- Draft -- for UTC/L2 Consideration only --

L2/98-053

February 22, 1998

Analysis and Proposed Changes in General Category Property Assignments for Quotation Characters

Submitted by: Asmus Freytag,
Version 0.02

Introduction

In the Unicode Character Database a generic property called Ps (opening punctuation) and Pe (closing punctuation) are assigned. This was done in an effort to provide information that is needed for implementing a kinsoku-style form of linebreaking. (See also Chapter 5, section). Upon closer analysis, the current property assignments are problematic, as they conflict with standard European usage in several languages.

While traditional Kinsoku style linebreaking is specific to East Asian typography, experience has shown that it is possible to use a generalized mechanism to handle both the Far Eastern and the Western case in a common manner. Making the Unicode *general category* character property consistent, makes it a more useful input to implementers of such systems.

I will present here only the analysis of quotation characters. Analysis of other punctuation characters' property assignment is in progress.

Background

Kinsoku style linebreaking

Kinsoku style linebreaking is needed where spaces are not used to separate words, but instead, line break opportunities can occur *in principle* after any character, but are prohibited in some contexts.

Opening and Closing

The simplest form of context is to divide punctuation into opening and closing (or leading and following) and prohibit a break after a closing or before an opening punctuation. Thinking of punctuation as operators, this ensures that operator and operand are on the same line.

Global **robustness**

It is a strong requirement that any such scheme be robust in the presence of any Unicode character. Equally motivated is the desire to be able to have such an algorithm act on plain text (that is without constant recourse to language tagging).

Language based usage of quotation characters

Czech, German, and Slovak use the LOW-9 style of quotation mark for opening instead of the standard open quotes. (The Unicode Standard claims in a comment in the names list "usually opening, sometimes closing" for these. Since, no language information is given, this is assumed to be a mistake and this comment should be removed).

Czech, German, and Slovak use the LEFT QUOTATION MARK style of quotation mark for closing instead of the more common RIGHT QUOTATION MARK forms. (The Unicode Standard calls the common forms "preferred". This is unnecessarily prejudicial toward local usage).

Danish, Finnish, and Swedish use the *same* RIGHT QUOTATION MARK character for both oppugn and closing quotation character. This is true for both office automation usage as well as books (which sometimes use the *guillemets*

or RIGHT POINTING DOUBLE ANGLE quotation marks for both opening and closing).

Hungarian and Polish follow the Scandinavian languages for the single quote and German for the double quotes.

French, Greek Russian and Slovenian use the *guillemets* , but Slovenian uses the opposite directionality. (Of these languages at least French inserts space between text and quotation marks, if NBSP are not used, most line breaking algorithms will fail.)

Consequences for semantics

The semantics of U+00AB, U+00BB (double *guillemets*) and U+201D (right double quotation) are context dependent. If adjacent to a space on the left they act as opening, otherwise as closing quotation. In the French case, NBSP must be used to distinguish the space that is enclosed between quotation mark and text.

The semantics of U+201A and U+201B (low-9 quotation marks) appear to be always oppugn, contrary to the statement in the book.

Special role of RIGHT SINGLE QUOTATION MARK as Apostrophe

It is pervasive practice to use RIGHT SINGLE QUOTATION MARK as apostrophe character. The semantics of U+2019 are therefore context dependent. If surrounded by text, it behaves as an in text punctuation character (does not separate words or lines). If bordered by space on one side, it is a quotation character.

Mistaken property in the book

U+301F (low double prime) is a closing punctuation character, contrary to it's assignment of Ps. This is a mistake in the book.

Proposal to handle case where Pe and Ps are ambiguous

Since some quotation characters do not posses an unambiguous Ps or Pe property, it is proposed to introduce the weaker Qs and Qe, where Qs and Qe, unlike Ps and Pe are ambiguous in some contexts, with the ambiguity resolved by taking into account adjacent space characters.

Proposal 1: Split the existing Ps and Pe properties into a new set of Pe, Ps, Qe and Qs, with the understanding that Qs may be treated as equivalent to Ps and Qe may be treated as equivalent to Pe, but that implementations that wish to have more sophisticated handling of quotation characters may treat Qs and Qe as ambiguous or context dependent.

Proposal 2: Split Pe further by introducing a property Qa and assigning it to U+2019, with the understanding that Qa may be treated as equivalent to Pe, but hat implementations that wish to have more sophisticated handling of the apostrophe may treat Qa as context dependent.

Proposed property assignments by character

00AB	Qs	1	
00BB	Qs	1	
2018	Qs	1	change 'preferred' wording
2019	Qa	2	change 'preferred' wording add comment in nameslist that is is used for apostro
201A	Ps	unchanged	remove comment in names list, or provide example
201B	Qs	1	remove comment in names list, or provide exampl
201C	Qs	1	change 'preferred' wording

201D	Qs	1	change 'preferred' wording
201E	Ps	unchanged	
201F	Qs	1	
301D	Ps	unchanged	
301E	Pe	unchanged	
301F	Pe	erratum	

All other Ps and Pe characters, independent of possible use as quotes have consistent and unambiguous semantics and no changes are proposed for them here.