

The Makor System for Typesetting Hebrew

Alan Hoenig

Department of Mathematics, John Jay College, City University of New York

ahoenig@suffolk.lib.ny.us

Introduction

I'd like to describe this morning a new \TeX -based system for typesetting Hebrew. I call this system *Makor*, from the Hebrew word for 'source'. I don't know how many Hebrew speakers there are here today—I'm not one myself—so I'll describe this project from the point of view of coping with yet another new font subject to some curious rules. The worth of this project—insofar as there is any—lies in a few areas:

1. First of all, I hope it will be of use not only for documents in Hebrew, but also in languages also using Hebrew fonts, such as Yiddish and Ladino.
2. Secondly, I hope this project serves as a model for adapting the \TeX engine for typesetting foreign languages. The problems associated with Hebrew are quite different from ones associated with European-language typesetting, and I hope this inspires people who want to typeset, say, Devanagari or Arabic efficiently with \TeX .
3. Finally, this project would have been orders of magnitude more difficult to implement without the convenience and power of virtual fonts. Virtual font technology is now over ten years old, and yet this area remains *terra incognita* for most \TeX users. Perhaps this project can be interpreted as additional instruction in virtual fonts.

Prior work Any work on Hebrew typesetting must acknowledge Yannis Haralambous's great *Tiqwah* system. It is described in several places [1], and the last illustration in that paper must rank as one of the milestones in \TeX typesetting history. Not only does it include two systems of diacritic vocalizations (one for pronouncing—see below—and one for liturgical chanting), but it does so so that the final product is a thing of great beauty. His METAFONT-created font is especially noteworthy. It's not clear, though, how to extend Yannis's work to apply to other fonts.

More recently, Sivan Toledo [2] has developed a PostScript system which cooperates with \TeX for Hebrew. The illustrations that appear in his article are intriguing, but I have not been successful in

using it myself. This article contains a useful list of sources for additional information.

The CTAN archives contain a generous collection of Hebrew meta-fonts. Generally speaking, though, these fonts are not of sufficiently-high caliber for use with a comprehensive typesetting system. Most of them, for example, do not contain dotted letter variants or the vowel marks (see below for explanations). In the same way, the dozens of Hebrew fonts available on the Web are not \TeX -worthy. So far, the only acceptable and freely-available font I've been able to lay my hands on is the font that is part of the Omega system (due to Yannis Haralambous and John Plaice). If anyone can point me to high quality fonts in addition to `OmegaSerifHebrew`, I would be grateful.

Describing Hebrew

For typesetting purposes, we can consider Hebrew in the following terms. It is a caseless language read from right to left (but from top to bottom). Although it is caseless, most letters have two forms—with and without a (more-or-less centrally located) dot. There are a total of twenty-seven letterforms (not counting the dotted variants), but five of those occur only at word endings. You can see these characters in figure 2. There they are presented in three groups separated by asterisks '*'. The last group (read right-to-left!) are the word-final letters.

Actually, all these letters are consonants. Vowel sounds are represented by diacritical marks which appear below the consonant. However, these marks are not mandatory, most often being omitted entirely from adult reading matter. They appear in language texts, children's books, and on occasion to make clear how to pronounce a new word or a foreign name. These diacritics differ from European accents in one intriguing way. They are centered *not* with respect to a central axis but with respect to a particular axis whose location varies from letter to letter. Actually, the location of this axis should be specified by the font designer. (I am indebted to Yannis Haralambous for this crucial observation.)

Vocalized consonants appear in figure 3. The same vowel (best described as a simple dot under a

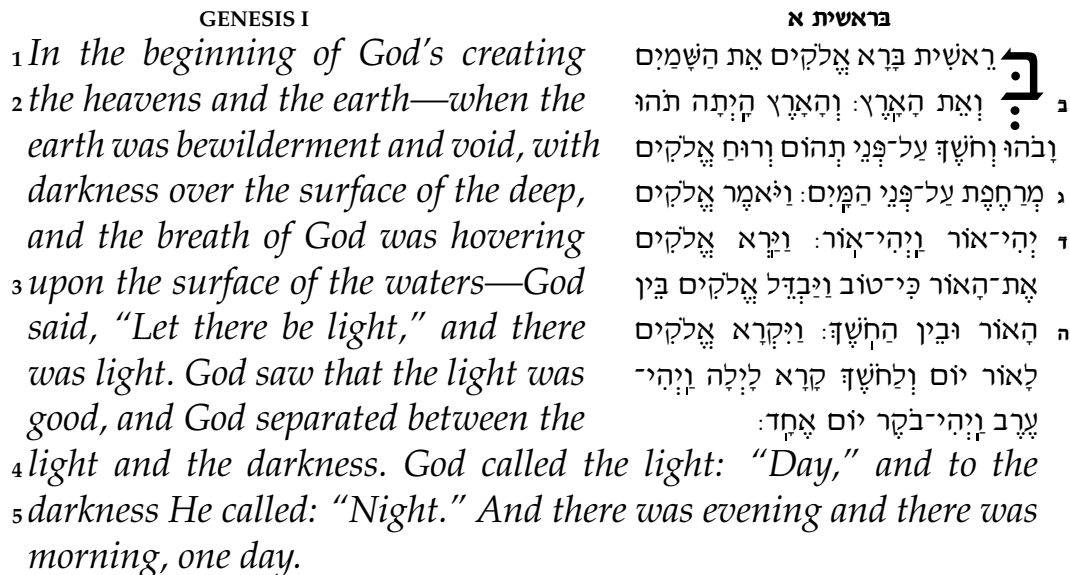


Figure 1: A proposed Biblical layout.

א ב ג ד ה ו ז ח ט י כ ל מ נ ס ע פ צ ק ר ש ט
* א ב ג ד ה ו ז ח ט י כ ל מ נ ס ע פ צ
ק ר ש ט * ך ם ן ף ץ

Figure 2: The base characters in the Hebrew alphabet.

consonant) appears under two different letters. As you see, in the left consonant, it appears under the right letter stroke—about as far away from centering as you can get! In the remaining letter, the vowel is closer to the center, but not exactly at it.

Further considerations pertain to keyboard entry. I wanted to be able to use a standard computer keyboard to type my Hebrew source. Most of the Hebrew letters correspond to Latin letters but there are complications. For various reasons, several Hebrew consonants represent the same sound, and there are a handful of Hebrew letters that have no English counterpart.

For those readers who have trouble delaying their gratification, I present in figure 4 a sample of fully-vocalized Makor output. This selection is drawn from the final few verses of Deuteronomy.

Goals and implementation Don't worry—I have no intention of describing the features of Makor that properly belong in the User manual. But even without these details, there is enough interesting stuff going on that I hope you won't be bored, or at least *too* bored. On occasion, I suppose I will have to refer to details of the system, but I will try to keep those references to a minimum.

I set myself the goal of designing a system that uses a reasonable keyboard entry mechanism to typeset Hebrew according to the rules set out above and consistent to the high quality of which T_EX is capable. You can decide how well I succeeded.

Several of these considerations are easily dealt with. To get right-to-left text, we demand that you use Makor with any of the extended T_EX's that do this, such as Omega or eT_EX. I noticed some minor but distinctive differences in the typeset output depending on which of these programs you use, so try both of them and use the best. The differences all appear in spacing before and/or after a bit of right-to-left text.

T_EX's virtual font mechanism easily handles end-of-word glyphs, so that's not a problem.

T_EX's ligature mechanism is quite helpful in this project. Although several Hebrew sounds are foreign to English, an English (or at least American) set of conventions has arisen for their representation. For example, Hebrew contains the guttural throat-clearing sound that is present in some German words, such as in the name of the composer J. S. Bach. We agree in Makor to type *ch* whenever we



Figure 3: Hebrew vowel diacritics are not centered.

וַיִּמַּת שֵׁם מֹשֶׁה עֶבֶד־יְהוָה בְּאֶרֶץ מִזְרַח עַל־פִּי
 יְהוָה: וַיִּקְבֹּר אֹתוֹ בְּגִי בְּאֶרֶץ מִזְרַח מִבַּיִת פְּעוֹר
 וְלֹא־יָדַע אִישׁ אֶת־קְבֻרָתוֹ עַד הַיּוֹם הַזֶּה: וּמֹשֶׁה
 בְּן־מֵאָה וָעֶשְׂרִים שָׁנָה בָּמָתוֹ לֹא־כָּהָתָה עֵינָיו
 וְלֹא־נָס לְחָה: וַיִּבְכּוּ בְּנֵי יִשְׂרָאֵל אֶת־מֹשֶׁה בְּעֶרְבַת
 מִזְרַח שְׁלֹשִׁים יוֹם וַיָּתֵמוּ יָמֵי בְּכִי אֲבָל מֹשֶׁה:
 וַיְהוֹשֻׁעַ בְּן־נוּן מֵלֵא רֹחַ חֲכָמָה קִי־סַמֵּךְ מֹשֶׁה
 אֶת־יָדָיו עָלָיו וַיִּשְׁמְעוּ אֵלָיו בְּנֵי־יִשְׂרָאֵל וַיַּעֲשׂוּ
 כְּאֲשֶׁר צִוָּה יְהוָה יְהוָה יְהוָה אֶת־מֹשֶׁה: וְלֹא־קָם נָבִיא עוֹד
 בְּיִשְׂרָאֵל כְּמֹשֶׁה אֲשֶׁר יָדְעוּ יְהוָה יְהוָה אֱלֹהֵינוּ:
 לְכֹל־הָאֵתָה וְהַמּוֹפְתִים אֲשֶׁר שָׁלַח יְהוָה לַעֲשׂוֹת
 בְּאֶרֶץ מִצְרַיִם לְפָרְעוֹ וְלְכֹל־עַבְדָּיו וְלְכֹל־אֶרֶץ:
 וְלְכֹל הַיָּד הַחֲזָקָה וְלְכֹל הַמּוֹרָא הַגָּדוֹל אֲשֶׁר עָשָׂה
 מֹשֶׁה לְעַיְנֵי כָּל־יִשְׂרָאֵל:

Figure 4: Fully vocalized Hebrew output.

want that gutteral to appear in text. In a Hebrew font, `ch` is a special ligature that selects that letter.

Vowels The nexus of the project involved the representation and typesetting of vowels. Deciding on input conventions was straightforward—first of all, I let English consonants and only these consonants represent Hebrew consonants, and the flip side of that coin is that English vowels and only these vowels represent Hebrew vowels. However, the conventions of reading Hebrew demand that the vowel—which is a diacritic accent, remember, not a distinct letter—*follows* the consonant, a convention contrary to normal `TeX` conventions. At the same time, each time a consonant is typeset, we need to access somehow the location of its visual axis to know where to place the diacritic accent.

It's too bad that there is no `\lastchar` primitive in `TeX`, the way there is a `\lastbox` and a few other `\last...` things. To keep track of the most recent character, I had to coopt the italic correction, one of the few signposts that `TeX` erects to mark a recently set character. Since there really is no such thing as a Hebrew italic, at least in my fonts, its loss for marking genuine italic corrections is not missed

by anyone. In my special fonts, I redefined the italic corrections to equal the character code of the glyph. Italic corrections are just kerns, and they can be removed with `\lastkern` and examined to see how big they are, and thus which character has just been placed in the `.dvi` file.

Once I know which character I've just set, I can get the location of the optical axis. I previously made sure to store the distance of this optical axis from the central axis as a *kern* between the character in question and a special character which, although present in the font, is not used otherwise to typeset. It's possible to measure this kern, and so to know where and how to center a vowel. The Makor macros do all this for us.

There's one other aspect of vowels that I wanted to take into account. I mentioned earlier on that vowels are often omitted in actual Hebrew typesetting. Nevertheless, one might want to include them in the source document, if for no other reason than to make reading of the source file less of a bother. For example, there is a (reasonably) well-known Jewish holiday that occurs usually in December. According to my Makor input conventions, you would typeset it, fully vocalized, as

`chanookauh`.

To eliminate the vowels in the output, you eliminate them in the input, so the source should now read `chnkh`.

Does any reader doubt that this input is ugly, unpleasant, and difficult to read? In Makor, there is a companion font which maps the vowels to null characters, so you could de-vocalize the output while maintaining ease of reading in the input by typing something like

```
{\CXLV chanookauh}
```

Here, the command `\CXLV` is not a Roman numeral, but rather the instruction to cancel the vowels. But it's really just a font selection command, and you can play with it as you would any font selection command.

Active characters The only way I could differentiate between consonants and vowels in the source is by making all vowels active. For those still breaking their teeth on \TeX , you should know that making a character active makes that character eligible to receive macro treatment—you can assign it a definition as you would to a control sequence.

This dual categorization leads to far few problems than I expected, but of course one immediate consequence is that virtually all of our familiar arsenal of \LaTeX and \TeX commands appear to become useless. A simple command like `\hspace` is no longer—at least from the point of view of the \TeX engine—an escape character followed by a sequence of letters. Here, we have an escape character (the backslash) followed by the letters `hsp`, followed by an active character `a` whose definition \LaTeX tries to plug in, followed by a `c` (which is typeset in the document), followed finally by an active `e`. There are straightforward ways out of this impasse, but it is a sticky wicket to be sure.

Numbers Numbers introduce an amusing problem as well. Although it's true that all Hebrew is typeset right to left, the only exception is for (ironically named) Arabic numerals when they appear in a Hebrew document—they are to appear left to right. How then do we typeset them?

If you think about it just a bit, you would be tempted (as I was) to simply come out of Hebrew mode to typeset the numerals and return to Hebrew mode for the remainder of the text. If `\[` and `\]` are the toggles for entering and exiting Hebrew mode, then you might typeset

אבגדה12345קרשת

schematically as

```
\[ first bit of Hebrew \]%
12345%
```

```
\[ second bit of Hebrew \]
```

If you do that, you get instead

קרשת12345אבגדה

which, for those not proficient in Hebrew reading, is schematically the same as

```
\[ second bit of Hebrew \]%
12345%
```

```
\[ first bit of Hebrew \]
```

If you think hard about this for a few moments, you see why this is so, and why it's necessary to introduce a `\NUM` macro to typeset the numbers. `\NUM` involved a recursive macro, easier to create than you might think since I plundered it wholesale from a timely example in the *TeXbook*.

Fleshing out the font Hebrew fonts tend to be sparse. What you get for your money is generally just the special Hebrew glyphs. If you're lucky, you get matching numerals, but almost never do you get a full complement of punctuation and special symbols. By virtue of the magic of virtual fonts, it's easy to flesh out the virtual Hebrew fonts with characters from an existing Latin font, but from which one?

It seems to me that the best match, at least for the `OmegaSerifHebrew` I used for my development work, is Palatino regular. However, it's a proprietary PostScript font, and although common (it's one of the fonts typically resident in any PostScript-enabled printer), I still wasn't comfortable assuming that everyone had access to it, and access to it under the fontname scheme that most modern \TeX 's adhere to.

Consequently, I used my second choice, although it's still a good choice. I chose the Computer Modern Fibonacci font `cmfib8` as the font-flesher-outer. Although `cmfib8` is a favorite of mine, I have heretofore not been able to find a legitimate use for it. Recall that this quirky font has `META-FONT` parameter values chosen from the sequence of Fibonacci numbers, an interesting group, at least from a mathematician's point of view. I used `cmfib8` scaled by 699, where 699 is the sum of the eleventh and fifteenth Fibonacci numbers (if that's at all significant!).

The two examples above show the Hebrew letters in combination with Fibonacci numerals.

Further examples I'd like to present now just a few examples of Hebrew-English typesetting with Makor that at least look interesting. In figure 5 you

Rabbinic Hebrew (RH) does not differ greatly from Biblical Hebrew (BH) in its inflection of the noun, although the neutralization of final *mem* and *nun* means that the masculine plural is often, as in Aramaic, ן- . Apart from the more frequent use of the archaic feminine suffix -ת as in כֹּהֲנֵת ‘priest’s wife’ and אִקְמָת ‘dumb woman’, RH also employs the suffixes -ית and -ות for example אַרְמִית ‘Aramaic’ and עֲבָדוֹת ‘servitude’. RH developed distinctive feminine plural suffixes in -אות (Babylonian) or -יות (Palestinian), for example $\text{מְרַחְצֵי־אוֹת}/\text{מְרַחְצֵי־יֹת}$ ‘bath-houses’ and -יות , as in מְלְכֵי־יֹת ‘kingdoms’ for BH מְלְכֵי־ת , for nouns ending in -ות in the singular. Masculine plural forms sometimes differ from those that would be expected, or are normally found, in BH, for example, נִזְקֵי from נִזֵּק ‘damage’, שְׁוֹרִים from שׁוֹר ‘ox’, שְׁוֹקִים from שׁוּק ‘market’, צְדָדִים from צֶדֶד ‘side’, חֲצֵי־אֵין from חֲצֵי ‘half’, and שְׁלֵחֵי from שְׁלִיחַ ‘envoy’. The same is true of feminine nouns, for example אוֹתֵי־יֹת from אוֹת ‘letter (of alphabet)’, בְּרִיתוֹת from בְּרִית ‘covenant (without plural in BH)’, and אִמֵּהוֹת from אִם ‘mother’.

Some masculine nouns take the feminine plural suffix -ות , for example, חַן־וֹת from חֵן ‘favour’, כְּלָלוֹת from כָּלַל ‘rule’, תִּינֻקוֹת from תִּינֻק ‘baby’, חֵילוֹת from חֵיל ‘army’, עִירוֹת from עִיר ‘city’, and מֵי־מַיִם from מַיִם ‘water’. Similarly, there are some feminine nouns which take the masculine plural suffix -ים — יוֹנֵי־ים from יוֹנָה ‘dove’, נְמֵלֵי־ים from נְמֵלָה ‘ant’, and בֵּיצֵי־ים from בֵּיצָה ‘egg’, for example. Occasionally, both types of plural are evidenced, as with יָמֵי־יָמֹת from יָוֵם ‘day’ or שְׁנֵי־שָׁנֹת from שָׁנָה ‘year’, with each form having a slightly different shade of meaning and the ‘feminine’ variant only used with suffixes. In RH we sometimes find plurals of nouns only attested in the singular in BH, for example אֲבָרִים from אֲבָר ‘limb’, דְּשָׁאֵין from דְּשָׁא ‘grass’, and תְּמִידִים from תְּמִיד ‘daily sacrifice’. Likewise, there are singular forms of nouns only attested in the plural in BH, for example אֲלָמוֹג ‘coral-wood’, בֵּיצָה ‘egg’, and בָּצֵל ‘onion’. The dual is used more than in BH, with existing forms retained and new ones created, for example מְסַפְרֵי־ים ‘scissors’ and בְּנֵתֵי־ים ‘meanwhile’. (1993: Saenz-Badillos, *A History of the Hebrew Language*, Cambridge University Press, pp.188-89.)

Figure 5: Mixing Hebrew and English together.

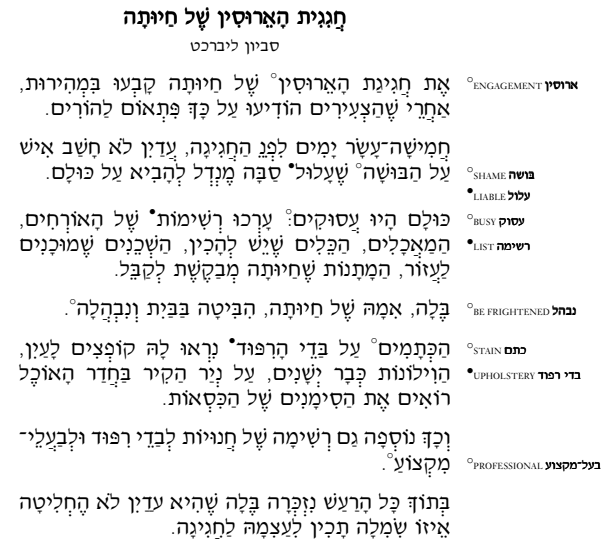


Figure 6: Page from a proposed Hebrew primer.

see how easy it is to mix lots of Hebrew with lots of English.

Although you do have to keep aware that vowels are active, it's surprising how few restrictions there actually are on Makor typesetting. In figure 6, we see the page of a book as it might be typeset for beginning students in the language.

Liturgical books frequently demand interesting layouts. In figure 1, verse numbering is automatic. The look of figure 7 mimics that of certain books of legal exposition.

Future work and further comments

Several important tasks remain in this project. Of course, as in all such projects, it's important to eliminate remaining bugs, continue macro support and so on; that goes without saying. The user manual has to be greatly expanded. Eventually, the Makor package will consist of a Perl script in addition to fonts. The script will greatly aid users who want to convert their own fonts to the form the Makor macros expect.

I also plan to add additional fonts—fonts for the typesetting of Yiddish and Ladino, and fonts which allow users to type their input from an Israeli keyboard. The Perl script I just mentioned will facilitate this task immeasurably. There are apparently a few other dialects of Hebrew plus whatever that use the Hebrew alphabet, and if I can find out more about them, I would like to support them as well. (Most people know what Yiddish is—sort of a creole between medieval German and Hebrew. Ladino stands in the same relation to medieval Spanish and Hebrew.)

Extended T_EX's I've been out of the loop for a long time—not that I was really ever *in* the loop—as far as the development of extended T_EX goes. The two such projects I know about are Omega, under the direction of Yannis Haralambous and John Plaice, and a larger team for eT_EX under the stewardship of Phil Taylor. There were several facilities that are missing in T_EX whose presence would have made my life much easier.

First off, I would have loved a facility that gives me the character number of the most recently set character, perhaps to be called `\lastchar`. Because of the presence of ligatures in a font, even knowing the last character in the source file is not sufficient for knowing the last typeset character.

I would have liked the possibility of having an entire series of character dimensions allotted for each character, in the same way that font dimensions are allotted for each font. The first four character dimen's would coincide with the width, height, depth, and italic correction of the character, but any additional `\chardimen`'s would have meanings given to them by the font designer. For example, a fifth and sixth `chardimen` for these Hebrew fonts might correspond to the location of the upper and lower visual axes. Some letters might have additional `\chardimen`'s giving the location of forbidden portions, areas where diacritical marks may never appear. I feel sure that this would be useful concept for other languages, for reasons I could not begin to guess. Of course, I have no idea how easy that would be to implement, nor what it would do to the size of the resulting `.tfm` file.

Finally, I would have liked additional category codes available to me. Classifying vowels as active, though useful, is fraught with some peril. Moreover, I suspect that for other languages, there are several categories of glyphs that I would not have been able to handle without additional `\catcode`'s.

Getting the software

The current version of Makor software is always available in the CTAN archives, in the directory

`languages/hebrew/makor`

Plea to the reader and user

If any reader or user could point me to additional high-quality Hebrew fonts, I'd be most grateful. As of this writing, I am unable to test my macros and information-encoding schemes on other fonts.

Please don't hesitate to forward additional suggestions, queries, requests for clarification, and so on to me; `email` is best. Thanks for your interest.

