L2/04-074

BT N 6530
(Draft Resolution BT C119/2001)
Issue date : 2001-10-31

Target Date : 2001-12-12

BT - TECHNICAL BOARD

1 TO DECIDE

---

2 SUBJECT: CEN/TC 304 - Publication of a CEN Report

---

3 BACKGROUND:

CEN/TC 304 "Information and communications technologies - European localization requirements" provided CMC with the enclosed document

> *"European ordering rules – Ordering for Latin, Greek, Cyrillic, Georgian and Armenian scripts"* (WI 00304031)

with the request to submit it to the CEN/BT for publication as a CEN Report.

CMC note:

The deliverable initially foreseen for this work item was a CEN Report. Although CEN/TC 304 decided some months ago to prepare an ENV, it now presents the document as a draft CEN Report in order to stick to the EC demand (this item is covered by the Order Voucher BC/CEN/97/26.16).

---

4 PROPOSAL:

See draft Resolution.

---

5 RESP: / /M. Balfroid

ORIGINATOR: CEN/TC 304 (IST)

BT N 6530
(Draft Resolution BT C119/2001)
Issue date : 2001-10-31

Target Date : 2001-12-12

REPLY FORM

SUBJECT: CEN/TC 304 - Publication of a CEN Report

BT authorizes the publication of the CEN/TC 304 "Information and communications technologies - European localization requirements" document entitled

> *"European ordering rules – Ordering for Latin, Greek, Cyrillic, Georgian and Armenian scripts"* (WI 00304031)

as a CEN Report, as included in document BT N 6530.

This resolution is applicable as from : 2001-12-12

ANSWER FROM THE MEMBER ON BT C119/2001:

BT Member	- agrees	□
	- disagrees with comments	□
	- disagrees fundamentally	□
	- abstains	□

COMMENTS / QUESTIONS :

Member Body :

Signature :

Date :

PLEASE RETURN IN DUE TIME TO BT SECRETARIAT

Foreword

This CEN report is intended to facilitate cross border communications and data exchange and to ensure that European cultural requirements are safeguarded in the increasingly interconnected world of today. It provides rules for ordering multilingual European texts and data into a single sequence. These rules come into effect if data from different languages must be brought into a predictable order that makes it easy for users to find information, which is often the case in pan-European applications.

This CEN reports extends the repertoire which is specified in ENV 13710:2000 *European Ordering Rules – Ordering of characters from the Latin, Greek and Cyrillic scripts.*

This CEN report does not intend to influence, let alone substitute itself for, national standards or customs in this field. Nevertheless, national standards have the opportunity to adapt this CEN report by declaring a formalized set of deviation rules ("delta") if they so wish.

Sorting assists users by presenting information in a structured way. This may include the subdivision of information by subject matters, e. g. by having several registers in a book, by splitting a phone book into several sections, one for each town that falls into its purview or by having multiple indices in a library. *Ordering* — the arrangement of information in alphabetical sequence — is in most circumstances an integral part of this procedure.

This CEN report must cater for two mutually exclusive demands: Implementers need clear guidelines and data which can readily be used in existing and future ordering applications. This can best be done by defining a European default ordering table in the syntax of the ordering standard ISO/IEC 14651:2001, of which the present document is a "profile". Users with no specific IT-background, however, need an explanation of the principles in a form more in line with existing national ordering standards or relevant practice. As tailoring tables in the syntax of ISO/IEC 14651 can be difficult to read for human readers, an explanation of the principles behind that table is given in the informative annexes. They are written in a more general style and users not familiar with the formal syntax of the tailoring table are advised to consult those annexes first. A web site on this subject is hosted by the Icelandic Standards Organization STAÐ LAR for further reference.[1]

1 Scope

This CEN report specifies the sequence to be established by alphabetical ordering of multilingual data composed of characters comprised in the *Multilingual European Subset Number 3* or subsets thereof. This collection is defined in CWA 13783.

NOTE The *Multilingual European Subset Number 3* is usually termed MES-3. A predecessor was known as the *Extended European Subset* (EES). Cf. ENV 1973:1995. MES-3 covers the Latin, Greek, Cyrillic, Armenian, and Georgian letters needed in European data interchange as well as symbols which are needed in Europe. MES-3 comes in two versions: MES-3A is an open collection whereas the fixed collection MES-3B is a snapshot of MES-3A against the repertoire of ISO/IEC 10646-1:1993 with

[1] At present STAÐ LAR can be accessed under `http://www.stadlar.is`

amendments 1 to 31. A CEN workshop agreement on the *Multilingual European Subsets of ISO/IEC 10646* has been published as CEN ISSS CWA 13873.

The ordering rules given here are only intended for data in more than one European language. They are not meant to influence, let alone replace existing national standards or practices.

The main part of this CEN report specifies letter-by-letter ordering of character strings. Informative Annex A presents equivalent information in a more readily accessible way. Informative Annex B deals with word-by-word ordering as a special form of ordering with multiple keys. Informative Annex C explains the use of further ordering criteria. Informative Annex D presents a widely used alternative to the main part, namely the amalgamation of several scripts in one index via implicit transliteration. Informative Annex F, finally, presents the information inherent in section 6 of the body of this CEN report in a formally equivalent, though condensed, form.

Following the practice of ISO/IEC 14651 characters are referenced as UXXXX where *X* stands for any hexadecimal digit and refers to the value of that character in ISO/IEC 10646-1:2000.This convention is used throughout this CEN report.

2 Normative references

This CEN report incorporates by dated or undated reference provisions from other publications. These normative references are quoted at the appropriate places in the text, and the publications are listed hereafter.

All standards are subject to revision. Dated references do not always refer to subsequent amendments of the publication in question. Undated references always refer to the latest edition.

ISO/IEC 10646-1:2000, Information Technology — Universal Multi-Octet Coded Character set (UCS). Second edition.

ISO 12199:2000, Alphabetical ordering of multilingual terminological and lexicographical data represented in the Latin alphabet.

ISO/IEC 14651:2001, International string ordering and comparison — Method for comparing character strings and description of the common template tailorable ordering.

NOTE An amendment to ISO/IEC 14651 is currently under preparation. It will extend the repertoire which is covered in the common template table of ISO/IEC 14651 to the full repertoire of ISO/IEC 10646-1:2000. It is expected that the table of the amendment will be a true superset of the table in ISO/IEC 14651:2001.

ENV 13710:2000, European Ordering Rules – Ordering of characters from the Latin, Greek and Cyrillic scripts.

3 Definitions

For the purpose of this CEN report the following definitions of ISO/IEC 10646-1 and of ISO/IEC 14651 apply:

3.1

character

A member of a set of elements used for the organization, control, or representation of data. [ISO/IEC 10646-1]

NOTE For the purpose of this CEN report a character is always a member of the MES-3.

3.2

character string

A sequence of characters. [ISO/IEC 14651]

3.3

delta

Differences from a given collation table. The given collation table, together with a given delta, forms a new collation table. Unless otherwise specified in this CEN report, the term " delta" always refers to differences from the Common Template Table as defined in ISO/IEC 14651. [ISO/IEC 14651]

3.4

ordering

A process by which two strings are determined to be in exactly one of the relationships of less than, greater than or equal to another. [ISO/IEC 14651]

4 Conformance

In order to be conformant to this CEN report an application shall meet the requirements prescribed in section 6 of ISO/IEC 14651 and use the default table of section 6 or an equivalent description of the information contained therein.

5 Tailorability

The European Ordering Rules defined in this CEN report can be taken as a template which can be tailored to the needs of any European country in the manner specified by ISO/IEC 14651.

6 Default Table

NOTE For the syntax of the table please consult ISO/IEC 14651:2001.´

NOTE The repertoire on which this delta table is based is the intersection of MES-3 and the repertoire of table 1 of ISO/IEC 14651:2001 with the addition of the following characters which are already in MES-2 and ENV 13710:2000:

- the modifier letter double apostrophe (U02EE);
- the Greek small letters digamma (U03DD), stigma (U03DB), koppa (U3DF) and sampi (U03E1);
- the Greek kai symbol (U03D7);
- the Cyrillic letters IE with grave (U0400, U0450) and I with grave (U040D, U045D).

```
%% EOR's EORDeltaTable
%
%% European Ordering Rules.
%
% EOR delta for MES-3 from ISO/IEC 14651:2000's CTT (ISO14651_2000_TABLE1).
%
% This delta gives only the actual changes from the first edition of the CTT.
%


reorder-after <BASE> % Introduce the LIG weight.
collating-symbol <LIG>
<BASE>
<LIG>

reorder-after <VRNT3> %Introduce more variants
collating-symbol <VRNT4>
collating-symbol <VRNT5>
collating-symbol <VRNT6>
<VRNT4>
<VRNT5>
<VRNT6>

collating-element <U000D_U000A> from "<U000D><U000A>"

reorder-after <SFFFF> % The only place where we can put the order_start line.

order_start forward;forward;forward;forward

% Reweighted non-alphanumeric characters (including some modifier letters):


% Currency signs (DRACHMA SIGN is not in ISO14651_2000_TABLE1):
<U0024> IGNORE;IGNORE;IGNORE;<U0024> % DOLLAR SIGN
<U00A2> IGNORE;IGNORE;IGNORE;<U00A2> % CENT SIGN
<U00A3> IGNORE;IGNORE;IGNORE;<U00A3> % POUND SIGN
<U00A4> IGNORE;IGNORE;IGNORE;<U00A4> % CURRENCY SIGN
<U00A5> IGNORE;IGNORE;IGNORE;<U00A5> % YEN SIGN
<U20A1> IGNORE;IGNORE;IGNORE;<U20A1> % COLON SIGN
<U20A2> IGNORE;IGNORE;IGNORE;<U20A2> % CRUZEIRO SIGN
<U20A3> IGNORE;IGNORE;IGNORE;<U20A3> % FRENCH FRANC SIGN
<U20A4> IGNORE;IGNORE;IGNORE;<U20A4> % LIRA SIGN
<U20A5> IGNORE;IGNORE;IGNORE;<U20A5> % MILL SIGN
<U20A6> IGNORE;IGNORE;IGNORE;<U20A6> % NAIRA SIGN
<U20A7> IGNORE;IGNORE;IGNORE;<U20A7> % PESETA SIGN
<U20A8> IGNORE;IGNORE;IGNORE;<U20A8> % RUPEE SIGN
<U20A9> IGNORE;IGNORE;IGNORE;<U20A9> % WON SIGN
<U20AA> IGNORE;IGNORE;IGNORE;<U20AA> % NEW SHEQEL SIGN
<U20AB> IGNORE;IGNORE;IGNORE;<U20AB> % DONG SIGN
<U20AC> IGNORE;IGNORE;IGNORE;<U20AC> % EURO SIGN
<U20AD> IGNORE;IGNORE;IGNORE;<U20AD> % KIP SIGN
<U20AE> IGNORE;IGNORE;IGNORE;<U20AE> % TUGRIK SIGN
<U20AF> IGNORE;IGNORE;IGNORE;<U20AF> % DRACHMA SIGN

% General category Lm (M.L. DOUBLE APOSTROPHE is not in ISO14651_2000_TABLE1):
<U02BB> IGNORE;IGNORE;IGNORE;<U02BB> % MODIFIER LETTER TURNED COMMA
<U02BD> IGNORE;IGNORE;IGNORE;<U02BD> % MODIFIER LETTER REVERSED COMMA
<U02BC> IGNORE;IGNORE;IGNORE;<U02BC> % MODIFIER LETTER APOSTROPHE
<U02BF> IGNORE;IGNORE;IGNORE;<U02BF> % MODIFIER LETTER LEFT HALF RING
<U02D1> IGNORE;IGNORE;IGNORE;<U02D1> % MODIFIER LETTER HALF TRIANGULAR COLON
<U02D0> IGNORE;IGNORE;IGNORE;<U02D0> % MODIFIER LETTER TRIANGULAR COLON
<U02D1> IGNORE;IGNORE;IGNORE;<U02D1> % MODIFIER LETTER HALF TRIANGULAR COLON
<U02EE> IGNORE;IGNORE;IGNORE;<U02EE> % MODIFIER LETTER DOUBLE APOSTROPHE
<U0559> IGNORE;IGNORE;IGNORE;<U0559> % ARMENIAN MODIFIER LETTER LEFT HALF RING

reorder-after <U0061>   % After LETTER A, just to make the 4th level
            % weights heavier than for punctuation
```

```
%% Latin
% Almost all changes here result from CEN/TC304's resolution for the
% Latin script part of MES-3 to treat only the letters a to z and
% thorn as distinct on the first level and have LETTER AE treated
% as a ligature, similar to how LIGATURE OE is treated in the CTT.
% Note that H WITH CARON is not in ISO14651_2000_TABLE1.
% Phonetic letters have not been included in this delta

<U00E6> "<S0061><S0065>";"<LIG><LIG>";"<MIN><MIN>";<U00E6> % LATIN SMALL LETTER AE
<U00C6> "<S0061><S0065>";"<LIG><LIG>";"<CAP><CAP>";<U00E6> % LATIN CAPITAL LETTER AE
<U01FD> "<S0061><S0065>";"<LIG><LIG><AIGUT>";"<MIN><MIN><BLK>";<U00E6> % LATIN SMALL
LETTER AE WITH ACUTE
<U01FC> "<S0061><S0065>";"<LIG><LIG><AIGUT>";"<CAP><CAP><BLK>";<U00E6> % LATIN
CAPITAL LETTER AE WITH ACUTE
<U01E3> "<S0061><S0065>";"<LIG><LIG><MACRO>";"<MIN><MIN><BLK>";<U00E6> % LATIN SMALL
LETTER AE WITH MACRON
<U01E1> "<S0061><S0065>";"<LIG><LIG><MACRO>";"<CAP><CAP><BLK>";<U00E6> % LATIN
CAPITAL LETTER AE WITH MACRON

<U0180> <S0062>;"<BASE><VRNT1>";"<MIN><BLK>";<U0180> % LATIN SMALL LETTER B WITH
STROKE
<U0253> <S0062>;"<BASE><VRNT2>";"<MIN><BLK>";<U0253> % LATIN SMALL LETTER B WITH HOOK
<U0181> <S0062>;"<BASE><VRNT2>";"<CAP><BLK>";<U0181> % LATIN CAPITAL LETTER B WITH
HOOK
<U0183> <S0062>;"<BASE><VRNT3>";"<MIN><BLK>";<U0183> % LATIN SMALL LETTER B WITH
TOPBAR
<U0182> <S0062>;"<BASE><VRNT3>";"<CAP><BLK>";<U0182> % LATIN CAPITAL LETTER B WITH
TOPBAR

<U0188> <S0063>;"<BASE><VRNT1>";"<MIN><BLK>";<U0188> % LATIN SMALL LETTER C WITH HOOK
<U0187> <S0063>;"<BASE><VRNT1>";"<CAP><BLK>";<U0187> % LATIN CAPITAL LETTER C WITH
HOOK

<U0111> <S0064>;"<BASE><VRNT1>";"<MIN><BLK>";<U0111> % LATIN SMALL LETTER D WITH
STROKE
<U0112> <S0064>;"<BASE><VRNT1>";"<CAP><BLK>";<U0112> % LATIN CAPITAL LETTER D WITH
STROKE
<U0189> <S0064>;"<BASE><VRNT2>";"<CAP><BLK>";<U0189> % LATIN CAPITAL LETTER AFRICAN D
<U0257> <S0064>;"<BASE><VRNT3>";"<MIN><BLK>";<U0257> % LATIN SMALL LETTER D WITH HOOK
<U018A> <S0064>;"<BASE><VRNT3>";"<CAP><BLK>";<U018A> % LATIN CAPITAL LETTER D WITH
HOOK
<U018C> <S0064>;"<BASE><VRNT4>";"<MIN><BLK>";<U018C> % LATIN SMALL LETTER D WITH
TOPBAR
<U018B> <S0064>;"<BASE><VRNT4>";"<CAP><BLK>";<U018B> % LATIN CAPITAL LETTER D WITH
TOPBAR
<U00F0> <S0064>;"<BASE><VRNT5>";"<MIN><BLK>";<U00F0> % LATIN SMALL LETTER ETH
<U00D0> <S0064>;"<BASE><VRNT5>";"<CAP><BLK>";<U00D0> % LATIN CAPITAL LETTER ETH
<U018D> <S0064>;"<BASE><VRNT6>";"<MIN><BLK>";<U018D> % LATIN SMALL LETTER TURNED
DELTA

<U0259> <S0065>;"<BASE><VRNT1>";"<MIN><BLK>";<U0295> % LATIN SMALL LETTER SCHWA
<U018F> <S0065>;"<BASE><VRNT1>";"<CAP><BLK>";<U018F> % LATIN CAPITAL LETTER SCHWA
<U018E> <S0065>;"<BASE><VRNT2>";"<CAP><BLK>";<U018E> % LATIN CAPITAL LETTER REVERSED
E
<U01DD> <S0065>;"<BASE><VRNT3>";"<MIN><BLK>";<U01DD> % LATIN SMALL LETTER TURNED E

<U0192> <S0066>;"<BASE><VRNT1>";"<MIN><BLK>";<U0192> % LATIN SMALL LETTER F WITH HOOK

<U01E5> <S0067>;"<BASE><VRNT1>";"<MIN><BLK>";<U01E5> % LATIN SMALL LETTER G WITH
STROKE
<U01E4> <S0067>;"<BASE><VRNT1>";"<CAP><BLK>";<U01E5> % LATIN CAPITAL LETTER G WITH
STROKE
<U0260> <S0067>;"<BASE><VRNT2>";"<MIN><BLK>";<U0260> % LATIN SMALL LETTER G WITH HOOK
<U0193> <S0067>;"<BASE><VRNT2>";"<CAP><BLK>";<U0193> % LATIN CAPITAL LETTER G WITH
HOOK
<U0263> <S0067>;"<BASE><VRNT3>";"<MIN><BLK>";<U0263> % LATIN SMALL LETTER GAMMA
<U0194> <S0067>;"<BASE><VRNT3>";"<CAP><BLK>";<U0194> % LATIN CAPITAL LETTER GAMMA
```

```
<U021F> <S0068>;"<BASE><CARON>";"<MIN><BLK>";<U021F> % LATIN SMALL LETTER H WITH
CARON
<U021E> <S0068>;"<BASE><CARON>";"<CAP><BLK>";<U021E> % LATIN CAPITAL LETTER H WITH
CARON
<U0127> <S0068>;"<BASE><VRNT1>";"<MIN><BLK>";<U0127> % LATIN SMALL LETTER H WITH
STROKE
<U0126> <S0068>;"<BASE><VRNT1>";"<CAP><BLK>";<U0126> % LATIN CAPITAL LETTER H WITH
STROKE
<U0195> "<S0068><S0076>";"<BASE><BASE>";"<MIN><MIN>";<U0195> % LATIN SMALL LETTER HV

<U0131> <S0069>;"<BASE><VRNT1>";"<MIN><BLK>";<U0131> % LATIN SMALL LETTER DOTLESS I
<U0197> <S0069>;"<BASE><VRNT2>";"<CAP><BLK>";<U0197> % LATIN CAPITAL LETTER I WITH
STROKE
<U0196> <S0069>;"<BASE><VRNT3>";"<CAP><BLK>";<U0196> % LATIN CAPITAL LETTER IOTA
<U0133> "<S0069><S006A>";"<LIG><LIG>";"<MIN><MIN>";<U0133> % LATIN SMALL LIGATURE IJ
<U0132> "<S0069><S006A>";"<LIG><LIG>";"<CAP><CAP>";<U0132> % LATIN CAPITAL LIGATURE
IJ
<U0192> <S0066>;"<BASE><VRNT1>";"<MIN><BLK>";<U0192> % LATIN SMALL LETTER F WITH HOOK
<U0191> <S0066>;"<BASE><VRNT1>";"<CAP><BLK>";<U0191> % LATIN CAPITAL LETTER F WITH
HOOK

<U0199> <S006B>;"<BASE><VRNT1>";"<MIN><BLK>";<U0199> % LATIN SMALL LETTER K WITH HOOK
<U0198> <S006B>;"<BASE><VRNT1>";"<CAP><BLK>";<U0198> % LATIN CAPITAL LETTER K WITH
HOOK
<U0138> <S006B>;"<BASE><VRNT2>";"<MIN><BLK>";<U0138> % LATIN SMALL LETTER KRA

<U0142> <S006C>;"<BASE><VRNT1>";"<MIN><BLK>";<U0142> % LATIN SMALL LETTER L WITH
STROKE
<U0141> <S006C>;"<BASE><VRNT1>";"<CAP><BLK>";<U0141> % LATIN CAPITAL LETTER L WITH
STROKE
<U0140> <S006C>;"<BASE><VRNT2>";"<MIN><BLK>";<U0140> % LATIN SMALL LETTER L WITH
MIDDLE DOT
<U013F> <S006C>;"<BASE><VRNT2>";"<CAP><BLK>";<U013F> % LATIN CAPITAL LETTER L WITH
MIDDLE DOT
<U019A> <S006C>;"<BASE><VRNT3>";"<MIN><BLK>";<U019A> % LATIN SMALL LETTER L WITH BAR
<U026B> <S006C>;"<BASE><VRNT4>";"<MIN><BLK>";<U026B> % LATIN SMALL LETTER L WITH
MIDDLE TILDE
<U019B> <S006C>;"<BASE><VRNT5>";"<MIN><BLK>";<U019B> % LATIN SMALL LETTER LAMBDA WITH
STROKE

<U019C> <S006C>;"<BASE><VRNT1>";"<CAP><BLK>";<U019C> % LATIN CAPITAL LETTER TURNED M

<U0149> <S006E>;"<BASE><VRNT1>";"<MIN><BLK>";<U0149> % LATIN SMALL LETTER N PRECEDED
BY APOSTROPHE
<U019E> <S006E>;"<BASE><VRNT2>";"<MIN><BLK>";<U019E> % LATIN SMALL LETTER N WITH LONG
RIGHT LEG
<U019D> <S006E>;"<BASE><VRNT2>";"<CAP><BLK>";<U019D> % LATIN CAPITAL LETTER N WITH
LEFT HOOK
<U014B> <S006E>;"<BASE><VRNT3>";"<MIN><BLK>";<U014B> % LATIN SMALL LETTER ENG
<U014A> <S006E>;"<BASE><VRNT3>";"<CAP><BLK>";<U014A> % LATIN CAPITAL LETTER ENG

<U00F8> <S006F>;"<BASE><VRNT1>";"<MIN><BLK>";<U00F8> % LATIN SMALL LETTER O WITH
STROKE
<U00D8> <S006F>;"<BASE><VRNT1>";"<CAP><BLK>";<U00D8> % LATIN CAPITAL LETTER O WITH
STROKE
<U01FF> <S006F>;"<BASE><VRNT1><AIGUT>";"<MIN><BLK><BLK>";<U01FF> % LATIN SMALL LETTER
O WITH STROKE AND ACUTE
<U01FE> <S006F>;"<BASE><VRNT1><AIGUT>";"<CAP><BLK><BLK>";<U01FE> % LATIN SMALL LETTER
O WITH STROKE AND ACUTE
<U026B> <S006F>;"<BASE><VRNT2>";"<CAP><BLK>";<U026B> % LATIN CAPITAL LETTER O WITH
MIDDLE TILDE
<U0186> <S006F>;"<BASE><VRNT3>";"<CAP><BLK>";<U0186> % LATIN CAPITAL LETTER
OPEN O

<U0153> "<S006F><S0065>";"<LIG><LIG>";"<MIN><MIN>";<U0153> % LATIN SMALL LIGATURE OE
<U0152> "<S006F><S0065>";"<LIG><LIG>";"<CAP><CAP>";<U0152> % LATIN CAPITAL LIGATURE
OE
<U01A3> "<S006F><S0069>";"<BASE><BASE>";"<MIN><MIN>";<U01A3> % LATIN SMALL LETTER OI
```

```
<U01A2> "<S006F><S0069>";"<BASE><BASE>";"<CAP><CAP>";<U01A2> % LATIN CAPITAL LETTER
OI

<U01A5> <S0070>;"<BASE><VRNT1>";"<MIN><BLK>";<U01A5> % LATIN SMALL LETTER P WITH HOOK
<U01A4> <S0070>;"<BASE><VRNT1>";<CAP><BLK>";<U01A4> % LATIN CAPITAL LETTER P WITH
HOOK

<U027C> <S0072>;"<BASE><VRNT1>";"<MIN><BLK>";<U027C> % LATIN SMALL LETTER R WITH LONG
LEG
<U01A6> <S0072>;"<BASE><VRNT2>";"<CAP><BLK>";<U01A6> % LATIN LETTER YR

<U00DF> "<S0073><S0073>";"<LIG><LIG>";"<MIN><MIN>";<U00DF> % LATIN SMALL LETTER SHARP
S

<U01A9> <S0073>;"<BASE><VRNT1>";"<CAP><BLK>";<U01A9> % LATIN CAPITAL LETTER ESH
<U01AA> <S0073>;"<BASE><VRNT2>";"<MIN><BLK>";<U01AA> % LATIN LETTER REVERSED ESH LOOP

<U0167> <S0074>;"<BASE><VRNT1>";"<MIN><BLK>";<U0167> % LATIN SMALL LETTER T WITH
STROKE
<U0166> <S0074>;"<BASE><VRNT1>";"<CAP><BLK>";<U0166> % LATIN CAPITAL LETTER T WITH
STROKE
<U01AD> <S0074>;"<BASE><VRNT2>";"<MIN><BLK>";<U01AD> % LATIN SMALL LETTER T WITH HOOK
<U01AC> <S0074>;"<BASE><VRNT2>";"<CAP><BLK>";<U01AC> % LATIN CAPITAL LETTER T WITH
HOOK
<U01AB> <S0074>;"<BASE><VRNT3>";"<MIN><BLK>";<U01AB> % LATIN SMALL LETTER T WITH
PALATAL HOOK
<U01AE> <S0074>;"<BASE><VRNT4>";"<CAP><BLK>";<U01AE> % LATIN CAPITAL LETTER T WITH
RETROFLEX HOOK

<U01B2> <S0076>;"<BASE><VRNT1>";"<CAP><BLK>";<U01B2> % LATIN CAPITAL LETTER V WITH
HOOK

<U01BF> <S0077>;"<BASE><VRNT1>";"<MIN><BLK>";<U01BF> % LATIN LETTER WYNN

<U01B4> <S0079>;"<BASE><VRNT1>";"<MIN><BLK>";<U01B4> % LATIN SMALL LETTER Y WITH HOOK
<U01B3> <S0079>;"<BASE><VRNT1>";"<CAP><BLK>";<U01B3> % LATIN CAPITAL LETTER Y WITH
HOOK
<U028A> <S0079>;"<BASE><VRNT2>";"<MIN><BLK>";<U028A> % LATIN SMALL LETTER UPSILON

<U01B6> <S007A>;"<BASE><VRNT1>";"<MIN><BLK>";<U01B6> % LATIN SMALL LETTER Z WITH
STROKE
<U01B5> <S007A>;"<BASE><VRNT1>";"<CAP><BLK>";<U01B5> % LATIN CAPITAL LETTER Z WITH
STROKE
<U0292> <S007A>;"<BASE><VRNT2>";"<MIN><BLK>";<U0292> % LATIN SMALL LETTER EZH
<U01B7> <S007A>;"<BASE><VRNT2>";"<CAP><BLK>";<U01B7> % LATIN SMALL LETTER EZH
<U01EF> <S007A>;"<BASE><VRNT1><CARON>";"<MIN><BLK><BLK>";<U01EF> % LATIN SMALL LETTER
EZH WITH CARON
<U01EE> <S007A>;"<BASE><VRNT1><CARON>";"<CAP><BLK><BLK>";<U01EE> % LATIN CAPITAL
LETTER EZH WITH CARON
<U01B9> <S007A>;"<BASE><VRNT2>";"<MIN><BLK>";<U01B9> % LATIN SMALL LETTER EZH
REVERSED
<U01B8> <S007A>;"<BASE><VRNT2>";"<CAP><BLK>";<U01B8> % LATIN CAPITAL LETTER EZH
REVERSED
<U01BA> <S007A>;"<BASE><VRNT3>";"<MIN><BLK>";<U01BA> % LATIN SMALL LETTER EZH WITH
TAIL


% Greek (KAI SYMBOL and the small forms for the Greek letters
% DIGAMMA, STIGMA, KOPPA, and SAMPI
% are not in ISO14651_2000_TABLE1):

<U00B5> <S03BC>;<BASE>;<MIN>;<U00B5> % MICRO SIGN
<U03DD> <S03DC>;<BASE>;<MIN>;<U03DD> % GREEK SMALL LETTER DIGAMMA
<U03DB> <S03DA>;<BASE>;<MIN>;<U03DB> % GREEK SMALL LETTER STIGMA
<U03DF> <S03DE>;<BASE>;<MIN>;<U03DF> % GREEK SMALL LETTER KOPPA
<U03E1> <S03E0>;<BASE>;<MIN>;<U03E1> % GREEK SMALL LETTER SAMPI
<U03D7> <S03BA>;"<BASE><VRNT1>";"<MIN><BLK>";<U03D7> % GREEK KAI SYMBOL
```

```
% Full conformance with GOST requirements for Cyrillic letters (in addition,
% IE WITH GRAVE I and I WITH GRAVE are not in ISO14651_2000_TABLE1):

<U0453> <S0452>;"<BASE><VRNT1>";"<MIN><BLK>";<U0453> % CYRILLIC SMALL LETTER GJE
<U0403> <S0452>;"<BASE><VRNT1>";"<CAP><BLK>";<U0403> % CYRILLIC CAPITAL LETTER GJE

<U0450> <S0435>;"<BASE><GRAVE>";"<MIN><BLK>";<U0450> % CYRILLIC SMALL LETTER IE WITH GRAVE
<U0400> <S0435>;"<BASE><GRAVE>";"<CAP><BLK>";<U0400> % CYRILLIC CAPITAL LETTER IE WITH GRAVE

<U045D> <S0438>;"<BASE><GRAVE>";"<MIN><BLK>";<U045D> % CYRILLIC SMALL LETTER I WITH GRAVE
<U040D> <S0438>;"<BASE><GRAVE>";"<CAP><BLK>";<U040D> % CYRILLIC CAPITAL LETTER I WITH GRAVE

<U045C> <S045B>;"<BASE><VRNT1>";"<MIN><BLK>";<U045C> % CYRILLIC SMALL LETTER KJE
<U040C> <S045B>;"<BASE><VRNT1>";"<CAP><BLK>";<U040C> % CYRILLIC CAPITAL LETTER KJE

% Georgian: Identical to ISO14651_2000_TABLE1

% Armenian: Identical to ISO14651_2000_TABLE1

reorder-end %% for EOR's EORDeltaTable
```

Annex A (informative) Exposition of relevant principles

A.0 Introduction

This annex aims to present the information inherent in section 6 in a more accessible form for those who are interested in the principles guiding the composition of the table. Those readers not concerned with implementation details may take this more traditional treatment of the matter as an authoritative interpretation of the body of this CEN report.

A.1 Definitions

For the purpose of this annex, the following definitions apply in addition to those in the body of this CEN report (see section 3).

A.1.1

digit

One of the characters 0 1 2 3 4 5 6 7 8 9.

A.1.2

letter

Character used to represent (either alone or in combination) sounds or sequences of sounds of a natural language in writing. Here equivalent to all characters of MES-3 whose name contains one of the words `LETTER` or `LIGATURE`.

A.1.3

first level letter

Character that is a member of the following list of letters:

Latin script:

a A b B c C d D e E f F g G h H i I j J k K l L m M n N o O p P q Q r R s S t T u U v V w W x X y Y z Z þ Þ

Greek script:

α Α β Β γ Γ δ Δ ε Ε Ϝ Ϛ ζ Ζ η Η θ Θ ι Ι κ Κ λ Λ μ Μ ν Ν ξ Ξ ο Ο π Π Ϙ ρ Ρ σ Σ τ Τ υ Υ φ Φ χ Χ ψ Ψ ω Ω Ϡ

NOTE Ϛ , Ϙ and Ϡ are archaic letters that are currently used to designate numerals. Ϝ is not used in any modern language.

NOTE MES-3 contains also a number of Coptic letters. Their order is specified in ISO/IEC 14651:2001.

Cyrillic script:

а А ӑ Ӑ ӓ Ӓ ә Ә ӛ Ӛ ӕ Ӕ б Б в В г Г ғ Ғ ђ Ђ д Д ђ Ђ ҙ Ҙ е Е ӗ Ӗ є Є ж Ж ӝ Ӝ җ Җ з З ӟ Ӟ ѕ Ѕ ӡ Ӡ и И й Й і І ї Ї й Й ј Ј к К қ Қ ķ Ķ ҟ Ҟ ҝ Ҝ ҡ Ҡ л Л љ Љ м М н Н ң Ң ӈ Ӈ ҥ Ҥ њ Њ о О ӧ Ӧ ө Ө ӫ Ӫ п П ҧ Ҧ ҁ Ҁ р Р с С ҫ Ҫ т Т ҭ Ҭ ћ Ћ у У ў Ў ӱ Ӱ ӳ Ӳ ү Ү ұ Ұ ѹ Ѹ ф Ф х Х ҳ Ҳ һ Һ ѡ Ѡ ѿ Ѿ ѽ Ѽ ѻ Ѻ ц Ц ҵ Ҵ ч Ч ӵ Ӵ ҷ Ҷ ӌ Ӌ ҹ Ҹ ҽ Ҽ ҿ Ҿ џ Џ ш Ш щ Щ ъ Ъ ы Ы ӹ Ӹ ь Ь ѣ Ѣ э Э ю Ю я Я ѥ Ѥ ѧ Ѧ ѫ Ѫ ѩ Ѩ ѭ Ѭ ѯ Ѯ ѱ Ѱ ѳ Ѳ ѵ Ѵ ѷ Ѷ ҩ Ҩ Ӏ

Georgian script:

ა ბ Ⴁ გ Ⴂ დ Ⴃ ე Ⴄ ვ Ⴅ ზ Ⴆ ჱ Ⴡ თ Ⴇ ი Ⴈ კ Ⴉ ლ Ⴊ მ Ⴋ ნ Ⴌ ჲ Ⴢ ო Ⴍ პ Ⴎ ჟ Ⴏ რ Ⴐ ს Ⴑ ტ Ⴒ ჳ Ⴣ უ Ⴓ ფ Ⴔ ქ Ⴕ ღ Ⴖ ყ Ⴗ შ Ⴘ ჩ Ⴙ ც Ⴚ ძ Ⴛ წ Ⴜ ჭ Ⴝ ხ Ⴞ ჴ Ⴤ ჯ Ⴟ ჰ Ⴠ ჵ Ⴥ

Armenian script:

ա Ա բ Բ գ Գ դ Դ ե Ե զ Զ է Է ը Ը թ Թ ժ Ժ ի Ի լ Լ խ Խ ծ Ծ կ Կ հ Հ ձ Ձ ղ Ղ ճ Ճ մ Մ յ Յ ն Ն շ Շ ո Ո չ Չ պ Պ ջ Ջ ռ Ռ ս Ս վ Վ տ Տ ր Ր ց Ց ւ Ւ փ Փ ք Ք օ Օ ֆ Ֆ և

A.1.4

diacritical mark

Any of a number of recurring graphical structures placed over, under or next to a first level letter which does not significantly modify the shape of the first level letter itself and which in combination with that first level letter is a valid letter. These structures modify meaning or pronunciation or some other feature of the first level letter. The diacritical marks which are relevant to this CEN report are listed in section A.8.1.1.

A.1.5

letter with diacritical marks

Letters which can be seen as equivalent to the combination between a first level letter and one or more diacritical marks.

NOTE Some *letters with diacritical marks* are treated as *first level letters* in some languages, e.g. ä in Swedish and ñ in Spanish. However, these are subject to national standards or local practices which are outside the scope of this CEN report.

NOTE Some Latin letters such as ǻ (U01FB) have more than one diacritical mark. A considerable number of Greek letters have more than one diacritical mark.

A.1.6

equivalent letter form

Character created by joining two or more distinct first level letters or two or more letters with diacritical marks or any combination of these.

A.1.7

second level letter

Letter that is neither a first level letter nor an equivalent letter form nor a letter with diacritical marks. The second level letters which are relevant to this CEN report are listed in A.8.2.

A.1.8

capital letter

Letter which has the string CAPITAL in its name in ISO/IEC 10646-1.

NOTE This definition works for the repertoire of MES-3, but not necessarily for the full repertoire of the UCS.

NOTE For the first level letters these are:

Latin script:

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z Þ

Greek script:

Α Β Γ Δ Ε Ζ Η Θ Ι Κ Λ Μ Ν Ξ Ο Π Ρ Σ Τ Υ Φ Χ Ψ Ω

Cyrillic script:

А Ӑ Ӓ Ә Ӛ Ӕ Б В Г Ғ Ҕ Д Ђ Ҙ Е Ӗ Є Ж Ӂ Җ З Ӟ Ѕ Ӡ И Ӣ І Ї Й Ј К Ӄ Қ Ҡ Ҡ̆ Ҟ Л Љ М Н
Ң Ӈ Ҥ Њ О Ӧ Ө Ӫ П Ҧ Ҁ Р С Ҫ Т Ҭ Ћ У Ў Ӱ Ӳ Ү Ұ ОуФ Х Ҳ Һ Ѡ Ѿ Ѽ Ѻ Ц Ҵ Ч Ӵ Ҷ Ӌ
Ҹ Ҽ Ҿ Џ Ш Щ Ъ Ы Ӹ Ь Ѣ Э Ю Я Ѥ Ѧ Ѫ Ѩ Ѭ Ѯ Ѱ Ѳ Ѵ Ѷ Ҩ

Georgian script:

Ⴀ Ⴁ Ⴂ Ⴃ Ⴄ Ⴅ Ⴆ Ⴡ Ⴇ Ⴈ Ⴉ Ⴊ Ⴋ Ⴌ Ⴢ Ⴍ Ⴎ Ⴏ Ⴐ Ⴑ Ⴒ Ⴣ Ⴓ Ⴔ Ⴕ Ⴖ Ⴗ Ⴘ Ⴙ Ⴚ Ⴛ Ⴜ Ⴝ Ⴞ Ⴤ Ⴟ Ⴠ Ⴥ

NOTE The function of capital letters in Georgian differs significantly from the function of capital letters in the other four scripts.[2] Capitalization routines that are common to Latin, Greek, Cyrillic and Armenian scripts should therefore not be applied to Georgian letters without careful consideration.

Armenian script:

Ա Բ Գ Դ Ե Զ Է Ը Թ Ժ Ի Լ Խ Ծ Կ Հ Ձ Ղ Ճ Մ Յ Ն Շ Ո Չ Պ Ջ Ռ Ս Վ Տ Ր Ց Ւ Փ Ք Օ Ֆ

A.1.9

small letter

Letter which is not a capital letter.

A.1.10

special character

Character that is neither a letter nor a digit.

NOTE *Special characters* are often called *symbols*, but also include punctuation marks, apostrophes, mathematical operators, monetary symbols and others.

[2] The Georgian letters which ISO/IEC 10646-1:2000 calls `GEORGIAN CAPITAL LETTER` make up the *asomtavruli* script that is primarily used in Old Georgian texts. The remaining letters (classified simply as `GEORGIAN LETTER` in ISO/IEC 10646-1:2000) are usually identified with the *mxedruli* or military script that is used almost exclusively for writing modern Georgian.

A.1.11

spacing character

one of the characters SPACE, NO-BREAK SPACE,EN QUAD, EM QUAD, EN SPACE, EM SPACE, THREE-PER-EM SPACE, FOUR-PER-EM SPACE, SIX-PER-EM SPACE, FIGURE SPACE, PUNCTUATION SPACE, THIN SPACE, HAIR SPACE, and LINE SEPARATOR.

NOTE There are a number of different types of "spaces" which are not part of MES-3 , but which are used very often in various fields of application. These may be understood as spacing characters for the purposes of this annex.

A.2 Preparatory procedures

A.2.1 Purpose

Most ordering tasks require more than simply the ordering of strings. In a telephone directory, for example, one might want to order by names first, followed by addresses and phone numbers, recurring to addresses only when ordering by names fails to establish a unique sequence and to phone numbers only if both names and addresses are identical.

Each of these units is called a key and the approach is called the multiple ordering key approach.

A.2.2 Methodology

More rigorously expressed, the multiple ordering key approach implies the preprocessing of the data in the following steps, any or all of which may be omitted, especially in the case of a single ordering key:

1. subdivision of data into multiple ordering keys through the introduction of a higher level protocol

2. establishing a hierarchy between these keys

3. extracting the keys from the data

4. subjecting the keys to some form of normalization

NOTE This normalization might include, but is not limited to: changing capital letters to small letters where it is considered appropriate (e. g. in the case of sentence initial capitals or capitals for emphasis), lemmatization (especially for inflected languages), expansion of abbreviations, or reduction of blanks between words to one throughout the data. It can also be left out entirely.

NOTE An especially important step is usually the correct treatment of numeral strings where leading zeroes might have to be introduced to ensure proper comparisons between corresponding decimals. Failure to do so may result in faulty ordering.

Starting with the keys highest in the hierarchy equivalent keys which were thus obtained are compared with the aid of the ordering rules as established in this CEN report. As soon as a unique sequence is established, further keys are ignored.

A.2.3 Further preprocessing

Further preprocessing of some kind may or may not be necessary, but is not within the scope of this CEN report.

This CEN report assumes that the user has already performed these preparatory procedures which are left entirely at his or her discretion and are thus out of its scope. It is concerned exclusively with the ordering of strings which belong to one key and which have undergone those preparatory procedures.

A.3 The multilevel ordering procedure

A.3.1 General principles

This CEN report defines in this annex a multilevel ordering procedure whose results are identical to those produced by the application of the rules of the body of this report.

Multilevel ordering procedure means that the input strings are first compared on the *first ordering level*. Only when the procedure described for this level fails to establish a unique and determined sequence for the strings the different parts of the *second ordering level* are taken into consideration. If this likewise fails to produce a unique sequence the *third ordering level* is invoked, and after this the *fourth ordering level*. If this also cannot establish a unique sequence, two strings are regarded as equivalent.

Each level compares two strings in the following manner: The first non-ignored characters are compared. If the ordering rules for that level specify a unique and determined sequence for these characters then this determines the sequence of the strings. If not, the second non-ignored characters are compared, and so forth until one of the following conditions is met. If more than one of the conditions are true, only the first one which is fulfilled is applicable:

1. the ordering rules for that level define a unique sequence for the two non-ignored characters which is then also the ordering sequence for the strings;

2. one of the strings has no more non-ignored characters whereas the other has. Then the string without more characters precedes the other one;

3. both strings have no more non-ignored characters. Then the next ordering level, if existing, is invoked. If there are no more levels, the two strings are deemed equivalent.

A.3.2 Assumptions and aims

This CEN report acts according to certain assumptions:

— access to information must be facilitated as much as possible;

— the user is not assumed to know details of ISO/IEC 10646-1:2000;

— the rules are derived from standardized rules and common practice in a large number of European languages without giving preference to the rules of any language or languages in particular;

NOTE These assumptions motivate a set of principles which underlie these European Ordering Rules and help to clarify the decisions taken:

— second level letters are ordered according to their visual appearance, not according to their pronunciation or meaning unless user-expectation demands something else;

NOTE For the details for the treatment of second level letters please cf. table A.8.2.

— forms which the user perceives as more basic should precede special or combined ones. Forms used primarily for emphasis should likewise follow after more basic forms.

A.3.3 Rules (valid throughout)

A.3.3.1 Ordering by script

Digits precede letters. Letters are ordered by scripts, putting Latin letters before Greek ones before Cyrillic ones before Georgian ones before Armenian ones.

A.3.3.2 Equivalent letter forms

Equivalent letter forms are decomposed into the letters out of which they are formed.

A.4 First ordering level

A.4.1 Validity

All of the following rules are valid for the first ordering level only.

A.4.2 Equivalent or ignored characters

A.4.2.1 Capital and small letters

Capital and small forms of the same letter are treated as equivalent.

A.4.2.2 Second level letters

Second level letters are treated as equivalent to one or more first level letters as specified in section A.8.2.

A.4.2.3 Letters with diacritical marks

Letters with diacritical marks are treated as equivalent to their corresponding first level letters.

NOTE For the definition of first level letters please cf. section A.1.3.

A.4.2.4 Special characters

Special characters are ignored.

A.4.3 Ordering sequences

A.4.3.1 Digits

Digits are to be ordered in the following sequence:

0 1 2 3 4 5 6 7 8 9

A.4.3.2 Latin script

Latin first level letters are to be ordered in the following sequence:

a b c d e f g h i j k l m n o p q r s t u v w x y z þ

A.4.3.3 Greek script

Greek first level letters are to be ordered in the following sequence:

α β γ δ ε ϝ ϛ ζ η θ ι κ λ μ ν ξ ο π ϙ ρ σ τ υ φ χ ψ ω ϡ

A.4.3.4 Cyrillic script

Cyrillic first level letters are to be ordered in the following sequence:

а ӑ ӓ ә ӛ ӕ б в г ғ ђ д ђ ҙ е ӗ є ж ӝ җ з ӟ ѕ ӡ и ӥ і ї й ј к қ ӄ ҡ ҟ к л љ м н
ң ӈ ҥ њ о ӧ ө ӫ п ҧ ҁ р с ҫ т ҭ ћ у ў ӱ ӳ ү ұ ѹ ф х ҳ һ ѡ ѽ ѿ ѻ ц ҵ ч ӵ ҷ ӌ
ҹ ҽ ҿ џ ш щ ъ ы ӹ ь ѣ э ю я ѥ ѧ ѫ ѩ ѭ ѯ ѱ ѳ ѵ ѷ ҩ Ӏ

NOTE This sequence is based on pan-Cyrillic requirements as specified by Г О С Т . It was officially communicated to the editor of this CEN report by Г О С Т 's designated expert in the field and maximally facilitates the process of finding information in pan-Cyrillic texts.

A.4.3.5 Georgian script

Georgian first level letters are to be ordered in the following sequence:

ა ბ გ დ ე ვ ზ ჱ თ ი კ ლ მ ნ ჲ ო პ ჟ რ ს ტ ჳ უ ფ ქ ღ ყ შ ჩ ც ძ წ ჭ ხ ჴ ჯ ჰ ჵ ჶ

A.4.3.6 Armenian script

Armenian first level letters are to be ordered in the following sequence:

ա բ գ դ ե զ է ը թ ժ ի լ խ ծ կ հ ձ ղ ճ մ յ ն շ ո չ պ ջ ռ ս վ տ ր ց ւ փ ք օ ֆ և

A.5 Second ordering level

A.5.1 No unique sequence after the first ordering level

If the first ordering level does not result in an unique sequence, the second ordering level is invoked. It is distinguished from the first ordering level by no longer treating letters with diacritical marks and second level letters as equivalent to first level letters.

The second ordering level is divided into two parts: second level letters and diacritical marks. If the treatment of second level letters alone results in a unique sequence, diacritical marks are to be ignored.

A.5.2 Equivalent or ignored characters

A.5.2.1 Capital and small letters

Capital and small forms of the same letter are treated as equivalent.

A.5.2.2 Special characters

Special characters are ignored.

A.5.3 Ordering sequences

A.5.3.1 Second level letters

Second level letters are to be ordered after their corresponding first level letter. In the case of multiple second level letters with the same first level letter they are to be ordered in the sequence specified by A.8.2.

A.5.3.2 Letters with diacritical marks

Letters with diacritical marks which have only one diacritical mark are to be ordered with respect to their diacritical mark in the sequence indicated in section A.8.1.1. For letters with more than one diacritical mark, the diacritical mark shall be considered in the following order: Inside the character before outside; below the character before above; working from bottom to top, then from left to right.

NOTE Some European countries, notably France, treat diacritics differently from this CEN report, and parse diacritics backwards within each word. For applications targeted for this market this must be taken into consideration by the declaration of a suitable delta.

A.6 Third ordering level

A.6.1 No unique sequence after the second ordering level

If the second ordering level also does not result in a unique sequence of strings, the *third ordering level* is invoked. It no longer treats capital and small letters as equivalent.

A.6.2 Ignored characters

Special characters are ignored.

A.6.3 Ordering sequences

A.6.3.1 Capitalization

Small letters are ordered before the corresponding capital ones.

A.7 Fourth ordering level

A.7.1 No unique sequence after the *third ordering level*

If the third ordering level likewise does not result in a unique sequence of strings, the fourth ordering level is invoked. It takes special characters into account.

A.7.2.Sequence of special characters

The special characters of the MES-3 are ordered in the sequence of the default tailorable template of ISO/IEC 14651. For most special characters this is the order in which they are listed in ISO/IEC 10646-1 and relevant appendices. However, for a number of special characters ISO/IEC 14651 defines a divergent sequence in line with the specification of the Canadian standard CAN/CSA Z243.230-1996.

NOTE It is advised to pay particular attention to special characters which may have the role of structuring entries in some manner. These include punctuation marks, hyphens, apostrophes and brackets.

A.7.3 Equivalence

Two strings between which after the fourth ordering level no unique sequence can be established are considered to be equivalent.

NOTE For further options to break the deadlock in certain circumstances please cf. the informative annex C: *Ordering by position and by style*.

A.8 Specific ordering sequences

A.8.1 Diacritical marks

A.8.1.1 Diacritical marks

This form of presentation has been chosen to enable the unification of diacritical marks across scripts without modifying the resulting sequence of strings.

Shape[8]		Diacritical mark[9]	Alternative names[10]
᾿	U1FBF	PSILI	spiritus lenis
῾	U1FFE	DASIA	spiritus asper
´	U1FFD	OXIA	
`	U1FEF	VARIA	
˘	U0306	COMBINING BREVE	VRACHY
͂	U0342	COMBINING GREEK PERISPOMENI	
΄	U0384	TONOS	
(ι	U1FBE	PROSGEGRAMMENI	iota adscriptum[11])
	U0345	COMBINING GREEK YPOGEGRAMMENI	iota subscriptum[12]

[8] Shapes may vary according to fonts and styles

[9] If possible, combining diacritical marks are referenced. If no corresponding combining diacritical mark exists, the table lists non-combining variants. Diacritical marks are unified for Cyrillic and Latin but not for Greek and Latin. This reflects prevalent usage and user-expectations

[10] Names in lowercase letters are only an informative selection of some of the most common alternative names. Names in capitals are normative.

[11] The iota adscriptum is unified with the iota subscriptum.

[12] Exists only in combination with α , η , ω as ᾳ , ῃ , ῳ .

¨	U0308	COMBINING DIAERESIS	DIALYTICA
¯	U0304	COMBINING MACRON	Greek macron, length
´	U0301	COMBINING ACUTE ACCENT	
`	U0300	COMBINING GRAVE ACCENT	
˘	U0306	COMBINING BREVE	
ˆ	U0302	COMBINING CIRCUMFLEX ACCENT	
ˇ	U030C	COMBINING CARON	
˚	U030A	COMBINING RING ABOVE	
¨	U0308	COMBINING DIAERESIS	umlaut, trema[15]
˝	U030B	COMBINING DOUBLE ACUTE	
˜	U0303	COMBINING TILDE	
˙	U0307	COMBINING DOT ABOVE	
¸	U0327	COMBINING CEDILLA	
	U0326	COMBINING COMMA BELOW[16]	
’	U0313	COMBINING COMMA ABOVE	psili

[15] Strictly speaking, umlaut and trema can be two typographically slightly different phenomena, but the distinction is increasingly becoming obsolete.

[16] The letters sometimes referred to as small g with comma above and capital g with comma below are to be ordered as small g with cedilla and capital g with cedilla respectively.

̨	U0328	COMBINING OGONEK	
̄	U0304	COMBINING MACRON	

A.8.2 Second level letters

Shape		Position and name of second level letter in ISO/IEC 10646-1	Equiv. FOL[18]
æ	U00E6	LATIN SMALL LETTER AE	ae
Æ	U00C6	LATIN CAPITAL LETTER AE	Ae
ǽ	U01FD	LATIN SMALL LETTER AE WITH ACUTE	áe
Ǽ	U01FC	LATIN CAPITAL LETTER AE WITH ACUTE	ÁE
ǣ	U01E3	LATIN SMALL LETTER AE WITH MACRON	ā e
Ǣ	U01E2	LATIN CAPITAL LETTER AE WITH MACRON	Ā E
ƀ	U0180	LATIN SMALL LETTER B WITH STROKE	b
ɓ	U0253	LATIN SMALL LETTER B WITH HOOK	b
Ɓ	U0181	LATIN CAPITAL LETTER B WITH HOOK	B
ƃ	U0183	LATIN SMALL LETTER B WITH TOPBAR	b
Ƃ	U0182	LATIN CAPITAL LETTER B WITH TOPBAR	B
ƈ	U0188	LATIN SMALL LETTER C WITH HOOK	c
Ƈ	U0187	LATIN CAPITAL LETTER C WITH HOOK	C
đ	U0111	LATIN SMALL LETTER D WITH STROKE	d
Đ	U0110	LATIN CAPITAL LETTER D WITH STROKE	D
Ɖ	U0189	LATIN CAPITAL LETTER AFRICAN D	D

[18] Equivalent on First Ordering Level

ɗ	U0257	LATIN SMALL LETTER D WITH HOOK	d
Ɗ	U018A	LATIN CAPITAL LETTER D WITH HOOK	D
ƌ	U018C	LATIN SMALL LETTER D WITH TOPBAR	d
Ƌ	U018B	LATIN CAPITAL LETTER D WITH TOPBAR	D
ð	U00F0	LATIN SMALL LETTER ETH	d
Ð	U00D0	LATIN CAPITAL LETTER ETH	D
ƍ	U018D	LATIN SMALL LETTER TURNED DELTA	d
ə	U0259	LATIN SMALL LETTER SCHWA	e
Ə	U018F	LATIN CAPITAL LETTER SCHWA	E
Ǝ	U018E	LATIN CAPITAL LETTER REVERSED E	E
ǝ	U01DD	LATIN SMALL LETTER TURNED E	e
ǥ	U01E5	LATIN SMALL LETTER G WITH STROKE	g
Ǥ	U01E4	LATIN CAPITAL LETTER G WITH STROKE	G
ɠ	U0260	LATIN SMALL LETTER G WITH HOOK	g
Ɠ	U01E4	LATIN CAPITAL LETTER G WITH HOOK	G
ɣ	U0263	LATIN SMALL LETTER GAMMA	g
Ɣ	U0194	LATIN CAPITAL LETTER GAMMA	G
ħ	U0127	LATIN SMALL LETTER H WITH STROKE	h
Ħ	U0126	LATIN CAPITAL LETTER H WITH STROKE	H
ƕ	U0195	LATIN SMALL LETTER HV	hv
ı	U0131	LATIN SMALL LETTER DOTLESS I	i
Ɨ	U0197	LATIN CAPITAL LETTER I WITH STROKE	I
Ɩ	U0196	LATIN CAPITAL LETTER IOTA	I
ĳ	U0133	LATIN SMALL LIGATURE IJ	ij

Ĳ	U0132	LATIN CAPITAL LIGATURE IJ	IJ
ƒ	U0192	LATIN SMALL LETTER F WITH HOOK	f
Ƒ	U0191	LATIN CAPITAL LETTER F WITH HOOK	F
ƙ	U0199	LATIN SMALL LETTER K WITH HOOK	k
Ƙ	U0198	LATIN CAPITAL LETTER K WITH HOOK	K
ĸ	U0138	LATIN SMALL LETTER KRA	k
ł	U0142	LATIN SMALL LETTER L WITH STROKE	l
Ł	U0141	LATIN CAPITAL LETTER L WITH STROKE	L
ŀ	U0140	LATIN SMALL LETTER L WITH MIDDLE DOT	l
Ŀ	U013F	LATIN CAPITAL LETTER L WITH MIDDLE DOT	L
ƚ	U019A	LATIN SMALL LETTER L WITH BAR	l
ƛ	U019B	LATIN SMALL LETTER LAMBDA WITH STROKE	l
Ɯ	U019C	LATIN CAPITAL LETTER TURNED M	M
ⁿ	U207F	SUPERSCRIPT LATIN SMALL LETTER N	n
ŉ	U0149	LATIN SMALL LETTER N PRECEDED BY APOSTROPHE	n
ƞ	U019E	LATIN SMALL LETTER N WITH LONG RIGHT LEG	n
Ɲ	U019D	LATIN CAPITAL LETTER N WITH LEFT HOOK	N
ŋ	U014B	LATIN SMALL LETTER ENG	n
Ŋ	U014A	LATIN CAPITAL LETTER ENG	N
ø	U00F8	LATIN SMALL LETTER O WITH STROKE	o
Ø	U00D8	LATIN CAPITAL LETTER O WITH STROKE	O
ǿ	U01FF	LATIN SMALL LETTER O WITH STROKE AND ACUTE	o
Ǿ	U01FE	LATIN CAPITAL LETTER O WITH STROKE AND ACUTE	O
Ɵ	U019F	LATIN CAPITAL LETTER O WITH MIDDLE TILDE	O

Ɔ	U0186	LATIN CAPITAL LETTER OPEN O	O
œ	U0153	LATIN SMALL LIGATURE OE	oe
Œ	U0152	LATIN CAPITAL LIGATURE OE	OE
ƣ	U1A3	LATIN SMALL LETTER OI	o
Ƣ	U01A2	LATIN CAPITAL LETTER OI	O
ƥ	U01A5	LATIN SMALL LETTER P WITH HOOK	p
Ƥ	U01A4	LATIN CAPITAL LETTER P WITH HOOK	P
ɼ	U027C	LATIN SMALL LETTER R WITH LONG LEG	r
Ʀ	U01A6	LATIN LETTER YR	R
ſ	U017F	LATIN SMALL LETTER LONG S	s
ß	U00DF	LATIN SMALL LETTER SHARP S	ss
Ʃ	U01A9	LATIN CAPITAL LETTER ESH	S
ƪ	U01AA	LATIN REVERSED ESH LOOP	S
ŧ	U0167	LATIN SMALL LETTER T WITH STROKE	t
Ŧ	U0166	LATIN CAPITAL LETTER T WITH STROKE	T
ƭ	U01AD	LATIN SMALL LETTER T WITH HOOK	t
Ƭ	U01AC	LATIN CAPITAL LETTER T WITH HOOK	T
ƫ	U01AB	LATIN SMALL LETTER T WITH PALATAL HOOK	t
Ʈ	U01AE	LATIN CAPITAL LETTER T WITH RETROFLEX HOOK	T
ư	U01B0	LATIN SMALL LETTER U WITH HORN	u
Ư	U01AF	LATIN CAPITAL LETTER U WITH HORN	U
Ʋ	U01B2	LATIN CAPITAL LETTER V WITH HOOK	V
ƿ	U01BF	LATIN LETTER WYNN	w
ƴ	U01B4	LATIN SMALL LETTER Y WITH HOOK	y

Ƴ	U01B3	LATIN CAPITAL LETTER Y WITH HOOK	Y
Ʊ	U01B1	LATIN CAPITAL LETTER UPSILON	Y
ƶ	U01B6	LATIN SMALL LETTER Z WITH STROKE	z
Ƶ	U01B5	LATIN CAPITAL LETTER Z WITH STROKE	Z
ʒ	U0292	LATIN SMALL LETTER EZH	z
Ʒ	U01B7	LATIN CAPITAL LETTER EZH	Z
ǯ	U01EF	LATIN SMALL LETTER EZH WITH CARON	z
Ǯ	U01EE	LATIN CAPITAL LETTER EZH WITH CARON	Z
ƹ	U01B9	LATIN SMALL LETTER EZH REVERSED	z
Ƹ	U01B8	LATIN CAPITAL LETTER EZH REVERSED	Z
ƺ	U01BA	LATIN SMALL LETTER EZH WITH TAIL	z
ς	U03C2	GREEK SMALL LETTER FINAL SIGMA	σ
ґ	U0491	CYRILLIC SMALL LETTER GHE UPTURN	г
Ґ	U0490	CYRILLIC CAPITAL LETTER GHE UPTURN	г
ѓ	U0453	CYRILLIC SMALL LETTER GJE	ђ
Ѓ	U0403	CYRILLIC CAPITAL LETTER GJE	Ђ
ќ	U045C	CYRILLIC SMALL LETTER KJE	ћ
Ќ	U040C	CYRILLIC CAPITAL LETTER KJE	Ћ

Annex B (informative): Word-by-word ordering

B.1 Modified terminology

For the purpose of this appendix a **special character** shall be a character that is neither a letter nor a digit nor a diacritical mark nor a spacing character.

NOTE For the purpose of this annex, a spacing character can include all characters which are usually considered to divide words. Typical examples of these might be hyphens, apostrophes and brackets. Cf. also note to A.1.11.

B.2 Principles

Word-by-word ordering is a frequently used alternative to letter-by-letter-ordering. It is a special case of multiple-key ordering which treats space characters as key separators. The maximal string is thus a set of characters enclosed by space characters.

NOTE The string can well be smaller if further keys so demand.

The sets of strings thus obtained are ordered following the European Ordering Rules as specified in the main part of this CEN report.

B.3 Example of Word-by-word vs. letter-by-letter ordering

Letter-by-letter ordering	Word-by-word-ordering
in-	in-
inability	in absentia
in absentia	in extenso
inadvisable	in medias res
in extenso	in memoriam
in medias res	inability
in memoriam	inadvisable

B.4 Simplified word-by-word ordering

If the text to be ordered word by word contains only few second level letters, letters with diacritical marks, or special characters, the following method will in most cases produce the same result as the method that is specified above.

In the *ordering by script* section (A.3.3.1) spacing characters precede digits and letters. The space character is then removed from the table of special characters. The other ordering rules remain unchanged.

Annex C (informative): Ordering by position and by style

C.1 Background

In some cases it is desirable to differentiate further on the *third ordering level*, e. g. in the case where different usages of a word are distinguished solely by the application of some form of internal tagging. This tagging usually takes in print the form of a formatting style. Especially in lexicography it is also often thought to be desirable to distinguish between loan words and native words in such a manner.

This formatting can be expressed by changing the position to the baseline, e. g. in mathematical or chemical formulae, or by highlighting it with certain typographic features, e. g. italic typeface, that serves to indicate some property of the word.

C.2 Recommended rules

This CEN report recommends that, if the implementer deems it necessary to make this differentiation, she or he modify (A.9.2.1) (Capitalization) on the *third ordering level* in the following manner:

Letters are to be arranged in the sequence indicated in this list:

1.small letter on baseline

2. capital letter on baseline

3. small letter above baseline

4. capital letter above baseline

5. small letter below baseline

6. capital letter below baseline

If this does not result in a unique sequence, typographic styles are to be taken into consideration in the sequence listed:

1. roman	abcde
2. boldface	**abcde**
3. italic	*abcde*
4. boldface italic	***abcde***
5. others	

Annex D (informative): Mixed-script ordering with one predominant script

D.1 Background

Many publications — often of the encyclopaedia type — handle scripts differently from this CEN report, especially if they cover predominantly one script with a few entries from other scripts interspersed. They implicitly transliterate strings from other scripts into the predominant one and order according to the rules for that script. For printing the strings are then rendered in their original form. This has the advantage for the user to find related articles e. g. on λόγος and logic near to each other.

D.2 Suggested steps

This may involve the following steps:

— extraction of the strings to be ordered from the relevant data. All preparatory procedures described in the main part of this CEN report may be relevant here;

— implicit transliteration into the predominant script;

— ordering of the strings thus obtained as specified in the main part of this CEN report;

— rendering of strings in their original form, but in the order thus obtained.

D.3 Explicit transliteration

A different, likewise common method is the method of explicit transliteration which selects the transliterated word - e. g. logos – and adds the original rendering in brackets.

Annex E (informative) Repertoire of the Multilingual European Subset No. 3 (MES-3A and MES-3B)

The CEN workshop agreement CEN ISSS CWA 13873 on the *Multilingual European Subsets of ISO/IEC 10646* defines the following repertoires for the open collection MES-3A and the fixed collection MES-3B. They are reproduced for ease of reference.

No.........	Collection name	hex range
1	BASIC LATIN	0020-007E
2	LATIN-1 SUPPLEMENT	00A0-00FF
3	LATIN EXTENDED-A	0100-017F
4	LATIN EXTENDED-B	0180-024F
5	IPA EXTENSIONS	0250-02AF
6	SPACING MODIFIER LETTERS	02B0-02FF
7	COMBINING DIACRITICAL MARKS	0300-036F
8	BASIC GREEK	0370-03CF
9	GREEK SYMBOLS AND COPTIC	03D0-03FF
10	CYRILLIC	0400-04FF
11	ARMENIAN	0530-058F
27	BASIC GEORGIAN	10D0-10FF
30	LATIN EXTENDED ADDITIONAL	1E00-1EFF
31	GREEK EXTENDED	1F00-1FFF
32	GENERAL PUNCTUATION	2000-206F
33	SUPERSCRIPTS AND SUBSCRIPTS	2070-209F
34	CURRENCY SYMBOLS	20A0-20CF
35	COMBINING DIACRITICAL MARKS FOR SYMBOLS	20D0-20FF
36	LETTERLIKE SYMBOLS	2100-214F
37	NUMBER FORMS	2150-218F
38	ARROWS	2190-21FF
39	MATHEMATICAL OPERATORS	2200-22FF
40	MISCELLANEOUS TECHNICAL	2300-23FF
42	OPTICAL CHARACTER RECOGNITION	2440-245F
44	BOX DRAWING	2500-257F
45	BLOCK ELEMENTS	2580-259F
46	GEOMETRIC SHAPES	25A0-25FF
47	MISCELLANEOUS SYMBOLS	2600-26FF
63	ALPHABETIC PRESENTATION FORMS	FB00-FB4F
65	COMBINING HALF MARKS	FE20-FE2F
70	SPECIALS	FFF0-FFFD

Rows	Positions (Cells)
00	20-7E A0-FF
01	00-FF
02	00-1F 22-33 50-AD B0-EE
03	00-4E 60-62 74-75 7A 7E 84-8A 8C 8E-A1 A3-CE D0-D7 DA-F3
04	00-86 88-89 8C-C4 C7-C8 CB-CC D0-F5 F8-F9
05	31-56 59-5F 61-87 89-8A
10	D0-F6 FB
1E	00-9B A0-F9
1F	00-15 18-1D 20-45 48-4D 50-57 59 5B 5D 5F-7D 80-B4 B6-C4 C6-D3 D6-DB DD-EF F2-F4 F6-FE
20	00-46 48-4D 6A-70 74-8E A0-AF D0-E3
21	00-3A 53-83 90-F3
22	00-F1
23	00-7B 7D-9A
24	40-4A
25	00-95 A0-F7
26	00-13 19-71
FB	00-06 13-17
FE	20-23
FF	F9-FD