Hebrew input encodings for use with LaTeX 2ε

Boris Lavva

Printed February 2, 2013

1 Hebrew input encodings

Hebrew input encodings defined in file `hebinp.dtx`[1] should be used with `inputenc` LaTeX 2ε package. This package allows the user to specify an input encoding from this file (for example, ISO Hebrew/Latin 8859-8, IBM Hebrew codepage 862 or MS Windows Hebrew codepage 1255) by saying:

`\usepackage[`*encoding name*`]{inputenc}`

The encoding can also be selected in the document with:

`\inputencoding{`*encoding name*`}`

The only practical use of this command within a document is when using text from several documents to build up a composite work such as a volume of journal articles. Therefore this command will be used only in vertical mode.

The encodings provided by this package are:

- `si960` 7-bit Hebrew encoding for the range 32–127. This encoding also known as "old-code" and defined by Israeli Standard SI-960.
- `8859-8` ISO 8859-8 Hebrew/Latin encoding commonly used in UNIX systems. This encoding also known as "new-code" and includes hebrew letters in positions starting from 224.
- `cp862` IBM 862 code page commonly used by DOS on IBM-compatible personal computers. This encoding also known as "pc-code" and includes hebrew letters in positions starting from 128.
- `cp1255` MS Windows 1255 (hebrew) code page which is similar to 8859-8. In addition to hebrew letters, this encoding contains also hebrew vowels and dots (nikud).

Each encoding has an associated `.def` file, for example `8859-8.def` which defines the behaviour of each input character, using the commands:

[1] The files described in this section have version number v1.1b and were last revised on 2004/02/20.

```
\DeclareInputText{slot}{text}
\DeclareInputMath{slot}{math}
```

This defines the input character *slot* to be the *text* material or *math* material respectively. For example, `8859-8.def` defines slots `"EA` (letter hebalef) and `"B5` (μ) by saying:

```
\DeclareInputText{224}{\hebalef}
\DeclareInputMath{181}{\mu}
```

Note that the *commands* should be robust, and should not be dependent on the output encoding. The same *slot* should not have both a text and a math declaration for it. (This restriction may be removed in future releases of inputenc).

The `.def` file may also define commands using the declarations: `\providecommand` or `\ProvideTextCommandDefault`. For example, `8859-8.def` defines:

```
\ProvideTextCommandDefault{\textonequarter}{\ensuremath{\frac14}}
\DeclareInputText{188}{\textonequarter}
```

The use of the 'provide' forms here will ensure that a better definition will not be over-written; their use is recommended since, in general, the best defintion depends on the fonts available.

See the documentation in `inputenc.dtx` for details of how to declare input definitions for various encodings.

1.1 Default definitions for characters

First, we insert a `\makeatletter` at the beginning of all `.def` files to use `@` symbol in the macros' names.

```
⟨-driver⟩\makeatletter
```

Some input characters map to internal functions which are not in either the `T1` or `OT1` font encoding. For this reason default definitions are provided in the encoding file: these will be used unless some other output encoding is used which supports those glyphs. In some cases this default defintion has to be simply an error message.

Note that this works reasonably well only because the encoding files for both `OT1` and `T1` are loaded in the standard LaTeX format.

```
⟨*8859-8 | cp862 | cp1255⟩
\ProvideTextCommandDefault{\textdegree}{\ensuremath{{^\circ}}}
\ProvideTextCommandDefault{\textonehalf}{\ensuremath{\frac12}}
\ProvideTextCommandDefault{\textonequarter}{\ensuremath{\frac14}}
⟨/8859-8 | cp862 | cp1255⟩
⟨*8859-8 | cp1255⟩
\ProvideTextCommandDefault{\textthreequarters}{\ensuremath{\frac34}}
⟨/8859-8 | cp1255⟩
⟨*cp862 | cp1255⟩
\ProvideTextCommandDefault{\textflorin}{\textit{f}}
```

```
⟨/cp862 | cp1255⟩
⟨*cp862⟩
\ProvideTextCommandDefault{\textpeseta}{Pt}
⟨/cp862⟩
```

The name `\textblacksquare` is derived from the AMS symbol name since Adobe seem not to want this symbol. The default definition, as a rule, makes no claim to being a good design.

```
⟨*cp862⟩
\ProvideTextCommandDefault{\textblacksquare}
   {\vrule \@width .3em \@height .4em \@depth -.1em\relax}
⟨/cp862⟩
```

Some commands can't be faked, so we have them generate an error message.

```
⟨*8859-8 | cp862 | cp1255⟩
\ProvideTextCommandDefault{\textcent}
   {\TextSymbolUnavailable\textcent}
\ProvideTextCommandDefault{\textyen}
   {\TextSymbolUnavailable\textyen}
⟨/8859-8 | cp862 | cp1255⟩
⟨*8859-8⟩
\ProvideTextCommandDefault{\textcurrency}
   {\TextSymbolUnavailable\textcurrency}
⟨/8859-8⟩
⟨*cp1255⟩
\ProvideTextCommandDefault{\newsheqel}
   {\TextSymbolUnavailable\newsheqel}
⟨/cp1255⟩
⟨*8859-8 | cp1255⟩
\ProvideTextCommandDefault{\textbrokenbar}
   {\TextSymbolUnavailable\textbrokenbar}
⟨/8859-8 | cp1255⟩
⟨*cp1255⟩
\ProvideTextCommandDefault{\textperthousand}
   {\TextSymbolUnavailable\textperthousand}
⟨/cp1255⟩
```

Characters that are supposed to be used only in math will be defined by `\providecommand` because LaTeX 2_ε assumes that the font encoding for math fonts is static.

```
⟨*8859-8 | cp1255⟩
\providecommand{\mathonesuperior}{{^1}}
\providecommand{\maththreesuperior}{{^3}}
⟨/8859-8 | cp1255⟩
⟨*8859-8 | cp862 | cp1255⟩
\providecommand{\mathtwosuperior}{{^2}}
⟨/8859-8 | cp862 | cp1255⟩
⟨*cp862⟩
\providecommand{\mathordmasculine}{{^o}}
\providecommand{\mathordfeminine}{{^a}}
⟨/cp862⟩
```

1.2 The SI-960 encoding

The SI-960 or "old-code" encoding only allows characters in the range 32–127, so we only need to provide an empty `si960.def` file.

1.3 The ISO 8859-8 encoding and the MS Windows cp1255 encoding

The `8859-8.def` encoding file defines the characters in the ISO 8859-8 encoding.

The MS Windows Hebrew character set incorporates the Hebrew letter repertoire of ISO 8859-8, and uses the same code points (starting from 224). It has also some important additions in the 128–159 and 190–224 ranges.

```
⟨*cp1255⟩
\DeclareInputText{130}{\quotesinglbase}
\DeclareInputText{131}{\textflorin}
\DeclareInputText{132}{\quotedblbase}
\DeclareInputText{133}{\dots}
\DeclareInputText{134}{\dag}
\DeclareInputText{135}{\ddag}
\DeclareInputText{136}{\^{}}
\DeclareInputText{137}{\textperthousand}
\DeclareInputText{139}{\guilsinglleft}
\DeclareInputText{145}{\textquoteleft}
\DeclareInputText{146}{\textquoteright}
\DeclareInputText{147}{\textquotedblleft}
\DeclareInputText{148}{\textquotedblright}
\DeclareInputText{149}{\textbullet}
\DeclareInputText{150}{\textendash}
\DeclareInputText{151}{\textemdash}
\DeclareInputText{152}{\~{}}
\DeclareInputText{153}{\texttrademark}
\DeclareInputText{155}{\guilsinglright}
⟨/cp1255⟩
```

```
⟨*8859-8 | cp1255⟩
\DeclareInputText{160}{\nobreakspace}
\DeclareInputText{162}{\textcent}
\DeclareInputText{163}{\pounds}
⟨+8859-8⟩\DeclareInputText{164}{\textcurrency}
⟨+cp1255⟩\DeclareInputText{164}{\newsheqel}
\DeclareInputText{165}{\textyen}
\DeclareInputText{166}{\textbrokenbar}
\DeclareInputText{167}{\S}
\DeclareInputText{168}{\"{}}
\DeclareInputText{169}{\textcopyright}
⟨+8859-8⟩\DeclareInputMath{170}{\times}
\DeclareInputText{171}{\guillemotleft}
\DeclareInputMath{172}{\lnot}
\DeclareInputText{173}{\-}
```

```
\DeclareInputText{174}{\textregistered}
\DeclareInputText{175}{\@tabacckludge={}}
\DeclareInputText{176}{\textdegree}
\DeclareInputMath{177}{\pm}
\DeclareInputMath{178}{\mathtwosuperior}
\DeclareInputMath{179}{\maththreesuperior}
\DeclareInputText{180}{\@tabacckludge'{}}
\DeclareInputMath{181}{\mu}
\DeclareInputText{182}{\P}
\DeclareInputText{183}{\textperiodcentered}
⟨+8859-8⟩\DeclareInputText{184}{\c\ }
\DeclareInputMath{185}{\mathonesuperior}
⟨+8859-8⟩\DeclareInputMath{186}{\div}
\DeclareInputText{187}{\guillemotright}
\DeclareInputText{188}{\textonequarter}
\DeclareInputText{189}{\textonehalf}
\DeclareInputText{190}{\textthreequarters}
⟨/8859-8 | cp1255⟩
```

Hebrew vowels and dots (nikud) are included only to MS Windows cp1255 page and start from the position 192.

```
⟨*cp1255⟩
\DeclareInputText{192}{\hebsheva}
\DeclareInputText{193}{\hebhatafsegol}
\DeclareInputText{194}{\hebhatafpatah}
\DeclareInputText{195}{\hebhatafqamats}
\DeclareInputText{196}{\hebhiriq}
\DeclareInputText{197}{\hebtsere}
\DeclareInputText{198}{\hebsegol}
\DeclareInputText{199}{\hebpatah}
\DeclareInputText{200}{\hebqamats}
\DeclareInputText{201}{\hebholam}
\DeclareInputText{203}{\hebqubuts}
\DeclareInputText{204}{\hebdagesh}
\DeclareInputText{205}{\hebmeteg}
\DeclareInputText{206}{\hebmaqaf}
\DeclareInputText{207}{\hebrafe}
\DeclareInputText{208}{\hebpaseq}
\DeclareInputText{209}{\hebshindot}
\DeclareInputText{210}{\hebsindot}
\DeclareInputText{211}{\hebsofpasuq}
\DeclareInputText{212}{\hebdoublevav}
\DeclareInputText{213}{\hebvavyod}
\DeclareInputText{214}{\hebdoubleyod}
⟨/cp1255⟩
```

Hebrew letters start from the position 224 in both encodings.

```
⟨*8859-8 | cp1255⟩
\DeclareInputText{224}{\hebalef}
\DeclareInputText{225}{\hebbet}
```

```
\DeclareInputText{226}{\hebgimel}
\DeclareInputText{227}{\hebdalet}
\DeclareInputText{228}{\hebhe}
\DeclareInputText{229}{\hebvav}
\DeclareInputText{230}{\hebzayin}
\DeclareInputText{231}{\hebhet}
\DeclareInputText{232}{\hebtet}
\DeclareInputText{233}{\hebyod}
\DeclareInputText{234}{\hebfinalkaf}
\DeclareInputText{235}{\hebkaf}
\DeclareInputText{236}{\heblamed}
\DeclareInputText{237}{\hebfinalmem}
\DeclareInputText{238}{\hebmem}
\DeclareInputText{239}{\hebfinalnun}
\DeclareInputText{240}{\hebnun}
\DeclareInputText{241}{\hebsamekh}
\DeclareInputText{242}{\hebayin}
\DeclareInputText{243}{\hebfinalpe}
\DeclareInputText{244}{\hebpe}
\DeclareInputText{245}{\hebfinaltsadi}
\DeclareInputText{246}{\hebtsadi}
\DeclareInputText{247}{\hebqof}
\DeclareInputText{248}{\hebresh}
\DeclareInputText{249}{\hebshin}
\DeclareInputText{250}{\hebtav}
⟨/8859-8 | cp1255⟩
```

Special symbols which define the direction of symbols explicitly. Currently, they are not used in LaTeX.

```
⟨*cp1255⟩
\DeclareInputText{253}{\lefttorightmark}
\DeclareInputText{254}{\righttoleftmark}
⟨/cp1255⟩
```

1.4 The IBM code page 862

The `cp862.def` encoding file defines the characters in the IBM codepage 862 encoding. The DOS graphics 'letters' and a few other positions are ignored (left undefined).

Hebrew letters start from the position 128.

```
⟨*cp862⟩
\DeclareInputText{128}{\hebalef}
\DeclareInputText{129}{\hebbet}
\DeclareInputText{130}{\hebgimel}
\DeclareInputText{131}{\hebdalet}
\DeclareInputText{132}{\hebhe}
\DeclareInputText{133}{\hebvav}
\DeclareInputText{134}{\hebzayin}
\DeclareInputText{135}{\hebhet}
```

```
\DeclareInputText{136}{\hebtet}
\DeclareInputText{137}{\hebyod}
\DeclareInputText{138}{\hebfinalkaf}
\DeclareInputText{139}{\hebkaf}
\DeclareInputText{140}{\heblamed}
\DeclareInputText{141}{\hebfinalmem}
\DeclareInputText{142}{\hebmem}
\DeclareInputText{143}{\hebfinalnun}
\DeclareInputText{144}{\hebnun}
\DeclareInputText{145}{\hebsamekh}
\DeclareInputText{146}{\hebayin}
\DeclareInputText{147}{\hebfinalpe}
\DeclareInputText{148}{\hebpe}
\DeclareInputText{149}{\hebfinaltsadi}
\DeclareInputText{150}{\hebtsadi}
\DeclareInputText{151}{\hebqof}
\DeclareInputText{152}{\hebresh}
\DeclareInputText{153}{\hebshin}
\DeclareInputText{154}{\hebtav}

\DeclareInputText{155}{\textcent}
\DeclareInputText{156}{\pounds}
\DeclareInputText{157}{\textyen}
\DeclareInputText{158}{\textpeseta}
\DeclareInputText{159}{\textflorin}
\DeclareInputText{160}{\@tabacckludge'a}
\DeclareInputText{161}{\@tabacckludge'\i}
\DeclareInputText{162}{\@tabacckludge'o}
\DeclareInputText{163}{\@tabacckludge'u}
\DeclareInputText{164}{\~n}
\DeclareInputText{165}{\~N}
\DeclareInputMath{166}{\mathordfeminine}
\DeclareInputMath{167}{\mathordmasculine}
\DeclareInputText{168}{\textquestiondown}
\DeclareInputMath{170}{\lnot}
\DeclareInputText{171}{\textonehalf}
\DeclareInputText{172}{\textonequarter}
\DeclareInputText{173}{\textexclamdown}
\DeclareInputText{174}{\guillemotleft}
\DeclareInputText{175}{\guillemotright}

\DeclareInputMath{224}{\alpha}
\DeclareInputText{225}{\ss}
\DeclareInputMath{226}{\Gamma}
\DeclareInputMath{227}{\pi}
\DeclareInputMath{228}{\Sigma}
\DeclareInputMath{229}{\sigma}
\DeclareInputMath{230}{\mu}
\DeclareInputMath{231}{\tau}
\DeclareInputMath{232}{\Phi}
\DeclareInputMath{233}{\Theta}
```

```
\DeclareInputMath{234}{\Omega}
\DeclareInputMath{235}{\delta}
\DeclareInputMath{236}{\infty}
\DeclareInputMath{237}{\phi}
\DeclareInputMath{238}{\varepsilon}
\DeclareInputMath{239}{\cap}
\DeclareInputMath{240}{\equiv}
\DeclareInputMath{241}{\pm}
\DeclareInputMath{242}{\ge}
\DeclareInputMath{243}{\le}
\DeclareInputMath{246}{\div}
\DeclareInputMath{247}{\approx}
\DeclareInputText{248}{\textdegree}
\DeclareInputText{249}{\textperiodcentered}
\DeclareInputText{250}{\textbullet}
\DeclareInputMath{251}{\surd}
\DeclareInputMath{252}{\mathnsuperior}
\DeclareInputMath{253}{\mathtwosuperior}
\DeclareInputText{254}{\textblacksquare}
\DeclareInputText{255}{\nobreakspace}
⟨/cp862⟩
```

`\DisableNikud` A utility macro to ignore any nikud character that may appear in the input. This allows you to ignore cp1255 nikud characters that happened to appear in the input.

```
⟨*8859-8⟩
\newcommand{\DisableNikud}{%
  \DeclareInputText{192}{}%
  \DeclareInputText{193}{}%
  \DeclareInputText{194}{}%
  \DeclareInputText{195}{}%
  \DeclareInputText{196}{}%
  \DeclareInputText{197}{}%
  \DeclareInputText{198}{}%
  \DeclareInputText{199}{}%
  \DeclareInputText{200}{}%
  \DeclareInputText{201}{}%
  \DeclareInputText{203}{}%
  \DeclareInputText{204}{}%
  \DeclareInputText{205}{}%
  \DeclareInputText{206}{}%
  \DeclareInputText{207}{}%
  \DeclareInputText{208}{}%
  \DeclareInputText{209}{}%
  \DeclareInputText{210}{}%
  \DeclareInputText{211}{}%
  \DeclareInputText{212}{}%
  \DeclareInputText{213}{}%
  \DeclareInputText{214}{}%
}
```

268 ⟨/8859-8⟩

Finally, we reset the category code of the `@` sign at the end of all `.def` files.

269 ⟨-driver⟩`\makeatother`