Re: Apparent Sentence_Break miscategorizations (PRI #372)

From: Mark Davis

Date: 2019-07-23

Regarding the following action:

155-A84	Andy Heninger, Mark Davis, Deborah Anderson, Chris Chapman	Investigate PRI #372 feedback from fantasai on "Apparent Sentence_Break miscategorizations" (Thu Mar 8 02:36:01 CST 2018) and make suggestions for Unicode 13.0, for UTC #158. (January 2019)

Andy did an investigation, and had the following recommendations (interspersed in boldface). (I agree with these recommendations.)

CSS is trying to rely on these categorizations, it would be helpful if they were rigorous or if we understood why they are idiosyncratic like this.

1. Semicolons are all categorized under Other, whereas colons and commas are categorized under SContinue. It seems to make more sense that semicolons be categorized under Scontinue.

   Semicolons ⇒ SContinue

2. The Greek Question Mark is categorized as Other rather than with the other question marks in STerm.

   Greek also uses the semicolon, and that is canonically equivalent. So the class has to work with both. But SContinue does, so move:

   Greek Question Mark ⇒ SContinue

3. Old Nubian punctuation (COPTIC OLD NUBIAN) seems to have not been categorized at all, and is filed under Other.

   This appears to mean the following, but they are not "Other"

   - U+2CF9 COPTIC OLD NUBIAN FULL STOP lb=EX
   - U+2CFA COPTIC OLD NUBIAN DIRECT QUESTION MARK lb=BA
   - U+2CFB COPTIC OLD NUBIAN INDIRECT QUESTION MARK lb=BA
   - U+2CFC COPTIC OLD NUBIAN VERSE DIVIDER lb=BA

   [TBD: get recommendations from Debbie].

4. Vertical Forms (PRESENTATION FORM FOR VERTICAL) punctuation is also not categorized with its canonical equivalents, and should be.

   [\p{Block=Vertical%20Forms}-[:cn:]] ⇒ sb(toNFKC(x))

See

https://unicode.org/cldr/utility/list-unicodeset.jsp?a=%5B%3AGeneral_category%3DPo%3A%5D&g=Sentence_Break&i=