[*Translating and the Computer 8*. Proceedings of a conference… 13-14 November 1986, ed. Catriona Picken (London: Aslib, 1987)]

Recent developments in IT Standards of interest to translators

Peter Fenwick

SUMMARY

'National variant' character codes limited to 94 printable characters and catering for a single language are obsolete. There is an International Standard character code which provides a repertoire of over 300 characters; forthcoming Greek and Cyrillic-alphabet additions to that Standard will make it cater for *all* modern European languages. Equipment capable of printing or displaying almost 500 different characters will be available, but not cheaply.

For those with a limited budget and more limited requirements, hardware and software handling multilingual 'regional variant' sets of 188 characters will soon be available. One such set will cater for all the major languages of Western Europe, and will become a de facto as well as a formal Standard far sooner than many people expect.

The multilingual character-code cited in all the new generation of 'open systems' and 'electronic office' Standards and protocols for communication between information technology (IT) equipments is specified in International Standard ISO 6937/2 'Coded character sets for text communication', published in 1983 and since adopted as a national Standard in several countries including the United Kingdom. (British Standard BS 6692: Part 2 is cheaper to buy than the ISO version, and has some misprints corrected.)

Subsets of ISO 6937/2 (omitting a few non-alphabetic characters) were adopted by the Comité Consultatif International des Télégraphes et Téléphones (CCITT) in their 1980 recommendations for Teletex ('super Telex') and Videotex, and confirmed in 1984.

ISO 6937/2 is used exactly in the CEPT (club of European PTTs) Videotex Standard towards which UK Prestel, French Télétel and German Bildschirmtext are supposed to be converging. The North American Presentation Level Protocol (NAPLPS) videotex system developed in Canada and embraced by the American Telephone and Telegraph Company (AT&T) in the United States uses ISO 6937/2 with a few characters added.

ISO 6937/2 is cited in the International Standards for office document architecture (ODA) and message handling systems (MHS) due to be finalised next year. In short, it has been widely accepted for international communication, where many languages must be catered for.

ISO 6937/2 only claims to cater for 41 languages (39 European ones plus Esperanto and Afrikaans) which use the Latin alphabet, but it can cover many more including languages from Oceania, Africa and South America. Earlier this year, I tracked down alphabets for Romano lil and Romsko pismo agreed by the 1981 International Gipsy Convention: ISO 6937/2 caters for them too.

Work on a revised version of ISO 6937/2 is under way. It will have ‘dot under’ and ‘line under’ as additional diacritical marks to make provision for the languages such as Yoruba which use one or other of those marks and for Latin-alphabet transliterations of Arabic and Indian sub-continent scripts.

In developing ISO 6937/2 (see Table 1), we catered for some relatively obscure languages such as Greenlandic, Lappish, and Sorbian, but only for current, not obsolete orthographies. An annex to ISO 6937/2 lists which letters are used in which language, with a careful disclaimer that the list is intended as a justification for including those letters in the character code, not as specifying what should or should not be used in a particular language.

In retrospect, we made some mistakes: we recognised that some people would always use two letters for a currency abbreviation such as Fr so an alternative single symbol ₣ would only make life complicated, but we yielded to people who had seen ij, œ and ŉ on Dutch, French and Afrikaans typewriter keyboards and insisted they were language-specific single characters. It was not Dutch or Flemish speakers who insisted upon ij, or Frenchmen who demanded œ. Spaniards who did not speak Catalan claimed that ‘L-dot’ would be essential on videotex terminals in Barcelona.

The essential feature of ISO 6937/2 is that it includes ‘floating’ diacritical marks, and codes the 150-plus accented letters it provides not as single characters but as accent-plus-letter: to be transmitted and stored in a computer as two bytes. This caters for many languages in a single code, but that accent-plus-letter coding is incompatible with a great deal of existing software.

Even completely new software will be simpler if it does not have to cope

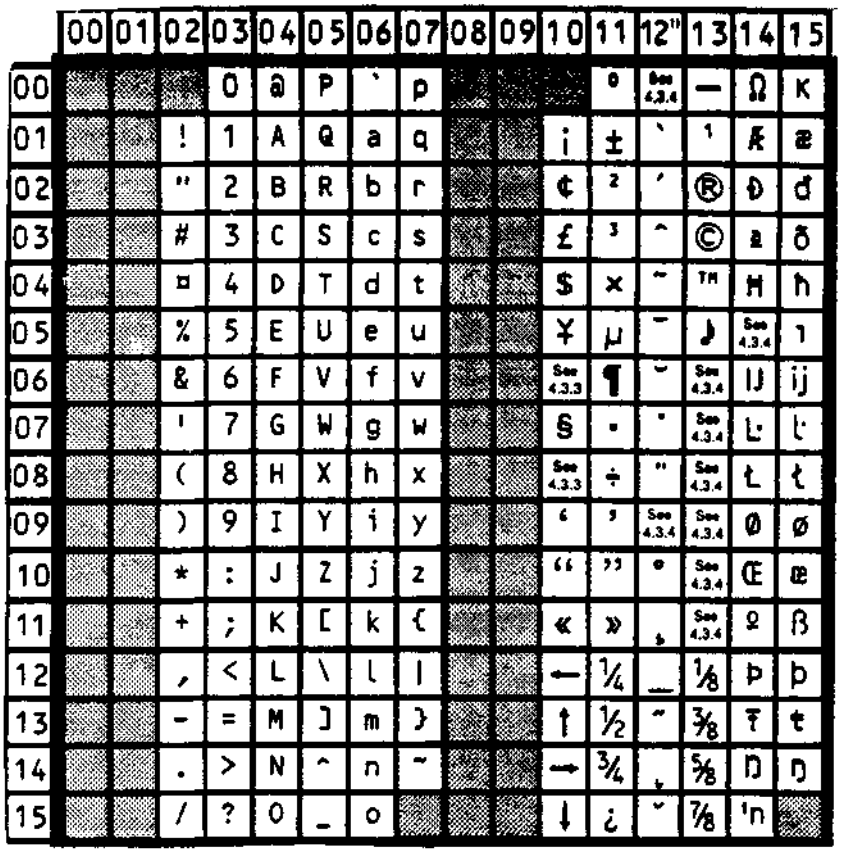
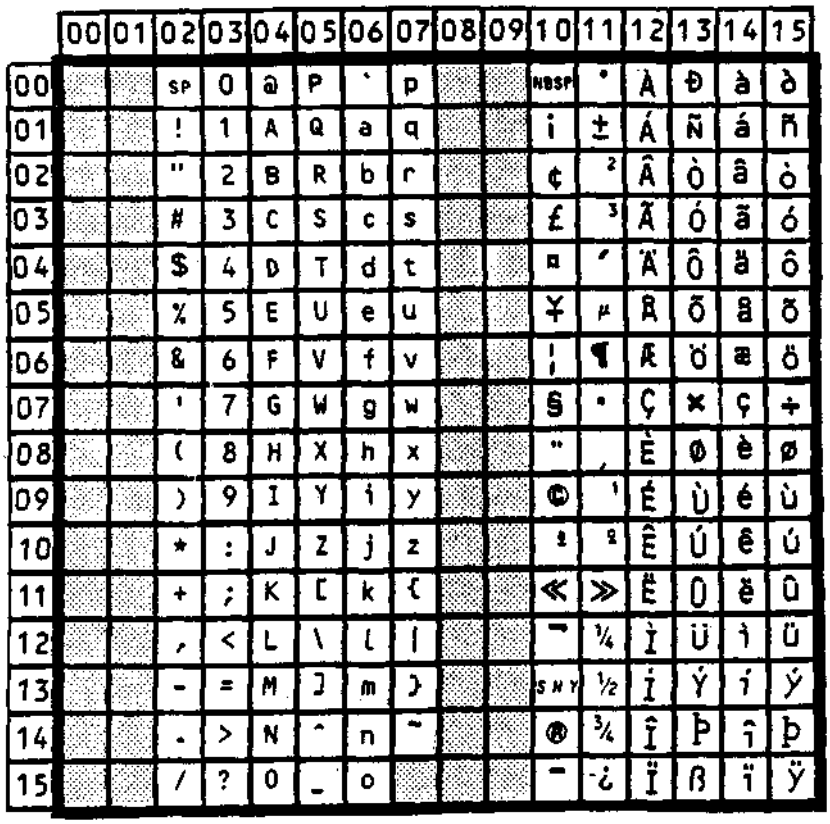


Table 1. ISO 6937/2

with having some characters occupy one byte of memory and others two. The large repertoire and accent-plus-letter coding of ISO 6937/2 are not liked by vendors whose hardware and software cover only 94 printable characters and cannot easily be modified to cope with more than twice that number: they would prefer to supply simpler (and cheaper) hardware and software catering for fewer languages.

Largely as a result of initiatives by the European Computer Manufacturers' Association (ECMA) (many of whose members have factories in Europe but head offices in the United States or Japan), a family of new character sets is being introduced, each of which can handle English and one or more other languages by including all the 94 printable characters of the old American Standard Code for Information Interchange (ASCII) and up to 94 further characters.

One such set (Table 2) will be known in the United States as 'ASCII-8 and supersede the present ASCII: in the terminology of American sup-



	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15
00			SP	0	@	P	`	p			NBSP	°	À	Ð	à	ð
01			!	1	A	Q	a	q			¡	±	Á	Ñ	á	ñ
02			"	2	B	R	b	r			¢	²	Â	Ò	â	ò
03			#	3	C	S	c	s			£	³	Ã	Ó	ã	ó
04			$	4	D	T	d	t			¤	´	Ä	Ô	ä	ô
05			%	5	E	U	e	u			¥	µ	Å	Õ	å	õ
06			&	6	F	V	f	v			¦	¶	Æ	Ö	æ	ö
07			'	7	G	W	g	w			§	·	Ç	×	ç	÷
08			(	8	H	X	h	x			¨	¸	È	Ø	è	ø
09			)	9	I	Y	i	y			©	¹	É	Ù	é	ù
10			*	:	J	Z	j	z			ª	º	Ê	Ú	ê	ú
11			+	;	K	[	k	{			«	»	Ë	Û	ë	û
12			,	<	L	\	l	|			¬	¼	Ì	Ü	ì	ü
13			-	=	M	]	m	}			SHY	½	Í	Ý	í	ý
14			.	>	N	^	n	~			®	¾	Î	Þ	î	þ
15			/	?	O	_	o				¯	¿	Ï	ß	ï	ÿ

Table 2. ASCII-8

pliers, 'foreign language' characters will then be catered for by 'domestic' equipment.

ASCII-8 will not be published as an American Standard until 1987, but it was adopted by ECMA in their Standard ECMA-94 in 1985, and should appear in an International Standard in 1986: ISO 8859/1 is in the queue at the printers in Geneva. ASCII-8 will not come as a surprise to most IT vendors, but some of them may be surprised by the rapidity of its acceptance.

Rumours in the industry say IBM will unveil a new range of personal computers codenamed Renegade 'early in 1987'. Before Easter 1986, the principal suppliers of software for the personal computer (PC) had been informed that IBM PCs sold in the Americas and Western Europe from 1987 onwards would support ASCII-8.

Few people anticipated the impact the original IBM PC would have upon IBM itself, other established computer manufacturers and the new sup-

pliers who a few years ago were better known for their cameras, photocopiers or sewing machines. Now the computer trade press says 'compatible' without bothering to add 'with the IBM PC' and no-one will doubt that where the IBM PC leads, everyone else's PCs (and 'PC compatible' printers, etc.) will follow.

A year hence, relatively inexpensive IT equipment will cater for all the printable characters used in Danish, Dutch, Finnish, French, German, Icelandic, Italian, Norwegian, Portuguese, Spanish, and Swedish (and Albanian, Basque, Breton, Corsican, Faroese, Frisian, Galician, Irish and Scots Gaelic, Occitan, and Rheato-Romance too). Whether the accompanying software and manuals will also cater for many of those languages is quite a different matter, more to do with what translators can sell to IT suppliers than what they might buy from them.

At least this time everyone will know what lead they are following: most vendors will be glad to follow a well-defined Standard, to abandon other character sets which are almost but not quite like ASCII-8, and especially to abandon the odd mixture which is the present IBM PC character set.

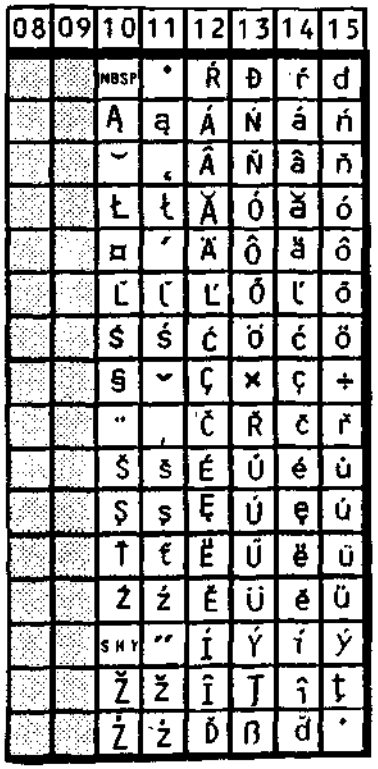
That set was defined in Boca Raton, Florida, apparently without advice from further afield than Miami or Palm Beach. To the chagrin of IBM product planners almost everywhere else, it had some curious inclusions and omissions: for example Æ/æ, Å/å and something which looked like Ø were included, but small ø was not; Ä, Ö, Ü, ä, ö and ü were there, but ß was not; on the other hand there was something that looked like it between alpha and gamma in the code table ....

People trying to make PC-compatible equipment often did not know what some characters were supposed to be: more than one 'PC-clone' has shown the letters FE squeezed together in place of æ, had a small c-cedilla looking like the old Greek koppa and the capital C-cedilla shown as overlapping letters C and L: perhaps someone genuinely thought it was supposed to be the sign which botanists use to denote a clone.

Such nonsenses should not recur. Manufacturers of dot-matrix printers, VDUs and cheap portable terminals/word processors will not stop using their own slightly different 'multilingual' character sets overnight, or stop offering odd characters for compatibility with old PCs, but new equipment will support ASCII-8 and there will be no uncertainty about what characters are or are not supported.

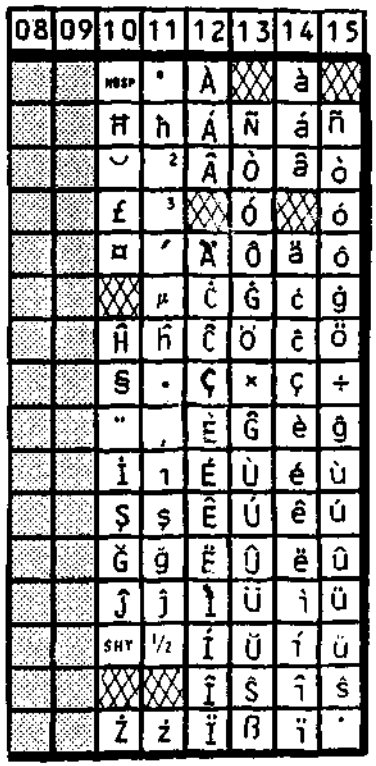
The major languages of Eastern Europe which are written in the Latin alphabet are catered for by the set of ISO 8859/2 which is also in the queue at the printers in Geneva, and is already published in ECMA-94. Table 3 shows the right half of the code table which can replace the right half of ASCII-8 in equipment with only 188 printable characters.

That set and ASCII-8 will be mutually-exclusive alternatives, not both available at once. Cheap equipment will be aimed at mass markets. Dot



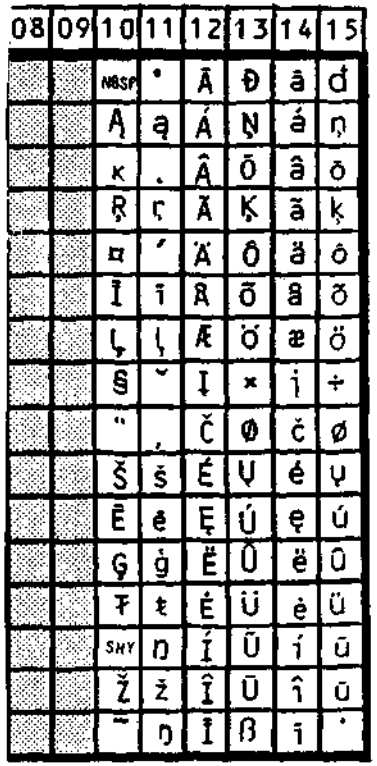
08	09	10	11	12	13	14	15
		NBSP	˚	Ŕ	Đ	ŕ	đ
		Ą	ą	Á	Ń	á	ń
		˘	˛	Â	Ň	â	ň
		Ł	ł	Ă	Ő	ă	ő
		¤	´	Ä	Ô	ä	ô
		Ľ	ľ	Ĺ	Ő	ĺ	ő
		Ś	ś	Ć	Ö	ć	ö
		§	ˇ	Ç	×	ç	÷
		¨	¸	Č	Ř	č	ř
		Š	š	É	Ú	é	ù
		Ş	ş	Ę	Ú	ę	ú
		Ť	ť	Ë	Ü	ë	ü
		Ź	ź	Ě	Ü	ě	ü
		SHY	˝	Í	Ý	í	ý
		Ž	ž	Î	Ţ	î	ţ
		Ż	ż	Ď	ß	ď	˙

Table 3. East European languages table (right half)



08	09	10	11	12	13	14	15
		NBSP	˚	À		à	
		Ħ	ħ	Á	Ñ	á	ñ
		˘	²	Â	Ò	â	ò
		£	³		Ó		ó
		¤	´	Ã	Ô	ã	ô
			µ	Ċ	Ġ	ċ	ġ
		Ĥ	ĥ	Ĉ	Ö	ĉ	ö
		§	·	Ç	×	ç	÷
		¨	¸	È	Ĝ	è	ĝ
		İ	ı	É	Ù	é	ù
		Ş	ş	Ê	Ú	ê	ú
		Ğ	ğ	Ë	Û	ë	û
		Ĵ	ĵ	Ì	Ü	ì	ü
		SHY	½	Í	Ŭ	í	ŭ
				Î	Ŝ	î	ŝ
		Ż	ż	Ï	ß	ï	˙

Table 4. Mediterranean languages table (right half)



08	09	10	11	12	13	14	15
		NBSP	˚	Ā	Đ	ā	đ
		Ą	ą	Á	Ņ	á	ņ
		ĸ	¸	Â	Ō	â	ō
		Ŗ	ŗ	Ã	Ķ	ã	ķ
		¤	´	Ä	Ô	ä	ô
		Ī	ī	Å	Õ	å	õ
		Ļ	ļ	Æ	Ö	æ	ö
		§	ˇ	Ī	×	į	÷
		¨	˙	Č	Ø	č	ø
		Š	š	É	Ų	é	ų
		Ē	ė	Ę	Ú	ę	ú
		Ģ	ģ	Ë	Û	ë	û
		Ŧ	ŧ	Ė	Ü	ė	ü
		SHY	ŋ	Í	Ũ	í	ũ
		Ž	ž	Î	Ū	î	ū
		¯	ŋ	Ï	ß	ī	˙

Table 5. Baltic and Scandinavian languages table (right half)

matrix printers selling for a few hundred pounds (and laser printers selling for a few thousand) may have repertoires of 600 characters, but are more likely to provide three styles of 200 or so characters than two styles of 300 — italic and bold face are more attractive to most customers than more characters for more languages.

There are two other ASCII-plus Latin alphabets in ECMA-94; their right halves are Tables 4 and 5. Table 4 caters for Mediterranean languages including Turkish and Maltese, with Esperanto thrown in for good measure. Table 5 covers Baltic and Scandinavian languages, including Lappish and Greenlandic, but not Faroese or Icelandic. Proposals to adopt those sets as International Standards are stalled, partly because of doubt about the need to cope with those particular combinations of languages rather than others.

ECMA aimed to have all the languages catered for by ISO 6937/2 catered for by one or other of the sets in Tables 2 to 5, but they missed Welsh, and only foresaw the more obvious combinations of languages. Basically, they were thinking of English and one other language as the need to be met, and it just happened that it was easy to make all four sets cater for German too.

If you do want a combination of languages not provided in one of Tables 2 to 5, it is unlikely that you will find inexpensive equipment to meet those needs. French plus Polish, Czech plus Spanish, or Turkish plus Swedish will to equipment vendors be ‘exotic’ combinations. If you want to write (say) both French and Polish, more versatile hardware and software implementing ISO 6937/2 will be the only answer.

There will be variant ‘188 character’ equipments which cater for Cyrillic or Greek as well as the old ASCII. For monotoniko Greek, Table 6 shows the right half of the ‘ASCII-plus’ character set which should be adopted as a Greek National Standard and an ECMA Standard before the end of 1986, and as an ISO Standard in 1987. Table 7 is the right half of an ASCII-plus set which caters for all the Slavic languages (and Moldavian and most of the Caucasian languages) written in Cyrillic; it is already adopted as an ECMA Standard, and will be an ISO Standard next year. Those sets will provide for English and Greek or Cyrillic, but not for accented Latin-alphabet letters at the same time.

Again, needs for more than 188 characters will be satisfied by ISO 6937. Minus the non-alphabetic characters (already provided in ISO 6937/2) the sets of Tables 6 and 7 are those of ISO Draft Proposals DP 6937/7 and DP 6937/8, which should reach Draft International Standard status before the end of 1986, and be ISO Standards before the end of 1987.

Parts 2, 7 and 8 of ISO 6937 will be able to be used together in a single character-code catering for all European languages. Use of Part 7 or Part 8 in conjunction with Part 2 of ISO 6937 will probably be adopted by the CCITT as ‘Standard Options’ for the international Teletex service in 1988.

The ASCII-plus-Arabic character set (Table 8) is again already an

Table 6. Greek language table (right half)

Table 7. Cyrillic Slavic languages table (right half)

Table 8. Arabic/Latin scripts table (right half)

ECMA Standard expected to become an ISO Standard in 1987. It provides for communication of the information content of Arabic text, not simple instructions for its presentation, which needs all the variant shapes of the letters: displaying or printing English and Arabic with reasonable quality is beyond the capability of simple variants of '188-character' equipment.

Proposals will be made in 1987 to add this set to ISO 6937 to cater for mixed Arabic and accented Latin-alphabet script, for example Arabic and French (and/or Spanish for Morocco). The set of Table 8 could be adopted as it is, but either immediately or subsequently, further characters could be added to it (without displacing or moving characters already there) to provide for Farsi, Urdu, Sindhi, Malay and many of the other languages which use an extended Arabic alphabet.

Looking a few years ahead, there will be one ISO character code which caters for *all* modern languages, including Chinese, Japanese and Korean. Work on this 'two-byte' code has been under way in an ISO Working Group for two years now.

At a meeting in Kyoto last month, we finalised two alternative proposals: one an 'evolutionary' approach providing for the coding of 36,000 characters, and a more radical design with scope for 64,000. The result of a ballot of national Standards bodies to choose between the two will be known by Easter 1987, and before the end of that year, we should have a first Draft Proposal for an International Standard character set catering for all languages written in the Latin and Cyrillic alphabets, classical as well as modern Greek, and at least the first 13,000 characters for Chinese and Japanese.

A possible spin-off of this work will be a sub-set code of about 3,000 characters omitting Chinese, Japanese and Korean, but catering for all other modern languages. In a few years time, it may be that there will be relatively inexpensive printers and displays which have not 188 but those 3,000 characters: after all, there are now more than a dozen different portable word processors to be seen in Japanese department stores which display and print more than twice that number.

AUTHOR

Peter W. Fenwick, information technology consultant, London, UK.