Title: Finalized Mapping between Characters of ISO 5426 and ISO/IEC 10646-1 (UCS)
Source: The Research Libraries Group, Inc.
Status: L2 Member Contribution
References: ISO/TC46/SC4/WG1 N 240, ISO/TC46/SC4/WG1 N 227 (revised 1998-05-29) and ISO/IEC JTC1/SC2 N3125 (Application for Registration No. 212) subsequently revised as replacement version of ISO-IR 53
Authors: Joan M. Aliprand
Action: For information
Date: 2000-07-02

1. Background

On May 9, 2000, ISO/TC 46/SC 4/WG 1 reviewed documents that compared mappings of the characters of TC 46 character set standards and character sets which were sources for these standards to ISO/IEC 10646. Based on these comparative analyses, WG1 agreed on final mappings for eight of the TC 46 character set standards.

This document is a modification of *Analysis of Proposed Mapping between Characters of ISO 5426 and ISO/IEC 10646-1 (UCS)* prepared for the WG1 meeting (ISO/TC46/SC4/WG1 N 240) and available at http://www.niso.org.

2. Essential Requirement for Mapping

Many of the characters in ISO 5426 also occur in ANSI/NISO Z39.47-1993, *Extended Latin Alphabet Coded Character Set for Bibliographic Use* (NISO Press, 1993), known by the abbreviation *ANSEL*. ANSEL is based on the original character sets for bibliographic use published by the Library of Congress in 1968, and is specified as the default extended Latin set for the MARC 21 format for library data (published by the Library of Congress). ISO 5426 is specified as the default G1 set for the UNIMARC format (published by the International Federation of Library Associations). **Characters common to these two standards should have identical mappings.**

3. Mapping of Characters

The following table shows the final WG1 decisions and (for information) the two key mapping efforts:

- N 227 is the mapping developed by ISO/TC46/SC4/WG1 (revised 1998-05-29)
- LC/ALA: Unicode/UCS mappings were established by the Library of Congress and the MARBI Committee of the American Library Association for the MARC 21 Extended Latin character set (equivalent to ANSEL). The mappings are available on the Library of Congress' Web site (http://lcweb.loc.gov/marc/specifications/speccharlatin.html). This column contains a value only when there are differences between the mappings. "n/e" = no equivalent in ANSEL to this ISO 5426 character.

ISO 5426		ANSEL equivalent		Mappings			Status
Code	**Character name**	**Code**	**Character name**	**N 227**	**LC/ALA**	**Final**	
2/0	Space	2/0	Reserved	0020		0020	
2/1	Inverted exclamation mark	4/6	inverted exclamation mark	00A1		00A1	
2/2	Double open quote			201E	n/e	201E	
2/3	Pound sign	3/9	British pound	00A3		00A3	
2/4	Dollar sign			0024	Note 1	0024	
2/5	Yen sign			00A5	n/e	00A5	
2/6	Single dagger			2020	n/e	2020	
2/7	Section or paragraph mark			00B6	n/e	00A7	Correction
2/8	Prime			2033	n/e	2032	Correction
2/9	Single open quote			2018	n/e	2018	
2/10	Double open quote			201C	n/e	201C	
2/11	Angle quote			00AB	n/e	00AB	
2/12	Musical flat	2/9	musical flat	266D		266D	
2/13	Copyright	4/3	copyright mark	00A9		00A9	
2/14	Sound recording copyright statement	4/2	phono copyright mark	2117		2117	
2/15	Registered trade mark	2/10	patent mark	2122	00AE	00AE	Correction

ISO 5426		ANSEL equivalent		Mappings			Status
Code	**Character name**	**Code**	**Character name**	**N 227**	**LC/ALA**	**Final**	
3/0	Ayn	3/0	‘ayn	02BD	02BF	02BB	Accepted
3/1	Alif/Hamzah	2/14	alif	02BE	02BE	02BC	Accepted
3/2	Single open quote			201A	n/e	201A	
3/3-3/5	*Unassigned*						
3/6	Double dagger			2021	n/e	2021	
3/7	Middle dot	2/8	middle dot	00B7		00B7	
3/8	Double prime			2033	n/e	2033	
3/9	Single close quote			2019	n/e	2019	Comment
3/10	Double close quote			201D	n/e	201D	
3/11	Angle quote			00BB	n/e	00BB	
3/12	Musical sharp	4/4	musical sharp	266F		266F	
3/13	Mjagkij znak	2/7	soft sign (miagkii znak)	02B9		02B9	
3/14	Tverdyj znak	3/7	hard sign (tverdyi znak)	02BA		02BA	
3/15	Inverted question mark	4/5	inverted question mark	00BF		00BF	
4/0	Low rising tone mark	6/0	low rising tone mark	0309		0309	Comment
4/1	Grave accent	6/1	grave accent	0300		0300	
4/2	Acute accent	6/2	acute accent	0301		0301	
4/3	Circumflex accent	6/3	circumflex accent	0302		0302	
4/4	Tilde	6/4	tilde	0303		0303	
4/5	Macron	6/5	macron	0304		0304	
4/6	Breve	6/6	breve	0306		0306	
4/7	Dot above	6/7	dot above	0307		0307	
4/8	Trema, Diaeresis			0308	Note 2	0308	Comment
4/9	Umlaut			0308	Note 2	0308	Comment

ISO 5426		ANSEL equivalent		Mappings			Status
Code	**Character name**	**Code**	**Character name**	**N 227**	**LC/ALA**	**Final**	
4/10	Circle above	6/10	circle above (angstrom)	030A		030A	
4/11	High comma off centre	6/13	high comma, off center	0315		0315	
4/12	High inverted comma centred	7/14	high comma, centered	0312	0313	0312	Affirmed
4/13	Double acute accent	6/14	double acute accent	030B		030B	
4/14	Horn			031B	n/e	031B	
4/15	Hacek	6/9	hacek (caron)	030C		030C	
5/0	Cedilla	7/0	cedilla	0327		0327	
5/1	Rude	7/8	right cedilla	031C	031C	031C	Comment
5/2	Hook to left	7/7	left hoof [sic]	0326	0326	0326	Comment
5/3	Hook to right or ogonek	7/1	right hook	0328		0328	
5/4	Circle below	7/4	circle below	0325		0325	
5/5	Half circle below	7/9	half circle below (upadhmaniya)	032E		032E	
5/6	Dot below	7/2	dot below	0323		0323	
5/7	Double dot below	7/3	double dot below	0324		0324	
5/8	Underline	7/6	underscore	0332		0332	
5/9	Double underline	7/5	double underscore	0333		0333	
5/10	Vertical bar			0329	n/e	0329	
5/11	Circumflex below			032D	n/e	032D	
5/12	*Unassigned*						
5/13	Left half of ligature sign and of double tilde			FE20	Note 3	FE20	Comment
5/14	Right half of ligature sign	6/12	ligature, right half	FE21	FE21	FE21	Comment
5/15	Right half of double tilde	7/11	double tilde, right half	FE23	FE23	FE23	Comment

ISO 5426		ANSEL equivalent		Mappings			Status
Code	**Character name**	**Code**	**Character name**	**N 227**	**LC/ALA**	**Final**	
6/0	*Unassigned*						
6/1	Ligature AE – Capital letter	2/5	ligature AE – uppercase	00C6		00C6	
6/2	Slash D– Capital letter	2/3	slash D– uppercase	0110		0110	Comment
6/3-6/5	*Unassigned*						
6/6	Digraph IJ– Capital letter			0132	n/e	0132	
6/7	*Unassigned*						
6/8	Slash L – Capital letter	2/1	slash L – uppercase	0141		0141	
6/9	Slash O– Capital letter	2/2	slash O– uppercase	00D8		00D8	
6/10	Ligature OE– Capital letter	2/6	ligature OE– uppercase	0152		0152	
6/11	*Unassigned*						
6/12	Thorn– Capital letter	2/4	thorn– uppercase	00DE		00DE	
6/13-7/0	*Unassigned*						
7/1	Ligature ae – Small letter	3/5	ligature ae – lowercase	00E6		00E6	
7/2	Slash d – Small letter	3/3	slash d– lowercase	0111		0111	
7/3	eth – Small letter	3/10	eth	00F0		00F0	
7/4	*Unassigned*						
7/5	Dotless i – Small letter	3/8	dotless i -- lowercase	0131		0131	
7/6	Ligature ij – Small letter			0133	n/e	0133	
7/7	*Unassigned*						
7/8	Slash l – Small letter	3/1	slash l – lowercase	0142		0142	
7/9	Slash o – Small letter	3/2	slash o – lowercase	00F8		00F8	
7/10	Ligature oe – Small letter	3/6	ligature oe – lowercase	0153		0153	

ISO 5426		ANSEL equivalent		Mappings			Status
Code	**Character name**	**Code**	**Character name**	**N 227**	**LC/ALA**	**Final**	
7/11	Ligature sz – Small letter			00DF	n/e	00DF	
7/12	Thorn – Small letter	3/4	thorn– lowercase	00FE		00FE	
7/13-7/14	*Unassigned*						

4. Notes on ANSEL Characters

Note 1: The dollar sign is in ASCII, used as the Basic Latin set with ANSEL as the Extended Latin set.

Note 2: ANSEL contains a single unified character used as both diaeresis/trema and umlaut.

Note 3: ANSEL has separate characters for the left half of the ligature sign and the left half of the double tilde.

5. WG1 Action: Correction of Errors

2/7 SECTION OR PARAGRAPH MARK

The mapping to U+00B6 PILCROW SIGN "¶" (in both N227 and R53) is wrong. The graphic at position 2/7 is the "double S" symbol §. Therefore the correct mapping is to U+00A7 SECTION SIGN "§".

2/8 PRIME

The graphic at position 2/8 is the single stroke *prime* character. The proposed mapping is to U+2033 DOUBLE PRIME (in both N227 and R53). The correct mapping is to U+2032 PRIME.

2/15 REGISTERED TRADE MARK

The mapping to U+2122 TRADE MARK SIGN (in both N227 and R53) is wrong. The graphic at position 2/15 is a capital R in a circle, and the character's name is *REGISTERED trade mark*. Therefore the correct mapping is to U+00AE REGISTERED SIGN.

6. WG1 Action: Affirmation of N227 Mapping

4/12 HIGH INVERTED COMMA CENTRED

In ISO 5426, this character is annotated "used in Latvian." Possible mappings are to U+0312 COMBINING TURNED COMMA ABOVE and U+0313 COMBINING COMMA ABOVE. The Unicode Standard notes *cedilla above* as an alternative name for U+0312 with use in Latvian.

ISO 5426:1983 (E) shows a comma-like glyph at position 4/12 in the code table (Clause 4 on page 2). However, in the list of diacritics (Clause 5.2 on page 4), the graphic and example of use show an inverted comma. Given the name of this character, and its designated use in Latvian (where the inverted comma form is conventionally used), it is assumed that the glyph in the code table is incorrect, that is, that ISO 5426 and ANSEL differ with respect to this character. The corresponding ANSEL character (in ANSI/NISO Z39.47-1993) is "high comma, centered" encoded at 7/14. It is consistently shown with a comma-like glyph, and is used only in Latvian.

ISO 5426 and ANSEL use different glyphs to represent this character, and the mappings to U+0312 or U+0313 preserve the differences. The recommended mapping in document L2/98-285 to U+03013 was "for consistency with character 07/14 in [SC2] N3138" (the proposed registration for ANSEL). Although this is functionally the same character in ISO 5426 and ANSEL, the long-established glyphic differences dictate separate mappings (contrary to the requirement in Section 2).

The comma above is a typographical convention substituted for the orthographic cedilla. U+0123 LATIN SMALL LETTER G WITH CEDILLA is rendered as a lower case g with inverted comma above. Its canonical decomposition is U+0067 U+0327 (lower case g, combining cedilla). Neither the ISO 5426 encoding not the ANSEL encoding (when converted to Unicode/UCS equivalents) will match the canonical decomposition.

7. WG1 Action: Acceptance of Proposed Mappings

The following mappings were accepted for coordination with the Library of Congress' mappings for ANSEL characters:

3/0 AYN

The options for mapping this character are U+02BD MODIFIER LETTER REVERSED COMMA (annotated as "weak aspiration" and "spacing clone of Greek rough breathing mark" in the Unicode Standard), U+02BF MODIFIER LETTER LEFT HALF RING ("transliteration of Arabic *ain* (voiced pharyngeal fricative)"), or U+02BB MODIFIER LETTER TURNED COMMA ("typographical alternate for 02BD or 02BF").

In document L2/98-285, NCITS/L2 recommended mapping this character to U+02BF (taking the designation of U+02BD in document SC2 N3125 to be a typographical error). CHASE also maps to U+02BF, as does the Library of Congress for the corresponding character in ANSEL.

The essential issue is whether to map based on character shape or character name. This character is not used exclusively to represent the letter *ain* in transliteration, but serves additional purposes:

- The example of use for this character in ISO 5426 is romanized Chinese.
- For the equivalent character in ANSEL, Table A1 shows its use in Hawaiian (which is written in Latin script). Table B2 shows its use for transliteration (according to ALA-LC romanization) of many languages, including Amharic, Armenian, Georgian, and Lao.

This character has diverse uses; it is not used exclusively for the transliteration of letters in Semitic language alphabets. Therefore, it is recommended that mapping be based on shape rather than name. That is, 3/0 Ayn should be mapped to U+20BB MODIFIER LETTER TURNED COMMA.

3/1 ALIF/HAMZAH

The options for mapping this character are U+02BC MODIFIER LETTER APOSTROPHE or U+02BE MODIFIER LETTER RIGHT HALF RING. U+02BE is annotated "transliteration of Arabic *hamzah* (glottal stop.)" U+02BC is annotated: "glottal stop, glottalization, ejective," "spacing clone of Greek smooth breathing mark," and "many languages use this as a letter of their alphabets."

The essential issue is whether to map based on character shape or character name. This character is not used exclusively to represent *alif* or *hamzah* in transliteration, but serves additional purposes:

- The example of use for this character in ISO 5426 is romanized Japanese.

- For the equivalent character in ANSEL, Table B2 shows its use in Indonesian and Turkish, as well as for the ALA-LC romanization of some other languages, including Burmese, Khmer and Korean. ("Turkish" may not refer to modern Turkish written in Latin script, but to Ottoman Turkish which was written in Arabic script.)

This character has diverse uses; it is not used exclusively for the transliteration of letters in Semitic language alphabets. Therefore, it is recommended that mapping be based on shape rather than name. That is, 3/1 Alif/Hamzah should be mapped to U+20BC MODIFIER LETTER APOSTROPHE.

8. Comments on Specific Characters (for information)

3/9 SINGLE CLOSE QUOTE

The mapping to U+2019 RIGHT SINGLE QUOTATION MARK is correct. The CHASE mapping to U+0027 APOSTROPHE is incorrect, but this mapping may reflect a problem with legacy data at the British Library since the name of the character is annotated "(misused as apostrophe)".

4/0 LOW RISING TONE MARK

ISO 5426 designates this character as a diacritic "used with other characters' and notes that it is "used in Vietnamese." The mapping to U+0309 COMBINING HOOK ABOVE, a combing diacritical mark that functions as a "Vietnamese tone mark," is therefore correct. The CHASE mapping to U+02CF MODIFIER LETTER LOW ACUTE ACCENT is wrong. U+02CF is annotated "low rising tone" in the Unicode Standard, but it is a spacing character.

4/8 TREMA, DIAERESIS & 4/9 UMLAUT

Library practice differs on the encoding of trema, diaeresis, and umlaut. ISO 5426 has separate characters for Trema, Diaeresis (at 4/8) and for Umlaut (at 4/9); ANSEL has a single unified character (as does ISO/IEC 10646 and the Unicode Standard). The distinction made between the two characters is said to be to support German filing conventions.

The advantage of having two discrete characters is that German collation practice can be supported easily. The disadvantage of having two characters is that they can be confused in both data input and searching, thus obviating the benefit of the distinction. The two characters are unified via mapping, since both are mapped to U+0308 COMBINING DIAERESIS.

N227 aimed to preserve the distinction, but (through mapping) unified this character with the ISO 5426-2 character COMBINING LATIN SMALL LETTER E ABOVE. N227 maps character 4/8 Trema, Diaeresis to U+0308 COMBINING DIAERESIS and

character 4/9 Umlaut to a proposed new character, COMBINING LATIN SMALL LETTER E ABOVE. This is also the proposed mapping for the ISO 5426-2 character COMBINING LATIN SMALL LETTER E ABOVE (at position 46). This mapping to the same (proposed) character eliminates the ability to identify the source character set after data conversion.

5/1 RUDE

The character “rude” (“right cedilla” in MARC 21) is specious. Its only use (according to ANSEL and ISO 5426 – although one may draw on the other) is in the romanization of Thai (specifically, according to the ALA-LC Romanization Tables). Ms. Aliprand discovered that this character only assumed its cedilla-like form in later versions of the ALA-LC romanization table for Thai. In the first published version of this table, the mark was simply an arc, similar to U+031C, COMBINING LEFT HALF RING BELOW (annotated “IPA: open variety of vowel” in the Unicode Standard). The choice of U+031C to represent the rude/right cedilla was confirmed by a professor of Thai language.

5/2 HOOK TO LEFT

In ISO 5426, this character is annotated “used in Latvian, Romanian.” Because of this use, the most appropriate mapping is to U+0326 COMBINING COMMA BELOW (annotated as “variant of the following” [combining cedilla] in the Unicode Standard).

5/13 LEFT HALF OF LIGATURE SIGN AND OF DOUBLE TILDE

ISO 5426 unifies characters which are separately encoded in ANSEL. ISO/IEC 10646 and the Unicode Standard encode the four halves as well as “double diacritic” whole forms that extend over two base characters. U+FE20 COMBINING LIGATURE LEFT HALF rather than U+FE22 COMBINING DOUBLE TILDE LEFT HALF is chosen as the mapping in Registration 53, possibly because the ligature occurs more frequently than the double tilde in bibliographic data.

A more exact mapping method that uses the base character is given in the following section.

5/14 RIGHT HALF OF LIGATURE SIGN

The CHASE mapping to U+0360 COMBINING DOUBLE TILDE is incorrect. The correct mapping is to U+0361 COMBINING DOUBLE INVERTED BREVE (if the intent was to map the two halves, 5/13 plus 5/14, to the single whole form).

5/15 RIGHT HALF OF DOUBLE TILDE

The CHASE mapping to U+0361 COMBINING DOUBLE INVERTED BREVE is incorrect. The correct mapping is to U+0360 COMBINING DOUBLE TILDE (if the intent was to map the two halves, 5/13 plus 5/14, to the single whole form).

6/2 SLASH D – CAPITAL LETTER

Both ISO 5426 and ANSEL encode only one *D with stroke* "used in Croatian, Icelandic, etc." (annotation in ISO 5426). Possible mappings are to U+00D0 LATIN CAPITAL LETTER ETH ("Icelandic, Faroese, old English, IPA"), to U+0110 LATIN CAPITAL LETTER D WITH STROKE ("Croatian, Vietnamese, Lappish"), or to U+0189 LATIN CAPITAL LETTER AFRICAN D. (Note that ISO 6438 (African letters) includes a D with stroke at position 2/5.) U+0110 is preferred (because of frequency of occurrence?).

9. Options for Mapping Certain Characters

4/8 TREMA, DIAERESIS & 4/9 UMLAUT

These two characters are unified in this mapping. If the distinction between the characters must be preserved for a particular application, U+0308 should be used for one and a Private Use value for the other.

5/13 LEFT HALF OF LIGATURE SIGN AND OF DOUBLE TILDE

This character is mapped to U+FE20 COMBINING LIGATURE LEFT HALF. Two alternative mappings are possible: a more exact mapping that takes the base character into account, and one (described under the following characters) that maps to the whole form "double diacritic" character instead of to the compatibility "halves."

The left half of the double tilde is only correctly used with the letter "n" (upper or lower case) in the ligature "ng with tilde" of Tagalog (ANSEL standard, Table B1, p, 17). The left half of the ligature sign is used with various letters in transliterations of Slavic languages (Table B1, p. 18).

Rules for Mapping based on Base Character

Base character (follows *left half* in the source data)	**Map 5/13 to**
N|n	U+FE22
I|i|K|k|O with macron|o with macron|P|p|T|t|Z|z	U+FE20
Any other base character (*error condition*)	U+FE20

If any non-spacing marks occur between 5/13 and the base character (*error condition*) map the non-spacing marks according to the table in Section 3, and map 5/13 according to the table above.

5/14 RIGHT HALF OF LIGATURE SIGN

The CHASE project specifies use of the combining double form instead of the compatibility "halves" but does not gives details. To yield the character sequence specified for *Diacritics Positioned Over Two Base Characters* (*The Unicode Standard, Version 2.0*, p. 6-14), the source string from the left half of the ligature sign/double tilde sign through the base character of the right half of the ligature sign would have to be rearranged. (Non-spacing marks *precede* the base character in library data.) That is:

> When character 5/14 is encountered, map and reorder the string beginning with 5/13 through the base character of 5/14 as follows ("+" is used as a separator to distinguish characters in the result):
>
> <base character of 5/13> + U+0361 + <any other non-spacing marks preceding base character of 5/13> +<base character of 5/14> + <any non-spacing marks other than 5/14 preceding base character of 5/14>

Example: 5/13 I 5/14 A ➔ I U+0361 A

5/15 RIGHT HALF OF DOUBLE TILDE

The CHASE project specifies use of the combining double instead of the compatibility "halves" but does not give details. Mapping and rearrangement would be:

> When character 5/15 is encountered, map and reorder the string from 5/13 through the base character of 5/15 (normally *lowercase g*) as follows ("+" is used as a separator to distinguish characters in the result):
>
> <base character of 5/13> + U+0360 + <any other non-spacing marks preceding base character of 5/13> +<base character of 5/15> + <any non-spacing marks other than 5/15 preceding base character of 5/15>

Example: 5/13 n 5/15 g ➔ n U+0360 g

6/2 SLASH D – CAPITAL LETTER

This character is mapped to U+0110 LATIN CAPITAL LETTER D WITH STROKE. If more precise language-based mapping is needed, coded language information in the bibliographic record may be used to map 6/2 to the appropriate character: U+00D0 LATIN CAPITAL LETTER ETH, U+0110 LATIN CAPITAL LETTER D WITH STROKE, or U+0189 LATIN CAPITAL LETTER AFRICAN D.