Advanced Unicode handling

Many Paratext supporters are aware of how Paratext can help you manage Unicode characters:

- You can type a Unicode character code value, then press Alt-X to switch to the character. Or if you press ALT-X when the cursor is to the right of a character, Paratext will change that character into its Unicode code value. (If you cannot edit the text, it will show the code in a popup).
  Example: type **0257**, press ALT-x, you will see **ɗ** in the text. Move the cursor to the right of a **ɗ** and press ALT-x, you will see **0257**. (This also works in the language properties lists, the wordlist filter box, and other places, not just in a text window).
- The characters inventory lists all the characters in a project and shows their Unicode code values. You can sort by the Unicode code values.
- You can specify a Unicode character in a autocorrect.txt file with a code \uXXXX where the 4 Xs represent the Unicode value. For example d/-->\u0257 will generate a **ɗ** when you type d then /.
- In RegExPal you can specify a Unicode character by its code value with \uXXXX, for example \u0257 for **ɗ**.

But what about Unicode characters off the Basic Multilingual Plane (BMP), whose code value is more than four hex digits? Do these tools still work?

- To work with ALT-X, enter 8 digits for any code greater than 4 digits. For example, U+11010 *Brahmi Letter Ai* you would enter the code as **00011010.** (Despite having only 1's and 0's, this is not a binary number but a hexadecimal number)
- In the character inventory, any Unicode value greater than four digits will also appear as 8 digits. So the character inventory will show **00011010** as the Unicode value for Brahmi Letter AI.
- For an autocorrect file, you have to encode these characters as a pair of four digit characters called a surrogate pair. Our Brahmi Letter AI can be written like this: **ai-->\ud804\udc10.** You can find the surrogate pair values for Unicode characters by Googling the code value (**"U+11010")**, then looking on the entry on compart.com or fileformat.info for the UTF16 value.
- In RegExPal, you specify the surrogate pair to specify a non-BMP character. **\ud804\udc10** will match the *Brahmi Letter Ai*.