L2/11-321

Subject:	Property Overview for Selected Odd Characters
From:	Mark Davis
Date:	2011-01-01

While looking at some of the property issues with UCA, I took a fresh look at the Lm characters and enclosed characters. Attached are overviews of these properties, with some inconsistent cases marked by white-on-black.

Because people trip over these inconsistencies fairly often, I think we should document these issues somewhere.

- See http://goo.gl/s4jpy for the spreadsheet view, which allows filtering and resorting.
- See http://goo.gl/34Axd for the HTML view

There are two separate sheets:

- "Letter Modifiers"
  - gc=Lm + dt=super + dt=sub
  - restricted to only 'cased' scripts "Zyyy", "Armn", "Copt", "Dest", "Glag", "Cyrl", "Grek", "Geor", "Latn"
- "Enclosed" Characters
  - block=/enclosed/
  - + name=/CIRCLED|PARENTHESIZED|SQUARED/
  - - \p{Cn} - \p{Sm} - \p{Lo} - \p{Po}

Key

- Count - number of characters
- Dt - decomposition type
- SC - script code (Zyyy = Common)
- dSc - script code* of NFKD form
- GC - general category
- CWUC - changes when uppercased
- L/U/A - '-' for non-Alphabetic; otherwise 'l' for Lowercase, ow 'U' for Uppercase, ow 'A' for other Alphabetic
- dL/U/A - L/U/A value* for NDKD form

*Details

For the script of a string, I collect the scripts for each character. If there is one script, or one script + common, I return that script. Otherwise I return Common.

Similarly, for the LUA of a string, I collect the LUA for each character. If I hit '-', I return it. Otherwise, if there is one case value or one case value + "A", I return it. Otherwise I return "A"