L2/13-215

Title: Inventory of duplicate and absent "@missing" directives in the UCD files
Source: Ken Whistler (SAP AG) and Laurențiu Iancu (Microsoft Corporation)
Status: Individual contribution
Action: For consideration by the Unicode Technical Committee
Sequel to: L2/13-169
Date: 2013-11-01

1. Background

In the UCD files, the specially formatted "@missing" comment lines are directives that provide default property values in a machine-readable form. The "@missing" directives are not uniformly defined across the UCD files that use them. Document L2/13-169 [1] summarized the inconsistencies and proposed options to address them. Following Consensus 136-C4 of the UTC [2], the inconsistencies affecting the then-draft BidiBrackets.txt data file were resolved and documentation was added in a new Section 4.2.10 *@missing Conventions* in UAX #44 of Unicode 6.3 [3].

To further rationalize the "@missing" directives, the UTC assigned Action Item 136-A20 [2] to identify the "@missing" directives defined in more than one UCD file for the same properties as well as identify the properties that do not have "@missing" directives anywhere. This document fulfils that action item and proposes modifications to the "@missing" lines to reduce both duplication and the number of "@missing" conventions, and thus also simplify the documentation.

2. Duplicate "@missing" directives

As of Unicode 6.3, the properties in the following table have "@missing" directives defined in more than one UCD file. Two files with unusual "@missing" lines are highlighted in yellow.

Properties	UCD 6.3 files	UCD 6.3 "@missing" lines
Bidi_Mirroring_Glyph	PropertyValueAliases.txt	# @missing: 0000..10FFFF; Bidi_Mirroring_Glyph; <none>
	BidiMirroring.txt	# @missing: 0000..10FFFF; <none>
Case_Folding and Simple_Case_Folding	PropertyValueAliases.txt	# @missing: 0000..10FFFF; Case_Folding; <code point>
		# @missing: 0000..10FFFF; Simple_Case_Folding; <code point>
	CaseFolding.txt	# @missing: 0000..10FFFF; C; <code point>
Line_Break	LineBreak.txt	# @missing: 0000..10FFFF; XX
	DerivedLineBreak.txt	# @missing: 0000..10FFFF; Unknown
NFKC_Casefold	PropertyValueAliases.txt	# @missing: 0000..10FFFF; NFKC_Casefold; <code point>
	DerivedNormalizationProps.txt	# @missing: 0000..10FFFF; NFKC_CF; <code point>
Numeric_Value	PropertyValueAliases.txt	# @missing: 0000..10FFFF; Numeric_Value; NaN
	DerivedNumericValues.txt	# @missing: 0000..10FFFF; NaN; ; NaN
Script_Extensions	PropertyValueAliases.txt	# @missing: 0000..10FFFF; Script_Extensions; <script>
	ScriptExtensions.txt	# @missing: 0000..10FFFF; <script>

Note that for all properties with duplicate "`@missing`" directives except Line_Break, one of the participating UCD files is PropertyValueAliases.txt. In the case of Line_Break, the "`@missing`" directives are in LineBreak.txt and DerivedLineBreak.txt (in the UCD subdirectory "extracted"). In that particular case, one conceivable way to resolve the duplication would be to delete the "`@missing`" directives from both files and add it to PropertyValueAliases.txt, as the ordinary home for "`@missing`" directives.

Duplicate "`@missing`" directives do not necessarily pose problems so long as the multiply-defined directives are consistent. The duplicate "`@missing`" directives tabulated earlier are all consistent between the instances corresponding to the same properties. The highlighted cases, CaseFolding.txt and DerivedNumericValues.txt, are unusual because of their "`@missing`" syntactic patterns. Section 4.2.10 *@missing Conventions* of UAX #44 documents four "`@missing`" line patterns [1, 3]. Most UCD files use two of them, while only three files (CaseFolding.txt, DerivedNumericValues.txt, and SpecialCasing.txt) use the other two:

1. code_point_range; default_prop_val
2. code_point_range; default_prop_val; default_prop_val; default_prop_val
    - used only in DerivedNumericValues.txt and SpecialCasing.txt
3. code_point_range; property_name; default_prop_val
4. code_point_range; status; default_prop_val
    - used only in CaseFolding.txt

2.1. "`@missing`" directive in CaseFolding.txt

The "`@missing`" directive in CaseFolding.txt contains a field that represents case-mapping status C (common), which participates in both Simple_Case_Folding and full Case_Folding. Thus that directive forms a duplicate with two individual directives from PropertyValueAliases.txt for the two case-folding properties. The "`@missing`" line in CaseFolding.txt has a unique pattern and corresponding dedicated documentation in Section 4.2.10 *@missing Conventions* in UAX #44. Deleting the "`@missing`" line in CaseFolding.txt would eliminate both the duplicate and the "`@missing`" syntactic pattern #4.

2.2. "`@missing`" directive in DerivedNumericValues.txt

The "`@missing`" directive in DerivedNumericValues.txt is one of the two out of a total of 62 "`@missing`" directives across the UCD files which list three default property values per line (the other file being SpecialCasing.txt) [1]. Deleting the "`@missing`" line in DerivedNumericValues.txt would eliminate another duplicate. If the "`@missing`" directive in SpecialCasing.txt is also moved (as proposed in the next section), then the documentation in UAX #44 would be further reduced by eliminating the "`@missing`" pattern #2.

3. "`@missing`" directive in SpecialCasing.txt

The full case-mapping properties Lowercase_Mapping, Titlecase_Mapping, and Uppercase_Mapping (all non-binary) have a combined "`@missing`" directive in SpecialCasing.txt:

```
# @missing: 0000..10FFFF; <slc>; <stc>; <suc>;
```

The format of that line is unusual for multiple reasons:

- It is one of the two out of a total of 62 "`@missing`" directives across the UCD files which list three default property values per line (the other file being DerivedNumericValues.txt) [1]
- It uses placeholder tags that are not used in any other "`@missing`" line, which therefore require dedicated documentation in Section 4.2.10 *@missing Conventions* in UAX #44
- It has a trailing semicolon (to match the line format of SpecialCasing.txt), also unlike other "`@missing`" lines

The "`@missing`" directive in SpecialCasing.txt is neither a duplicate nor an absent "`@missing`" issue. However, given that PropertyValueAliases.txt already includes "`@missing`" directives for the simple case-mapping properties (Simple_Lowercase_Mapping etc.) and given the unusual "`@missing`" line format and its dedicated documentation in UAX #44, it would be an improvement to both the UCD file data and the UAX #44 documentation to remove the "`@missing`" directive from SpecialCasing.txt and add its three equivalents (for Lowercase_Mapping etc.) to PropertyValueAliases.txt.

4. Absent "`@missing`" directives for non-binary properties

To identify the properties with no corresponding "`@missing`" directives, one can survey the properties listed in Table 9 *Property Table* of UAX #44 and cross-reference them with a dump of all "`@missing`" directives across the UCD files. An "`@missing`" directive is not absent if it appears in at least one UCD file, specific or primary, derived, or PropertyValueAliases.txt. For example, the properties Joining_Type and Joining_Group are not absent because they have "`@missing`" lines in derived files, even though there is no "`@missing`" directive in the primary file for those two properties, ArabicShaping.txt. Binary properties are excluded from this search because, according to Section 4.2.9 *Default Values* in UAX #44, the default value of binary properties is No (False) and is always omitted from the UCD files.

The comparison above returns a single non-binary property which does not have an "`@missing`" directive defined anywhere: General_Category. The General_Category property does not have an "`@missing`" directive in PropertyValueAliases.txt in UCD 6.3 and it cannot have one in UnicodeData.txt because of the format of that file. The fix for that problem would be to add an "`@missing`" directive for General_Category in PropertyValueAliases.txt.

5. Recommendations

To complete the coverage of "`@missing`" directives for non-binary properties, the proposal is to add an "`@missing`" directive for General_Category to PropertyValueAliases.txt.

Additionally, to eliminate those duplicate "`@missing`" directives which also have unusual formats compared to the majority, the proposal is to delete the "`@missing`" directives in CaseFolding.txt, DerivedNumericValues.txt and SpecialCasing.txt, and add to PropertyValueAliases.txt the three equivalents of the "`@missing`" directive from SpecialCasing.txt:

```
# Lowercase_Mapping (lc)
# @missing: 0000..10FFFF; Lowercase_Mapping; <code point>
# Titlecase_Mapping (tc)
# @missing: 0000..10FFFF; Titlecase_Mapping; <code point>
# Uppercase_Mapping (uc)
# @missing: 0000..10FFFF; Uppercase_Mapping; <code point>
```

The latter recommendation would also allow to simplify the documentation in UAX #44, by:

- Discarding two of the four "`@missing`" syntactic patterns documented in Section 4.2.10 *@missing Conventions* (in both body text and table), which in turn simplifies parsers
- Eliminating the irregular "`@missing`" patterns that have the special placeholder tags `<slc>`, `<stc>`, `<suc>`, and the status field with value C
- Eliminating the indirect way in which the defaults for the full case-mapping properties (`lc`, `tc`, `uc`) are defined, in SpecialCasing.txt, in terms of the values of the simple case-mapping properties (`slc`, `stc`, `suc`), which can be elusive to end-users or more difficult to implement.

6. References

[1] Ken Whistler and Laurențiu Iancu, *Inconsistencies between "`@missing`" directives in UCD files*, L2/13-169, 2013-07-29, http://www.unicode.org/L2/L2013/13169-at-missing-dir.pdf.

[2] The Unicode Consortium, *Draft minutes of UTC meeting #136*, revised 2013-10-15, http://www.unicode.org/L2/L2013/13132.htm, containing

Consensus 136-C4: Remove the "`@missing`" line from BidiBrackets.txt; add "`@missing`" for Bidi_Paired_Bracket_Type to PropertyValueAliases.txt; add documentation to UAX #44 [...]

Action Item 136-A20 for Laurențiu Iancu: Produce a list of the duplicate and missing (non-binary) "`@missing`"s.

[3] Mark Davis, Laurențiu Iancu and Ken Whistler, Editors, *Unicode Standard Annex #44, Unicode Character Database*, an integral part of *The Unicode Standard*, Version 6.3.0, 2013-09-25, http://www.unicode.org/reports/tr44/tr44-12.html.