Subject: **Dependency Graph for Segmentation**
From: **Mark Davis (Google, Inc.), Laurențiu Iancu (Microsoft Corporation)**
Date: **2016-11-01**

We had an action of identify those other Unicode Properties that the segmentation algorithms explicitly depend on. That is, a change to those other properties can cause the results of segmentation to change. In some cases, the results only depend on a few particular values of the property, not all or most of the values.

148	**A001**	**Mark Davis, Laurentiu Iancu**		**Create a dependency graph for derived properties where the derivation is either explicit or implicit, for the November 2016 UTC meeting. The initial focus is on the four segmentation properties.**

There are other (implicit) dependencies, but the explicit ones are the focus of this doc. (An example of an implicit dependency is the GCB=SM. The excluded marks are gc=Mc from SE Asian scripts (Myanmar, Tai Tham etc.) which are post-base vowels or tone marks (approx. InSC=Vowel_Dependent, InPC=Right), but are ultimately hand-tuned based on feedback from experts.)

Grapheme Cluster Break

1. General_Category=Unassigned, Control, Format, Surrogate, Line_Separator, Paragraph_Separator, Spacing_Mark, Nonspacing_Mark, Enclosing_Mark
2. Default_Ignorable_Code_Point
3. Grapheme_Extend
4. Regional_Indicator*
5. Hangul_Syllable_Type
6. Indic_Syllabic_Category = Consonant_Preceding_Repha, Consonant_Prefixed
7. Prepended_Concatenation_Mark
8. Emoji_Modifier_Base
9. Emoji_Modifer
10. Glue_After_Zwj*
11. E_Base_GAZ*

** The items marked with a star don't currently have UCD properties. There are, however, proposals to have UCD properties for them.*

Word Break

1. General_Category=Other_Letter, Spacing_Mark, Format, Connector_Punctuation
2. Script=Hebrew, Katakana, Hiragana
3. Grapheme_Extend
4. Regional_Indicator*
5. Line_Break=Complex_Context, Infix_Numeric, Numeric
6. Alphabetic
7. Ideographic
8. Emoji_Modifier_Base
9. Emoji_Modifer
10. Glue_After_Zwj*
11. E_Base_GAZ*

Sentence Break

1. Grapheme_Extend
2. General_Category=Spacing_Mark, Format, Titlecase_Letter, Open_Punctuation, Close_Punctuation
3. White_Space
4. Lowercase
5. Uppercase
6. Alphabetic
7. Line_Break=Numeric, Quotation
8. Sentence_Terminal

Line Break

Line_Break is a primary property whose values are not expressed in terms of of other properties. They are assigned by some unwritten heuristics, but not formally. We do have that, for instance, lb=EM is equivalent with Emoji_Modifier=Yes (because it would make no sense otherwise), but there is no formal dependency.

The original assignments had started from some rules, but by now we have probably accumulated irregularities or deviations from those rules. For example, see http://www.unicode.org/L2/L1999/99179.pdf

That being said, were we to have explicit dependencies (with perhaps exceptions), they would probably be the following:

1. General_Category
2. Hangul_Syllable_Type
3. Emoji_Modifier_Base
4. Emoji_Modifer
5. Regional_Indicator
6. East_Asian_Width