written symbols that were created to represent *miqra* could not exist on their own, because *miqra* itself did not exist on its own. Rather, the written tradition of the scribes and the oral "calling out" of the readers existed *in parallel*. What was needed, therefore, was a single text that would *combine* the written tradition of the scribes with the oral tradition of the readers. This was to be the work of the masoretes.

2019) [Hebrew]. For a short, balanced summary of the topic in modern scholarship see Mira Balberg, *Gateway to Rabbinic Literature* (Raanana: Open University of Israel, 2013), pp. 33–38 (Hebrew). One of the many issues analyzed by Sussmann is how medieval Talmudists viewed the orality of the Oral Torah. Medieval scholars in the Islamic world—which revered the written word, and possessed a rich and sophisticated written culture in which Jews avidly participated—often described the Oral Torah as having been in written form from early on. The idea that Oral Torah was transmitted in writing may also have been advantageous to them given the challenge from Karaites, who dismissed oral transmission as unreliable. Quite different was the attitude of Talmudists in Christian Europe, who typically described tannaitic and amoraic activity as *oral* activity. Although they did possess manuscripts of the Talmud, its study still seems to have remained largely oral in Germany and northern France during the era of Rashi and the Tosafists.

[11] Sussmann, ibid. There seem to have been some exceptions that were earlier. For instance, there is Talmudic evidence of written aggadic texts (with reluctant expression of rabbinic approval) and of written blessings and prayers (with harsh expression of rabbinic disapproval). Yet even if there were sporadic written texts, especially for private use, most activity even in the areas of *aggadah* and prayer remained oral during the rabbinic era.

2. *Miqra* as Written Torah

Ḥazal speak of twenty-four *sefarim* ("books") which are called out in public.[12] In fact, the very term "twenty-four books" seems to be the closest thing they had to a title for the Bible as a whole.[13] These twenty-four "books" were scrolls: Torah scrolls, and scrolls of the books of *Nevi'im* and *Ketuvim*. The production of scrolls for public use was governed by the tradition of the scribes, and its details were eventually given expression as formal halakhic rules. To this day, the text in such scrolls still consists of letters along with blank spaces between words and sections. Thus, its public vocalization (*miqra*) depends upon the reader.

The material format of choice in the Middle Ages for the reduction of oral traditions to writing was not the scroll but rather the codex. This was true not only for the vocalization of the books of the Bible, but also for traditions that previously existed in fully oral forms, such as Mishnah and Talmud. The codex was superior to the scroll as technology: Pages were bound together on one side, and text was efficiently written on both sides of each leaf of parchment (much like a modern book). One could flip through a codex to reach a desired page immediately (unlike a scroll that had to be rolled). For the Bible, a codex had the added advantage of being a non-halakhic format, i.e., it was ungoverned by the rules of the ancient scribal tradition. This allowed the creators of biblical codices to add symbols representing the oral tradition of *miqra*, and to further supply the text with notations—neither of which was allowed in a scroll.[14] A codex of the entire Hebrew Bible or a major part of it (*Torah, Nevi'im*, or

[12] For instance: "Just as a bride is decorated with twenty-four kinds of decorations, so too must a Torah scholar be sharp and fluent in twenty-four books" (*Midrash Tanḥuma*, *Ki Tissa* 16). The twenty-four books are listed, and their order and authorship are discussed, in *Bava Batra* 14b.

[13] This is reflected in printed Bibles, some of which are simply entitled "twenty-four" (ארבעה ועשרים); for example, see the title page of *Miqra'ot Gedolot*, first edition (Venice, 1518).

[14] Geoffrey Khan, *The Tiberian Pronunciation Tradition of Biblical Hebrew* (Cambridge: OpenBook Publishers, 2020), vol. I, p. 20. Karaites, who rejected the halakhic tradition, actually read directly from masoretic codices in their synagogues. On the Karaite contribution to the masoretic project, and on the claim that some of the Tiberian masoretes (including Aharon ben Asher himself) were themselves Karaites, see Khan, ibid., pp. 30–33. The available evidence leads to a somewhat different conclusion, in terms of Karaite masoretic scholarship in the Land of Israel from the late 10th century: "All this suggests that Karaite scholars joined forces with an existing stream of tradition of 'Bible scholarship' in Rabbanite Judaism, enhancing it and developing it" (ibid., p. 31).

Ketuvim) was reverentially called a "crown" (*keter* in Hebrew or *taj* in Judeo-Arabic).[15] A magnificent *keter* might be a community's proudest possession.[16]

The process of reducing the oral tradition of *miqra* into a useful written form took generations, and apparently centuries. The two centers of this activity were Babylon and *Erez Yisra'el.* The activity itself became known as *masorah* and its practitioners as masoretes (*ba'alei ha-masorah*), and the vocalized text that they produced is commonly called the masoretic text (MT). In rabbinic Hebrew the verb מסר has the sense of "transmission" or "handing over." However, it may also retain its primary biblical sense as a "bond" (Ezekiel 20:37), along with the idea of "counting" (the masoretic project is replete with the counting of words).[17]

Since a masoretic codex combined two traditions—of the scribes and of the readers—it was most often prepared by two different experts: First a scribe penned the letter text, and then a masorete supplied vocalization and added masoretic notes.[18] Ideally, following this the codex would be further proofread and corrected in all of its aspects, but that did not always take place.

In each geographic area of masoretic activity, vocalization symbols were developed to represent the spoken form of *miqra* according to the received local pronunciation. Many of these symbols serve the most basic and crucial function, namely, to represent vowels. Yet there is far more to *miqra* as vocalization than consonants (letters) and vowels. Of all the versions of the masorah, the most nuanced and precise by far was produced in Tiberias; besides vowels it included many other marks for vocal details. For instance, dots were used to distinguish between two different sounds of the letter ש, to indicate when the letters א and ה are to be silent or spoken and when the letters בגדכפת are voiced as plosives or fricatives, or

[15] Khan, ibid., p. 19.

[16] Heshey Zelcer reports in the name of Professor Elazar Hurvitz (Yeshiva University faculty: Dr. Samuel Belkin Chair in Judaic Studies; Emeritus Professor of Bible) that there is a manuscript of the *She'iltot* which bears the title *Keter*, and that a particularly good (or important) version of a manuscript might also be called by that name.
In terms of the *She'iltot* specifically, perhaps it might also be thought of as a *Keter Torah* in the sense that it teaches halakhah and aggadah didactically in connection to the Torah's weekly portions, thus unifying the Written Torah with the Oral Torah.

[17] Khan, ibid., pp. 14–15 and n. 12. Note that "counting" is from the same root as "scribe" in Hebrew.

[18] Khan, ibid., 21–22.

to indicate the lengthening of consonants. Other occasional marks indicated the sounding of *sheva* or the shortening of a vowel within a syllable, when two adjacent words are read as if they are a single word, and secondary stress within a word. The primary stress marks are musical, and they further serve as building blocks in a complex hierarchy of conjunctive and disjunctive accents which maps out divisions and subdivisions in every verse, and simultaneously indicates the degree of conjunction or disjunction between every pair of adjacent words in the entire Hebrew Bible.

The addition of these vocalization symbols to the letters produced a written text with clearly meaningful words (as opposed to plain letter-sequences that could be read in more than one way). Masoretic notes were then added, which mostly dealt with how one must spell a particular word according to the scribal tradition in each place that it occurs throughout the Bible. Such notes were impossible without prior vocalization, and thus oral *miqra* came *before* correction of the written text in the production of masoretic codices.[19] Once complete, a well-executed *keter* served as a model to determine both the correct spelling of words within scrolls and their correct vocalization for readers.

It may seem odd that most masoretic notes point out the correct spelling of words (i.e., they focus on the scribal tradition), rather than dealing with vocalization (the tradition of the readers). After all, anyone who looks at a vocalized Hebrew Bible—whether a manuscript or a printed version—sees relatively large, clear Hebrew letters surrounded by tiny dots and circles and lines, whose exact positions and angles are critical. It must surely be easier to reproduce the letters with accuracy than to copy all those tiny symbols correctly! If the masoretic notes are designed to prevent errors, then should they not focus on vocalization as well?

The answer to this question may be counterintuitive to the modern mind. It turns out the Tiberian masoretes did not *copy* the vocalization symbols from codex to codex. Rather, they called out the *miqra* orally as they transcribed it via written symbols. When a masorete added vocalization to a codex, he literally transformed an oral tradition into a written one. Furthermore, the oral tradition transcribed by the masoretes was

[19] This point has been made with great clarity by Breuer, *The Aleppo Codex and the Accepted Text of the Bible* (Jerusalem: Mossad Harav Kook, 1976), pp. 91–94; and later again in "The Letter Tradition and the Reading Tradition," *Iyyunei Miqra u-Parshanut* 7 (2005), pp. 25–32 (both Hebrew). He eloquently stresses, as a fundamental starting point, that the work of the masoretes is a nuanced synthesis between the two parallel traditions of the scribes and the readers, and that the latter tradition comes first in their work. But his apt expressions of this point are embedded within highly technical contexts rather than standing on their own.

highly cohesive: There are surprisingly few discrepancies between the Tiberian codices in terms of the vowels and accents. The differences that do exist are mostly about secondary or tertiary aspects of vocalization. This indicates that the oral tradition known to these masoretes was a solid one. In fact, when the Tiberian codices are compared to one another, they exhibit much greater cohesiveness in terms of the oral tradition than in terms of the scribal tradition! That is why the masoretes developed an error-correction system which focuses almost exclusively on the latter.

That masoretic vocalization directly transformed the oral into the written is indicated by the *types* of vocalization errors which appear on occasion in some of the oldest Tiberian codices. For instance, when two similar verses have different cantillation, one codex (B) shows that the masorete sometimes "heard" the cantillation of one verse when he accented another. Another codex (S1) shows that its masorete sometimes mistakenly followed common musical sequences of cantillation even in verses where they are inappropriate. Such errors indicate that these two masoretes were transcribing what they knew *orally*, even if imperfectly. They were *not* copying from another codex.[20]

A further indication of the Tiberian vocalization's striking overall cohesiveness may be found in masoretic lists of vocal discrepancies. While it is true that masoretic *notes* in the Tiberian codices focus on the scribal letter-text, there are also masoretic *treatises*—a separate literature, as opposed to the notes in the margins of the biblical text—which focus to a significant degree on vocalization. One genre of this literature is called *ḥillufim* ("discrepancies"), which are lists of differences between different "schools" within the masorah. For instance, there are lists of differences between the tradition of "Ben Asher" and that of "Ben Naphtali" (both c. 900). In the *Sefer ha-Ḥillufim* of Mishael ben Uzziel (c. 1000 but seemingly based on earlier sources), there is first a list of seven "global" differences between these two masoretes. Then there is a detailed list of 867 disagreements between them regarding individual words, along with 406 places where they agree (presumably in contrast to another tradition). This detailed list follows the order of the biblical books from beginning to end, paying special attention to its division into liturgical units for public reading.

What is striking is not the *number* of recorded differences between Ben Asher and Ben Naphtali (which is small given the number of words in the Bible), but rather their *type*: Nearly all of them deal with matters of secondary stress within a word, or with small disparities regarding conjunctive accents, especially in the poetic books. (These are matters of which

[20] This striking and important point was first demonstrated by Rabbi Mordecai Breuer; see his conclusion in *The Aleppo Codex*, p. 67.

many modern readers are barely cognizant, and which most do not enunciate in their reading.) In other words, at the time of Ben Asher and Ben Naphtali, there were hardly any disputes between the Tiberian masoretes about the fundamentals of vowels and accents throughout the Bible. Mishael ben Uzziel's list seems to have been meant to serve expert readers who needed clarification only about the most minute points of disagreement within the oral tradition of *miqra*.

Not only is the Tiberian vocalization highly nuanced and consistent, it is also ancient. To be sure, it betrays medieval linguistic influence in certain ways. Yet as a whole it represents a living oral tradition that was transmitted intact from Second Temple times, no less so than the written letter-text of the scribes. Even though its *symbols* are clearly medieval, what they were designed to *represent* is not.[21] In this light, the common attitude which dismisses the masoretic vocalization as a medieval invention, with no greater claim to authority than a medieval commentary, is in error. On the contrary, the medieval commentators and grammarians strove to interpret *both* of the two parallel, ancient transmissions of the biblical text that they received: the tradition of the scribes and the tradition of the readers.[22]

What made it so hard for the scribes to achieve consistency in the letter-text, as opposed to the oral transmission? The problem is that a particular word might be spelled in more than one way (e.g., with or without a *mater lectionis*) wherever it occurred. This was compounded by the further difficulty in identifying the occurrences of a particular word, since an unvocalized letter-sequence can be read in multiple ways. This led to a reality in which the *oral* transmission was *more stable* than the *scribal* one. As Rabbi Mordecai Breuer describes it: "The Talmudic sages were not experts in deficient and *plene* (full) spelling (*Qiddushin* 30a). But we have no evidence that they lacked expertise in the vocalization of the words. The opposite is true. In many places it is said: 'Do not read… but rather…'. Yet we only say, 'Do not read' to someone who already possesses a received tradition for reading."[23]

In the masoretic age, this meant codices that were highly consistent with one another in terms of vocalization, but not in terms of spelling. It also meant the development of a complex apparatus of masoretic notes

[21] Yosef Ofer, *The Masora on Scripture and Its Methods* (Berlin: De Gruyter, 2019), pp. 3–4; Khan, I:0:8, pp. 56–85.

[22] On exegesis and grammar within the masoretic project itself, see Ofer, ibid., pp. 221–263.

[23] Breuer, ibid., p. 91 (ג.10).

to guide masoretes as they checked and corrected the spelling in their codices. This, however, was done with varying degrees of success. The culmination of the process was reached in a single model codex that reached an unparalleled level of accuracy, clarity, and consistency in the written and oral traditions alike. This was the great *taj* of the entire Tanakh edited by Aharon ben Asher, which later became known as the Aleppo Codex (AC).

The extraordinary perfection of the AC is reflected in the way that its tiny vocalization signs are clear, unambiguous, and follow highly consistent patterns.[24] It is further reflected in the way that its letter-text accords with the masoretic apparatus—both its own masoretic notes and those found in parallel manuscripts—nearly 100 percent of the time.[25] It is reflected yet again in how both its spelling and vocalization match that of the majority of close parallel codices throughout the Bible (even in places where an anomaly, an ambiguity, or an outright error exists in one or two of them). When we consider that there are roughly three million (!) orthographic signs in the masoretic Bible (letters and vocalization),[26] such perfection is almost a superhuman achievement.

Maimonides describes the AC as follows: “All relied upon it, since Ben Asher corrected it and examined it meticulously for many years, and corrected it many times according to tradition…”.[27] What Maimonides

[24] This perfection is especially evident in two areas: (1) *ga'yot* and (2) conjunctive accents in the poetic books; see the eloquent description of the AC’s uniqueness in its vocalization, especially in these two specific areas, by Moshe Goshen-Gottstein in his foreword to Israel Yeivin, *The Aleppo Codex of the Bible: A Study of its Vocalization and Accentuation* (Jerusalem: Magnes, 1968), pp. v–vii. It is because of this perfection that the vocalization of the missing parts of the Aleppo Codex, along with its spelling, can be reconstructed with little difficulty and at a very high degree of certainty (except perhaps for *ga'yot*, which are to some degree a function of probability).

[25] Chapter 5 of the online Hebrew introduction to *Miqra According to the Masorah* (MAM) contains a chart listing all possible exceptions in the extant part of the Aleppo Codex (most of the items in the chart are not definite errors but rather places where there is a degree of doubt, no matter how slight, regarding a certain letter). See <https://he.wikisource.org/wiki/משתמש:Dovi/מקרא על פי המסורה/מידע על מהדורה זו/פרק ה>. Also cf. Breuer, p. 140 (ד.7); Ofer, pp. 34–48.

[26] See Yosef Ofer, “Proofreading the Biblical Text for the Jerusalem Crown Edition,” *Leshonenu* 64 (2002), p. 199 (Hebrew).

[27] Maimonides, *Laws of Tefillin and Mezuzah and a Torah Scroll* (8:4). For alternative translations see Ofer, *The Masora on Scripture*, p. 67; *The Code of Maimonides, Book Two: The Book of Love*, trans. Menachem Kellner (New Haven: Yale University Press, 2004), p. 100. The translation of the final phrase, כמו שהעתיקו (“according

writes is well-evident from a careful, systematic examination of the codex.[28] We might imagine that after his initial pass, in which he vocalized the codex, Aharon Ben Asher read it aloud repeatedly, from beginning to end, making corrections along the way. By the end of his lifetime, he brought his codex as close to perfection as a human being could possibly achieve.

Once this near-perfect, model codex existed, and its quality became known to Jews around the world, subsequent transmission of the masorah became a matter of *written* transmission. Accuracy could be reached in two different ways: (1) By consulting the Ben Asher codex itself, either directly or through testimony about it. (2) By consulting the text and the masoretic apparatus within other presumably accurate manuscripts, and utilizing that data to bring the text as near as possible to perfection. The degree of success was mixed: The letter-text of the Torah—but not that of *Nevi'im* and *Ketuvim*—was brought very close to perfection in most communities via the second method, based on the rulings of Rabbi Meir ha-Levi Abulafia ("Ramah," c. 1170–1244) in his work *Masoret Seyag la-Torah*. It was brought to absolute perfection via the first method by the Jews of Yemen.[29] Thus, when it comes to Torah scrolls, the scribal tradition as

to tradition"), follows Maimonides' consistent use of similar formulations to indicate accurate transmission of the Torah, usually of the Oral Torah but also of the text of the written Torah (e.g., *Laws of Tefillin and Mezuzah and a Torah Scroll* just above in 7:8); see the commentary *Yad Peshutah* by Rabbi Dr. Nachum L. Rabinovitch (Jerusalem: Maaliot Press, 1994). Also see Mordechai Glatzer, "The Aleppo Codex: Codicological and Paleographical Aspects," *Sefunot* 19 (1988), p. 226 and n. 6 (Hebrew). The verb להעתיק meant "to transmit" in medieval Hebrew, as opposed to the narrow sense of "to copy" in modern Hebrew. Even a translator was called a מעתיק. Ofer notes that if Maimonides' phrase refers to traditions that reached him regarding Ben Asher's expertise, then it should be translated "as people have transmitted." But this is less likely, given the way Maimonides uses similar formulations in other contexts.

[28] Meticulous examinations of the Aleppo Codex by Israel Yeivin and Mordecai Breuer in their respective books bear this out. Yeivin's book (above, n. 24) describes the vocalization of the Aleppo Codex in exhaustive detail, while Breuer's book (above, n. 19) uses the perfection of the extant parts of the codex to prove that both the scribal text and the vocalization in its missing parts can be reconstructed with a very high degree of certainty.

[29] Breuer, *The Aleppo Codex*, pp. 87–89. Yet the scribes in Yemen may have further corrected the letter-text of the Torah based on masoretic notes, leading to a text which may be slightly more accurate than the Aleppo Codex itself! Cf. Yosef Ofer, "Cassutto's Notes on the Aleppo Codex," *Sefunot* 19 (1989), p. 339 and Jordan Penkower, *New Evidence for the Pentateuch Text in the Aleppo Codex* (Ramat-Gan: Bar-Ilan University, 1992 [Hebrew]) pp. 67–71; 77–80, 90. When it comes

transmitted by the Tiberian masorah was ultimately adopted by all of Israel throughout the lands of its exile.

When it comes to vocalization, dissemination of the masoretic project involved a dramatic shift in principle: The Tiberian masoretes once *wrote* what they knew *orally*. But now Jews around the world took the *written* masoretic system of vocalization and expressed it *orally*, interpreting the signs based upon their own local traditions for pronouncing the biblical text and singing it. The written signs for the vowels and accents were interpreted anew in the communities of Israel around the world, from Yemen to Spain and from Germany to India. This reality led to the Tiberian vocalization signs being adapted in certain small ways to conform to the local vocalization, in addition to errors in transcription (since the signs were now often copied rather than written from oral memory). In general, this moderate trend affected the vowels more than the accents.[30] But even given such changes, the vocalization found in later manuscripts and printed versions is still remarkably similar to the Tiberian masoretic codices from which they ultimately derive. The differences between all of these texts are about such small details that they are only apparent to those who know what to look for.[31]

In the age of printing, the dual tradition of the scribes and the readers was mass-produced in published versions, which for the very first time allowed numerous people and communities to possess texts that were identical to each other in every detail (yet less accurate than the Tiberian codices). This meant that efforts to correct the text could begin with a single agreed-upon edition. The second edition of *Miqra'ot Gedolot* (Venice, 1524–1525), in which the biblical text was based upon manuscripts and corrected via a corpus of masoretic notes compiled by its editor, served that purpose. About a century later, its letter-text of the Torah was critiqued by Rabbi Menaḥem di Lonzano (*Or Torah*), and the entire

to *Nevi'im* and *Ketuvim*, the examination of codices by Jewish scribes in Yemen similarly shows that they tend to be very close to the spelling of the Aleppo Codex in its extant parts (and thus to what it presumably contained in its missing parts). In the infrequent places where their spelling differs from that of the Aleppo Codex, there is nearly always an explicit masoretic note in the Yemenite codex. It is thus possible that the entire text of *Nevi'im* and *Ketuvim* was brought close to masoretic accuracy via a careful application of the masoretic apparatus (second method). But it seems more reasonable that the text was initially based on the Aleppo Codex itself (first method), and then changed in certain places in adherence to a masoretic note.

[30] Breuer, ibid., pp. 67, 91–94.

[31] Ofer, *The Masora on Scripture*, pp. 203–204.

Tanakh including vocalization by Rabbi Yedidyah Norẓi (*Minḥat Shai*). Together, their works influenced the spelling in Torah scrolls and the vocalized text in subsequent published editions of the Hebrew Bible.[32]

In the 20th century, several scholars of the masorah published the Hebrew Bible based directly on Tiberian manuscripts, rather than relying on much later sources. The first was Paul Kahle,[33] who tried to purchase the AC for this purpose, but the Jewish community of Aleppo rejected his offer. Instead, he was forced to use the Leningrad Codex (LC), another monumental Tiberian manuscript which was written and vocalized "on the basis of books corrected by the instructor Aharon ben Moshe ben Asher" according to its colophon. Its vocalization is remarkably close indeed to that of the AC.[34] Kahle used the LC as the base text for the 3rd edition of *Biblia Hebraica Kittel* (*BHK3*, completed 1937),[35] which was later revised as *Biblia Hebraica Stuttgartensia* (*BHS*, completed 1977), and is now being revised yet again as *Biblia Hebraica Quinta* (*BHQ*). Although *BHS* enjoys a strong academic reputation as a tool for textual study of the Hebrew Bible, its Hebrew text is actually riddled with errors as a transcription of the LC.[36] A better edition of the LC was published by Aharon Dotan in 1973.[37]

The AC was brought to Israel in 1958, but it suffered extensive damage; about 60 percent of it remains.[38] It soon became available to scholars

[32] Ofer, *The Masora on Scripture*, pp. 186–187.

[33] The orientalist Paul Kahle (1875–1964) and his family were persecuted by the Nazis for helping Jews, and fled to England just before World War II. Rabbi Yeḥiel Ya'akov Weinberg was Kahle's doctoral student and assistant until the war.

[34] For the most part, to copy the vocalization of the Leningrad Codex is the same as to copy the vocalization of the Aleppo Codex (in its extant parts). This is true despite intermittent anomalies which are easy to identify, since they stand out clearly in contrast to the consistent rules of vocalization in the Aleppo Codex, and in contrast to what is actually found in the Aleppo Codex and other Tiberian manuscripts. Therefore, if used with care, the Leningrad Codex serves as a solid basis for reconstruction of the vocalization of the Aleppo Codex in its missing parts. For analysis see Breuer, *The Aleppo Codex*, pp. 46–51, 67, 91–94.

[35] *Biblia Hebraica*, 3rd edition, was edited by Rudolph Kittel (1853–1929) and Paul Kahle; it was published in installments in Leipzig from 1929–1937.

[36] See below regarding errors in the *Westminster Leningrad Codex* that were inherited from *BHS*.

[37] Tel Aviv: Adi, 1973 (in cooperation with the Judaic Studies department at Tel Aviv University); reprinted many times (with a short commentary) by the IDF rabbinate from 1975 through the 1990s. Later republished as *Biblia Hebraica Leningradensia* (Peabody, MA: Hendrickson Publishers, 2001).

[38] On the missing parts of the Aleppo Codex see Matti Friedman, *The Aleppo Codex: In Pursuit of One of the World's Most Coveted, Sacred, and Mysterious Books* (Chapel

and later to the public. Among the earliest scholars who examined it methodically and demonstrated its unique value were Professor Israel Yeivin and Rabbi Mordecai Breuer. The former described its vocalization in exhaustive detail and frequently compared it to closely related texts, thus demonstrating its unparalleled quality and unique features. The latter published a monumental study proving how the spelling and vocalization in its missing parts can be reconstructed with near certainty based on an objective method. He further published three sequential editions of the entire Tanakh based upon it and upon his method (1982, 1996, 2000).[39] Breuer's method was largely adopted—with certain reservations and minor consequential changes—by the *Miqra'ot Gedolot Haketer* project of Bar-Ilan University (1992–2019), which is also based upon the AC and a reconstruction of its missing parts. The same thing is true of the *Hebrew University Bible Project* (HUBP).

While Jews across the ages traditionally focused on determining details of the masoretic text down to its tiniest nuances and learning to vo-

Hill, NC: Algonquin Books, 2012); a thoughtful rebuttal to Friedman is offered by Ofer, *The Masora on Scripture*, pp. 136–144. Towards the beginning of the documentary film *The Lost Crown: The Mystery of the Lost Pages of the Aleppo Codex, the Most Important Bible in the World* (Israel television *Kan 11*, 2019) there is a short interview with Michael Magen, manuscripts preservation expert at the Israel Museum. Magen remarks (5:15 ff.): "Two types of people deal with the Crown: those who occupy themselves with what there *is*, and those who occupy themselves with what is *lost*. I am one of the people who occupy themselves with what there *is*" (our emphasis). The film is available online; see <https://youtu.be/MFyQrH7WWdA?t=303> (Hebrew). We side with Magen's emphasis on "what there *is*." Despite the loss of much of the Aleppo Codex, the true Crown—namely, the dual tradition of the scribes and the readers—was never lost. It is true that the Aleppo Codex is a critical link in the chain of transmission for the dual tradition, and that link was tragically damaged. Yet it is precisely what *remains* of the Aleppo Codex, along with the combined efforts of scribes, readers, and masoretic scholars across the generations, which thankfully enables us to mend that single link via reconstruction of the lost parts.

[39] Besides Breuer's original method, his later editions also take into account further evidence about the missing parts of the Aleppo Codex which has been uncovered by several scholars. The most important studies which present such evidence and evaluate it are: Ofer, "Cassutto's Notes on the Aleppo Codex," *Sefunot* 19 (1989), pp. 277–344 and "The Aleppo Codex and the Bible of R. Shalom Shachna Yelin" in *Rabbi Mordecai Breuer Festschrift: Collected Papers in Jewish Studies*, ed. M. Bar-Asher, volume 1, pp. 295–353 (both Hebrew); Penkower, *New Evidence*; Raphael Zer, "Rabbi Jacob Sappir's *Me'orot Natan* (ms JTS L 729)," *Leshonenu* 50 (1989), pp. 151–182 (Hebrew).

calize them, we might suppose that modern academic publishing had different priorities. The *HUBP* (based on AC) and the *Biblia Hebraica* series (based on LC) are both, after all, tools for the textual criticism of the Bible. They provide the full MT, whether AC or LC, as a basis or starting point. Then they supplement it with a critical apparatus which suggests alternative readings based on ancient versions or other evidence. However, since the apparatus is provided as a supplement to the main text—namely the MT—these editions have the psychological effect of placing the MT at the center, and its tiniest details become the object of intense editorial focus. As Emanuel Tov puts it: "Remarkably, although in principle the critical editions remove our thinking from MT by discussing other versions in the apparatus, in practice they make MT even more central than before because they compete with each other in producing ever more precise versions of the Leningrad or Aleppo codex."[40]

Yet there may be good reasons for the centrality of the MT (as opposed to other ancient witnesses to biblical text such as the Septuagint or the Dead Sea Scrolls), even within scholarly projects which accord to it no special status in principle. The most obvious suggestion is that the MT is, overall, the best available Hebrew text of the Bible. Indeed, according to Tov, the MT is an "excellent" representative of most (but not all) books of the Bible and especially the Torah. He concludes: "Overall, compared with the other known texts, MT is generally the best text available. By 'generally' we mean that this is not the case in all words or all verses, nor in all books."

And yet, this important evaluation does not suffice to explain why the MT serves as the central Hebrew text even for those who do *not* revere it as tradition. It is quite possible, after all, to present different versions of ancient texts in parallel columns and draw attention to their differences. Such an arrangement may be viewed in the first volume of Benjamin Kennicott's critical Bible (18th century). In this edition, the MT's Torah (unvocalized) is presented parallel to the Samaritan Torah, with other variants listed below. A similar method would presumably work for certain biblical books using some of the fuller texts from Qumran, or for the whole Tanakh using an ancient translation like the Septuagint. However, Kennicott published the MT alone for the rest of the Hebrew Bible (volume two), since the Samaritans have only the Torah in their canon; a similar problem would exist for other ancient versions that contain only part of the Bible. Even if the masoretic Hebrew text is shown alongside a full

[40] "Editions and Translations of MT," part 10 of *The (Proto-)Masoretic Text: A Ten-Part Series* at <https://www.thetorah.com/article/editions-and-translations-of-mt>.

ancient translation like the Septuagint, the juxtaposition is primarily designed to reveal textual variants or ancient exegetical traditions via the latter in order to enhance our understanding of the former. Only the MT transmits a full ancient version in Hebrew, and quite a good one at that. It thus remains central even when compared to partial Hebrew witnesses or full ancient translations.

Beyond this, however, something crucial is lacking in *every* ancient version besides the MT: Not one of them preserves a systematic vocalization for the Hebrew. At best, sporadic aspects of ancient vocalization may be gleaned from them, but it is impossible to "call them out" uninterruptedly, in a highly nuanced way, based upon a firm oral tradition, because they lack full vocalization. Any "calling out" that might be done from them today is nothing more than a reader's best interpretation of the scribal text, or a scholar's informed reconstruction. But it is not an ancient vocalization that has been transmitted and transcribed. Only the MT provides an authentic oral tradition that is nuanced and continuous in every verse and for every word of the Bible. The nuanced vocalization of every other ancient version is long lost.

In other words, no alternative to the MT exists that can be properly called *miqra*. Even if it is possible to recover an alternative *scribal* tradition directly from ancient Hebrew scrolls, or attempt to derive it from an ancient translation (e.g., the presumed *Vorlage* of the Septuagint), it is still impossible to recover an alternative *oral* tradition. Scribal traditions survive if writings survive, but oral traditions are lost unless they are transmitted or somehow recorded. The work of the masoretes is indeed the starting point for any critical work, not just because it is complete and of better overall quality than any of the alternatives, but also for a reason that is absolute and objective: The MT—and *only* the MT—contains a full, ancient vocalization that was transmitted and recorded with extraordinary care. No other source simultaneously provides the parallel traditions of the scribes and the readers in all their fullness. No other source is or can possibly be *miqra*.

As such—critical scholarship aside—the work of the masoretes is the sole foundation for the literary tradition and spiritual culture of Israel. It is nevertheless true that there are educated Jews today, including traditional or Orthodox ones, who find that to study alternative textual versions of the books of the Bible can be intellectually stimulating, fascinating and rewarding.[41] At times they may find the results of such study to

[41] On the transmission of the biblical text as understood in the Jewish tradition, and the problem of textual variants in traditional Jewish thought and *halakhah*,

be critical for understanding a particular biblical passage. And yet, all such engagement remains on the level of theory alone. No suggested emendation to the MT, no matter how convincing, can ever be "called out" according to an ancient oral tradition. To do so would mix apples and oranges: The masoretes did not *ask* how the text *should* be vocalized, but rather transmitted their received vocalization. Any alternative to the MT, in contrast—if it is vocalized—hypothesizes a "calling out" for which no such tradition exists. This can of course be done, but in doing so a scholar engages in the *opposite* of masoretic activity: To weave such a change into the body of the masoretic Hebrew is not only foreign to the Jewish tradition but also a glaring historical anachronism.[42]

In terms of Jewish life today, as in the past, textual criticism provides less direct value than producing "ever more precise" and more useful editions of the Tiberian masorah (especially now in digital online formats), so that it can be properly "called out" (*miqra*) in study and public reading. And when it comes to modern Bible scholarship and scholarly editions of the Bible, they too take the MT as their starting point. This is because only the MT provides us with a full, ancient tradition of vocalization. Whether as the bearers of tradition or as scholars, only editions of the MT can serve us as *miqra*.

3. *Miqra* as Digital Torah (I): The Transcription

In the computer age, the dual tradition of the scribes and the readers, as synthesized by the masoretes, took on a new form as digital text. It might

see B. Barry Levy, *Fixing God's Torah* (Oxford University Press, 2001). On the place of the masoretic project within the wider story of how the biblical text was transmitted, and the limited yet fascinating degree of its relevance to textual criticism, see Ofer, *The Masora on Scripture*, pp. 170–188. A demand for ancient textual witnesses to the biblical text on the part of Torah scholars is reflected in the "Mikraot Gedolot for Scholars" provided as a supplement to AlHaTorah.org under a separate URL <https://mgs.alhatorah.org/>.

[42] This is a principled problem in the approach taken by *The Hebrew Bible: A Critical Edition* (HBCE), published by the Society of Biblical Literature. Its text weaves fully vocalized emendations (with vowels and accents) directly into the masoretic Hebrew. Without reading the apparatus and commentary, it is not immediately clear whether the suggested emendation modifies the scribal tradition or the oral tradition (or both): It is as if the MT constitutes a single, linear text—which has been emended—rather than a synthesis of two parallel traditions. HBCE further employs masoretic vocalization signs, but unlike the masoretes it does so in order to suggest how a word *ought to be* called out theoretically, rather than transcribe how it *is* called out in practice. An eclectic text like HBCE is an anachronism which suggests a basic misunderstanding of the masoretic project.

seem at first glance that to produce a digital version of the Tiberian masorah is simply a matter of *transcription*: The MT's millions of orthographic signs—every single letter and each vocalization mark—must be typed accurately on a computer keyboard to convert the text into digital form. This activity may be viewed as a step not unlike that of the medieval scribes who copied and edited masoretic codices, or that of the printers who first typeset the Tiberian text with vocalization in the late fifteenth century.[43] Digital masoretic typing was further enabled by the development of specialized fonts to show all these symbols on screens and in printouts. Eventually, a Unicode standard for Hebrew emerged, enabling digital masoretic transcriptions to be represented in a font-independent way on multiple word processors and websites. But this is no different in principle from the creation and development of written symbols in the manuscripts, or of typefaces in printing.

Indeed, the early efforts towards creating digital masoretic Bibles were basic transcription projects, and they saw themselves as nothing more than that. Their explicit goal was to transcribe the vocalized Hebrew text of *BHS*, following the completion of its publication in 1977. The "Preface to the 1999 Hebrew-English Edition" of the JPS Tanakh records these efforts:

> [A]t the University of Michigan, H. Van Dyke Parunak and Robert Eckert devised computer-readable codes for the biblical text's characters and main features; Parunak oversaw the transcription of BHS into three megabytes of data (1982). Soon thereafter, Richard E. Whitaker of the Claremont Graduate Schools coordinated revisions. Finally, J. Alan Groves of Westminster Theological Seminary (Philadelphia) with Emanuel Tov of The Hebrew University (Jerusalem) directed a proofreading team (1987), a project that JPS helped to fund.

[43] The first publication to be successfully vocalized was a *ḥumash* printed by Abraham ben Ḥayyim dei Tintori in Bologna in 1482. This edition also contains the Targum (unvocalized) alongside the biblical text in a small Sephardic semicursive font, and the commentary of Rashi above and below in the same font. The first vocalized edition of the entire Tanakh was that of Soncino (1488). Other early editions of biblical texts gave up on vocalization midway, because of the difficulty, or made no attempt to include it at all. For instance, in the very first published edition of a biblical text in Hebrew, which was an edition of Psalms with the commentary of Rabbi David Qimḥi (Bologna, 1477), the publisher quit vocalizing after the first few Psalms. Some early editions of the *ḥumash* (such as Híjar, 1486 and Híjar, 1490) lack vocalization altogether, although their owners often supplied vocalization by hand.

> The result is called the Michigan-Claremont-Westminster (MCW) electronic BHS. It has provided JPS with a text nearly identical to the Leningrad Codex Manuscript. Each round of revision has corrected previous typographical errors and misreadings while introducing a smaller number of other typos and mistakes. Its machine-readable format has nearly precluded new typos in our own production process. Meanwhile, BHS notes have provided vital supporting documentation.[44]

The entire focus here is on accurate transcription and the elimination of errors. Yet the textual basis for this project was not the Leningrad Codex itself, but rather BHS, which itself is highly imperfect as a transcription. Nevertheless, that basis was sufficient for the project's intended purpose: It meant to provide an electronic text for linguistic analysis which might prove useful to academic scholars of religion, as well as to translators of the Bible. For that BHS was sufficient, and even advantageous given its academic reputation.

The electronic transcription of the masoretic text was *not* intended for the traditional purposes of the masorah, namely, the correction of the letter-text in scrolls and the oral vocalization of the biblical books ("reading" or *miqra*). For such purposes the Leningrad Codex is inadequate due to the many hundreds of errors in its letter-text[45] and the numerous errors

[44] *JPS Hebrew-English Tanakh: The Traditional Hebrew text and the New JPS Translation* (Philadelphia: JPS, 1999), pp. xii–xiii. On this electronic transcription project also see the J. Alan Groves Center website. Robert Kraft (University of Pennsylvania) is mentioned there as a predecessor of Groves (along with Tov); similar credits may be found in Norman L. Geiser and William E. Nix, *From God to Us: How We Got Our Bible*, 2nd edition (Chicago: Moody Publishers, 2012), pp. 193–4. Christopher Kimball, long-time editor of the Unicode version of the *Westminster Leningrad Codex* (now an independent edition called the Unicode/XML Leningrad Codex or UXLC), informed us in private correspondence that Kraft was responsible for the release of the electronic text, when it was deemed complete, to the Oxford Text Archive <https://ota.bodleian.ox.ac.uk/repository/xmlui/handle/20.500.12024/0525> on February 7, 1987; this made it available for the first time to scholarly projects. Over the years the text was variously called MCWT (Michigan-Claremont-Westminster Text), CCAT (Center for Computer Analysis of Text at the University of Pennsylvania, still available at <https://ccat.sas.upenn.edu/gopher/text/religion/biblical/mbhs/> with an explanation at <https://ccat.sas.upenn.edu/gopher/text/religion/biblical/mbhs/readme.txt>), and eBHS (=electronic BHS). We are grateful to Chris for his help in hunting down these details about how the Tanakh first became a digital text.

[45] In most of these cases, the spelling in LC contradicts its own masoretic notes. This raises the question as to what a "diplomatic edition" of a masoretic codex

or anomalies in its vocalization. BHS adds a great many further inaccuracies. Yet these errors and anomalies rarely affect the *meaning* of the Hebrew text. Thus, the Hebrew text of BHS, in digital form, was *good enough* for religion departments in the universities. Moreover, it was beneficial for those who sought electronic tools to make Bible translation an easier task. The latter group included Christians who found a calling to translate the Bible into every human language. All of these users—who were by and large the intended market for an electronic transcription of the Hebrew Bible—supported the elimination of errors in *transcription* so that the text would be accurate and reliable in terms of its own stated goals. But the exact *basis* for transcription was less critical to them. The LC is after all the oldest complete masoretic Bible in existence, and an excellent (if imperfect) representative of its genre. And BHS is a respected representative of the LC. These two facts were quite enough for them.

It is likely that other electronic transcriptions of the masoretic Bible were attempted in the early decades of the computer age.[46] Yet the MCW electronic BHS is the first one based (even indirectly) upon a bona fide Tiberian masoretic manuscript. Furthermore, it is this electronic transcription which became the basis for other important projects down the line. Evidence shows that *Mikraot Gedolot Haketer* (the first attempt to publish a critical edition of the entire *Miqra'ot Gedolot*) and Mechon Mamre's Tanakh (the very first online, digital Tanakh corrected according to the masorah for use by Jews) both began with a version of the MCW electronic BHS.[47]

actually means: Should a transcription of the vocalized text in the LC be considered diplomatic, i.e., as completely loyal to a single textual witness, if it ignores other masoretic materials within the very same document which bear upon the transcribed text? The masorete himself, Samuel ben Jacob, made an effort to correct errors based upon his own notes and other masoretic traditions that he knew; for those errors that he missed it seems wrong to ignore the larger masoretic project in which he saw himself as taking part, including his very own notes.

[46] In correspondence, Christopher Kimball relates: "When I was a boy at the University of Michigan, there were some strange (to me) fellows with *kippot* typing stuff onto punch cards. That was ca 1962. So 1987 is a little late to be the start of a digital Tanach, I think." It is reasonable to suspect that sporadic attempts to transcribe the Tanakh were made by numerous Jews who had computer access at the time and were tempted to transcribe the Torah in this new medium.

[47] The evidence is found in minute transcription errors which occur in one project or the other, and are also found in MCW-derived text. Some of these errors have since been corrected and are no longer visible (since neither of these projects provides public documentation for changes). Several of them have nevertheless

Originally, the MCW text was in a special coding which transformed pointed Hebrew text into characters found on existing, English keyboards. For example, the first word of Genesis was "B.:/R")$I73YT" (=בְּ/רֵאשִׁית).[48] This was suitable for machine processing or specialized typesetting programs, but not for general use. It was only in 2003 that an online version of the MCW transcription was made available to the public by Christopher Kimball, at a website (tanach.us) which was then called the *Westminster Leningrad Codex* (WLC).[49] This was revolutionary because the website was in conventional Hebrew characters, and also because Chris generously released the entire digital text and its ongoing revisions into the public domain.[50] This allowed the WLC to be corrected, enhanced and reused for multiple purposes. And that is exactly what happened: The WLC quickly became ubiquitous, the most widely used text of the Hebrew Bible in the online world. It is now maintained independently and called the "Unicode/XML Leningrad Codex" (UXLC).[51]

There is an "open-source programming" adage that says: "Given enough eyeballs, all bugs are shallow." This means that if a program's source code is visible and available to any programmer, then errors

been recorded for posterity in the documentation notes to *Miqra According to the Masorah* (see below).

[48] From the Oxford Text Archive <https://ota.bodleian.ox.ac.uk/repository/xmlui/handle/20.500.12024/0525> where the file "biblheb-0525.txt" may be downloaded.. The slash ("/") is a morphological division marker included within the OTA text.

[49] Christopher Kimball relates: "The earliest tanach.us of 7 Feb 2003 was produced from a Unicode/XML text provided by Alan Groves. The text had Unicode Hebrew characters, was formatted in XML, but contained many errors (e.g. Gen 1:4 was missing). I have no idea how or where he obtained these files (which are still available under 'PreviousVersions'). All subsequent text from the Groves Center was in MCW coding. Starting in 2004, I produced web pages in a Unicode/XML format based on the current MCW text provided by the Westminster Hebrew Institute, now the Groves Center. I don't know what else was available at the time. To my knowledge, the Westminster Hebrew Institute/Groves Center never offered online versions of the WLC." Regarding the basis of the text, Chris adds: "The big point for me is that the original keystrokes of the WLC (and hence UXLC) were entering the BHS, not LC."

[50] As per the website's license: <https://tanach.us/License.html>.

[51] Until April 2020, the Groves Center updated the WLC in MCW coding and Chris published the corresponding Unicode Hebrew text online. After April 2020 the text at the website became the Unicode/XML Leningrad Codex (UXLC) at <https://hcanat.us/Tanach.xml>, with new versions based on published corrections suggested by the website's users.

(“bugs”) will be found and fixed, enhancements will be suggested and implemented, and the program itself will grow and improve continuously. What is true for open-source programs proved true of the WLC/UXLC as well: Once the public was able to see the digital transcription and use it freely, error-correction became rapid and effective. The text at the WLC website matured into an ever-better reflection of the LC’s text (including the correction of numerous errors inherited from BHS). And it was simultaneously used and enhanced in various ways at other websites. For the past two decades, thanks to the WLC and its liberal license, Jewish users of the internet became used to having online, digital versions of the Tanakh with vowels and cantillation, instantly and freely available at numerous websites. The WLC was the initial version of the Tanakh at Hebrew Wikisource (2004), AlHaTorah.org (2011), and Sefaria (2013).

In the summer of 2013, the first full draft of *Miqra According to the Masorah* (MAM) was completed at Hebrew Wikisource.[52] This began with previous digital transcriptions but revised them from scratch, working directly from the manuscripts, in order to provide a vocalized, online Tanakh under an open license (CC-BY-SA), based on the Aleppo Codex and related sources. Unlike the WLC/UXLC, MAM does not replicate the thousands of anomalies, idiosyncrasies, and outright errors in the Leningrad Codex. And unlike Mechon Mamre, MAM is free to use and develop for any purpose, so long as attribution is given and all adaptations or derivative works are themselves freely licensed. Finally, unlike any previous edition of the Tanakh in handwritten, published, or digital form, MAM provides complete transparency about every detail of its text: The reasons and sources for all global or local editorial decisions are fully explained and documented. This is accomplished via a thorough general introduction in Hebrew as well as a local documentation note about every specific point of concern.[53] Since its appearance, MAM has become the default text of the Tanakh with vocalization at all three of the above-mentioned websites.

[52] The Torah was completed more than a year earlier, with the publication of *Parashat Yitro* in the Hebrew year 5772 (2012).

[53] The full introduction to MAM may be found at <https://he.wikisource.org/wiki/MAM:MAVO> and an English abstract at <https://en.wikisource.org/wiki/User:Dovi/Miqra_according_to_the_Masorah#About_this_Edition_(English_Abstract)>. The documentation notes to MAM may be viewed in a convenient format at <https://bdenckla.github.io/phonetic-hbo/>. About two thousand details are documented in the places where the base text of MAM is the Leningrad Codex. Where the base text is the Aleppo Codex, there are far fewer points within the text that require documentation.

4. *Miqra* as Digital Torah (II): The Dataset

To transcribe and proofread the entire Tanakh is an important and even monumental task. Yet it results in a digital text which is entirely linear. The letters and diacritical marks must be typed in one after another. This forces the editor to decide upon a single, exact textual sequence even when the manuscript is ambiguous, unclear, or problematic in some other way. In such cases the editor must decide which of the possible readings to encode and discard the others. In other words, transcription—even of a single manuscript—is not as simple as it sounds. It involves a significant amount of interpretation and there are many judgment calls. As a result, while the user of a digital, masoretic Bible is likely to assume that the underlying text and its transcription are identical, some valid concern may remain about whether they fully match.

A deeper, less obvious concern derives from the very nature of the masoretic text itself as the object of transcription. It is in a sense *not* linear because it is, as we have seen, a meticulous synthesis of two traditions that were originally independent—that of the scribes and that of the readers.

This means that at its core, every word of the MT is not really one word but two: It is a word from the scribal tradition and a parallel word from the oral tradition. Thus, for every word we may ask whether these two traditions match. According to the masoretes, the answer to this question is "yes" for most of the words in the Bible. But for a significant minority they answered "no," and those are the words in which they noted a *qeri*.[54] Similarly, in other places, they noted a *mater lectionis* in the scribal text that was either superfluous, lacking, or unusual.

In fact, the very question as to whether the two traditions match can be subjective. For instance, the word pronounced אָהֳלוֹ (*ʾohŏ**lô***, "**his** tent") is normally spelled with the vowel letter ו at the end, which indicates that the word ends with an "o" sound. But in Genesis 12:8 it is spelled אָהֳלֹה with the vowel letter ה at the end according to the scribal tradition, which implies, but does not mandate, a different vocalization and a somewhat different meaning, namely אָהֳלָהּ (*ʾohŏ**lāh***, "**her** tent"). According to the tradition of the readers, however, in Genesis 12:8 it is nevertheless pronounced with an "o" sound at the end (meaning "his tent"), even though the final vowel letter is ה and not ו. In Genesis 12:8 the vowels for אָהֳלוֹ

[54] A masorete noted a *qeri* when, in his judgment, the oral tradition of the readers reflected a different letter-text than the one preserved in the scribal tradition. The *ketiv* (i.e., the spelling tradition of the scribes) is not noted, but is rather provided in the letters of the biblical text, with the vocalization of the *qeri* superimposed upon it.

are superimposed over the written letters אהלה, and the result leaves no ambiguity about how the word should be read: אָהֳלֹה ends with an "o" sound and means "**his** tent" (even though it ends with ה).

Does the tradition of the readers match the spelling of the scribes in this case? The Leningrad Codex lacks an explicit *qeri* note on this word, yet it does note this as one of four unusual places where the word אָהֳלוֹ (*ʾohŏlô*, "**his** tent") is spelled with the vowel letter ה rather than the expected ו. The exquisite Sephardic manuscript known as the Lisbon Bible (1483) explicitly notes a *qeri* here. The Sephardic Catalan Bible (14th century) notes nothing at all (which is sufficient, since both the spelling and vocalization are clear nonetheless). If we look at some of the better-known printed editions from modern times, we find that Heidenheim (1818) notes an unusual vowel letter as in the Leningrad Codex, Baer (1869) and Ginsburg (1926) both note an explicit *qeri* as in the Lisbon Bible, while Letteris (1870) notes nothing at all as in the Catalan Bible. The more recent Koren (1962) and Breuer (1989, 1998, 2000) editions also note nothing here, presumably so as not to burden the reader with an extra note for a word whose vocalization is entirely clear without one (despite an unusual vowel letter).

Each of these three different options offers a legitimate answer to the question of whether the tradition of the readers matches the tradition of the scribes for a single word in Genesis 12:8. There are numerous other cases like this one, in which the answer to that question is subjective. In fact, even within a single manuscript or edition, similar cases may be dealt with in different ways. It is therefore impossible to count the exact number of instances of *qeri* in the Hebrew Bible: Estimates range from a minimum of about 800 absolute cases of *qeri* (less than one per chapter on average), where the traditions of the scribes and readers clearly do not match. However, there may be up to about 1500 cases (1–2 per chapter on average) if we include hundreds of borderline cases that call for subjective decisions by the scribe or the editor.[55] Therefore, the question as to whether the two traditions match for any given word has *three* possible answers: "yes," "no," and "maybe." The subjective element inherent in "maybe" means that if an edition is to be both consistent and transparent, then its editor must provide explicit criteria for when *qeri* is noted and when it is not. Furthermore, whatever is found in the base text (or texts) upon which the edition is based must be documented. And ideally, different types of *ketiv*/*qeri* pairs (including the ambiguous kinds) should be

[55] On this topic see Ofer, *The Masora on Scripture*, pp. 92–93.

classified and labeled, whether or not *qeri* is noted. That kind of information cannot be fully captured in a linear transcription.[56]

MAM was encoded as a *dataset* because of this consideration and many others like it. We found it impossible to transmit the masoretic Bible digitally via transcription alone. Therefore, it is important to explain what a dataset is, and then cite further examples to illustrate how and why a dataset is the proper tool if we aim to capture the spirit and content of the masoretic project in digital form.[57]

The difference between a "dataset" versus any particular "edition" of the masoretic Bible is first and foremost a matter of consumption: An edition of the Bible—whether in print or electronic—is designed to be consumed by a human being, usually visually. This is in contrast to a dataset, which is designed to be consumed by a computer program. Such a program might produce an edition of the Bible for human consumption from the dataset, or an analysis of the biblical text based upon its data. It could also produce yet another dataset with further features and capabilities.

When we attempt to capture information in a dataset, *the first rule of thumb is to capture it abstractly*. This allows the abstract information to be presented in a variety of concrete ways within different editions based upon the very same dataset. Here are examples of seven basic issues that must be decided one way or another in any edition of the Hebrew Bible:

[56] In MAM we have labeled cases that are a matter of a single vowel letter, and in which the vocalization is unambiguous. We have also labeled several unique forms of *ketiv/qeri* pairs so that they can be formatted appropriately. But ideally, each and every *ketiv* should be documented twice (with and without vocalization), and each *qeri* twice (with and without vocalization). Plus, the overall classification of *ketiv/qeri* pairs can and should be more nuanced than it currently is. Volunteers who would like to help implement this feature more fully are welcome to contact us.

[57] Many of the ideas from here until the end of this section are an outgrowth of our several years of work together on MAM. Initially, Avi found that the special nature of the MT made it necessary to encode many elements of it abstractly. He used MediaWiki templates for that purpose, since the text was published at Hebrew Wikisource. When Ben joined the project, he began to implement consistent abstract coding for each special element of the biblical text that went beyond simple, clear transcription, and worked the digital text into a dataset format that is more useful to programmers. Ben recently formulated the main ideas in this section, and some of the examples, in a presentation entitled "MAM and UXLC: Two Hebrew Bible Datasets." This took place on November 18, 2023, at the Society for Biblical Literature Annual Meeting, San Antonio, Texas, in an open session on "Digital Humanities in Biblical, Early Jewish, and Christian Studies."

1. How are the different kinds of section breaks (*parashot*) found in the masoretic Bible to be shown in a particular edition? Should the letters פ and ס be used (as in a great many printed editions)? Or should they appear only as whitespace, in a way that bears some relationship to the way they are shown in the manuscripts? If we adopt the latter approach, then how exactly should the whitespaces appear?
2. Should verses in the three poetic books (Psalms, Proverbs, and Job) be presented in a special format, as they are in most manuscripts? If so, how exactly should they be shown? Or should such special formatting be dropped entirely (as in most printed editions)?
3. How should verses be separated? In the absence of a section break, should one verse flow directly into the next, on the same line, if space permits?[58] Or should each verse start on a fresh line?
4. What numerals (if any) should be shown to label verses? E.g., טז or יו or 16? Or perhaps nothing at all (since the chapter and verse numbers are not part of the masoretic tradition)?
5. Is the beginning of each weekly Torah portion to be marked within an edition (as it is in the manuscripts)? Should the division of each such portion into seven parts (for those who are called to read from the Torah) be shown (as it is in nearly all printed editions)? And what about the division of the entire Bible into *sedarim* (which also appears in the manuscripts and in most printed editions)?
6. How should pairs of *ketiv* and *qeri* be formatted? Should the *qeri* appear within the primary text (as in some recent printed editions) or in the margin (as in the manuscripts)? Should the *qeri* be vocalized (unlike the masoretic practice which leaves it unvocalized)? Should the *ketiv* be vocalized (with the points of the *qeri*, as it is in the manuscripts and most printed editions)? If both *ketiv* and *qeri* are shown in the primary text, then which one should come first, and how exactly should they be formatted to distinguish between them? Should they always be shown in the same order, even when

[58] This is the choice of nearly all manuscripts. There is one early eastern manuscript in which each verse starts on a fresh line, in order to show Judeo-Arabic *Tafsir* after each verse (ms. St. Petersburg EVR II C 1). But in the vast majority of manuscripts the text flows, unbroken, from one verse to the next, even if it contains verse-by-verse translations. We are aware of no pre-modern, Hebrew-only text in which each verse starts on a fresh line, with the exception of special formatting in the poetic books (Psalms, Proverbs, and Job).

such consistency interrupts the flow of the text when the given word is joined to the word before it or after it via *maqaph*?

7. How should "special" letters be formatted (i.e., large letters, small letters, and "hung" letters)? Should they be formatted at all (especially given the fact that the Tiberian manuscripts largely fail to implement large and small letters, despite their presence in masoretic treatises)?

Abstraction leaves all of these questions open on purpose. This enables a variety of automatic options for concrete presentation in each edition based upon the dataset.

In a dataset, *the second rule of thumb is to capture variation in content, rather than squashing it.* The masoretes did not have datasets, yet the nature of the masoretic project reflects this state of mind. When they combined the two parallel traditions of the scribes and readers, the masoretes carefully preserved variation in the form of *ketiv* and *qeri*. This avoided "squashing" that variation by choosing one over the other. Yet another vivid example of this mindset is the "lower" and "upper" cantillations for the Decalogue: Here the masoretes knew two different traditions for how to divide the verses of the Ten Commandments—and hence two parallel vocalizations for them (in Exodus and Deuteronomy alike). They preserved these two oral traditions by carefully combining the symbols for both in the very same text.[59]

In MAM we strive to preserve and present the mature product of the Tiberian masoretes. This includes all of its internal variation, as it is found in the Aleppo Codex and related manuscripts. We even try to modestly extend it in ways that facilitate "calling out" the text (in the original sense of the word *miqra*). First, we thoroughly describe and document ambiguous or unexpected details in our edition's two base manuscripts—the Aleppo Codex, or the Leningrad Codex where the former is missing—and provide the data which explains the basis for the text that does appear when we do not follow them (even regarding the most minor issues). Second, in places where the masoretic corpus takes care to preserve internal variation (most notably *ketiv*/*qeri* pairs and double cantillation), we capture and label those variations in each of their forms and thus offer multiple options for clear comparison and concrete presentation. Third, in

[59] In terms of variation in content, we might therefore say that the verses of the Ten Commandments contain *four* kinds of data for each and every word: (1) the scribal tradition, (2) the first oral tradition, (3) the second oral tradition, and (4) the exact way in which a given manuscript or edition combines those three traditions in written form. A linear transcription cannot capture all of that data in a useful way.

places where the masoretic vocalization contains a degree of built-in ambiguity, we either provide clarity or else clearly provide and label the various possibilities.

The third and final element (clarifying built-in ambiguities) may be viewed as an extension of masoretic practice. The most prevalent examples of it in our dataset are: (1) The use of a special diacritical mark to indicate *qamaẓ qatan*; (2) the addition of "stress helper" accents in the cantillation of the 21 books;[60] and (3) a distinction between vertical lines in the text which indicate either *legarmeih* or *paseq* (depending on their context). Each of these examples has a certain amount of masoretic precedent, which we have extended to consistent application.

(1) *Qamaẓ qatan*: In the Tiberian system of vocalization, the diacritical mark called *qamaẓ* usually indicates a long vowel, but sometimes indicates a shortened version of that vowel. The Tiberian masoretes assumed that the distinction between the long and short forms of the vowel was normally clear to the reader in context. However, in certain ambiguous cases where the reader might not know which form was intended, the masoretes purposely added two small dots to the right of the vowel (i.e., they used the mark that we call *ḥataph qamaẓ*). This mark was used to clearly indicate that a short vowel was intended. The practice was sporadic: In the Aleppo Codex it exists in 4–5 ambiguous contexts (Jeremiah 2:12, 49:28; Ezekiel 15:4 [?], 32:20; 2 Chronicles 6:42), and in other manuscripts it can be found in other places (e.g. Exodus 37:25 in the Leningrad Codex).

MAM uses the Unicode code point for *qamaẓ qatan* consistently throughout the entire Tanakh, in thousands of places.[61] This eliminates the ambiguity built into the codices (of which the masoretes were consciously aware but not overly concerned about). Specific words where *ḥataph* was explicitly marked for this purpose by the masorete in the base manuscript are noted.

An additional complication is a discrepancy between the Tiberian vocalization and that of the classic Sephardic grammarians. In hundreds of specific cases, the latter deem *qamaẓ* to be a long vowel, even where the

[60] On stress helpers in *Sifrei Emet* (Psalms, Proverbs, and Job), see below.

[61] In this we follow the example of many dozens of *siddurim*, *ḥumashim*, and complete editions of the Tanakh that have been published in Israel over the past 2–3 decades (and occasionally abroad). The trend seems to have begun with *Siddur Rinat Yisrael* (Jerusalem: Moreshet, 1970); in its 1973 reprint *qamaẓ qatan* was added even within its photo-offset reprint of the Torah readings. A designated Unicode code point for *qamaẓ qatan* has been available since 2004.

Tiberian masoretes apparently vocalized it as a short vowel.[62] Thus, an ambiguity already present in the Tiberian vocalization, which the masoretes made sporadic efforts to rectify, resulted in two conflicting rules for pronunciation in the centuries to come. In today's Israel, both of these systems can be heard when the Torah is read in synagogues. The only way to do them justice—and to make MAM useful for those who follow either tradition—was to capture the variation by providing and labeling both forms consistently throughout the entire Tanakh. This approach may be seen as a modest extension of the masoretic project, based on an element already present within that project, as well as an application of the dataset state of mind.

(2) "Stress helpers": In the masoretic codices, most accents (cantillation or *trope* marks) are placed in the syllable that is stressed. In fact, one major purpose of the accents is to indicate stress within a word. But several accents are either prepositive (i.e., they are always written in the first letter of a word) or postpositive (i.e., they are always written above the last letter of a word), regardless of which syllable is stressed. In most such cases (as for *qamaz qatan*), the Tiberian masoretes assumed that the syllable to be stressed was clear to the reader. But in some ambiguous cases the masoretes wrote the accent a *second* time over the first letter of the stressed syllable. We call these extra accents "stress helpers." For the postpositive accent *pashta*, the masoretes added a stress helper consistently.[63] In some later manuscripts (one of the earliest is Vatican Library Urb. ebr. 2, dated to around 1100), and in certain modern editions (such as those published by Heidenheim [1818], Baer [1869], Koren [1962], and most recently *Simanim* [2004]), stress helpers are added consistently for all prepositive and

[62] The sound of the long vowel also differs qualitatively from that of the short vowel in the Sephardic pronunciation (and in Israeli Hebrew): "a" versus "o". This exacerbates the problem by making the difference highly audible. On these two different systems, and considerations about how to deal with them in published *siddurim* and *ḥumashim*, see Chanan Ariel, "On Marking *Qamaz*, *Sheva*, and Stressed Syllables in the New *Siddur Koren*" (Hebrew). A version without notes was published as an appendix to the "Ashkenaz" and "Sepharad" versions of the new *Siddur Koren* (Jerusalem, 2011). On the identification of *qamaz qatan* in the Tiberian vocalization see Werner Weinberg, "The Qamāṣ Qāṭān Structures," *Journal of Biblical Literature* 87:2 (June, 1968), pp. 151–165. Weinberg's article is the primary basis for decisions about *qamaz qatan* in MAM.

[63] In most manuscripts, the "stress helper" is always written for *pashta* whenever the stress is not on the final syllable. In the Aleppo Codex, the stress helper is written consistently for *pashta* in cases where the stress is not on the final syllable *and* the first letter of the stressed syllable is not the second-to-last letter of the word.

postpositive accents throughout the 21 prose books. We have done the same in MAM, but also note when the stress helper was added explicitly by the masorete in the base manuscript. Here too, our approach may be seen as a modest extension of the masoretic project, based on an element already present within that project. In addition, we have recorded stress helpers in *Sifrei Emet*, the three poetic books—Psalms, Proverbs, and Job—which have their own distinct system of cantillation, for the accents *deḥi* (which is prepositive) and *ẓinnor* (which is postpositive). These do not actually appear yet in most editions of MAM, but they are available as an option in the dataset.[64]

(3) *Legarmeih* versus *paseq*: A small vertical line following a word in the biblical text may be either *legarmeih* (part of a disjunctive accent) or *paseq* (which warns the reader not to run two words together even though they are joined by a conjunctive accent). In general, masoretic Bibles expect the reader to be expert enough to distinguish *legarmeih* from *paseq*. But masoretic treatises also provide rules for this, as well as complete lists of *legarmeih* and *paseq* in the Bible. In later manuscripts, the vertical lines are sometimes explicitly labeled *legarmeih* or *paseq* in the margin. In MAM we mark each *legarmeih* and *paseq* abstractly, which allows them to be clearly differentiated in a concrete edition based upon the dataset. Once again, this may be seen as a modest extension of the masoretic project, as well as an application of the dataset state of mind. MAM is almost alone in making the *legarmeih*/*paseq* distinction: *Simanim*[65] is the only other edition of which we are aware that also makes this distinction.

These are three of the most vivid examples of how MAM provides full, transparent documentation about its base manuscripts, while simultaneously providing the reader with additional clarity and consistency via mature reading aids whose seeds are already contained in the work of the masoretes themselves. This is achieved by representing data abstractly and

[64] For the accent *deḥi* in particular, the manuscripts sometimes add *ga'ya* to indicate the stressed syllable (perhaps because an extra *deḥi* might be confused for the similar accent *tarḥa*). There is thus some masoretic precedent for supplying "stress helpers" in these three books, as well as a need for it. However, to the best of our knowledge, consistent application of stress helpers in *Sifrei Emet* was not done in manuscripts or in the classic published versions of the Tanakh. Here too, the best solution is to capture the variation abstractly by noting where the stress belongs for each case of *deḥi* and *ẓinnor*, thus allowing the user to choose whether or not to include these stress helpers in a concrete edition derived from the dataset. In the meantime, stress helpers for *Sifrei Emet* may be viewed in the phonetic transliteration of MAM at <https://bdenckla.github.io/phonetic-hbo/>.

[65] *Torah Nevi'im Ketuvim: Simanim* (Jerusalem: Feldheim, 2005).

striving to capture variation in content. Beyond these three examples, there are many other subtle ways in which our dataset provides clarity to those who want to properly vocalize verses as *miqra*.[66]

When it came to the scribal tradition, however, a different approach was necessary. The masoretes did *not* attempt to capture variation in the letter-text. In fact, the main point of masoretic notes was to carefully define a single, ideal letter-sequence from which manuscripts and scrolls could be corrected. Quite unlike the second dataset rule-of-thumb, here the goal was precisely to squash written variations from that ideal, i.e., to literally scratch a mistaken spelling out of the parchment and rewrite the text.

As we mentioned above, halakhic activity in the centuries that followed the Tiberian masorah eventually brought Torah scrolls into conformity with the masoretic ideal and the Aleppo Codex alike. Very few discrepancies remained. MAM therefore follows the letter-text of Torah scrolls, and the letters of the Aleppo Codex (or its reconstruction) for the books of *Nevi'im* and *Ketuvim*. In the infrequent places where there are scribal differences of any sort between the Torah scrolls of different communities, or between Torah scrolls and what is known about the Aleppo Codex, we provide actual footnotes within the text of the Torah to call attention to the discrepancy.[67] These footnotes are an attempt to capture variation—not variation in the Tiberian masorah itself, but rather variation in the results of the halakhic process that followed it.[68]

The greatest variation of all is found not in the written form of the masoretic text, but rather in the traditions which "call it out" in public reading (*miqra*). Ultimately, we would like to build high-quality audio implementations into the text of MAM, thus binding the oral practice back into the symbols that express it. These could be based on recordings of talented readers (as in the commercial Kol Kore program) as well as on

[66] Chapter 2 of the complete introduction to MAM is devoted to editorial policy regarding numerous issues of this sort, its reasons, and the technical details of its implementation. See <https://he.wikisource.org/wiki/משתמש:Dovi/מַקרא_על_פי_המסורה/מידע_על_מהדורה_זו/פרק_ב/> (Hebrew). The way in which full, transparent documentation of its base texts coexists with the actual text that appears in MAM, including its reading aids, is explained in chapters 3–5 of the introduction.

[67] We do the same for the Book of Esther.

[68] There is also documentation (hidden but available in most editions) whenever the letter-text does not follow the base manuscript. This occurs hundreds of times in the parts based on the Leningrad Codex, whose scribal text is riddled with errors that were never corrected. It is rare in the parts based directly on the Aleppo Codex.

mechanical renditions (as in the commercial Trope Trainer program). Clearly, the phonetic and musical traditions native to all communities would be welcome, as well as the individual style of each reader. There is limitless, rich variation to be captured in this area, and to do so is yet another way to preserve and continue the masoretic project. In terms of recordings, a small, initial stab in this direction has been made under the title *Vayavinu ba-Miqra*.[69] A functional transliteration program for MAM is also available, which could become the basis for high-quality audio renditions via speech-synthesis.[70]

In the meantime, MAM is currently maintained as text at Hebrew Wikisource, where the concrete formatting of abstract data is accomplished through native templates; the main elements of that text are simultaneously organized and backed up in a spreadsheet, which includes a log for changes and corrections. It is further maintained (in sync) at GitHub as parsed JSON files. This makes the dataset more useful to programmers and those who would like to format the text for use in other projects.[71]

5. *Miqra* as Digital Torah (III): The Dataset as Free Software

MAM is a dataset that is provided freely to the public. It is not just free as in "free beer" (i.e., free of charge), but also free as in "freedom"—you can copy it, adapt it, or enhance it as you see fit. This is called a "free content" license. The only obligations incumbent upon you are to give attribution, and share the results of your adaptation or enhancement under the very same free license, so that others can benefit from your work just as you benefitted from ours.

For example: MAM is the text of Tanakh at two superb Torah-study websites, namely Sefaria and AlHaTorah.org. Both of them have not just copied the text for the benefit of their users, but also suggested important corrections and improvements over the course of time (as well as receiving corrections from MAM). We are grateful for this cooperation. In ad-

[69] See <https://he.wikisource.org/wiki/MAM:VAYAVINU>. Volunteers who would like to contribute recordings to *Vayavinu ba-Miqra*, or help develop an audio implementation for MAM, are welcome to contact us.

[70] See <https://bdenckla.github.io/phonetic-hbo/>.

[71] The log for changes and corrections is found at <https://docs.google.com/spreadsheets/d/1mkQyj6by1AtBUabpbax-aZq9Z2X3pX8ZpwG91ZCSOEYs/edit?gid=953964633#gid=953964633>. The GitHub project is at <https://github.com/bdenckla/MAM-parsed>.

dition, AlHaTorah.org has recently enhanced the entire Tanakh by formatting *sheva na‘* and *dagesh ḥazaq* in a way that makes them visually distinct.[72] This has already been done in dozens of books published in Israel, but not in a digital online text. Since MAM is released under a "share alike" license, such corrections and enhancements become available to the public for any purpose in perpetuity.

We are told the following story in a *baraita* (*Bava Qamma* 50b):[73]

> **The Sages taught: A person should not throw stones from his property into the public domain. An incident** occurred **involving a certain individual who was throwing stones from his property into the public domain, and a certain pious man found him.** The latter **said to him: Lowlife [*reiqa*], for what** reason **are you throwing stones from property that is not yours into your property?** The man **mocked him**, as he did not understand what he meant, as the property from which he was throwing stones was his.
>
> **Some days** later, **he was forced to sell his field** from which he had thrown the stones. **And he was walking in the same public domain** into which he had thrown them, **and he stumbled on those same stones. He said: That pious man said it well to me when he said: For what** reason **are you throwing stones from property that is not yours into your own property**, since that property no longer belongs to me, and only the public domain remains mine to use.

An open content license is a way to share and thereby preserve valuable work in a public domain that belongs to us all. Regular copyrighted material dies—even if the owner puts it online. This happens whenever a personal, non-profit or commercial website becomes unavailable, or when it is no longer properly maintained, or when its file formats or website-specific programs are no longer supported, or when the organization which runs it ceases to function or closes down altogether. This happens

[72] There is still no Unicode standard for these two important marks, which makes implementation more difficult than it should be. Anyone willing to help push forward this process with the Unicode Consortium would be making a valuable contribution to the public.

[73] The translation is from the William Davidson digital edition of the Koren Noé Talmud, with commentary by Rabbi Adin Steinsaltz Even-Israel, as shown at Sefaria (slightly modified).

all the time, and in fact more information is being created and then subsequently lost in the computer age than ever before in human history.[74] None of us knows what the future will bring. However, it seems likely that many or most of our private websites, our copyrighted online projects (whether non-profit or commercial), and even our published books will become inaccessible and lost to future generations. Even that which is not lost is likely to become stale and irrelevant due to long-term legal restrictions on re-use: The current length of copyright is draconian, which means that any creative work we do is automatically out-of-bounds to others for several generations unless we grant explicit legal permission to re-use it. Therefore, if we possess material that has value to the public, it behooves us to house it in "property that is ours," and which will remain ours (and our neighbor's and our children's) in the truest sense. An open license can accomplish this.

The masoretes desired that their work be copied and disseminated. Digital technology is a powerful way to do so, and is yet another step in the masoretic story, as are open licenses. In MAM we have worked hard to give life to the masoretic project in a new form, as a digital dataset, to do so responsibly and remain faithful to its spirit. We are grateful for correction reports and other suggestions for improvement. We also welcome volunteers who want to work towards adding new functionality to the project. Anyone interested is welcome to contact us.

The work of the masoretes is the closest thing we now possess to what was once contained inside of the Ark of the Covenant. In a very real sense, it is the "holy of holies" of the People of Israel. It is the starting point for all Torah study. It is furthermore the direct basis for the Torah scrolls found in the Holy Arks of synagogues around the world, and for the public reading (*miqra*) that is done from them. This precious inheritance unites us as a people. As Rabbi Mordecai Breuer expressed it:[75]

> Over a thousand years have passed since the masoretic era. In the course of those many days Judah was exiled time and again. The tribes of Israel moved [further] away from each other and changed in terms of their customs and ways of life. Ashkenazim are different

[74] See Chris Freeland, "Vanishing Culture: A Report on Our Fragile Cultural Record," available at <https://blog.archive.org/2024/10/30/vanishing-culture-a-report-on-our-fragile-cultural-record/> (where the full study may be downloaded). A tragic example of this in terms of Torah study is the venerable website SeforimOnline.org (along with its valuable subprojects on Tanakh, Tosefta, Talmud Yerushalmi and Talmud Bavli). Some of its material is still available thanks to the Internet Archive's "Wayback Machine."

[75] From the first page of his introduction to the first edition of the Tanakh that he published based on the Aleppo Codex (Jerusalem: Mossad Harav Kook, 1989).

> from Sephardim, Sephardim from Yemenites—in the order of prayer, in the text and enunciation of the Mishnah and the Gemara. But the twenty-four books of the Bible are identical, or nearly identical, in all the communities—in both the spelling of the letters and in the vowels and accents. This is one of the great wonders in the annals of Israel, that at the crossroad on the way to exile the masoretic sages made a fence around the Torah. In their merit there is one Torah in the hands of all the tribes of Israel. And every man of Israel knows how to read that book, in which the Master of All enclosed the vision of all.

In fact, it is all of the scribes and readers and scholars who faithfully transmitted the written and oral traditions throughout *all* the generations—before, during, and after the masoretic era—who gave those traditions continued life. They preserved not just the twenty-four books, but also helped preserve Israel as a nation before its God. In particular, the living, spoken tradition of vowels and accents (*miqra*), which Jews knew fluently in the lands of their exile until the advent of modernity, gave continued life to the shared language of Israel and enabled its transmission from one generation to the next. This reality ultimately enabled the revival of the Hebrew language[76] and helped facilitate the re-establishment of Israel in its land.[77] The masoretic project is a precious treasure, an ancient

[76] Until emancipation, Jews in most times and places were able to use Hebrew to communicate. They did so when they encountered other Jews who spoke a different mother tongue, and sometimes even when there was no practical need. Hebrew was not their mother tongue, and they did not normally use it in everyday communication, yet they became intimately familiar with it via the activity of *miqra* along with blessings and prayers "at an extremely early and impressionable age"; see Cecil Roth, "Was Hebrew Ever a Dead Language?" in his *Personalities and Events in Jewish History* (Philadelphia: Jewish Publication Society, 1953), pp. 136–142 (at 136). It was precisely this familiarity and ability which enabled the revival of Hebrew as a mother tongue in modern times. As Roth is famously said to have put it, "*Before* Ben-Yehuda… Jews *could* speak Hebrew; after him, they *did*." The living reality of Hebrew in *both* of these stages was a remarkable achievement, and may well be considered miraculous, the former no less than the latter. (Note that the second quotation from Roth is found in Jack Fellman, *The Revival of a Classical Tongue: Eliezer Ben Yehuda and the Modern Hebrew Language* [The Hague: Mouton, 1973] p. 139, but does not appear as cited there in Roth's "Was Hebrew Ever a Dead Language?"; we have not been able to locate its source.)

[77] It would have been impossible to unify the Jews of Israel, before and after 1948, and give them a sense of common nationhood, without the Hebrew language as a shared inheritance and subsequently as their spoken tongue. For Jewish immigration flowed into the country from every part of Europe, every corner of the

crown that continues to bestow majesty on our people. Blessed is the Omnipresent, Who put His world into the hands of guardians.[78]

Afterword: We completed this article during wartime (winter 5784 [2023]). All of Israel, soldiers and civilians alike, are in harm's way, threatened by an axis of foes who seek our utter annihilation: "They plot against Your people, and take counsel against Your treasured ones. They have said: 'Come, and let us wipe them out as a nation, that the name of Israel no longer be remembered'" (Psalms 83:5). We pray to the God of Israel to grant wisdom, strength, and protection to the leaders, soldiers, and citizens of Israel. We ask that He heal our wounded, free our captives, comfort our bereaved, protect us and grant us victory over our enemies. "May the Lord give strength to His people; may the Lord bless His people with peace" (Psalms 29:11).[79] ☙

Arab-Muslim world, and beyond. Each immigration tide brought its own language, and Israel sounded like a Tower of Babel. The synthesizing factor was Hebrew, and its role in deepening the Jewish national spirit cannot be overestimated. (This paragraph is a restatement of Abram Leon Sachar, *A History of the Jews* [New York: Alfred A. Knopf, 1965], p. 410.)

[78] *Avodah Zarah* 40b; cf. Breuer, appendices to *Torah Nevi'im Ketuvim* (Jerusalem: Horev, 1997), p. 5; and similarly in *Keter Yerushalayim* (Jerusalem: N. Ben-Zvi Printing Enterprises, 2000), pp. 6–7.

[79] The authors are grateful to Aharon Cassel, Menachem Kellner, Christopher Kimball, and Daniel Holman for their valuable assistance with this article. Avi dedicates his part in this article to his teachers of Bible and Hebrew at Yeshiva University: Rabbi Dr. Moshe J. Bernstein, Dr. Barry Eichler, Prof. Nechama Leibowitz *z"l,* Rabbi Dr. Shnayer Leiman, Prof. Yeshayahu Maori *z"l*, Rabbi Dr. Mitchell Orlian *z"l*, Dr. Samuel Schneider, Rabbi Allen Schwartz, Dr. Richard Steiner, Rabbi Dr. David Sykes *z"l,* and Prof. Elazar Touitou *z"l*.