Perspectives on Technology

Hebrew in Bits and Bytes: An Introduction to Coding and Formatting of Hebrew Electronic Resources

Heidi Lerner

During the past two decades, many Hebraic language electronic resources became available in both Hebrew and Latin scripts. These texts, databases, and bibliographic tools either required researchers to transliterate (romanize) the text using Latin characters, or to use a proprietary and stand-alone software program that displays Hebrew characters using special fonts and add-on features.

Romanization of Hebrew is problematic at best: There are many different schemes in use today, and to provide correct vocalization, a strong knowledge of Hebrew grammar is required. Most of us have experienced great frustration in trying to locate Hebraica materials in library catalogs and periodical indexes that are in the Latin alphabet only. Academic journals and encyclopedias vary in their requirements for transliteration of Hebraica. Diacritics are required to represent certain Hebrew characters. Additional diacritics are employed to represent special consonants employed by the many and diverse Jewish languages written in Hebrew characters.

At the same time, users of Hebrew script software have been faced with difficulties when attempting to share their work with anyone, communicate via e-mail, or transfer files between programs and operating systems. Today's rapidly evolving research environment requires that Jewish studies specialists know the basics of multilingual and multiscript computing.

ASCII versus Unicode

A computer records text as a sequence of numbers in binary form. One of the earliest standards for numerically encoding the Latin alphabet was the American Standard Code for Information Interchange (ASCII). This 7-bit code (a "bit" is a single "binary digit" with a value of 0 or 1) only covered 128 characters, consisting of the English alphabet, numbers, punctuation and some symbols. ASCII was later extended to 8 bits to include accented characters for other Western European languages. Other international standards developed to include character sets such as Greek and Hebrew. However, 8-bit character encoding was limited to 256 characters, and these standards were usually inadequate when users wanted to work in more than one language at once. Applications using such standards are forced to switch between character sets to obtain characters or symbols not provided by the normally used default set. To make matters more complicated, earlier DOS and Apple programs used character sets that did not comply with international standards.

Unicode Version 4.0 encodes over 95,000 characters, covering most modern and historic scripts. Unicode has the *potential* to encode over a million characters. The Unicode Standard also gives specifications for the presentation of bi-directional text: Hebrew, Arabic, etc., are properly output as right-to-left.

Hebrew and MS Windows Unicode support is provided in Windows 2000 and Windows XP, and, in a more limited scope, in Windows 98, NT4, and ME. What this means is that users can create and disseminate documents that are directly readable, searchable, and printable in Hebrew and Latin scripts. Scholars can cut and paste Hebrew text directly from Unicode-based resources into Word, send Hebrew e-mail in Outlook and Outlook Express, and mix scripts within documents. Windows XP and Windows 2000 also support bi-directionality, allowing users to use *most* software both from left-to-right and right-to-left (provided that the programs allow for bi-directional use). In addition, Microsoft Proofing Tools offers special editing tools for Hebrew: thesauri, spelling and grammar checkers, a translation dictionary, and specialized fonts. The other more proprietary and often incompatible formats do not allow the same ease for interchanging data between databases and application software. Hebrew support is not yet available for Macintosh versions of Internet Explorer or Office. The Netscape 7 Web browser is Unicode-compatible.

Most of the Hebrew fonts bundled with Windows (Times New Roman, Arial, Tahoma, Courier New, Arial Unicode, Lucida Sans Unicode, David, and Miriam) do not support cantillation (*te'amim*) or even some of the non-standard *nikud*. For these characters, special Unicode fonts are required that support Hebrew fully and some that are already available include SIL Ezra Hebrew Unicode Fonts (freeware produced and distributed by the Summer Institute of Linguistics [www.sil.org/computing/catalog/show_software.asp?id=76]), and Code2000 and Code2001 fonts (shareware, produced by James Kass [home.att.net/~jameskass/code2001.htm]).

Hebraica Resources

A number of important electronic resources in Jewish studies now use Unicode. These include: the most recent editions of the *Bar-Ilan Judaic Library*, some publications from Mechon Mamre, the *Penn/Cambridge Genizah Fragment Project* based at the University of Pennsylvania's Schoenberg Center for Electronic Text and Image, and bibliographic databases including the *Eureka* interface to the RLIN database, the *Index to Hebrew Periodicals* (IHP), the *Index to Periodicals in Jewish Studies* (RAMBI), and the *Israel Union Catalog* (ULI) and *Union List of Serials* (ULS) in Israeli libraries. Unfortunately, many other electronic resources still rely on older 7-bit and 8-bit encoding. These include the *Historical Dictionary of the Hebrew Language*, *Otzar ha-Poskim*, *Takdin*, *Bibliography of the Hebrew Book*, *Dead Sea Scrolls Electronic Reference Library*, and the *Henkind Talmud Text Databank*. One hopes that publishers of these resources will adopt the Unicode standard, a step that would greatly enhance their scholarly utility.

DEFINITIONS:

Bi-directional Display (BIDI): The process or result of mixing left-to-right oriented text and right-to-left oriented text in a single line.
Character: The minimal unit of encoding for a character set. A character often corresponds to a single graphic sign of a writing system, e.g., a letter or a punctuation mark.
Character Set: A table that assigns codes to characters so that the characters can be stored and manipulated in computer applications.
Code point: A numerical index (or position) in an encoding table used for encoding characters.
Diacritic: A small mark added above, below, or after a base character to change its pronunciation.
Encoding: The process of assigning characters to available code points so that the characters can be represented in computer applications.
Font: A collection of glyphs used for the visual depiction of character data.
Glyph: An image used in the visual depiction of characters. Often, for a given font, there is a one-to-one relationship between an encoded character and a glyph. But in languages with complex writing, one character may correspond to several glyphs, or several characters to one glyph.
Logical order: Order in which characters are typed on a keyboard.
Nikud/Te'am : See "Diacritic."
Visual order: Order of characters as they are presented for reading.

FAQ'S:
What do I do if?

Hebrew appears backwards or displays as gibberish in Internet Explorer?: Open the **View** menu and choose **Encoding**. If **Hebrew (Windows)** or **Hebrew (ISO-Logical)** is selected, click on **Hebrew (ISO-Visual)**. If **Hebrew (ISO-Visual)** is selected, click on **Hebrew (Windows)**. You can also try **Hebrew (ISO-Logical)**.

Hebrew appears backwards or displays as gibberish in Netscape 7?: Open the **View** menu and choose **Character Coding**. If **Hebrew (Windows-1255)** or **Hebrew (ISO-8859-8-I)** is selected, click on **Hebrew Visual (ISO-8859-8)**. If **Hebrew Visual (ISO-8859-8)** is selected, click on **Hebrew (Windows-1255)**. You can also try **Hebrew (ISO-8859-8-I)**.

I need to insert a special diacritic or symbol in Word?: Select a Unicode font, all of which offer a full array of diacritics. You can **(A)** Click where in the document you want to insert the character. Open the **Insert** menu and click **Symbol**. Click the **Special Characters** tab and double-click the character you want to insert. **(B**) Use **Character Map** by opening Start menu/Programs/Accessories/System Tools/Character Map. Windows 2000/XP users may check **Advanced View**; set **Character Set** to **Unicode**; and group by **Unicode Subrange**. Next, choose **Hebrew** to display the full array of Hebrew characters in the selected font. Double-click on the selected character, or highlight a character and click on **Select**. The character(s) can then be pasted into Word.

Heidi Lerner is the Hebraica/Judaica Cataloger at Stanford University.

Resources:
1. Unicode home page: www.unicode.org
2. Hebrew Computing on Windows (Web site, maintained by Tsuguya Sasaki): www.jewish-languages.org/windows.html
3. Issues in the Representation of Pointed Hebrew in Unicode (3rd draft, Peter Kirk, August 2003): www.qaya.org/academic/hebrew/Issues-Hebrew-Unicode.html
4. Enabling International Support in Windows 2000: www.microsoft.com/globaldev/handson/user/2kintlsupp.mspx
5. Working with Non-Roman Script Text in MS Windows Applications: www.lib.umich.edu/area/Near.East/NonRomanDemo.pdf