L2/19-191

Title: **UAX#38 vs. UAX#42, and Unihan properties**
Author: Richard Cook
Action: For UTC Consideration
Date: 2019-02-02

UAX#38 (UniHan) and UAX#42 (UCD in XML) use somewhat different conventions to represent the syntax of Unihan properties.

UAX#38 uses a rather compressed style of PCRE, with non-ASCII (combining mark) character escapes, operating on NFD. Regex appearing in a given version of UAX #38 match the data for the corresponding published version of Unihan (without regard to backward compatibility with prior versions). In the current UAX#38 draft the syntax description has been formatted for improved legibility, though it is still rather compressed.

UAX#42 uses a more verbose style of regex (not clearly PCRE), without non-ASCII (combining mark, etc.) character escapes, operating on NFC. The regex appearing in a given version of UAX#42 seem to preserve backward compatibility with certain prior versions of Unihan data. The current UAX#42 draft expands the property prefixes, verbosely, with improved legibility.

In ad hoc discussion it has been suggested that it might be possible (and useful) to agree on standard representations of Unihan property syntax, to clarify the relation between (and guaranty the equivalence of) UAX#38 and UAX#42 practices. As such, the differences and benefits of each approach need be determined, in order to eliminate unnecessary differences and to converge on a single common representation (such as may be useful).