Developments in Hebrew TEX

Rama Porrat

The Computation Center
The Hebrew University
Jerusalem 91904
Israel
rama@noa.huji.ac.il

Abstract

Manipulating Hebrew texts by TEX requires bi-directional typesetting. The available implementations are discussed. Related macros packages as well as available Hebrew fonts are described. Various sources of information for the Hebrew TEX user - noa and ymir ftp archives, WWW database, IVRITEX mailing list - are described.

Key words: Hebrew, noa, WWW, IVRITEX.

1 Implementations

The Hebrew language is written from right to left; hence, typesetting Hebrew texts relates to the issue of typesetting bi-directional texts.

In the last few years there has been a constant development in this respect within the TeX environment. Typesetting bi-directional texts using TeX and/or LaTeX is becoming easier as newer packages are being developed.

In 1987 Knuth and MacKay described the complexity of typesetting bi-directional texts [1]. They explained the two necessary operations for typesetting text from right to left (assuming order of text insertion into the input file is always the same, "natural" order, no matter whether the final output is to be a left-to-right or a right-to-left text) - namely, re-arranging characters within words and then re-arranging words within lines. Re-arranging words involves placing a word to the left of its preceding word - or, if the line is full, placing the word at the right end of the next line. Those operations become more complex according to the complexity of the bidirectional text. Hyphenation problems are also involved.

Knuth and MacKay's article seems to have been an inspiration for later implementations.

1.1 Early implementation - Reversing text before TeXing

Prof. Jacues J. Goldberg of the Technion, Haifa, Israel has presented his solution for TeXing in Hebrew in 1988 [2].

In this implementation (and I more or less quote from Prof. Goldberg's article), markers are used within the original text, to indicate where words should or should not be individually reversed, as well as the beginning and end of a Hebrew segment. Markers actually used have been a vertical bar and a dollar sign. Context sensing, and Hebrew word parsing and reflecting, is done by a preprocessor to TeX, written in C, feeding TeX with a temporary intermediary file instead of the original. Then the task of building lines by extending the current box to the left, until its size exceeds the allowed line size (done in Hebrew mode only), is achieved by a set of simple TeX macros.

PRTEX is a "vanilla" version of the preprocessor. SUPRTEX is an extended version, modified by Stephen Woods so as to allow usage of any coding for the Hebrew letters that one likes (handy for those without Hebrew keyboards). [1]

1.2 Later implementation - TeX-XeT

Inserting the operations necessary for typsetting bi-directional text into TeX itself, that is, creating TeX-XeT, was the next step.

Knuth and MacKay have presented in [1] a detailed explanation of the changes that should be incorporated into TeX in order to turn it into TeX-XeT. They suggested an extension to TeX by defining four new primitive operations,
\beginL and \endL which will surround left-to-right text,
as well as
\beginR and \endR which will surround right-to-left text.

Each of these new operations, when used, will create a new sort of "whatsit node" on the current horizontal list.

Whatsits are mechanisms that have been provided in TeX, such that using them makes it possible to incorporate into TeX features which are not

[1]These packages can be obtained from ymir.claremont.edu in the directory [anonymous.tex.babel.hebrew.prtex] via bitnet.
Associated macros reside in [anonymous.tex.babel.hebrew.macros].
the files can also be obtained by sending the command GET IVRITEX PACKAGE to LISTSERV@TAUNIVM.BITNET

actually part of the program. Whatsits are not expanded while macros are being expanded; commands within them are deffered and executed later.

The TeXbook explains two sorts of whatsits which are part of TeX. One sort of whatsits has to do with writing output on text files. This is used for creating tables of contents and indexes via auxiliary files, as well as for other purposes. The second kind of whatsit nodes are those holding the content of \special, which is not expanded by TeX but written on the dvi file to be later used by the dvi driver.

The suggested new kind of whatsits are said to put newly defined operation codes 250 *('begin_reflection')* and operation codes 251 *('end_reflection')* into the dvi file, to be further processed.

As a result, the outcoming dvi file created is no longer a standard one. It has to be turned into a standard one by a program which knows what to do with operation codes 250 and 251; alternatively, it can be handled by specially "enlarged" dvi drivers which "know" what to do with the additional operation codes.

A change file has been created containing those additions suggested by Knuth and MacKay and is currently available for TeX3.14 [2]. It can be used in order to create a working TeX-XeT for Unix. This version of TeX behaves like an "ordinary" TeX for regular text, and it also recognizes the four additional control sequences, namely, \beginR \beginL \endR \endL .
Once any of these keywords is used, a non-standard dvi file is created.

Prof. Wayne Sullivan has incorporated the TeX-XeT extension into his SBTeX package for PC DOS. His TeX-XeT is also referred to as SBXeT or just XET. Like the Unix TeX-XeT, this XET also outputs a non-standard dvi file, having the extension ivd. [3]

[2] The change file can be obtained from noa.huji.ac.il where it exists in tex/install_in_unix

[3] SBTeX for the IBM PC which supports the TeX-XeT extensions is available from LISTSERV@TAUNIVM.BITNET in the file SB38XET ZIP. This is a binary archive file. If you can receive binary files over BITNET (you will need to be on a VM/CMS system connected directly to BITNET), you can retrieve the file with the command GET SB38XET ZIP and transfer it as a binary file from the mainframe to your PC. For those who cannot receive binary files over BITNET, it is possible to receive the file in UUENCODE format with the command GET SB38XET ZIP F=UUE in XXENCODE format with the command GET SB38XET ZIP F=XXE or hex format with the command GET SB38XET ZIP F=HEX

The file is available for anonymous FTP from hmcvax.claremont.edu in [anonymous.tex.babel.hebrew.tex-xet] in the file SB38XET.ZIP. It cannot be obtained via mailserv.

Alternatively, it can be ftp'ed from noa.huji.ac.il where it exists in tex/install_in_pc . If you use the noa archive, start be getting and reading the file tex/install_in_pc/readme.first

Prof. Sullivan's package is now very popular in Israel. Many users use SBTeX's XET as an addition to a previously installed TeX for PC DOS - PCTeX or EMTeX. Others use SBTeX for both regular and Hebrew TeX.

In order to facilitate usage of Hebrew within TeX-XeT, sets of TeX (and LaTeX) macros have been written, and the user is advised to use them. For details, see "Macros facilitating Hebrew Usage" below.

A pre-processor for the non-standard dvi files has been written by Larry Denenberg. It is called ivd2dvi, and it creates a standard dvi file out of the non-standard file. [4]

The archive noa.huji.ac.il located in Jerusalem, Israel contains files related to TeX-XeT, for both Unix and PC DOS. All files can be retrieved via anonymous ftp from the directory tex. A file called "information" is available, containing a description of the entire archive in detail. For more information, see the section "Archives holding Hebrew material" below.

1.3 Recent implementation - TeX--XeT

Recently Peter Breitenlohner has announced his TeX--XeT, a new implementation of bi-directional TeX which is inspired by, yet different from, TeX-XeT as presented in the above mentioned article. There are a few differences between the two, the most interesting one (in my opinion) lies in the fact that this TeX--XeT makes the necessary reversals of text within the shipout routine, that is, while a page is copied from TeX's memory onto the dvi file. The outcoming dvi file is thus a standard one, without any begin_reflection and/or end_reflection operation codes. [5]

2 Macros facilitating Hebrew usage

In order to facilitate typesetting Hebrew Texts by TeX-XeT, sets of macros have been prepared for the user. There are two sets; one set of macros for TeX-XeT and the other for LaTeX(-XeT). For the LaTeX user the Hebrew

[4] The program may be FTP'd from ymir.claremont.edu in the directory [anonymous.tex.babel.hebrew.ivd2dvi], or from noa.huji.ac.il in tex/ivd2dvi, or obtained via BITNET by sending the command GET IVD2DVI PACKAGE to LISTSERV@TAUNIVM.BITNET

[5] The package is available at present via anonymous ftp from ftp.uni-stuttgart.de (129.69.1.12) in the directory /soft/tex/tex-sources/tex--xet
or from Niord.SHSU.edu (192.92.115.8) in the directory [FILESERV.TEX--XET]
To retrieve the package of 2 files from the SHSU fileserver, include the command: SENDME TEX--XET in the body of a mail message to FILESERV@SHSU.BITNET (FILESERV@SHSU.edu).

macros are implemented by a style file used as an option with any LaTeX document style.

2.1 Macros for the Hebrew TeX user

The macros file has been written by me, based on Knuth and MacKay's suggestions and upon contributions of users. The most important macros are actually copied from the above mentioned article [1]. These are \L... and \R... used for insertion of left-to-right text in a right-to-left environment, and vice versa. Here are the definitions:

```
\let\next=\

\def\L{\afterassignment\moreL \let\next= }
\def\moreL{\bracetext \aftergroup\endL \beginL \rm}
\def\R{\afterassignment\moreR \let\next= }
\def\moreR{\bracetext \aftergroup\endR \beginR \hebrew}
\def\bracetext{\ifcat\next{\else\ifcat\next}\fi
\errmessage{Missing left brace has been substituted}\fi \bgroup}
```

The setting of the current direction is done using the commands

```
\sethebrew
\unsethebrew
```

Those keywords are implemented using \everypar; \everypar has the effect of prepending its argument to each coming paragraph in the following text.

The keywords \newif\ifheb \hebtrue \hebfalse are used for specifying different actions taken according to whether the current direction is left-to-right or right-to-left. They are used within the following definitions:

```
\newif\ifheb
\def\sethebrew{\everypar{\leavevmode\kern-\parindent\beginR
 \indent\the\font}
 \global\hebtrue}
\def\unsethebrew{\rm\everypar{}\hebcheck=0 \global\hebfalse}
```

heb_macros.tex contains various keywords which are aimed at removing small inconveniences in a "Hebrew" (right to left) environment. For example, the \item macro has to be modified in order to give the right indentation on the right side;

```
\def\hebitem#1{\item{#1}\hangindent=-\hangindent}
\def\hebitemitem#1{\itemitem{#1}\hangindent=-\hangindent}
```

Other macros are called for to solve small problems pertaining to right-to-left writing as opposed to the ordinary direction in TeX;

```
\def\openquote{\L{''}}
\def\closequote{\L{``}}
```

Still other macros are there to remove the inconvenience stemming from the fact that the available Hebrew fonts are less then perfect and symbols are missing;

```
\def\longdash{\L{--}}
\def\longerdash{\L{---}}
```

2.2 Macros for the Hebrew LaTeX user

For LaTeX, the Hebrew macros take the form of a style file (hebrew.sty) used as an option with any document style. The first version of this style file has been written by Sergio Fogel; it has later been modified by Yaniv Bargury, Yael Dubinsky and myself.

The default Hebrew fonts in LaTeX are (see next section for a short description of the Hebrew fonts available):

```
Jerusalem    denoted by \jm
TelAviv      denoted by \ta
DeadSea      denoted by \ds
OldJaffa     denoted by \oj
```

In order to work in Hebrew, the user types \sethebrew and the following setting comes into effect:

```
\jm replaces \rm   (The regular font)
\ds replaces \bf   (boldface)
\oj replaces \it   (italics)
\ta replaces \tt   (typewriter)
```

The basic macros in hebrew.sty have been written resembling the definitions in the standard file lfonts.tex, which defines the fonts used by LaTeX. Only a sample is included here, showing definitions for 10-point and for 12-point sizes.

```
......
\font\tenjm=Jerusalem
\font\twlvjm=Jerusalem scaled \magstep1
.......
.......
\font\tenoj=OldJaffa
\font\twlvoj=OldJaffa scaled \magstep1
.......
.......
\font\tenta=TelAviv
\font\twlvta=TelAviv scaled \magstep1
........
........
\font\tends=DeadSea
\font\twlvds=DeadSea scaled \magstep1
.......
.......
\def\jm{\protect\pjm}
\def\oj{\protect\poj}
\def\ta{\protect\pta}
\def\ds{\protect\pds}
......
......
\let\oldxpt\xpt
\def\xpt{\oldxpt\def\pjm{\tenjm}\def\pta{\tenta}%
  \def\pds{\tends}\def\poj{\tenoj}}

\let\oldxiipt\xiipt
\def\xiipt{\oldxiipt\def\pjm{\twlvjm}\def\pta{\twlvta}%
```

```
  \def\pds{\twlvds}\def\poj{\twlvoj}}
.......
.......
\let\oldbf\bf

\def\tohebrew
{\def\rm{\jm}\def\bf{\ds}\def\it{\oj}\def\tt{\ta} \jm}}

\def\toenglish{\def\rm{\protect\prm}\def\bf{\oldbf}
       \def\it{\protect\pit} \def\tt{\protect\ptt} \rm}

\newif\ifheb
\def\sethebrew{\hebtrue\beginR\tohebrew}
\def\unsethebrew{\hebfalse\beginL\toenglish}
```

\ifheb is used to distinguish between right-to-left and left-to-right environments, so that different actions can be taken according to the situation; for example, in a right-to-left environment headers should be in Hebrew, and Hebrew words for "Chapter" "Section" etc. should be used.

hebrew.sty is not bug-free. We still encounter instances where the reversal of directions does not work right. It is still being developed. Having a reliable hebrew.sty file is still a major task in the Hebrew TeX environment. Its completion might be easier in the newer version of LaTeX.

2.3 Regarding Hyphenation

As Prof. Goldberg has pointed out in his article [2], ignoring hyphenation when dealing with Hebrew texts turns out to be a reasonable technique. Hebrew words are short, as there are no vowels included. No hyphenation patterns have been devised, and the English patterns which are used by default give a good result. From time to time there arises the need to hyphenate a word by hand, especially when specific foreign terms are written in Hebrew characters.

The latest version of the hebrew option for LaTeX avoids hyphenation in Hebrew by setting \hyphenpenalty=@MM (@MM is set to 20000 in plain.tex).

The project of devising a hyphenation file for Hebrew might be undertaken by someone who can contribute both linguistic expertise and an inclination towards programming.

3 Hebrew Fonts

There are two complications pertaining to the Hebrew fonts. One complication lies in the fact that there are several arrangements used for Hebrew characters as regards the ASCII sequence.

Available Hebrew fonts are either 7-bit fonts or 8-bit fonts, sometimes referred to as Old Code and New Code, respectively.

In Old Code, each character is represented internally by 7 bits. This is an older system, but it is still much used. In this representation, only 128 characters are possible. Hence, Hebrew characters have to occupy ASCII locations regularly occupied by other characters; actually they take the place of the lower case English characters (ASCII locations 97 and on). However, since there are 27 Hebrew characters and only 26 English lower case characters, the first Hebrew character, Aleph, occupies ASCII location 96, which is the grave accent's location.

In New Code, each character is represented internally by 8 bits. This makes for 256 different characters. The most common code for Hebrew characters, ISO 8859-8, has Hebrew characters starting at ASCII 224.

IBM PCs also use Hebrew 8-bit fonts, but here the Hebrew characters are located in different ASCII locations; they start at ASCII 128.

Still another system for representing Hebrew characters is used in EBCDIC, IBM's collating sequence.

Those various possibilities give rise to confusion.

Another complication lies in the fact that Hebrew has "vowel marks"; those are marks which appear below, within or on top of characters and denote the "vowel" pertaining to the relevant character.

There are about 14 vowels, which can be combined with most of the Hebrew characters. The possibilities are numerous. Surely it is not practical to create fonts containing all the possible character-vowel combinations. It seems that use of vowels should be done via macros.

Most Hebrew texts are written without "vowel" marks, as people understand Hebrew text without them. However, certain kinds of texts - poetry, old texts, texts for beginners in Hebrew - use vowels.

3.1 Available fonts

Fonts with or without vowels are available, which enable creation of nice outputs in Hebrew, but better fonts are needed for high quality printouts.

A REDIS font family has been written in Metafont by Prof. Goldberg and his assistants. It has been written in a modular way, resembling Computer Modern. It has been created in a number of sizes, including a slanted and a boldface version.

Four separate fonts, Jerusalem, TelAviv, OldJaffa and DeadSea, of an unknown source have been converted from PostScript to Metafont using a utility called Typo; this is a font editor, meant for devising and modifying fonts, and it is accompanied by a set of conversion programs which can convert fonts between a number of formats. Typo accepts fonts in the following formats: Metafont, Ikarus, Frame maker, PostScript type 1, Fontographer packed PostScript, EPSF, BDF, PCX.

Typo is a commercial product, written by Prof. Jakob Gonczarowski (jakob@humus.huji.ac.il) from Jerusalem. It has been bought by the Computation Center of the Hebrew University in Jerusalem. TeX users are welcomed to use our services of converting their own fonts by sending their request to me. More information can be obtained from me via e-mail.

Mr. Joel Hoffman has written a "modernized classical Hebrew font", including all vowels. He has also supplied macros which facilitate usage of the vowels. The fonts are called hclassic and hcaption. I have added digits and symbols copied from the font DeadSea onto these two fonts, and have called the modified fonts dclassic and dcaption.

Mr. Jonathan Brecher has created another version of a font family he calls Shalom. He also supplied macros which facilitate using them. These fonts have been written in PostScript via Macintosh Fontographer, and have also been converted to Metafont with Typo. The fonts are called ShalomOldStyle10, ShalomScript10, ShalomStick10.

The Metafont sources of all the above mentioned fonts are available for anonymous ftp retreival from noa.huji.ac.il in tex/fonts. Explanations pertaining to the Shalom fonts can be found in tex/fonts in the file Shalom.readme . Explanations pertaining to hclassic and dclassic fonts can be found in the TeX guide classic_guide.tex, available in tex/tex_guides.

Here's a sample of the available fonts; all Hebrew fonts have been scaled by \magstep1 for readability.

redis10	שורה זו כתובה בעברית
rediss10	שורה זו כתובה בעברית
redisbx10	שורה זו כתובה בעברית
Jerusalem	שורה זו כתובה בעברית
TelAviv	שורה זו כתובה בעברית
OldJaffa	שורה זו כתובה בעברית
DeadSea	שורה זו כתובה בעברית
hclassic, dclassic	שוּרָה זוֹ כְּתוּבָה בְּעִבְרִית
hcaption, dcaption	שוּרָה זוֹ כְּתוּבָה בְּעִבְרִית
ShalomOldStyle10	שוּרָה זוֹ כְּתוּבָה בְּעִבְרִית
ShalomScript10	שוּרָה זוֹ כְּתוּבָה בְּעִבְרִית
ShalomStick10	שוּרָה זוֹ כְּתוּבָה בְּעִבְרִית

4 Creating a Hebrew text

This is a very important though technical issue. In Israel, where there is a demand for Hebrew text entry, manufacturers offer Hebrew glyphs in the display ROM of terminals and PCs, as well as keys which serve as toggles between different mappings of the keyboard. Hebrew fonts have also been written for X-Windows, and can be obtained from several sources. At the same time, a Hebrew supplement to Emacs already exists. Organizing the available information and data pertaining to this matter will hopefully be done soon, and it will be available in noa.huji.ac.il

5 Sources of information for the Hebrew TeX user

5.1 Archives holding Hebrew material: noa.huji.ac.il, ymir.claremont.edu

The archive noa.huji.ac.il [132.64.254.10], located at the Computation Center of the Hebrew University of Jerusalem, Israel, holds material pertaining to Hebrew TeX, TeX in general and related subjects. It is possible to retrieve files from the archive via anonymous ftp from the directory tex. Available subdirectories are:

install_in_unix	TeX-XeT for Unix
install_in_pc	EMTeX, SBXeT, dvips for PC DOS
fonts	Hebrew fonts - sources, pk, tfm, information
hebrew_stuff	Various Hebrew macros files
ivd2dvi	ivd2dvi program for conversion of dvi files
tex_guides	Various guides, mostly concerned with Hebrew usage
latex_guides	Guide for Hebrew Latex, and others.
www	TeX database. See below.

In the near future the following subdirectories will be added:

tex--xet	The new TeX--XeT
pubtex	PUBTeX: TeX--XeT for PC DOS
editing_hebrew	Information about editors, data entry etc.

The archive ymir.claremont.edu holds most of the Hebrew material, available for anonymous ftp, in [anonymous.tex.babel.hebrew]. Not all files are updated.

The available subdirectories in ymir.claremont.edu's hebrew directory are:

fonts-hclassic
fonts-huji
fonts-ps-shalom
fonts-redis
ivd2dvi
macros
prtex
tex-xet

5.2 TeX Database, Usable Via WWW

The subdirectory tex/www, in noa.huji.ac.il, serves as an information database for all TeX matters, including Hebrew. The database is built so that it can easily be accessed via WWW – WorldWideWeb, a public domain interactive line browser.

Files within the database have been and are being copied (and modified) from information lists pertaining to TeX; others have been written by me. Files are constantly being updated and added.

To both see how it works and to use the TeX database, you can telnet onto vms.huji.ac.il (currently 128.139.4.3) with the username WWW . The database itself can be copied and used by any system which has a working WWW. All you have to do is copy the files and make a local link to the top file, called top.html, which is included.

A "tarred" and compressed file, called tex_database.tar.Z, containing the latest version of all the files in the database, is located in noa.huji.ac.il in tex/www/compressed.

WWW itself can be retrieved from info.cern.ch in pub/www

5.3 The IVRITEX mailing list

IVRITEX is a discussion list dealing with all subjects related to typesetting texts via Hebrew TEX. It is maintained by Don Hosek (dhosek@hmcvax.claremont.edu). To join the list, send a one line message saying

SUB IVRITEX name familyname to

LISTSERV@TAUNIVM.BITNET or to

LISTSERV%TAUNIVM.BITNET@CUNYVM.CUNY.EDU

References

[1] Donald E. Knuth and Pierre MacKay, "Mixing right-to-left texts with left-to-right texts", *TUGboat*, Volume 8 (1987), No. 1, pp. 14--25.

[2] Jacques J. Goldberg, "Approximate TEX for Semitic Languages", in *1988 Conference Proceedings, TEX Users Group*, pp. 171–178.