L2/17-427

Re: Reducing singleton Joining_Group values
From: Mark Davis
Date: 2017-12-14

The Joining_Group property is specified in ArabicShaping.txt (for better formatting, see DerivedJoiningGroup.txt). Historically, I originated the property, whose purpose is to indicate when two or more characters are known to have the same "letter skeleton". The documentation, however, has never fully been spelled out: see Arabic Joining Groups for the little we have. It was intended as a guide to font developers, not a property that would be used at runtime — and that's how it is typically used.

Note that whether or not the character has joining **behavior** is determined by the Joining_Type, not by the Joining_Group.

Joining_Group was originally designed for Arabic, but extended afterwards to other scripts that have similar behavior. After the initial work, the subsequent values have been assigned in roughly the following way: No_Joining_Group used for characters that that are not in a cursive script, or for cursive scripts for which joining group analysis has not been done; and distinct Joining_Group property values for characters where joining group analysis has been done — even "singleton" values that don't have the same letter skeleton as another value.

Examples from Unicode 10.0:

`10AC1; MANICHAEAN BETH; D; MANICHAEAN_BETH` `10AC2; MANICHAEAN BETH WITH 2 DOTS ABOVE; D; MANICHAEAN_BETH`	2 characters with the same letter skeleton
`1820; MONGOLIAN A; D; No_Joining_Group`	implicit singleton (no joining group analysis)
`10AC0; MANICHAEAN ALEPH; D; MANICHAEAN_ALEPH`	distinct singleton (joining group analysis)

However, the distinct singleton groups are unnecessary, since the purpose of the property is to show when characters share the same letter skeleton — with another character — according to best available information. Assigning a distinct singleton value makes the property assignment more complicated, and unnecessarily multiplies the number of property values. There are **326** explicit No_Joining_Group values in Unicode 10.0 (such as U+1820), so they far outnumber the distinct-singleton cases (**64**).

> **Proposal:** *We add no new distinct-singleton values after Unicode 10.0.*
> This affects no property values from Unicode 10.0. But removes **54** *draft* explicit singleton property values from Unicode 11.0. The U11.0 draft characters having explicit singleton property values are changed to have No_Joining_Group.
>
> The Unicode Standard text is also updated to explain the situation. Here is rough draft text:
>
> "The No_Joining_Group value indicates that the character does not share a cursive (joining) letter skeleton with another character, according to best available information. Historically, some Joining_Group values have been assigned, such as Hamza_On_Heh_Goal, even though they do not share a letter skeleton with another character.
>
> Joining_Group values may be changed in successive versions, as more information becomes

available. For example, if it is later found that two characters with No_Joining_Group share the same letter skeleton, a new Joining_Group value will be created for them and assigned to them."

Alternative: Add new value Singleton to use when a character currently shares no letter skeleton with another character. The U11.0 draft characters having explicit singleton property values are changed to have Singleton. The rough draft TUS text above would take some wordsmithing for this option.

Implementations shouldn't be affected by this change (except a reduction in future property data and API constants).

Distinct-Singleton Joining_Group values

The following are the distinct-singleton JG property values in the draft Unicode 11.0 UCD.

Age	Count	List
11.0	54	Hanifi_Rohingya_A, Hanifi_Rohingya_Ba, Hanifi_Rohingya_Ca, Hanifi_Rohingya_Da, Hanifi_Rohingya_Dda, Hanifi_Rohingya_Ga, Hanifi_Rohingya_Ha, Hanifi_Rohingya_Ja, Hanifi_Rohingya_Ka, Hanifi_Rohingya_Kha, Hanifi_Rohingya_Kinna_Wa, Hanifi_Rohingya_La, Hanifi_Rohingya_Ma, Hanifi_Rohingya_Na, Hanifi_Rohingya_Nga, Hanifi_Rohingya_Nya, Hanifi_Rohingya_Ra, Hanifi_Rohingya_Rra, Hanifi_Rohingya_Sa, Hanifi_Rohingya_Sakin, Hanifi_Rohingya_Sha, Hanifi_Rohingya_Ta, Hanifi_Rohingya_Tta, Hanifi_Rohingya_Vowel_Sign_A, Hanifi_Rohingya_Vowel_Sign_O, Hanifi_Rohingya_Vowel_Sign_U, Hanifi_Rohingya_Wa, Hanifi_Rohingya_Ya, Hanifi_Rohingya_Za, Sogdian_Aleph, Sogdian_Ayin, Sogdian_Beth, Sogdian_Feth, Sogdian_Gimel, Sogdian_He, Sogdian_Heth, Sogdian_Kaph, Sogdian_Lamedh, Sogdian_Lesh, Sogdian_Mem, Sogdian_Nun, Sogdian_One, Sogdian_One_Hundred, Sogdian_Pe, Sogdian_Resh_Ayin, Sogdian_Sadhe, Sogdian_Samekh, Sogdian_Shin, Sogdian_Taw, Sogdian_Ten, Sogdian_Twenty, Sogdian_Waw, Sogdian_Yodh, Sogdian_Zayin
10.0	11	Malayalam_Bha, Malayalam_Ja, Malayalam_Lla, Malayalam_Llla, Malayalam_Nga, Malayalam_Nna, Malayalam_Nnna, Malayalam_Nya, Malayalam_Ra, Malayalam_Ssa, Malayalam_Tta
9.0	3	African_Feh, African_Noon, African_Qaf
7.0	21	Manichaean_Aleph, Manichaean_Daleth, Manichaean_Dhamedh, Manichaean_Five, Manichaean_Heth, Manichaean_Hundred, Manichaean_Lamedh, Manichaean_Mem, Manichaean_Nun, Manichaean_One, Manichaean_Resh, Manichaean_Sadhe, Manichaean_Samekh, Manichaean_Taw, Manichaean_Ten, Manichaean_Teth, Manichaean_Thamedh, Manichaean_Twenty, Manichaean_Waw, Manichaean_Yodh, Straight_Waw
6.1	1	Rohingya_Yeh
4.0	3	Fe, Khaph, Zhain
3.0	20	Alaph, E, Final_Semkath, He, Heth, Kaph, Lamadh, Mim, Nun, Pe, Qaph, Reversed_Pe, Sadhe, Semkath, Shin, Syriac_Waw, Taw, Yudh, Yudh_He, Zain
1.1	5	Heh, Nya, Swash_Kaf, Hamza_On_Heh_Goal, Yeh_With_Tail
total	**118**	

Non-Singleton Joining_Group values

The following 37 Joining_Group values each have more than one character sharing the property value.

ND	37	Ain, Alef, Beh, Beth, Burushaski_Yeh_Barree, Dal, Dalath_Rish, Farsi_Yeh, Feh, Gaf, Gamal, Hah, Hanifi_Rohingya_Kinna_Ya, Hanifi_Rohingya_Pa, Heh_Goal, Kaf, Knotted_Heh, Lam, Manichaean_Ayin, Manichaean_Beth, Manichaean_Gimel, Manichaean_Kaph, Manichaean_Pe, Manichaean_Qoph, Manichaean_Zayin, Meem, Noon, Qaf, Reh, Sad, Seen, Tah, Teh_Marbuta, Teth, Waw, Yeh, Yeh_Barree

Explicit No_Joining_Group values

The following 326 characters have explicit Joining_Group=No_Joining_Group and have joining behavior (that is, Joining_Type ≠ T, U):

Age	Codepoint(s)	Count	Type	Name
1.1	U+0640	1	Join_Causing	ARABIC TATWEEL
1.1	U+200D	1	Join_Causing	ZERO WIDTH JOINER
3.0	U+1807	1	Dual_Joining	MONGOLIAN SIBE SYLLABLE BOUNDARY MARKER
3.0	U+180A	1	Join_Causing	MONGOLIAN NIRUGU
3.0	U+1820 ..1877	88	Dual_Joining..	MONGOLIAN LETTER A.. MONGOLIAN LETTER MANCHU ZHA
3.0	U+1887 ..18A8	34	Dual_Joining..	MONGOLIAN LETTER ALI GALI A.. MONGOLIAN LETTER MANCHU ALI GALI BHA
5.0	U+07CA ..07EA	33	Dual_Joining..	NKO LETTER A.. NKO LETTER JONA RA
5.0	U+07FA	1	Join_Causing	NKO LAJANYALAN
5.0	U+A840 ..A872	51	Dual_Joining..	PHAGS-PA LETTER KA.. PHAGS-PA SUPERFIXED LETTER RA
5.1	U+18AA	1	Dual_Joining	MONGOLIAN LETTER MANCHU ALI GALI LHA
6.0	U+0840 ..0855	22	Right_Joining..	MANDAIC LETTER HALQA.. MANDAIC LETTER AT
7.0	U+10B80 ..10B91	18	Dual_Joining..	PSALTER PAHLAVI LETTER ALEPH.. PSALTER PAHLAVI LETTER TAW
7.0	U+10BA9 ..10BAE	6	Right_Joining..	PSALTER PAHLAVI NUMBER ONE.. PSALTER PAHLAVI NUMBER TWENTY
9.0	U+1E900 ..1E943	68	Dual_Joining..	ADLAM CAPITAL LETTER ALIF.. ADLAM SMALL LETTER SHA
	Total	**326**		