[From: *Translation and Communication: Translating and the Computer 6*. Proceedings of a conference ... 20-21 November 1984, ed. Catriona Picken (London: Aslib, 1985)]

Handling non-Roman character sets with computers

Hugh McGregor Ross

Data Systems Consultants, Painswick, UK

Translators and bibliographers work with texts, which are made of words, each of which is made of characters. Characters are therefore the basic elements of the crafts of translating and bibliography. For use in equipment, and especially for communication, they must be encoded into binary form.

Character coding involves:

1. identifying a specific application-area;
2. selecting a set of characters;
3. encoding them.

A coded character set for a particular application is a character code.

We speak of graphic characters - letters, numerals and signs - and of control characters or functions, embedded in the data string, to control equipments or to determine layout, emphasis etc, all of which contribute to the meaning the text is intended to convey.

Translators working entirely on their own, with only their own equipment, may use whatever character code they wish. If, however, they wish to communicate - or, more generally, to interchange text in machine-usable form - with any other person or equipment, there is no alternative but to use a standard character code. The situation is shown more specifically in Figure 1.

Between each local domain (e.g. a word processor) and the interchange domain - which may use disk, tape or telecommunications - the 'coding interface' is visualised; this is an imaginary surface through which the coded data

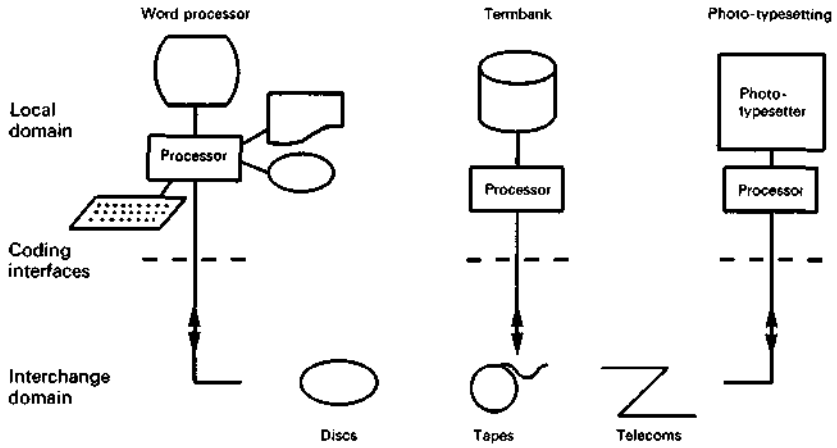

Figure 1. The coding interface in data interchange.

passes, and at which the standard character code is specified and is deemed to be used. This ensures reliable, meaningful, interchange between people and equipment.

If the character code within any one equipment differs from the standard, conversion on the local side of the coding interface is needed; this is usually simple with modern techniques, especially if the ISO sets (Ref. 10) are used. The preferred arrangement is to use the standard code within the local domain, thus avoiding conversion.

The groups that make standard character codes are:

1. ISO/TC97/SC2 - for the general principles of character coding, and the codes for general-purpose use. An active group with many projects of importance to translators in progress.
2. ISO/TC46/SC4 - bibliographic applications. Much excellent work done, based on the general principles from ISO/TC97/SC2, but very slow moving.
3. European Computer Manufacturers' Association - looking after manufacturers' interests; active, and publishes its standards quickly.
4. CCITT - the international standards-making body for services such as teletex and videotex.

The first two international groups have national equivalents within many countries.

Regrettably, there are few translators or bibliographers in ISO/TC97/SC2 and few translators in ISO/TC46/SC4.

This is a most serious situation for the interests of those attending this conference, and one which urgently needs to be rectified. You have a right to take part in this work, and your experience and judgement will be welcomed.

STANDARD CHARACTER CODES (Refs 1, 2, 4, 5)

ISO 646 was the first standard specifically for computers. It is a 7-bit code, providing 32 control characters, Space, and 94 graphic characters. Twelve of these may be adapted to suit the needs of each country. This has made it useful for national purposes, but makes international communication difficult, and translation (i.e. multilingual) work very complicated. Although it has served as the basis of all subsequent character codes, it is obsolescent for bibliographers and translators.

Many available equipments are based on ASCII, the American national variant of this code, or minor variants of it. These are likely to be most troublesome for translators.

For bilingual or multilingual work a character code of greater capacity is required. Figure 2 shows a proposal for the graphic characters of an 8-bit code that is currently being worked on (Ref. 11). Minor improvements have been

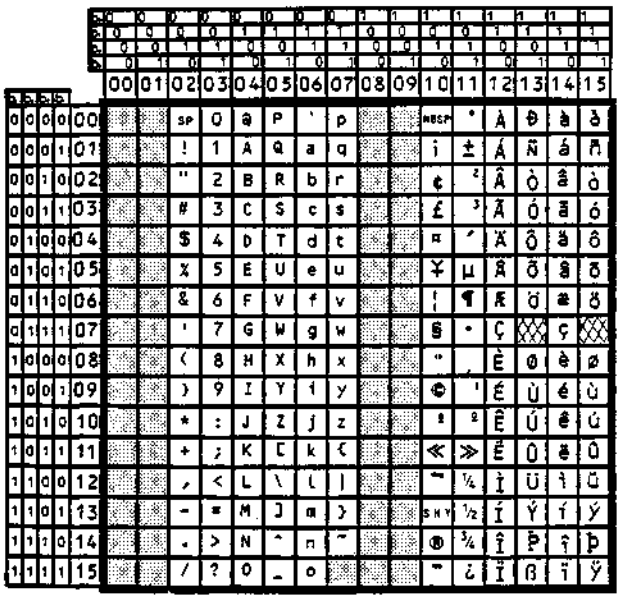

Figure 2. ECMA proposal for graphic characters of an 8-bit multilingual code

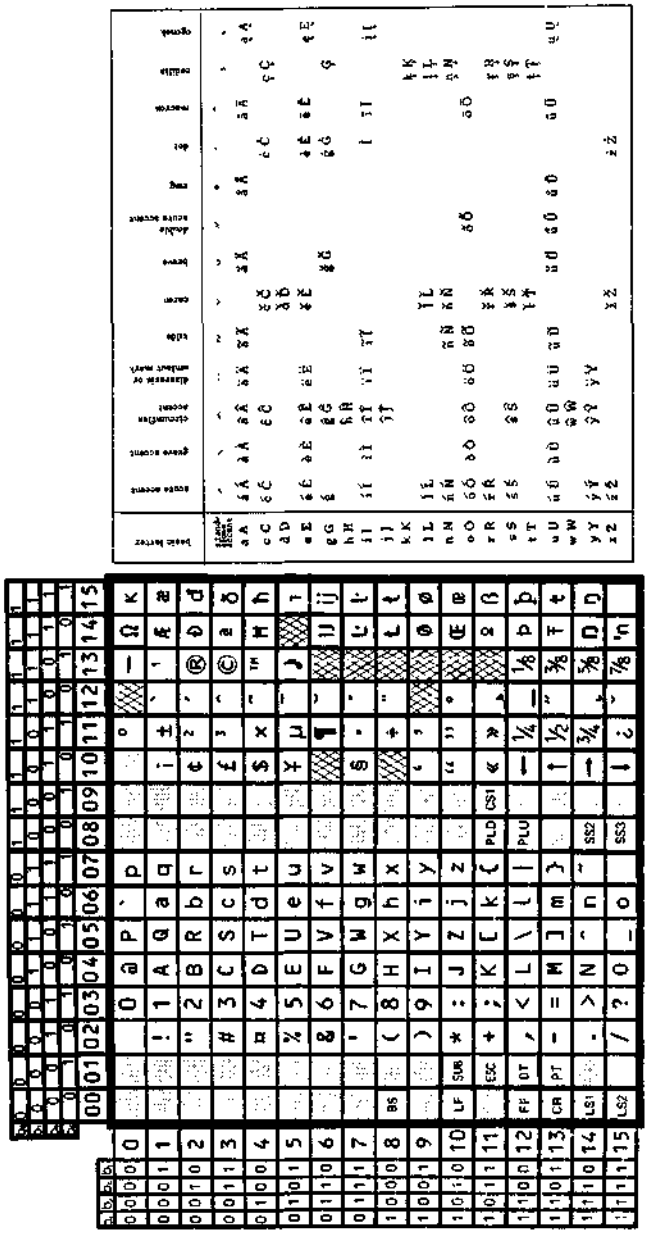

Figure 3. The individually coded characters and the repertoire of accented letters of ISO 6937 Part 2

called for, and the standard may become stable, with equipments based on it available, during 1985. This code provides for general-purpose text (e.g. that used in commerce and industry, and as produced on typical typewriters) in the main languages of the major western European countries and the Americas. It has the advantage that each graphic character is encoded within a single octet of bits, and so may be readily processed with the usual high-level programming languages.

The EBCDIC codes widely used in IBM-type equipment have a similar capacity. These form a very large family, which differ between one and another, and are not standardised. Users of such equipment are advised to request their IBM representative to propose means for any specific bilingual or interchange task.

Text of a literary nature, or in minority languages or those of countries other than western Europe, requires a larger repertoire of accented letters and signs than can be accommodated in the foregoing 8-bit type of code. For these ISO 6937 has been developed (Ref. 10). Part 2 of ISO 6937 covers the needs of all living European languages that use the Roman script, and the requirements of translators (so far as they were known) were honoured in preparing it (Ref. 7). It provides the characters shown in the usual code table in Figure 3 which include a set of 'non-spacing' accents. Every accented letter is encoded as a pair of characters: accent-then-letter; the standard specifies 168 of these, shown on the right of Figure 3. By this encoding technique a total of 344 graphic characters are provided.

Part 3 of the standard gives a comprehensive set of control functions to regulate the presentation of text, and even provides for right-to-left Arabic and top-to-bottom Japanese.

Work is currently in progress on further parts for Greek and Cyrillic scripts (Ref 12), and for technical and publishing applications, e.g. photo-typesetters.

The encoding technique of ISO 6937 (accent-then-letter pairs) is satisfactory for interchange, but is not easy to accommodate in processing, especially with the usual high - level programming languages. Special processing techniques are needed (the work of the Oxford University Computing Service is noteworthy). It also needs a relatively sophisticated technique to implement in displays; however, this is within the local domain.

The needs of bibliographers, and of some translators, may go beyond whatever can be provided in the foregoing types of codes of general applicability. These can be met by exploiting the code extension technique of ISO 2022 (Refs 3, 4/9). Put very briefly, this provides for:

1. An unlimited number of sets of 94 (or 96) graphic

characters, each identified by an Escape sequence, intelligible to humans and processors.
2. A register of such code sets (ISO 2375, Ref 7).
3. Means to 'designate' or select any particular set(s) required.
4. Means to 'invoke' or bring into action the set required at any moment. In effect, the newly invoked code set, or a single character from it, overlays and replaces the characters previously in the code table in use.
5. Means to go back to the original code sets.

Bibliographers in ISO/TC46/SC4 have made many standards exploiting this technique (Refs 1, 4, 5, 7). It is extremely powerful and comprehensive. It does require appropriate features in the processor software, but these are not difficult to provide so long as the technique is understood.

NON-ROMAN SCRIPTS

In any use of computers it is essential to identify the main requirements and the difficulties; when these are dealt with, the easy requirements largely look after themselves. This will be applied to the following review - presented in note form for brevity - of non-Roman scripts in relation to their character coding and their use by bibliographers and translators.

Greek (Refs 5, 6, 8, 16)

1. Three living forms of the Greek language - the new monotonic, Dhimotiki and Katharevousa. Also classical (for scholars) and Kione (for theologians).
2. Very complex use of accents and diacritical marks. These differ markedly in the three living forms of the language.
3. Many pairs of words have the same letters but with different accents, giving significantly different meanings.
4. No authoritative statement of the repertoires of accented letters needed by each form of the language.
5. The placement of the accents/diacritics in relation to their letter is very complex, and requires processing capability in order to give correct presentation.
6. ISO 5428 is a thoroughly-prepared bibliographic standard, but does not specify the repertoire of accented letters, nor provide for the monotonic form.
7. The Greek national standards body has recently issued a 7-bit code based on ISO 646 but with Greek

letters. To provide any accented letters would require the now deprecated technique of using Backspace.

8. The first proposal for Part 7 of ISO 6937 for Greek (Ref 12), in trying to follow ISO 646, clashes in every way with ISO 5428. To have two different international standards related to the same topic always leads to trouble. These clashes are unnecessary.
9. Several registered code sets (Refs 4, 7), none of which satisfy Greek as a living language.
10. No Greek authorities take part in international character coding standards-making.
11. There is no difficulty in making a Greek keyboard (for which there is a Greek standard), nor a dual Greek-Roman keyboard.

Cyrillic (Refs 5, 6, 8, 9, 17)

1. No Russian authorities take part in international standards-making, although there is a well-established and satisfactory 8-bit Russian national standard - GOST 19768-74, (see Ref 4/9 p. A1.2). It is dual Cyrillic-Roman.
2. Regrettably, some well-informed persons in the west are precluded from giving assistance on military grounds.
3. The major languages using the Cyrillic script are Russian, Byelorussian, Ukrainian, Bulgarian, Serbian, Macedonian. However, there are about eighty other languages that use it (Refs 6, 17) but there is no consensus over which should be considered for character coding work.
4. The major languages use a few accented letters, but the minor languages use many more (Refs 6, 9); no definitive information on this is available.
5. Romanian uses some special letters, which are not easy to include in any general scheme although they are available in the bibliographic standard ISO 6861.
6. There are several other bibliographic standards (Refs 1, 4/10) relating to Cyrillic.
7. The first proposal for Part 8 of ISO 6937 for Cyrillic (Ref 12) provides for the accented letters of the major languages (see 3 above). In doing so, it precludes the other accented letters (see 4 above), the special letters and accented letters of the countries that have close relations with Russia but use the Roman script (ISO 6937 Part 2), and translation to or from any language other than English.

8. There is no difficulty in making Cyrillic or dual Cyrillic-Roman keyboards.

Arabic (Refs 8, 14)

1. No Arabic representatives take part in international standards-making, although correspondence takes place.
2. There are some Arabic standards related to the 7-bit ISO 646 (some in Ref 7).
3. Arabic reads from right to left, but numbers left-to-right; two types of numerals are used.
4. The Arabic language takes three forms.
5. Many other important languages than Arabic use the Arabic script. Many of these use a significantly larger number of letters. There are no definitive statements about these (Ref 8).
6. Each consonant may take three forms - initial, medial or final. This requires significant processing capability for presentation, and it is difficult to attain the objective of a cursive script running throughout each word.
7. The vowels, from the point of view of meaning and encoding, have to be treated as independent letters. However, for presentation they have to be treated as 'vowel-signs' and be placed above or below a related consonant. The situation is made more complex in that they are omitted in some classes of text. All this requires processing capability.
8. Arabic text may contain words transliterated into Roman; these require 4 accents.

Hebrew (Refs 8, 15)

1. So far has been very little considered.
2. Reads from right to left.
3. Five of the consonants have different forms when at the end of words.
4. Vowels not usually written, but when they are they comprise a complex set of 'vowel-signs' above, below or within a consonant. Processing capability required.
5. Complex jots and tittles, which may be omitted in certain classes of text.

Devenagari (Refs 8, 15)

1. Some work is thought to be going on in India, but this has not reached the international scene.
2. Main uses in Sanskrit and Hindi.

3. Thirteen vowels have initial forms at start of words, but different medial forms; they then appear as 'vowel-signs' supported by a consonant - before, above, after or below. This would require processing capability for presentation.
4. Transliteration into Roman (especially the English language) is required. There are various well-established schemes but none are sufficiently authoritative to determine the repertoire of accents and accented-letters that is required. Anyway, those in ISO 6937 are inadequate.

Japanese Katakana (Refs 7, 15)

1. A phonetic alphabet or syllabary.
2. Requires 63 symbols.
3. Uses western-Arabic numerals.
4. Japanese national standard exists as 7-bit code (JIS C-6220).

Chinese Hanzi/Japanese Kanji

1. Over 40,000 ideographic symbols.
2. Standard character codes exist (GB 2312-80/JIS C-6226-1983) specifying 6763/6353 ideographic symbols. Many of these symbols are the same in both standards, but it is considered that the differences are such as to make it impractical to contemplate making a composite set.
3. These standards also identify subsets of 3755/2965 of these, being those most frequently used in simple texts.
4. Other groups of Chinese/Japanese characters and numerals are included.
5. Roman, Greek and Cyrillic scripts are included. The Chinese standard includes the small Roman accented letters used in Pinyin, but not the full set needed for translation to the European languages.
6. Japanese trade would require the full set of European-Roman and also Cyrillic accented letters.
7. Each character is encoded as a pair of 7-bit groups. This gives code tables of 94 rows and 94 columns. In the standards these are presented split up into 12 smaller tables.
8. Text is written downwards, with lines following from right to left; or may be written in lines from left to right, lines following downwards. Controls in ISO 6937 Part 3 Add 1 provide for this.
9. Display and printer equipment capable of presenting

these very large and complex character sets has already become available.

10. Several schemes for keyboards to input these character sets exist, but all seem to require substantial skill by the operator.

A SOLUTION FOR BIBLIOGRAPHERS' AND TRANSLATORS' NEEDS

The foregoing has illustrated how various applications, of increasing complexity, demand character sets of increasing size, and how, when any one critical size is exceeded, it becomes necessary to use a different technique of encoding.

It is also apparent that bibliographers and especially translators require large multilingual or even dual-script character sets. Although these can be provided for by 7-bit or 8-bit coded sets together with clever use of the code extension techniques (designation, invocation and shifts), this complicates programming and is difficult to implement in hardware designed for more modest uses.

Accordingly, work that has just started on a standard 16-bit character code will be of especial interest and value to bibliographers and translators (Refs 13, 14). The aim of this work is to provide a single coded character set, which will include all the living languages and scripts considered to be of sufficient importance. Each character, including every one of the accented letters, will be coded uniquely within a single 16-bit combination. This will make it easy to process this code with most of the high-level programming languages. It may also be used in any competent processsor that can put two octets of bits together. When it is implemented there should be very little need to employ the code extension shifting technique.

Such a 16-bit code will be conceived of as being in a code table of 256 rows (the most-significant octet) and 256 columns (the least-significant octet). Current work is concentrating on establishing the basic structure of this code table, to optimise the space available and to make it easy to slot in existing standardised code sets. An important point is that the accented letters currently attained in ISO 6937 and the bibliographic standards by accent-then-letter sequences, will all be encoded as single 16-bit characters.

This new 16-bit character code could provide an entirely satisfactory solution to the requirements of bibliographers and translators - provided you join in the work.

REFERENCES

1. H.McG. ROSS. Standards for computers and office automation. Chapter in Computers Users' Year Book, Computing Publications Ltd, London. A comprehensive and regularly updated listing of these standards, including those for bibliographers and translators.

2. Character codes as they affect the user. Ibid.

3. An outline of the code extension concept. Ibid. A simple explanation of the very complex standard ISO 2022, and an updated list of the registered code sets.

4. NCC Guides to Computing Standards:
No 9. Character sets and coding
No 10. Automation in bibliography
No 11. Transliteration
National Computing Centre, Manchester. Although getting out of date over details, these remain the only comprehensive discussion of the standards, and their inter-relation, in each field. Essential reading for anyone seeking to use these standards.

5. H.McG. ROSS. The application of standardized character sets to multilingual text communication and processing. Prepared for the Commission of the European Communities (CEC DG-XIII). 1980. A thorough discussion of the use of character codes within a multilingual environment with major translation requirements. Includes Greek and Cyrillic.

6. M. PIOLLE. Annexes to Report Etudes préalables à 1'élaboration d'un code universel pour la présentation et le traitement de textes multilingues dans un contexte européen. CREL-France, 1980. Done for CEC. A very comprehensive survey of the requirements, especially for accented letters, of the languages that use the Roman, Cyrillic and Greek scripts.

7. International register of coded character sets to be used with escape sequences. From European Computer Manufacturers' Association, Geneva. The formal Register of code sets produced under ISO 2375, introduced in Ref. 3 above.

8. H.McG. ROSS. Working documents ISO/TC97/SC2/WG4 N 230 (1980) and N 347 (1981). Although the detailed proposals of these papers were not taken up, they contain collated information on the character sets

required for languages that use the Greek, Cyrillic, Arabic, Hebrew and Nāgāri scripts.

9. Peter W. FENWICK. Working document ISO/TC97/SC2/WG4 N 432 (1983). Summary of characters required by minority languages that use the Cyrillic script.

10. ISO 6937: Coded character sets for text communication.
Part 1 General introduction
Part 2 Latin alphabetic and non-alphabetic graphic characters
Part 3 Control functions for document interchange.
International standards may be obtained from each national standards body.

11. ISO/DIS 8859: Part 1: 8-bit single byte coded character sets: Latin alphabet no. 1.

12. Working document ISO/TC97/SC2/WG4 N 521 (1984).

13. H.McG. ROSS. Working document ISO/TC97/SC2/WG2 N 370. Trial of fitting the Roman, Cyrillic and Greek scripts into the 16-bit character code. 1985.

14. Joseph D. BECKER. Multi-lingual word processing. Scientific American, 251 (1), July 1984. A paper which will contribute to the character sets that the 16-bit code will contain.

15. John CLEWS. World Scripts Paper for this conference.

16. H.McG. ROSS. Working document ISO/TC97/SC2/WG2 N 35. The requirements of the Greek language and script. 1985.

17. H.McG. ROSS. Working document ISO/TC97/SC2/WG2 N 34. The requirements of languages that use the Cyrillic script. 1985.

Author

Hugh McGregor Ross, Data Systems Consultants,
Simmondley, Queensmead, Painswick, Gloucestershire,
GL6 6XA, UK.