[*Translating and the computer 10*. Proceedings of a conference … 10-11 November 1988, ed. Pamela Mayorcas (London: Aslib, 1990)]

Multilingual word processing for translation

David C. Jackson

Managing Director, Vuman Computer Systems Ltd, UK

INTRODUCTION

This paper looks at the difficulties experienced by professional translators in the search for word processing system which allow full exchange of documents between different systems without the need for special file translation programs or utilities. It is the author's view that document preparation systems can be well represented by the pyramid shown in Figure 1 in which, with the exception of the typewriter, a document may begin its life at any point in the pyramid, although it would normally start near the bottom. The document will then pass through a series of edits, following which it may begin to percolate up the pyramid, stopping at the level most appropriate for the quality of output required. Thus many internal documents need never go beyond basic word processing, whereas documents for limited or internal promotion may go through to desktop publishing (DTP) while documents for high volume or external consumption would be sent to a professional typesetter. This ideal situation will only occur if computer-readable files can be transferred from one level to the next without loss of information. It should also be possible to transfer files in a horizontal direction between word processing systems made by different manufacturers. Regrettably this is not the case and this paper looks at the historical reasons for this and also suggests a way forward to achieve these aims.

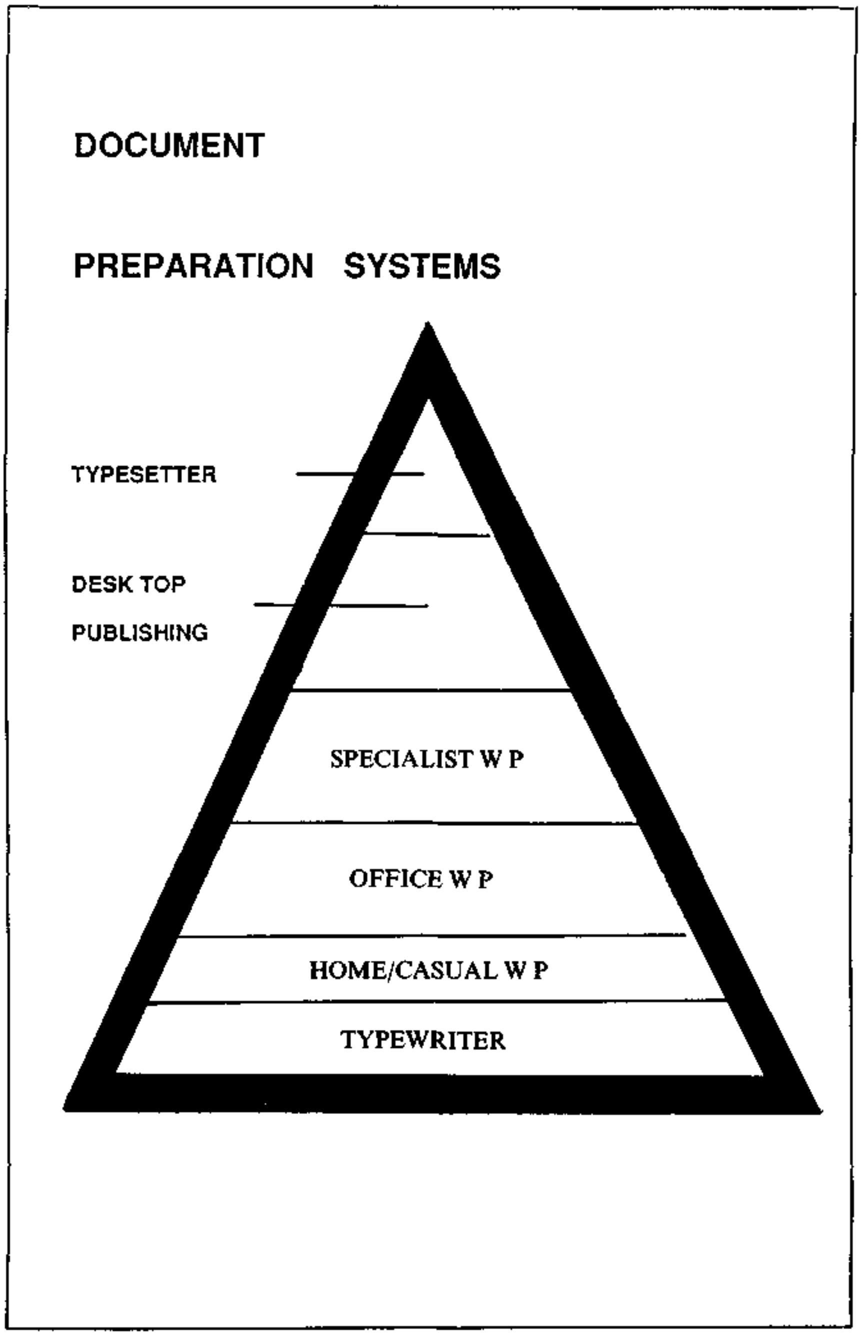
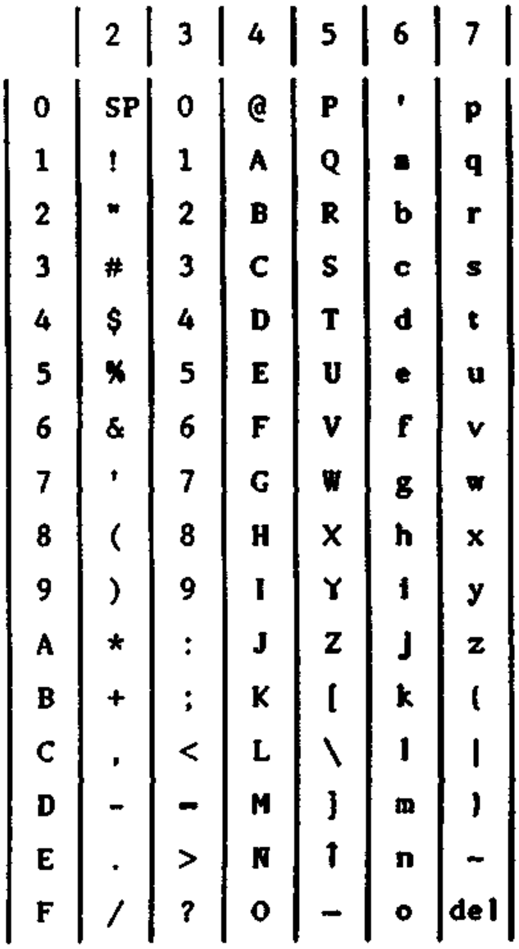


Figure 1. Document preparation pyramid

SIMPLE AND EXTENDED CHARACTER SETS

In the beginning there was WordStar, as the first word processing program for microcomputers (dedicated word processing systems will not be covered here because they are expensive and fast becoming obsolete). The WordStar character set was simple (Figure 2) and caused little confusion, using the American Standard Code for Information Interchange (ASCII). As a consequence of this almost any word processor could read WordStar files and compatibility problems were slight. For the translator, however, the WordStar character set was less than perfect since it did not provide for accented characters.



	2	3	4	5	6	7
0	SP	0	@	P	`	p
1	!	1	A	Q	a	q
2	"	2	B	R	b	r
3	#	3	C	S	c	s
4	$	4	D	T	d	t
5	%	5	E	U	e	u
6	&	6	F	V	f	v
7	'	7	G	W	g	w
8	(	8	H	X	h	x
9	)	9	I	Y	i	y
A	*	:	J	Z	j	z
B	+	;	K	[	k	{
C	,	<	L	\	l	|
D	-	=	M	]	m	}
E	.	>	N	↑	n	~
F	/	?	O	—	o	del

Figure 2. 7-bit simple ASCII character set

In the early 1980s a number of suppliers realised that if 8-bit codes were used for character set representation, these could provide approximately twice as many characters as were available in WordStar or ASCII. Vuman was one of the first companies in the UK to design a word processor offering an extended character set and to produce utility programs for generating character shapes for screen display and printing (Figure 3).

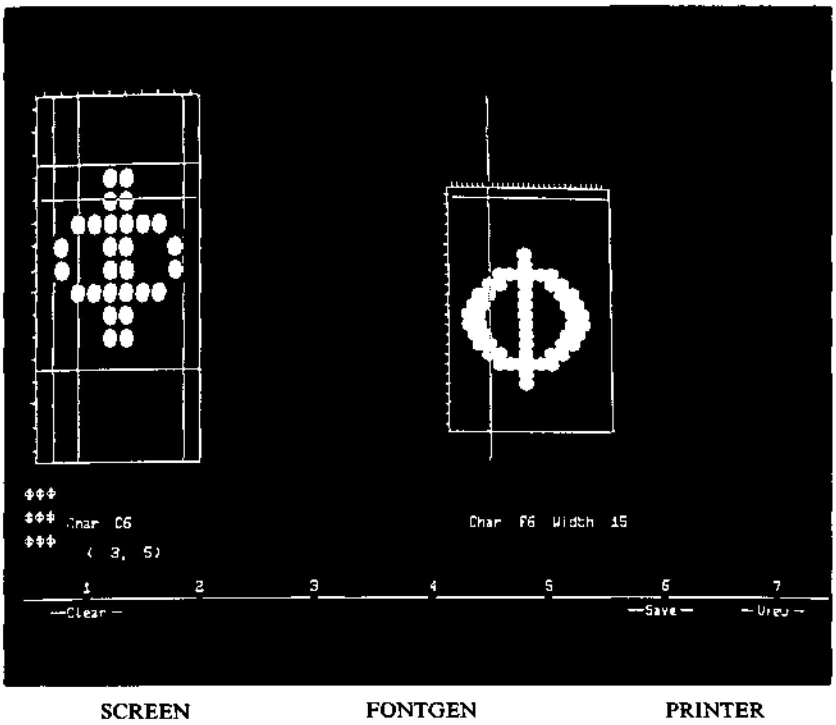
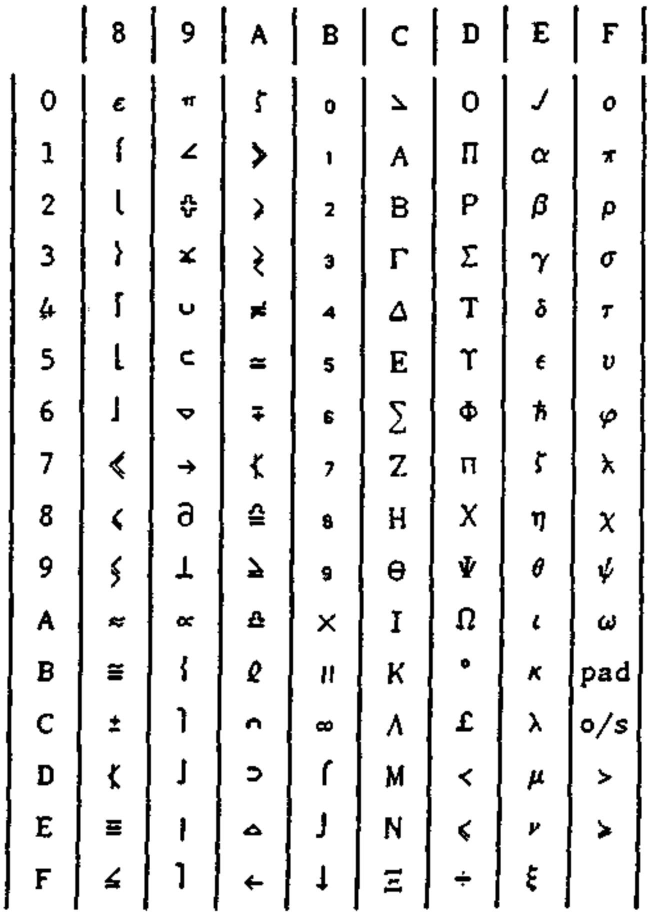


Figure 3. Screen and printer character editors

Unfortunately, as far as translators are concerned, Vuman chose to aim its first product, Vuwriter, at the scientist and so populated the additional character positions with scientific symbols (Figure 4).

In 1982 IBM launched its personal computer (PC) which in many ways transformed the market for small business computers. It also provided an extended character set (IBM code page 437) which is something of a compromise in that it contains a mixture of multilingual and scientific characters (see Figure 5). IBM code page 437 has been used as the basic character set for a large number of word processing systems and provides the user with sufficient characters to handle the main European languages. However, it does not adequately support floating accents on

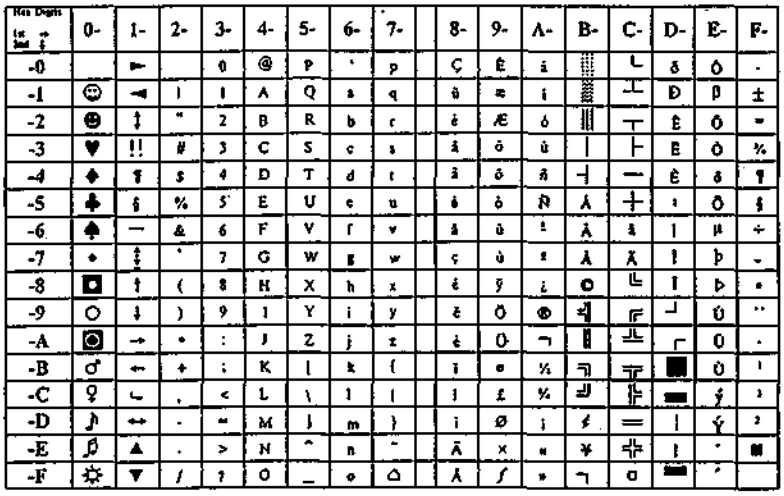


	8	9	A	B	C	D	E	F
0	ε	π	ʕ	0	≥	O	√	o
1	⎧	∠	≻	1	A	Π	α	π
2	⎩	⊕	≻	2	B	P	β	ρ
3	⎫	≭	≷	3	Γ	Σ	γ	σ
4	⌠	∪	≠	4	Δ	T	δ	τ
5	⎣	⊂	≃	5	E	Υ	ϵ	υ
6	⎦	∇	∓	6	∑	Φ	ℏ	φ
7	≪	→	⊀	7	Z	π	ζ	ƛ
8	‹	∂	≙	8	H	X	η	χ
9	⋚	⊥	≥	9	Θ	Ψ	θ	ψ
A	≉	∝	≜	×	I	Ω	ι	ω
B	≅	⎨	ℓ	∥	K	°	κ	pad
C	±	⎤	∩	∞	Λ	£	λ	o/s
D	⊀	⎦	⊃	⌠	M	<	μ	>
E	≡	⎪	△	⌡	N	⩽	ν	⩾
F	≨	⎤	←	↓	Ξ	÷	ξ	

Figure 4. 8-bit extended character sets

upper case versions of many common characters such as ‘é’. Nor is it a particularly good tool for scientists, having far too few of the special characters and symbols required for most scientific disciplines.

However, there was some progress and following the announcement of the new IBM PS series of personal computers in 1987, IBM also announced the existence of code page 850 (Figure 6). This represents a big improvement for the translator, since the scientific characters have been replaced by a more extensive language set, plus floating accents. With this character set it would be possible to handle most European languages and further, documents prepared on any word processor supporting code page 850 should, with minimal difficulty, be transferable to any other system, providing that also supports code 850.

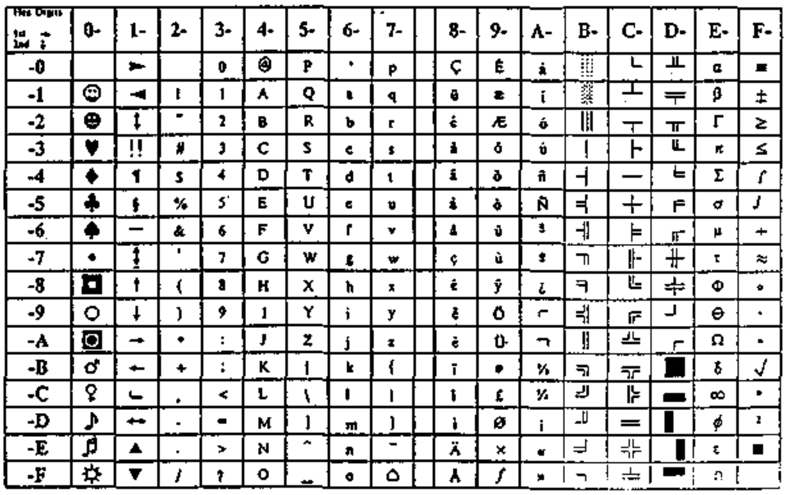


Hex Digits 1st → 2nd ↓	0-	1-	2-	3-	4-	5-	6-	7-		8-	9-	A-	B-	C-	D-	E-	F-
-0		►		0	@	P	`	p		Ç	É	á	░	└	ð	Ó	-
-1	☺	◄	!	1	A	Q	a	q		ü	æ	í	▒	┴	Ð	ß	±
-2	☻	↕	"	2	B	R	b	r		é	Æ	ó	▓	┬	Ê	Ô	‗
-3	♥	‼	#	3	C	S	c	s		â	ô	ú	│	├	Ë	Ò	¾
-4	♦	¶	$	4	D	T	d	t		ä	ö	ñ	┤	─	È	õ	¶
-5	♣	§	%	5	E	U	e	u		à	ò	Ñ	Á	┼	ı	Õ	§
-6	♠	▬	&	6	F	V	f	v		å	û	ª	Â	ã	Í	µ	÷
-7	•	↨	'	7	G	W	g	w		ç	ù	º	À	Ã	Î	þ	¸
-8	◘	↑	(	8	H	X	h	x		ê	ÿ	¿	©	╚	Ï	Þ	°
-9	○	↓	)	9	I	Y	i	y		ë	Ö	®	╣	╔	┘	Ú	¨
-A	◙	→	*	:	J	Z	j	z		è	Ü	¬	║	╩	┌	Û	·
-B	♂	←	+	;	K	[	k	{		ï	ø	½	╗	╦	█	Ù	¹
-C	♀	∟	,	<	L	\	l	|		î	£	¼	╝	╠	▄	ý	³
-D	♪	↔	-	=	M	]	m	}		ì	Ø	¡	¢	═	¦	Ý	²
-E	♫	▲	.	>	N	^	n	~		Ä	×	«	¥	╬	Ì	¯	■
-F	☼	▼	/	?	O	_	o	⌂		Å	ƒ	»	┐	¤	▀	´	

Figure 5. IBM code page 437

THE NEED FOR STANDARDISATION

This is a development which is very promising for translators working in European languages, but what about other languages? To take Russian as an example, there are a number of word processing systems offering Russian as an option (including Vuwriter). The problem at present is that each system uses its own character set coding and this precludes simple



Hex Digits 1st → 2nd ↓	0-	1-	2-	3-	4-	5-	6-	7-		8-	9-	A-	B-	C-	D-	E-	F-
-0		►		0	@	P	`	p		Ç	É	á	░	└	╨	α	≡
-1	☺	◄	!	1	A	Q	a	q		ü	æ	í	▒	┴	╤	ß	±
-2	☻	↕	"	2	B	R	b	r		é	Æ	ó	▓	┬	╥	Γ	≥
-3	♥	‼	#	3	C	S	c	s		â	ô	ú	│	├	╙	π	≤
-4	♦	¶	$	4	D	T	d	t		ä	ö	ñ	┤	─	╘	Σ	⌠
-5	♣	§	%	5	E	U	e	u		à	ò	Ñ	╡	┼	╒	σ	⌡
-6	♠	▬	&	6	F	V	f	v		å	û	ª	╢	╞	╓	µ	÷
-7	•	↨	'	7	G	W	g	w		ç	ù	º	╖	╟	╫	τ	≈
-8	◘	↑	(	8	H	X	h	x		ê	ÿ	¿	╕	╚	╪	Φ	°
-9	○	↓	)	9	I	Y	i	y		ë	Ö	⌐	╣	╔	┘	Θ	∙
-A	◙	→	*	:	J	Z	j	z		è	Ü	¬	║	╩	┌	Ω	·
-B	♂	←	+	;	K	[	k	{		ï	ø	½	╗	╦	█	δ	√
-C	♀	∟	,	<	L	\	l	|		î	£	¼	╝	╠	▄	∞	ⁿ
-D	♪	↔	-	=	M	]	m	}		ì	Ø	¡	╜	═	▌	φ	²
-E	♫	▲	.	>	N	^	n	~		Ä	×	«	╛	╬	▐	ε	■
-F	☼	▼	/	?	O	_	o	⌂		Å	ƒ	»	┐	╧	▀	∩	

Figure 6. IBM code page 850 (multilingual)

file transfer. However, the International Standards Organisation (ISO) has classified the majority of character sets available from all over the world and unambiguous codes have been assigned to individual characters (see Figure 7). The reason this coding mechanism has not been in widespread use is, in the opinion of the author, a direct result of the confusion and incompatibility of the base character set codings used in word processing systems. As a consequence software designers have squeezed additional language characters in wherever they could find space or felt they could afford to sacrifice existing characters. The result is a complete mess.

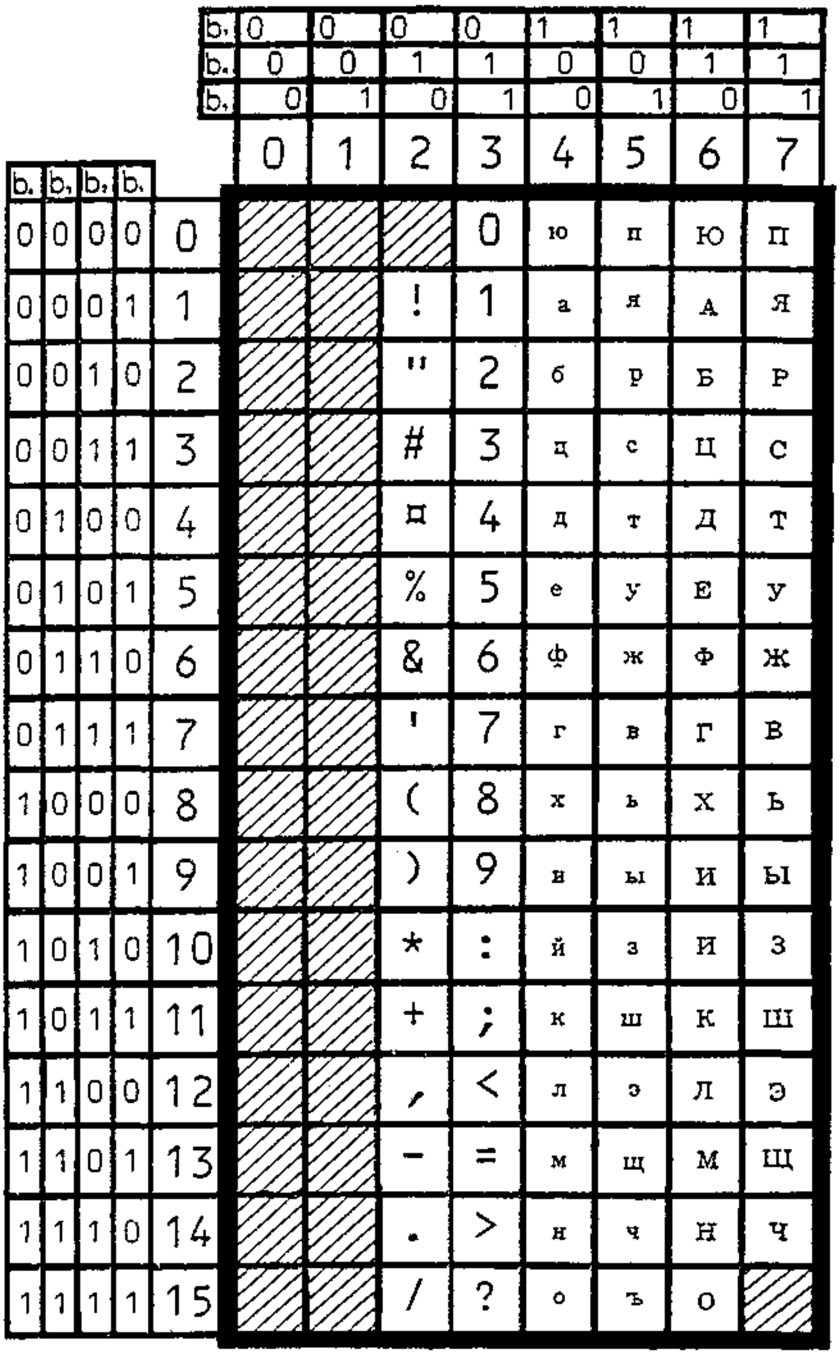
Nevertheless a real basis for standardisation does seem to be emerging. If IBM code page 850 becomes accepted as the base character for international document interchange – and it probably will in view of the influence IBM has on the rest of the market – then there are good reasons for implementing additional character sets in the ISO format. The principal reason is that IBM code page 850 is a full and useful character set for international communication with no 'wasted' characters which might be substituted for, say Hebrew or Russian. If software suppliers cannot provide alternative languages by substituting characters, then they will have to design software for switching to alternative sets. Having gone to these lengths, it is then a simple matter to implement the additional character sets in ISO form – in fact there is no reason not to.

Hence the author's solution for multilingual word processing file interchange is to adopt IBM page code 850 as the base character set with other character sets implemented in ISO format (Figure 8).

DTP FOR THE MULTILINGUAL ENVIRONMENT

To move onto desktop publishing (and typesetting) and to see how this relates to the previous discussion, we should first look at what DTP is. At its simplest it provides accurate control of page layout, it enables items to be 'placed' on a page either in a frame or with text flowing around, it offers a wide selection of typeface styles and sizes, and it allows text to be merged with a wide range of graphics. All this, is of course, prepared on the computer screen and corrections or adjustments are made before the document is printed (Figure 9).

It is at present a fundamental assumption of the two market leaders in DTP that authors will prepare the text for the document using their favourite word processing system. The text is checked for accuracy and edited using the word processor, but no attempt is made to organise the layout of the text, nor are headings treated in any particular way – this is the job of the DTP system. For a seamless transfer between word processor and DTP it is of course necessary for the character sets to be compatible. Ideally the character sets would be identical, although a



				b_7	0	0	0	0	1	1	1	1
				b_6	0	0	1	1	0	0	1	1
				b_5	0	1	0	1	0	1	0	1
b_4	b_3	b_2	b_1		0	1	2	3	4	5	6	7
0	0	0	0	0	▨	▨	▨	0	ю	п	Ю	П
0	0	0	1	1	▨	▨	!	1	а	я	А	Я
0	0	1	0	2	▨	▨	"	2	б	р	Б	Р
0	0	1	1	3	▨	▨	#	3	ц	с	Ц	С
0	1	0	0	4	▨	▨	¤	4	д	т	Д	Т
0	1	0	1	5	▨	▨	%	5	е	у	Е	У
0	1	1	0	6	▨	▨	&	6	ф	ж	Ф	Ж
0	1	1	1	7	▨	▨	'	7	г	в	Г	В
1	0	0	0	8	▨	▨	(	8	х	ь	Х	Ь
1	0	0	1	9	▨	▨	)	9	и	ы	И	Ы
1	0	1	0	10	▨	▨	*	:	й	з	И	З
1	0	1	1	11	▨	▨	+	;	к	ш	К	Ш
1	1	0	0	12	▨	▨	,	<	л	э	Л	Э
1	1	0	1	13	▨	▨	−	=	м	щ	М	Щ
1	1	1	0	14	▨	▨	.	>	н	ч	Н	Ч
1	1	1	1	15	▨	▨	/	?	о	ъ	О	▨

Figure 7. ISO Russian character set

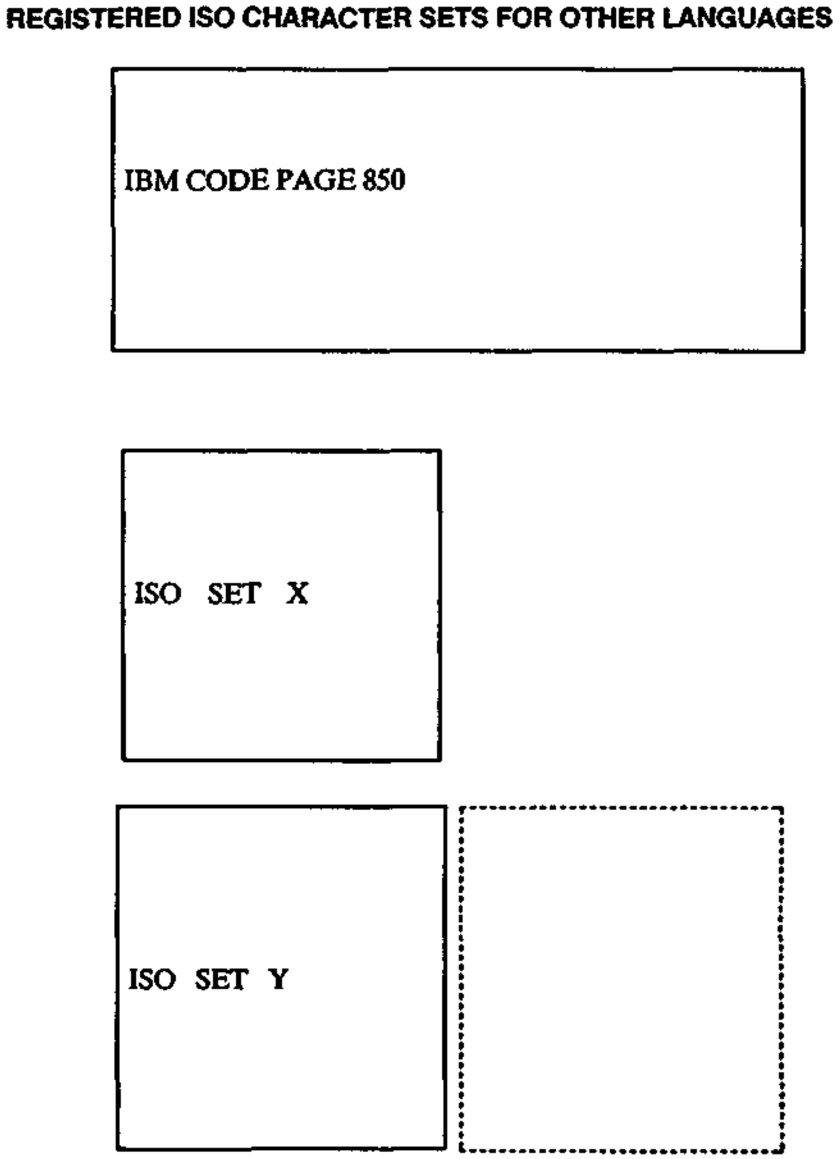
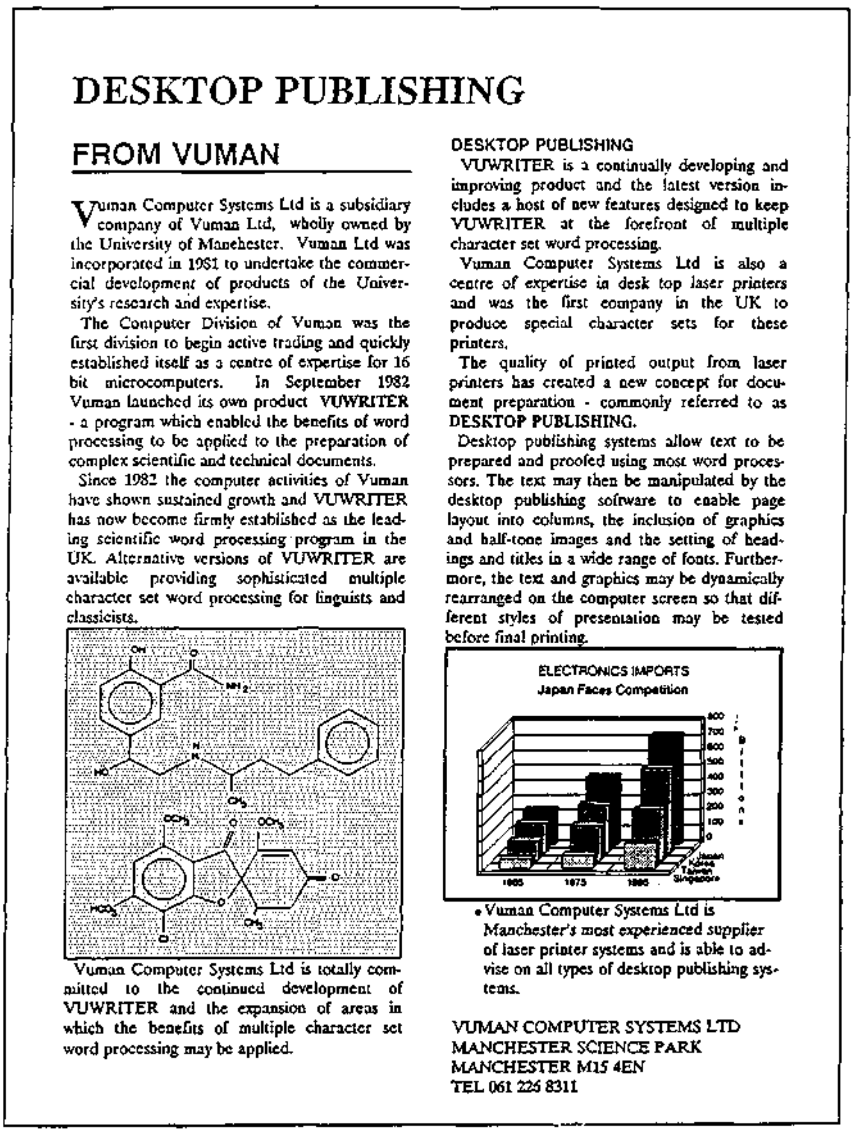


Figure 8. IBM code page 850 as the international base



DESKTOP PUBLISHING

FROM VUMAN

Vuman Computer Systems Ltd is a subsidiary company of Vuman Ltd, wholly owned by the University of Manchester. Vuman Ltd was incorporated in 1981 to undertake the commercial development of products of the University's research and expertise.

The Computer Division of Vuman was the first division to begin active trading and quickly established itself as a centre of expertise for 16 bit microcomputers. In September 1982 Vuman launched its own product VUWRITER - a program which enabled the benefits of word processing to be applied to the preparation of complex scientific and technical documents.

Since 1982 the computer activities of Vuman have shown sustained growth and VUWRITER has now become firmly established as the leading scientific word processing program in the UK. Alternative versions of VUWRITER are available providing sophisticated multiple character set word processing for linguists and classicists.

Vuman Computer Systems Ltd is totally committed to the continued development of VUWRITER and the expansion of areas in which the benefits of multiple character set word processing may be applied.

DESKTOP PUBLISHING

VUWRITER is a continually developing and improving product and the latest version includes a host of new features designed to keep VUWRITER at the forefront of multiple character set word processing.

Vuman Computer Systems Ltd is also a centre of expertise in desk top laser printers and was the first company in the UK to produce special character sets for these printers.

The quality of printed output from laser printers has created a new concept for document preparation - commonly referred to as DESKTOP PUBLISHING.

Desktop publishing systems allow text to be prepared and proofed using most word processors. The text may then be manipulated by the desktop publishing software to enable page layout into columns, the inclusion of graphics and half-tone images and the setting of headings and titles in a wide range of fonts. Furthermore, the text and graphics may be dynamically rearranged on the computer screen so that different styles of presentation may be tested before final printing.

- Vuman Computer Systems Ltd is Manchester's most experienced supplier of laser printer systems and is able to advise on all types of desktop publishing systems.

VUMAN COMPUTER SYSTEMS LTD
MANCHESTER SCIENCE PARK
MANCHESTER M15 4EN
TEL 061 226 8311

Figure 9. Example of a complex document created on a DTP system

certain amount of re-mapping is possible. A look at the base character set for Xerox Ventura shows that it corresponds to neither code page 437 nor code page 850, so there are clearly potential problems here. Nevertheless, it is the author's belief that DTP will be led by word processing and that if the majority of word processing suppliers support a common character set based on code page 850, then the DTP systems will follow. Exactly the same argument applies to typesetters and many typesetters are able to read files from word processors. At present the ability to transfer complex text is limited. However, with a greater user base supporting code page 850 one can look forward to a substantial degree of compatibility in the years to come.

AUTHOR

David Jackson, Vuman Computer Systems Ltd, Enterprise House, Manchester Science Park, Lloyd Street North, Manchester M15 4EN.