The Role of the Local Alphabet in Enhancing of the Development of Digital Communication and the Associated Problems

Zoran Gligorov and Filip Stojanovski
University of Saints Cyril and Methodius, Skopje, Republic of Macedonia
zglg@soros.org.mk, fstojan@soros.org.mk

Abstract

One of the often overlooked aspects of the today's electronic information spread is that not all people use the Latin alphabet for writing, and even those who use it do not use the standard English version of it. Since the standards and software applications of those standards are mainly made by people who use the Latin alphabet and speak English as a first language, the situation today is that people who do not use English or would like to use their own alphabet are somewhat deprived and as a result are sometimes quite reluctant to start using electronic communication. This paper concerns itself with the sources, consequences and means to solve this problem. The issue is discussed with regard to real-life examples involving the use of the Macedonian language and a comparative analysis of the solutions applied.

The examples concentrate on Internet publishing using the Macedonian language and the digital usage of the Macedonian Cyrillic Alphabet, which is different from any other Cyrillic alphabet and thus incompatible with the specific solutions applied in other countries. The paper provides an analysis of the current situation from which conclusions applicable to other cultures can be extracted.

Introduction

"Using computers" has almost become a synonym for the term "communications". And it will be even more obvious as the mass-media give up analog technology in favor of digital in the near future. Although the advent of totally digital TV, received via optical fiber (in order to preserve the limited scope of transmission frequencies available) is still far in the future, the electronic descendant of classical print is making giant steps in the direction of digitalization. Its use has now risen to a level when more standards are required in order to get the most from the new technology.

A sine qua non of publishing and, naturally, of all electronic publishing is the use of computers. Usage of computers makes it possible to create a publication with an efficiency and speed unthinkable in the past, no matter who the publisher is or what is the language used in the work.

Books are fine. Once printed, they will look the same everywhere in the world. However, the process of printing and distribution has big pitfalls because of the high cost and the environmental matters involved. Since today the technical side of the process of producing a book is almost entirely based on the use of computers, it would be logical to suggest that a book should not be printed on paper (the only analog part of the publishing process), but to distribute it digitally - on a disk, CD-ROM, or via Internet. Using the Internet seems the best alternative, since it does not require additional spending of material resources for making the material medium for this specific piece of information to be distributed. But is it that easy?

The main concerns of an electronic publisher are:
1. input of the information so that it can be processed (using keyboard, speech-recognition software or any other method);
2. the compatibility of the product - is their published material easily accessible and will it look the same on every computer/platform?

This paper reviews these concerns using the example of the use of the Macedonian language for digital communication/electronic publishing.

Background

The basic conditions are that the Macedonian language is not English and uses an alphabet different from English. This is not uncommon; there are more than 3000 languages today that are different from English and most of them use different systems for written representation. Thus, the situation and some of the problems/solutions from the example of the Macedonian language can be used with regard to other non-English languages.

The Macedonian language uses its own version of the Cyrillic alphabet, which has some characters different from the other Cyrillic Alphabets (Russian, Serbian, Bulgarian) and because of that had to undergo a process of development of separate solutions.

English is taken as a referral point because the standards for computer use, the software applications and the hardware itself is mainly designed by persons who use the English Latin alphabet and speak English as a first language. These people also compose the largest share of the electronic publishing market *today* (50% of all Internet users today are in North America), and one of the reasons for this is that the electronic publishing products have no difficulty to reach its *target audience in its own language*. Language is one of the foundations of the culture because it is a tool that people use for thinking, and the success of electronic publishing in non-English countries/cultures depends on the obvious, but often overlooked, fact that it has to be done in the local language.

Although the English language is the lingua franca of computer technology and more and more people learn it as a second language, an electronic publisher in a non-English country cannot afford to set as a target only those persons who know it - there are simply not enough of them. This situation is evident in most developing countries, and the countries of Central and Eastern Europe, including Republic of Macedonia. In order to buy a product, especially an intellectually related one, the user must be comfortable about using it with ease. This can be only accomplished if the local language is used.

This set of circumstances puts the speakers of Macedonian in a difficult position to make full use of this new technology. The majority of the problems appear to be in user-computer communication. That is, the information is entered using a keyboard that is not designed for that alphabet and the viewing software cannot show the appropriate characters.

This problem has been partially solved by the makers of the operating systems initially designed to communicate in English. Some of them evolved and became more international, by supporting other languages, but not all languages were supported and the support was not and still is not compatible between all programs.

Computing in Macedonia

The most common computers in Macedonia are Intel based PCs, and the review of the operating systems for them with regard to the problem is as follows:

MS-DOS originally used the 8-bit ASCII character set supporting 256 characters. The first 128 were mainly the English alphabet, while the rest were graphical ones. In order to accommodate the character set to the needs of different alphabets, the codes of the higher ASCII characters were used to substitute for the characters not identical with the English ones. Many countries registered their additions and changes with international code-pages.
The code page 850 contained the ex-Yugoslav (but not Cyrillic) characters, and then the problems began. The publications of that time were not compatible with all programs, so a document containing these characters transferred electronically had a high probability of being altered by the software.

The main engines that generated positive change were the Bulletin Board Systems and the need of their users to communicate using their own alphabet. Simply said, some people did not want to read and write in the Macedonian language using transliteration into the English Latin alphabet, where the characters who are not identical in appearance are replaced by a combination of two or more Latin characters.

Since communication with Bulletin Board Systems (BBS) is (still) based on the 8-bit ASCII set of characters, a group of volunteers inspired by Informa EIS (the biggest BBS in Macedonia, http://www.informa.mk) developed a program package called MAKOLR (consisting of video card drivers) that did some unofficial run-time changes of the set allowing Cyrillic characters to be used as well as Latin but:

- If the set works on one software under DOS, the chances are it will not work under other types of software. In some instances, since the program was based on altering the screen representation of the characters by changing the output of the video card (made possible when VGA cards appeared), these alterations did not affect the other aspects of the display process and such texts could not be printed without further customizations.
- If a reader does not have the appropriate software to change the ASCII set, the information is completely unreadable
- The conversion from Cyrillic to Latin is possible with special tools, but from 28-lettered Latin to Cyrillic is impossible without human interaction.

BBS users in Macedonia are still using this partial solution for some of the digital communication, but on a very local level. Recently, InForma EIS developed a web-driven interface which allows access to its data-base of messages via HTTP. The messages from the data base written using MAKOLR are unreadable when seen as HTML.

MS Windows gives much more freedom for electronic publishing. Electronic documents could use any True Type Font (TTF), and they are compatible on every machine that has Windows. But again, the published material needs to contain the appropriate font in order to show the characters properly. This is acceptable for most of the electronic publishers who distribute their work on CD-ROMs or floppy discs.

If every country had their own character set, and there are enough spaces in the set to be filled with the characters used in every country, one standard character set with enough characters would be enough for making the whole world happy. Microsoft tried to do this in their

international edition of Windows 95, but so far as much of Macedonian Cyrillic is concerned, Microsoft has failed so far. The background of this problem has two possible contexts:

1. Economic: Macedonia is a too small a market for Microsoft to think that it is profitable to apply a bit of extra effort to accommodate the needs for the Macedonian-language based version of Windows.
2. Political: It is unclear how international politics are connected with this particular instance of the problem, but some connections are suspected.

The good idea of using a code page that is compatible with all other similar (for example Cyrillic or East European) code-pages comes clear when other platforms such as UNIX, OS/2, Solaris or others are used.

Keyboard layout

The keyboard layout is important for electronic publishers. Since the writer is seen as an electronic publisher in the future, it is clear that electronic publishing must be simple enough for practically everybody.

While waiting for speech-recognition software, publishers today are still limited to keyboards that can make difficulties in their lives. The good side of the keyboard is that it can be reprogrammed, so it can be made to meet anyone's needs. The bad side is that not all the characters of the non-English alphabets can be seen on the keyboards. Having a specially designed keyboard for a different alphabet is fine. This is the case in many countries: The keyboards in the UK are different from those in France, Germany or USA. Since most computers are brought from the USA, it is common to find USA keyboards in Eastern Europe. These keyboards contain no characters of other alphabets and differ from the local standardized keyboards (used originally for analog typing machines). Some people find it useful to stick the characters of their alphabet to (re-programmed) keyboards with glue; this may be a very good solution.

Standards and their usefulness

During 1997 there was public discussion about the standard for the Macedonian Cyrillic alphabet, known as JUS I.K1.004 (1987). This standard needs 47 keys in order to contain all Macedonian Cyrillic characters. The 47 keys are enough to show all Macedonian Cyrillic letters, digits and other characters.

Figure 1. Graphical symbols defined in the JUS I.B1.004 standard.

Although the standard is acceptable, and all Cyrillic typewriters use it, it is somewhat different from the US standard concerning some of the graphical symbols and some of the Latin equivalents. It is much closer to the German standard, but this is of no use since most computer keyboards are US style. This standard is supported by official government institutions and several large electronic publishers, and is planned to be incorporated into the future versions of Windows on a system level.

The ISO standard proposal concerning Macedonian Cyrillic characters (ISO 3098/4, 1984) defines Cyrillic characters for a large number of countries, including ex-Yugoslavia. It is, however, no where near the Macedonian Cyrillic; 6 letters are missing. However, it was not officially declined by any country (including ex-Yugoslavia), although it has this enormous failure.

During the course of the last several years, unofficial keyboard layout standards also appeared, trying to get the most of the US-English keyboards and the graphical similarities of the characters. These standards are implemented by the use of custom-made True Type Fonts, and are most widely used in the Republic of Macedonia today. Being incompatible with the officially proposed standard, the spread of these standards demonstrates the influence of the keyboard layout on publishing. It is generally regarded that if enough support is put into the official standard, these standards will become obsolete.

Keyboard Layout & Standards For Macedonian Cyrillics

1. Arial 2. Mac C Swiss 3. DOS Compatible

1	—	б	ќ	46	.	.	.	91	[	Ш	[	136	ˆ	□	З	181	µ	Є	ѕ	226	â	Ò	в
2	—	б	ќ	47	/	/	/	92	\	Ґ	\	137	‰	ф	Є	182	¶	І	і	227	ã	Ъ	г
3	—	б	ѓ	48	0	0	0	93	]	Ќ	]	138	Š	□	и	183	·	□	ї	228	ä	Ў	д
4	*	S	C	49	1	1	1	94	^	Ч	^	139	‹	Ч	Ј	184	¸	К	ґ	229	å	є	ѓ
5	—	б	ќ	50	2	2	2	95	_	_	_	140	Œ	Ё	К	185	'	□	ѕ	230	æ	ж	е
6	—	б	ќ	51	3	3	3	96	`	ж	`	141		□	Л	186	°	М	і	231	ç	ц	ж
7	—	б	ќ	52	4	4	4	97	a	а	a	142		□	љ	187	»	Ч	ї	232	è	э	э
8	—	б	ќ	53	5	5	5	98	b	б	b	143		□	М	188	¼	□	ј	233	é	ѐ	ѕ
9	—	б	ќ	54	6	6	6	99	c	ц	c	144		□	Н	189	½	□	ј	234	ê	ѓ	и
10	*	в	п	55	7	7	7	100	d	д	d	145	'	д	Њ	190	¾	□	љ	235	ë	ё	ј
11	*	г	Р	56	8	8	8	101	e	е	e	146	'	е	О	191	¿	Ґ	ґ	236	ì	ы	к
12	—	б	ќ	57	9	9	9	102	f	ф	f	147	"	в	п	192	À	Ы	Ы	237	í	и	л
13	—	б	ќ	58	:	:	:	103	g	г	g	148	"	г	Р	193	Á	Ю	⊥	238	î	ѐ	љ
14	—	б	ќ	59	;	;	;	104	h	х	h	149	•	S	С	194	Â	Я	т	239	ï	ѓ	м
15	/	/	/	60	<	<	<	105	i	и	i	150	–	Ъ	Т	195	Ã	Ь	⊦	240	ð	□	н
16	—	б	ќ	61	=	=	=	106	j	ј	j	151	—	б	ќ	196	Ä	'	–	241	ñ	й	њ
17	—	б	ќ	62	>	>	>	107	k	к	k	152	˜	□	У	197	Å	Є	+	242	ò	Т	о
18	—	б	ќ	63	?	?	?	108	l	л	l	153	™	ђ	Ф	198	Æ	Ж	⊧	243	ó	ò	п
19	—	б	ќ	64	@	Ж	@	109	m	м	m	154	š	□	Х	199	Ç	Щ	⊪	244	ô	ò	р
20	—	б	ќ	65	A	А	A	110	n	н	n	155	›	Ћ	Ц	200	È	щ	⊫	245	õ	ѣ	с
21	—	б	ќ	66	B	Б	B	111	o	о	o	156	œ	ē	Ч	201	É	Å	⊩	246	ö	ў	т
22	—	б	ќ	67	C	Ц	C	112	p	п	p	157		□	Џ	202	Ê	ћ	⊫	247	÷	ґ	ќ
23	—	б	ќ	68	D	Д	D	113	q	љ	q	158		□	ш	203	Ë	ч	⊪	248	ø	П	у
24	—	б	ќ	69	E	Е	E	114	r	р	r	159	Ÿ	й	Є	204	Ì	э	⊫	249	ù	ю	ф
25	—	б	ќ	70	F	Ф	F	115	s	с	s	160			й	205	Í	ъ	=	250	ú	ý	х
26	—	б	ќ	71	G	Г	G	116	t	т	t	161	¡	ш	О	206	Î	Ъ	⊕	251	û	ѝ	ц
27	—	б	ќ	72	H	Х	H	117	u	у	u	162	¢	ћ	ѐ	207	Ï	Э	⊥	252	ü	я	ч
28	—	б	ќ	73	I	И	I	118	v	в	v	163	£	ч	й	208	Ð	□	⊤	253	ý	□	ѵ
29	—	б	ќ	74	J	Ј	J	119	w	њ	w	164	¤	Ж	ѐ	209	Ñ	Ё	⊤	254	þ	□	ш
30	—	б	ќ	75	K	К	K	120	x	џ	x	165	¥	ю	„	210	Ò	ё	⊤	255	ÿ	в	□
31	—	б	ќ	76	L	Л	L	121	y	ѕ	y	166	¦	□	ª	211	Ó	ю	⊩				
32				77	M	М	M	122	z	э	z	167	§	§	º	212	Ô	я	⊧				
33	!	!	!	78	N	Н	N	123	{	ш	{	168	¨	Ќ	г	213	Õ	Э	⊦				
34	"	"	"	79	O	О	O	124	|	ґ	|	169	©	Ш	⌐	214	Ö	Є	⊥				
35	#	#	#	80	P	П	P	125	}	ќ	}	170	ª	Л	¬	215	×	□	⊥				
36	$	$	$	81	Q	Љ	Q	126	~	ч	~	171	«	Ч	½	216	Ø	Џ	⊥				
37	%	%	%	82	R	Р	R	127	▒	□	⌂	172	¬	Т	¼	217	Ù	є	⊥				
38	&	&	&	83	S	С	S	128		□	А	173	-	□	¡	218	Ú	ђ	г				
39	'	'	'	84	T	Т	T	129		□	Б	174	®	я	«	219	Û	S	▉				
40	(	(	(	85	U	У	U	130	‚	т	В	175	¯	□	»	220	Ü	Й	▄				
41	)	)	)	86	V	В	V	131	ƒ	э	Г	176	°	Ё	▒	221	Ý	ь	▌				
42	*	*	*	87	W	Њ	W	132	„	і	Д	177	±	Б	▓	222	Þ	□	▐				
43	*	+	+	88	X	Џ	X	133	…	Щ	Ґ	178	ª	□	▓	223	ß	Ї	▀				
44	,	,	,	89	Y	S	Y	134	†	ъ	Е	179	³	□	|	224	à	я	а				
45	-	-	-	90	Z	З	Z	135	‡	–	Ж	180	´	Ћ	ѕ	225	á	а	б				

Common Keyboard Shortcuts

C=Control
S=Shift

CS+=	—
C+=	–
CS+[	•
CS+]	·
C+[	'
C+]	'
CS+8	•
CS+6	§

The Internet

Up to this point, it is clear that electronic publishing on hard media (CDs, floppy disks) and based on one operating system is fine for everyone because programming solutions can be applied, one way or the other, and there is enough space on the medium for them). But when publishing electronically, the whole idea is to distribute the information digitally as well.

The Internet is a growing medium for electronic publishing and very applicable to use by Macedonian-speaking users because it can serve as a direct connection bridge with the Macedonian ethnic community outside the borders of Macedonia, which can be reached via analog media (print, TV) only at great cost. Macedonians who live abroad, as ethnic minorities in the neighboring countries or large immigrant communities in America, Australia and Western Europe experience a lack of published material in their own language which has two main consequences:

- easing the assimilation in the culture of the country of citizenship, which is sometimes seen as a threat to the ethnic identity and leads into ghettoization;
- need for materials published in Macedonian in order to support the persistence of the original culture and identity. This can serve as a tool for incorporating the ghettoized communities into the mainstream culture. Internet-empowered Macedonians twhobenefit from electronic publishing in their native language would be able to share its Macedonian traits with the other ethnic groups and enhance the cultural wealth of their respective societies/countries as a whole.

Yet, at the moment, Internet is one of the worst mediums for supporting other alphabets. Serious problems are experienced here. Once the information (published material, document) is entered, it can be transmitted on the Internet as e-mail message, put on the Web, or sent using FTP or Gopher. The Internet has much lower bandwidth than CD-ROMs or floppy disks, and this is why it is impossible to send the content of the CD-ROM via the Internet to millions of potential readers. The publisher will have to rely on the viewer's software instead. That is the e-mail client or the Web browser.

E-mail is the most convenient way of communication. Yet, it is limited to 7-bit character encoding, and while it can transmit this document, it is impossible to send messages in Cyrillic. 7-bit ASCII means that only 127 characters are supported, and no Cyrillic character can enter the Internet. This seems to be no problem only to people who use computers for long time, they are just used to "converting" their Cyrillic to Latin.

There is still a way to view and type e-mail in Cyrillic. HTML enables the user to change the font to a Cyrillic one. But what the publisher/sender does not know is whether the receiving side has the same font.

A typical e-mail message is about 5-10Kb. The font is about 100Kb, so sending the font with every message written in Cyrillic is nonsense. Also the receiving side may use platform other than Windows. If this is DOS, the message would be readable only if there is a DOS-based converter (8-bit ASCII). If the receiving side is UNIX-based, then the message will not be readable at all.

So, when changing the font, it is necessary to make sure that:

1. The receiving side has the same font and compatible software,
2. If the same font is not found, it can be substituted with similar one, or
3. If there is no similar substitute, the message must be still readable by any software and with the standard character set - use of transliteration into the standard English Latin alphabet is applied.

The situation on the Web is very similar. A Web page made in Macedonian language would have to use one of the following solutions.

Solution 1: complete transliteration into English Latin Alphabet

Readable by all web browsers on all operating systems, this technology uses substituting of characters with combinations of English characters.
Example:

Macedonian cyrilic:
А Б В Г Д Ѓ Е Ж З Ѕ И Ј К Л Љ М Н Њ О П Р С Т Ќ У Ф Х Ц Ч Џ Ш

Transliteration standards:
A B V G D GJ E ZH Z DZ I J K L LJ M N NJ O P R S T KJ U F H C CH DZH SH
A B V G D GG E ZZ Z DJ I J K L LJ M N NJ O P R S T KK U F H C CC DZZ SS

But this solution does not provide the use of the Cyrillic alphabet, threatening the basic foundations of the Macedonian culture.

Solution 2: use of True Type Fonts

Browsers such as Netscape Navigator and Microsoft Internet Explorer version 3 and higher support including the name of the TTF in the tag that describes the properties of the font to display the text with. This is quite elegant solution, and all that the maker of the document is required to do is to generate a tag similar to this before the Cyrillic text:
<FONT FACE="name of appropriate Cyrillic TTFont">.
Butif the user does not have the font installed on the client machine, or is accessing the site with another (lower version of a) browser, or is doing it from a UNIX/Linux based system (such as many academic computers in the world) the properties of the font are lost and the text is shown in Latin alphabet with some characters substituted with the ones that take the place of the Cyrillic in the Latin set, making it if not just partially readable, but also definitely annoying.
Example:

Macedonian text:
Се што треба да се стори е да се види чија е желката, оти таа ќе сака да е со оној кој ја одгледал.

As shown on a Windows/Mac machine without the appropriate font, or by using a browser/OS that does not support use of True Type Fonts:
Se {to treba da se stori e da se vidi ~ija e `elkata, oti taa }e saka da e so onoj koj ja odgledal.

Solution 3: Code page usage

Some browsers have built-in code page support. This feature enables them to show texts written using the standard code pages, and most commonly used for Cyrillic texts is CP 1250. If the reader uses the *appropriate* browser, all additional requirements are to set the "encoding" preference to the code page from the menu. Limitations of this method are that

not all browsers support code page encoding (client side), and the making of such text requires additional software usage by the electronic publisher, which complicates the code generating and revising processes.

Solution 4: Dynamic HTML and downloadable fonts

Web browsing utilities such as Netscape Communicator 4 and Microsoft Internet Explorer 4 have a feature that enables them to use downloadable fonts. The principle is that the fonts come with the document(s) that uses them in real time. Although this solution is clumsy when e-mail is concerned, it can be applied to Web content since the downloadable font is not very big (less than 20k, because it is generated by reducing and compressing the True Type Font on the publisher's machine). This solution also addresses the copyright issues, because these fonts can be used for viewing of web pages that the maker of the fonts allows to be viewed (when generating the downloadable font file, the maker can set which documents can be viewed with it). This makes the downloadable fonts unusable for other applications on the client side (such as text processing) because they can not be taken out of the browser's environment. For more information on Dynamic HTML please refer to:
http://developer.netscape.com/one/dynhtml/
This solution solves the problems of most of the previous alternatives, but is limited to systems that can use the class 4 browsers (that comply to the Mozzila/4 standard). This would include computers on a performance level similar or higher with a Pentium 100 processor with at least 16MB RAM, thus excluding a fair amount of existing machines. If the requirements are not met, the problems of solution 3 remain.

Conclusions about the solutions

A sample statistical analysis of the visits of the Virtual Macedonia Web site (http://www.vmacedonia.com), one of the starting point for getting information about Macedonia on the web, which applies the latest in web development and is constantly updated to make the most of the usage of the newest browser generation shows the following results:
- 44 % of the visitors use generation 4 browser (solution 4 applies - downloadable fonts)
- 23 % of the visitors use a browser comparable to generation 3 browsers (solution 2 applies - HTML with TTF support *if they have the fonts installed*, or usage of code pages)
- 27% of the visitors use a browser on the level of Netscape 2 or lower for which transliteration is needed.

These numbers can be taken as a rough guide to the degree of usefulness when applying the solutions discussed above

Conclusion

These practical examples show that there is great potential for electronic publishing in the smaller developing countries, as well as an undiscovered potential customer base in countries with ethnic communities from these countries. This potential can be used if the needs of the people for preserving their cultural identity are accessed. Solving the alphabet problems on a global level would enhance the growth of trans-national communities based on similar cultural traits and in turn open new oportunities for electronic publishing.

Acknowledgments

The authors would like to thank the following persons/instituions for the support they provided during the research for this project: Virtual Macedonia's Boris Shoposki; Saso Mickov - Login Systems, Skopje; Elizabeta Goneska; Mira Polazarevska.