Language Support

Cyrillic encodings for LaTeX 2ε multi-language documents

A. Berdnikov, O. Lapko, M. Kolodin,
A. Janishevsky, A. Burykin

Abstract

This paper describes the X2 and the T2A, T2B, T2C encodings designed to support Cyrillic writing systems for the multi-language mode of LaTeX 2ε. The encoding X2 is the "Cyrillic glyph container" which can be used to insert into LaTeX 2ε documents text fragments from all modern Cyrillic writings, but it does not strictly obey all the rules of LaTeX 2ε. The encodings T2A, T2B and T2C are the "true" LaTeX 2ε encodings which satisfies all the requirements of the LaTeX 2ε kernel, but as a result three encodings are necessary to support the whole variety of languages based on the Cyrillic alphabet.

These restrictions of the LaTeX 2ε kernel, the specific features of Cyrillic writing systems and the basic principles used to create the encodings X2 and T2A, T2B, T2C are considered. This project supports all the Cyrillic writing systems known to us, although the majority of the accented letters need to be constructed using internal TeX tools. The X2 encoding was approved at *CyrTUG-97*—the annual conference of Russian-speaking TeX users and was previously presented at the *EuroTeX-98* Conference. The encodings X2, T2A, T2B and T2C were intensively discussed on the `cyrtex-t2` mailing list.

1 Introduction

The project (originally named T2) to produce encodings necessary to support modern Cyrillic languages in LaTeX 2$_\varepsilon$ multi-language mode was initiated at the *TUG-96* Conference in Dubna. this paper presents the results of that project. Although some minor corrections could still appear as new information about minor Cyrillic writings appear, the kernel of the project seems to be stable. The encoding support includes:

- LHfont font collection version 3.2 — the Computer Modern Cyrillic fonts and European Computer Modern Cyrillic fonts created by O. Lapko,
- T2enc macro package created by V. Volovich and W. Lemberg — the input and output encoding and font definitions necessary for the LaTeX 2$_\varepsilon$ packages fontenc and inputenc,
- the hyphenation patterns in encoding-independent style: ashyphen by A. Slepuhin, ruhyphen by D. Vulis, lvhyphen by M. Vorontsova and S. Lvovski, znhyphen by S. Znamenskii,
- rusbabel package created by V. Volovich and W. Lemberg to support Cyrillics (based on new encodings) in BABEL.

The β-versions of these packages are on the *TeXLive 3* CD-ROM distribution. The final versions are available on CTAN.[1]

Support for Cyrillics includes the following encodings:

- X2 — the Cyrillic glyph container which contains all the glyphs necessary to support modern Cyrillic writing. It does not obey all the specifications that LaTeX 2$_\varepsilon$ requires for an encoding with the prefix T, but as a result it is enough to have just one encoding to insert in LaTeX 2$_\varepsilon$ documents the characters, words, names, bibliography references, short sentences, citations, etc., specific for all modern Cyrillic writings without too large an increase in the number of fonts used for this purpose (i.e., this encoding is a tool which enables Latin-writing people to occasionally use Cyrillics in their documents). The price is that some Cyrillic letters do not exist as a separate glyphs but must be composed from pieces (accents and modifiers) contained in X2, and the user must obey some rules of safety (described in section 6) since X2 does not satisfy all the requirements obligatory for T$\langle n\rangle$ encodings;
- T2A, T2B, T2C — the encodings which *strictly* satisfy the requirements necessary for LaTeX 2$_\varepsilon$ multi-language mode. With these encodings it is safe to mix different languages inside your documents and to use large pieces of text without any problems. The price is that it is necessary to have three encodings (and an enormous number of fonts each in agreement with the encoding conventions of the European Computer Modern fonts) to support the whole variety of Cyrillic alphabets. Some Cyrillic languages are supported by one or two encodings from the T2$*$ encoding family, some are supported by all three encodings. The encodings are in agreement with the LaTeX 2$_\varepsilon$ multi-language mode and encoding paradigm. Although there are no obstacles to the use of these encodings for original Cyrillic texts, native users may have different preferences;
- LR0 and LR1 — the encodings which combine the OT1 encoding for 0 – 127 and Cyrillic letters and symbols from the leading languages of the Former Soviet Union for 128 – 255. (The encoding LR0 contains the Russian letters only, the encoding LR1 contains in addition the most frequently used national letters.) These encodings are as close as possible to X2 and T2A/T2B/T2C and are designed mainly for non-LaTeX formats based on the CM font family and the original *Plain* TeX. There is some hope (not confirmed at this moment) that LR0 and LR1 may become the inter-platform and inter-format standard for representing Russian letters inside TeX (as ASCII is the standard for English);
- LR$\langle n\rangle$ — local encodings necessary to support individual Cyrillic languages. The T2$*$ encodings are intended for the multi-language mode of LaTeX 2$_\varepsilon$ and for this reason may be not well suited to bilingual documents or to the preferences of the native users. It is definitely not the task of the T2 working group, but that of the national TeX User Groups, to organize local encodings that are useful for their language;
- TS2 — the encoding containing accents, nonletter symbols, etc., necessary for Cyrillic writings which are outside X2 and T2$*$ encodings. This encoding is under consideration, and with great probability all necessary additional glyphs could be added to TS1 encoding. (The latter case has the advantage that it prevents the increase of the number of fonts needed to support multi-language mode in LaTeX 2$_\varepsilon$);

[1] /fonts/cyrillic, /languages/hyphenation/ruhyphen, and /macros/latex/contrib/supported/t2.

- LWN — the encoding which generalizes the WNCyr font family by adding new Cyrillic letters and new substitution pairs (ligatures) based on ASCII Latin input. It is suitable for Latin-writing users who use Cyrillics only occasionally. (This encoding is still under development and is not described here.);
- T5 encoding(s) to support Old Slavonic, Glagolitic, Church Slavonic, etc., writings. The project to develop these encodings has just started and its discussion is outside the scope of this publication.

2 LaTeX 2$_\varepsilon$ system of encodings

The following types of encodings are recognised by the LaTeX Project:[2]

OT$\langle n\rangle$ — essentially 7 bit 'old' encodings. Typically these will be small modifications of the original TeX encoding, OT1 (for example, OT4, a variant for Polish).

T$\langle n\rangle$ — 8 bit Text Encodings. T$\langle n\rangle$ encodings are the main text encodings that LaTeX uses. They have some essential technical restrictions to enable multilingual documents with standard TeX: (a) they should have the basic Latin alphabet, the digits and punctuation symbols in the ASCII positions, (b) they should be constructed so that they are compatible with the lowercase code used by T1. Further discussion of the technical requirements for T$\langle n\rangle$ encodings is given in section 3.

X$\langle n\rangle$ — other 8 bit Text Encodings (eXtended, or eXtra, or X=Non Latin). Sometimes it may be necessary, or convenient, to produce an encoding that does not meet the restrictions placed on the T$\langle n\rangle$ encodings. Essentially arbitrary text encodings may be registered as X$\langle n\rangle$, but it is the responsibility of the maintainers of the encoding to clearly document any restrictions on the use of the encoding.

TS$\langle n\rangle$ — Text Symbol Encodings. Encodings of symbols that are designed to match a corresponding text encoding (for example, paragraph signs, alternative digit forms, etc.). The font style of fonts in TS$\langle n\rangle$ encoding will ordinarily be changed in parallel with that of the fonts in T$\langle n\rangle$ encoding using NFSS mechanisms. As a result, at any moment the TS$\langle n\rangle$ font style is compatible with the T$\langle n\rangle$ font and the glyphs from TS$\langle n\rangle$ font (accents, punctuation symbols, etc.) can be mixed with the glyphs from the corresponding T$\langle n\rangle$ font.

S$\langle n\rangle$ — Symbol encodings. The style of fonts in S$\langle n\rangle$ encoding need not be synchronized with that of T$\langle n\rangle$ fonts. These encodings are used for arbitrary symbols, 'dingbats', ornaments, frame elements, etc.

A$\langle n\rangle$ — Encodings for special Applications (not currently used).

E$*$ — Experimental encodings but those intended for wide distribution (currently used for the ET5 proposal for Vietnamese).

L$*$ — Local, unregistered encodings (for example, the LR0, LR1 and LR$\langle n\rangle$ encodings mentioned above).

OM$*$ — 7 bit Mathematics encodings.

M$*$ — 8 bit Mathematics encodings.

U — Unknown (or unclassified) encoding.

3 Specifications for T$\langle n\rangle$ and X$\langle n\rangle$ encodings

There are two main restrictions to be fulfilled before an encoding may be considered as an encoding with the prefix 'T' satisfying the requirements of the LaTeX 2$_\varepsilon$ kernel:

- the `\lccode`–`\uccode` pairs should be the same as they are in the LaTeX 2$_\varepsilon$ kernel (i.e., as they are in the T1 encoding);
- the Latin characters and symbols: `!`, `'`, `(`, `)`, `*`, `+`, `,`, `-`, `.`, `/`, `:`, `;`, `=`, `?`, `[`, `]`, `` ` ``, `|`, `@` (questionable), `0`–`9`, `A`–`Z`, `a`–`z` should be at the positions corresponding to ASCII, and the symbols produced by the ligatures `--`, `---`, ` `` `, `''` (at arbitrary positions).

If the encoding requires the redefinition of the values `\lccode`–`\uccode`, or if it does not contain the necessary Latin characters in the ASCII positions, it will produce undesirable effects in some situations inside LaTeX 2$_\varepsilon$ and will make use of the encoding incompatible with the general multi-language mode. The reasons for such restrictions are explained in detail in [1].

Although the LaTeX Team's technical specifications for X$\langle n\rangle$ encodings are less restrictive than those for 'ordinary' text encodings, there are corresponding restrictions on their use, and some *desirable* properties for them to have. In particular:

- If the encoding does not have Latin letters in ASCII slots then the users must take care not to enter such text, otherwise 'random' incorrect output will be produced, with no warning from

[2] The following text is slightly adapted from a post by David Carlisle to the mailing list `cyrtex-t2`.

the LaTeX system. Also, care must be taken with 'moving' text that is generated internally within LaTeX (such as cross references), which may fail if the encodings change;

- To reduce the problems with cross reference information, the LaTeX maintainers strongly recommend that at least the digits and 'common' punctuation characters are placed in their ASCII slots;
- If the encoding uses a lowercase table that is incompatible with the lowercase table of T1, then it is not possible to mix this encoding and a T$\langle n\rangle$ encoding within a single paragraph, and obtain correct hyphenation with standard TeX.

If the X$\langle n\rangle$ encoding does not use a lowercase table that is compatible with that of T1, the package supporting this encoding should ensure that encoding switches only happen between paragraphs (or that hyphenation is suppressed when temporarily switching to the new encoding). It should be noted that this restriction on the lowercase table applies *only* to systems using standard TeX (version 3 and later). Using ε-TeX version 2 will remove the need for this restriction as the hyphenation system has been improved — it will use a suitable lowercase table for each language (the table will be stored along with each language hyphenation table), and surely it deals not at all with the *Omega* system.

4 "Cyrillic glyph container" — the X2 encoding

The encoding X2 should include all the glyphs necessary to represent in LaTeX 2$_\varepsilon$ documents containing texts from stable Cyrillic languages. The basis of X2 is the Russian alphabet (since it is the main language used for publication in Cyrillic). Taking account of the variety of old Cyrillic texts, only those modern alphabets which are still in use are included in X2. The exceptions are the characters Ѣ/ѣ, Ѫ/ѫ, Ѵ/ѵ which were used in Russian and Bulgarian texts at the beginning of the 20th century.

The X2 encoding is designed so that by combining `"00 – "7f` from OT1 and `"80 – "ff` from X2 one can construct an encoding which is adequate to support the most common Cyrillic languages. This permits use of X2 as the base Cyrillic encoding for a variety of TeX formats (*Plain*, $\mathcal{AMS}$-TeX, *BLUE*TeX, LaTeX 2.09, etc.) as well as LaTeX 2$_\varepsilon$. (This local encoding is called LR1 below. The design aim for LR1 was to select glyphs required by the most widely-used languages and to put them into the 128 – 255 section of X2.)

Unfortunately the full set of glyphs including accented letters is too big to fit into 256 slots, especially taking into account the `\lccode`–`\uccode` restrictions. So it is necessary to accept some principles of selection which enable us to decrease the number of Cyrillic glyphs included in X2:

1. The X2 encoding follows the LaTeX 2$_\varepsilon$ agreements about `\lccode`–`\uccode` not to produce garbage for the headings, table of contents, hyphenations inside paragraphs, arguments of `\uppercase` and `\lowercase`;
2. All glyphs used in publishing for some language are included in X2 if they cannot be constructed as accented letters or letters with additional modifiers using TeX commands. Variant glyphs for Cyrillic alphabets are also included in X2 if there is some free space and if different languages use different variants;
3. The X2 encoding includes all punctuation symbols, digits, mathematical symbols, accents, hyphens, dashes, etc., needed to form the full set of symbols necessary for Cyrillic typography;
4. The additional Cyrillic letters which are used in the PC 866 and MS Windows 1251 code pages are included in X2 even if they are accented forms;
5. Glyphs which are not used now but which were used at some stage in the 20th century may be included if there are good reasons to do so (as, for example, with the old Russian and Bulgarian letters);
6. Glyphs which were used in old Cyrillic texts before 1900 (Old Slavonic, Church Slavonic, Glagolitic, old phonetic symbols, etc.) should be moved to a separate glyph container. There could also be an additional glyph container to collect the exotic glyphs used in some contemporary Cyrillic texts;
7. When jettisoning accented letters it is necessary to take into account that they may be necessary for hyphenation patterns for some languages (if such patterns have been created or if there is a chance that they will be created sometime). For example, accented letters for Russian, Ukrainian and Belorussian, Kazakh, Tatar, and Bashkir are included in X2;
8. When deciding whether to jettison an accented letter that is used in a language supported by LR1, one must keep in mind that only the CM accents are available in that encoding;
9. The following priorities are used when the accented letters or letters with simple modifiers are thrown away: (0) letters which are easily constructed by the internal command `\accent` (so that the letters using accents available in

CM fonts have lower significance); (1) letters which contain a centered diacritic below the letter (cedilla, ogonek, dot, macron) and are easily constructed using a command similar to `\c` in *Plain* TEX; (2) letters which contain a horizontal stroke positioned symmetrically; (3) letters which require special alignment of accents and modifiers;

10. Accents and modifiers used in Cyrillic are included in X2 even if all accented forms are included in X2 for some other reasons (an example is *Cyrillic breve* used for Й and Ў);
11. Latin letters or glyphs which are similar to some Latin letter (used in Macedonian, Kurdish, etc.) are placed at the same positions as the Latin letters are in ASCII. Among other things, this increases the number of languages supported by the LR1 encoding;
12. Whenever it is possible, glyphs (ASCII, accents, special symbols, etc.) are placed at the same positions as they are placed in T1 encoding.

The X2 encoding is shown in Fig. 1. The Russian letters А – Я, а – я (except Ё and ё) are placed in the only region in the encoding table where 32 consecutive letter positions are available — i.e., positions `"c0` – `"df` and `"e0` – `"ff`. The Russian letters Ё and ё are placed at the end of the block `"80` – `"9c` and `"a0` – `"bc` which simplifies the ordering of non-Russian letters. Latin letters and letters similar to Russian letters are placed as in ASCII. Letters used in other Cyrillic alphabets are grouped into the parts `"80` – `"ff` and `"00` – `"7f` of the encoding table according to the “popularity” of the corresponding languages (to satisfy the requirements of the LR1 encoding). They are placed in free positions reserved by LaTeX 2_ε for letters in some quasi-alphabetic order. The old Russian and Bulgarian letters are placed at the end of the block of letters in `"00` – `"7f`.

Accents and modifiers are placed in X2 at `"00` – `"1f`; those also used by T1 are placed at the same positions as in T1. The same is true for additional symbols produced by the ligatures `--`, `''`, etc. The punctuation symbols, digits, mathematical symbols, etc., are placed as they are positioned in ASCII. A special case is made of the symbols № ¤ § „ « » which are essential for Russian typography. These symbols are placed in `"80` – `"ff` at the positions reserved for symbols, to guarantee the correctness of the LR1 encoding.

Some accents (macron, dot) can be used as lower accents as well for transliteration systems. In some specific cases the upper comma (`"1b`) and lower comma are also used as accents. The lower accents will be constructed using TEX commands from the upper accents available in X2.

The accents ̑ (`"12`) and ̏ (`"13`) are used as stresses in Serbian; there is no letter in any Cyrillic language where these symbols are used as “normal” accents.

The quasi-letters ’ (apostrophe, `"27`), ” (double apostrophe, `"22`) and Ӏ (palochka, `"0d`) are used like letters in some languages but do not have uppercase and lowercase forms (i.e., for these letters the uppercase form is just the same as the lowercase form).

Single quotes are not used in Cyrillic writing, and for this reason there is no need to keep single French quotes. Instead, in their place, the angle brackets ⟨ (`"0e`) and ⟩ (`"0f`) are provided. Angle brackets *are* used in Cyrillic typography, and it is good if their style is changed in parallel with the style of other symbols.

The Cyrillic breve “˘” (`"14`) is a very famous glyph (it is even included in the Adobe and WordPerfect Cyrillic fonts). Although all letters with this accent (Й/й, Ў/ў) are included in X2, it is included as a special glyph as well.

Cedilla “¸” (`"0b`) and ogonek “˛” (`"0c`) are used by some letters already included in X2 (Ҙ, Ҫ, Ҿ). These letters have variant forms where *cedilla* could be oriented to the left or to the right depending on the user’s taste. Also, some applications use *ogonek* instead of *descender* for Қ, Ҳ, Ӷ, Ҭ, etc. The availability of *cedilla* and *ogonek* in X2 makes it possible to satisfy these needs.

Percentage zero “₀” (`"18`) is included as a useful idea borrowed from the T1 encoding and EC fonts: this symbol is used to convert ‘%’ into ‘‰’ and ‘‱’.

Punctuation ligatures, i.e., the symbols produced by the abbreviations `--` (endash, `"15`), `---` (emdash, `"16`), ` `` ` (opening English quotes, `"10`), `''` (closing English quotes, `"11`) are used in the same manner and are placed at the same position as in T1, as is `-` (the hyphen used for hanging Hyphenation, `"7f`). It is worth noting that the ligature `---` (emdash, `"16`) corresponds to *Cyrillic emdash* which (following the traditions of Russian typography) is much shorter than that glyph in Latin-encoded CM and EC fonts.

There are the special cross-modifiers: horizontal “-” at `"17`, grave-diagonal “⸌” at `"19` and acute-diagonal “⸍” at `"1a` which are used to construct from pieces the letters Ӻ/ӻ, Ӿ/ӿ, Ꞓ/ꞓ and Ᵽ/ᵽ used in some minor Cyrillic languages. (These letters are included as separate glyphs in T2A/T2B/T2C, but

	x0/x8	x1/x9	x2/xA	x3/xB	x4/xC	x5/xD	x6/xE	x7/xF	
0x	`	´	ˆ	˜	¨	˝	˚	ˇ	0x
	˘	¯	˙	¸	˛	I	⟨	⟩	
1x	“	”	⌢	˵	ᵕ	–	—	-	1x
	₀	ˎ	ˏ	’	δ	δ	Ԋ	ԋ	
2x	␣	!	"	#	$	%	&	’	2x
	(	)	*	+	,	-	.	/	
3x	0	1	2	3	4	5	6	7	3x
	8	9	:	;	<	=	>	?	
4x	@	Æ	Ђ	Ћ	Ҽ	Ҿ	Ҝ	Ҟ	4x
	Ӆ	I	J	Љ	Ӎ	Њ	Ҩ	Ҧ	
5x	Ҏ	Q	Ҭ	S	Ҵ	Џ	Ҹ	W	5x
	Ѣ	Ѫ	Ѵ	[	\	]	^	_	
6x	‘	æ	ђ	ћ	ҽ	ҿ	ҝ	ҟ	6x
	ӆ	i	j	љ	ӎ	њ	ҩ	ҧ	
7x	ҏ	q	ҭ	s	ҵ	џ	ҹ	w	7x
	ѣ	ѫ	ѵ	{	|	}	~	-	
8x	Ґ	Ғ	Ӷ	Ҕ	Һ	Җ	Ҙ	Ӡ	8x
	Ї	Қ	Ҡ	Ӄ	Ԓ	Ң	Ҥ	Ӈ	
9x	Ө	Ҫ	Ў̈	Ү	Ұ	Ҳ	Ӽ	Ӌ	9x
	Ҷ	Є	Ә	Ԑ	Ё	№	¤	§	
Ax	ґ	ғ	ӷ	ҕ	һ	җ	ҙ	ӡ	Ax
	ї	қ	ҡ	ӄ	ԓ	ң	ҥ	ӈ	
Bx	ө	ҫ	ў̈	ү	ұ	ҳ	ӽ	ӌ	Bx
	ҷ	є	ә	ԑ	ё	„	«	»	
Cx	А	Б	В	Г	Д	Е	Ж	З	Cx
	И	Й	К	Л	М	Н	О	П	
Dx	Р	С	Т	У	Ф	Х	Ц	Ч	Dx
	Ш	Щ	Ъ	Ы	Ь	Э	Ю	Я	
Ex	а	б	в	г	д	е	ж	з	Ex
	и	й	к	л	м	н	о	п	
Fx	р	с	т	у	ф	х	ц	ч	Fx
	ш	щ	ъ	ы	ь	э	ю	я	
	x0/x8	x1/x9	x2/xA	x3/xB	x4/xC	x5/xD	x6/xE	x7/xF	

Figure 1: The “Cyrillic glyph container” X2

unfortunately the limit of 256 characters prevents including them in X2.)

The positions `"1c/"1d` and `"1e/"1f` are used for the exotic letters Ꙃ/ꙃ and Ӈ/ӈ used by some minor Cyrillic alphabets. Although following LaTeX 2ε rules these positions should be used for *symbols*, not for *letters*, we have made them exceptions from the severe LaTeX 2ε requirements. Formally speaking, the LaTeX 2ε requirements are not violated since the `\lccode`–`\uccode` data for these positions conserve the original values. Instead of the explicit use of the `\lccode`–`\uccode` mechanism, the lowercase–uppercase conversion is performed by the LaTeX 2ε `\MakeUppercase` and `\MakeLowercase` transformations using the list `\@uclclist` of identifiers. As a result these "letters" cannot be used in hyphenation patterns and they break the automatic hyphenation whenever they appear in a word. The gain is that even exotic Cyrillic texts could be created (if necessary) using X2 only and without additional encodings and fonts.

5 "True" LaTeX 2ε encodings T2A, T2B, T2C

The base features of the T2∗ encodings are determined by the LaTeX 2ε specifications for T$\langle n\rangle$ encodings and by the already created X2 encoding. Some more requirements are added due to the necessity to keep the fonts in the T1, X2 and T2∗ encodings compatible. For example, it is necessary to keep similar glyphs at the same positions in all fonts whenever it is possible. As a result the following basic principles appear.

1. The set of T2∗ encodings supports the full set of modern Cyrillic languages, each Cyrillic language is supported at least by one encoding T2∗ so that there is no necessity to mix encodings for some languages;
2. Cyrillic letters occupy the positions reserved in LaTeX 2ε for letters, Cyrillic symbols — positions reserved for symbols, Cyrillic accents — positions reserved for accents;
3. Letters included in T2∗ follow the LaTeX 2ε convention about uppercase and lowercase letters (i.e., have the same `\uccode`–`lccode` assignments as in T1);
4. ASCII glyphs (Latin letters, digits, mathematical and punctuation symbols, etc.) are placed at `"20`–`"7f`;
5. The symbols produced by the ligatures `--`, `---`, ` `` `, `''` are included at the same positions as they are in T1 and X2 (it is worth noting that, as in X2, the *emdash* glyph is the *Cyrillic emdash* which is shorter than that in OT1 and T1);
6. The ff-ligatures (ff, fi, fl, ffi, ffl) are included at `"1b`–`"1f` as in T1 to keep the full set of Latin glyphs necessary for typography;
7. The standard accents and symbols, and Cyrillic-specific accents and symbols used in `"00`–`"1a` of X2, are reproduced in T2∗ at the same positions except the cross-modifiers which are not necessary (the letters with cross-modifiers are included as separate glyphs);
8. The Russian alphabet is reproduced similar to X2;
9. The symbols specific to Cyrillic typography (№ ¤ § „ « ») are reproduced in `"9d`, `"9e`, `"9f`, `"bd`, `"be` and `"bf`, similar to X2;
10. Positions `"80`–`"9b` and `"a0`–`"bb` are used for national Cyrillic letters (these are the only positions that differ from the T2∗ encodings since all other codes are already fixed as described above);
11. To prevent an increase in the number of encodings up to infinity, the accented letters for Cyrillic languages which do not have the hyphenation tables in TeX format (and rarely will have in future) are not included;
12. Equivalent letters occupy just the same positions in all the T2∗ encodings whenever possible (i.e., some glyphs may be absent in some encodings, but if the glyph is included in an encoding, it occupies the same position as in the other encodings).

The encodings can therefore be decomposed into the following regions:

- the accent, non-ASCII punctuation and ligature symbols (`"00`–`"1f`),
- the ASCII-encoded Latin letters, digits, punctuation and mathematical symbols, etc. (`"20`–`"7f`),
- non-letter symbols at `"9d`–`"9f` and `"bd`–`"bf` (№ ¤ § „ « »),
- specific (uppercase and lowercase) Cyrillic letters (`"80`–`"9b`, `"a0`–`"bb`),
- uppercase and lowercase Russian letters (`"c0`–`"df`, `"e0`–`"ff`, `"9c`, `"bc`).

The accent part (see Fig. 2) is copied from X2 with minor changes:

a) the exotic letters (`"1c`–`"1f`) and the upper comma accent (`"1b`) are substituted by the ff-ligatures "*ff*", "*fi*", "*fl*", "*ffi*", "*ffl*",

	x0/x8	x1/x9	x2/xA	x3/xB	x4/xC	x5/xD	x6/xE	x7/xF	
0x	`	´	ˆ	˜	¨	˝	˚	ˇ	0x
	˘	¯	˙	¸	˛	I	〈	〉	
1x	“	”	⌢	˵	̑̑	–	—		1x
	₀	ı	ȷ	ff	fi	fl	ffi	ffl	

a) Accents, ligatures, special symbols, etc.

	x0/x8	x1/x9	x2/xA	x3/xB	x4/xC	x5/xD	x6/xE	x7/xF	
2x	␣	!	"	#	$	%	&	’	2x
	(	)	*	+	,	-	.	/	
3x	0	1	2	3	4	5	6	7	3x
	8	9	:	;	<	=	>	?	
4x	@	A	B	C	D	E	F	G	4x
	H	I	J	K	L	M	N	O	
5x	P	Q	R	S	T	U	V	W	5x
	X	Y	Z	[	\	]	^	_	
6x	‘	a	b	c	d	e	f	g	6x
	h	i	j	k	l	m	n	o	
7x	p	q	r	s	t	u	v	w	7x
	x	y	z	{	|	}	~	-	

b) ASCII glyphs

	x0/x8	x1/x9	x2/xA	x3/xB	x4/xC	x5/xD	x6/xE	x7/xF	
Cx	А	Б	В	Г	Д	Е	Ж	З	Cx
	И	Й	К	Л	М	Н	О	П	
Dx	Р	С	Т	У	Ф	Х	Ц	Ч	Dx
	Ш	Щ	Ъ	Ы	Ь	Э	Ю	Я	
Ex	а	б	в	г	д	е	ж	з	Ex
	и	й	к	л	м	н	о	п	
Fx	р	с	т	у	ф	х	ц	ч	Fx
	ш	щ	ъ	ы	ь	э	ю	я	

c) Russian letters

Figure 2: Common parts in T2A/T2B/T2C

b) the cross-modifiers are substituted by the *compound-word-mark* symbol[3] (`"17`), the *dotless-i* (`"19`) and the *dotless-j* (`"1a`) like it is in T1.

The ASCII-encoded Latin letters, digits, punctuation and mathematical symbols at `"20`–`"7f` are placed exactly as in the T1 encoding as shown in Fig. 2.

The Russian alphabet letters and non-letter symbols at `"9d`–`"9f`, `"bd`–`"bf` are just copied from the corresponding positions in the X2 encoding (see Fig. 2).

The component which is different in T2A/T2B/T2C/ are the national letters at `"80`–`"9b` and `"a0`–`"bb`, which are shown in Fig. 3 and Table 1. Although we tried to fulfill the requirement that the equivalent letters should occupy the same positions in all encodings, it was impossible to completely fulfill this requirement. Fortunately there are only two exceptions: the letters Љ/љ and Њ/њ are placed at `"87`/`"a7` and `"9b`/`"bb` in T2A, but at `"88`/`"a8` and `"99`/`"b9` in T2B. All other letters and symbols in T2A, T2B and T2C (and in the accent part `"00`–`"1a`, symbol/digit part `"20`–`"3f`, `"5b`–`"5f`, `"7b`–`"7f` and lower part `"80`–`"ff` of X2) have fixed positions.

A summary of the languages covered by T2A/T2B/T2C is shown below. The encoding T2A contains the leading languages sorted by using statistical data on populations. The encoding T2B contains the majority of the remaining languages. Finally, the encoding T2C contains several languages with exotic letters which do not fit into T2A or T2B: Abkhazian, Orok (Uilta), Saam (Lappish), Old-Bulgarian, Old-Russian. Due to the intersections between Cyrillic alphabets some languages are supported by two or three encodings simultaneously:

T2A: Abaza, Avar, Agul, Adyghei, Azerbaidzan, Altai, Balkar, Bashkir, Belorussian, Bulgarian, Buryat, Gagauz, Dargin, Dungan, Ingush, Kabardino-Cherkess, Kazah, Kalmyk, Karakalpak, Karachaevskii, Karelian, Kirgiz, Komi-Zyrian, Komi-Permyak, Kumyk, Lak, Lezgin, Macedonian, Mari-Mountain, Mari-Valley, Moldavian, Mongolian, Mordvin-Moksha, Mordvin-Erzya, Nogai, Oroch, Osetin, Russian, Rutul, Serbian, Tabasaran, Tadjik, Tatar, Tati, Teleut, Tofalar, Tuva, Turkmen, Udmurt, Uzbek, Ukrainian, Hanty-Obskii, Hanty-Surgut, Gipsi, Chechen, Chuvash, Crimean-Tatar;

T2B: Abaza, Avar, Agul, Adyghei, Aleut, Altai, Balkar, Belorussian, Bulgarian, Buryat, Gagauz, Dargin, Dolgan, Dungan, Ingush, Itelmen, Kabardino-Cherkess, Kalmyk, Karakalpak, Karachaevskii, Karelian, Ketskii, Kirgiz, Komi-Zyrian, Komi-Permyak, Koryak, Kumyk, Kurdian, Lak, Lezgin, Mansi, Mari-Valley, Moldavian, Mongolian, Mordvin-Moksha, Mordvin-Erzya, Nanai, Nganasan, Negidal, Nenets, Nivh, Nogai, Oroch, Russian, Rutul, Selkup, Tabasaran, Tadjik, Tatar, Tati, Teleut, Tofalar, Tuva, Turkmen, Udyghei, Uigur, Ulch, Khakass, Hanty-Vahovskii, Hanty-Kazymskii, Hanty-Obskii, Hanty-Surgut, Hanty-Shurysharskii, Gipsi, Chechen, Chukcha, Shor, Evenk, Even, Enets, Eskimo, Yukagir, Crimean Tatar, Yakut;

T2C: Abkhazian, Bulgarian, Gagauz, Karelian, Komi-Zyrian, Komi-Permyak, Kumyk, Mansi, Moldavian, Mordvin-Moksha, Mordvin-Erzya, Nanai, Orok (Uilta), Negidal, Nogai, Oroch, Russian, Saam, Old-Bulgarian, Old-Russian, Tati, Teleut, Hanty-Obskii, Hanty-Surgut, Evenk, Crimean Tatar.

6 The Cyrillic glyph container X2 versus Cyrillic encodings T2A, T2B, T2C

As was already specified, there are two main requirements essential to the reliable working of LaTeX 2ε in multi-language mode:

- we must keep the `\lccode`–`\uccode` table as in T1,
- we must keep the ASCII encoding for positions 32–127.

Both requirements are fulfilled by the T2A/T2B/T2C encodings and hence they can be safely mixed with the Latin encodings OT1 and T1 inside a document. The encoding X2 conserves the `\lccode`–`\uccode` values but does not contain these ASCII glyphs. As a result it *may* cause problems and unexpected effects inside LaTeX 2ε documents if the user is not careful enough. So, why do we need X2 when we have T2A/T2B/T2C?

The reason is that the requirement to keep all the ASCII glyphs is very restrictive — it leaves only 61 positions for non-ASCII letters.[4] To fit all Cyrillic letters into T$\langle n\rangle$ encodings requires *three* tables

[3] The empty character with the zero thickness and the height equal to 1*ex* used in EC fonts and T1 encoding for special applications — such as hyphenating compound words, breaking down ligatures, creating accents to be placed over the invisible space between two letters.

[4] For Cyrillic encodings it is even more restrictive: it is necessary to keep 32 base Russian letters in each encoding as well since they are encountered in almost all Cyrillic alphabets.

	x0/x8	x1/x9	x2/xA	x3/xB	x4/xC	x5/xD	x6/xE	x7/xF	
8x	Ґ	Ғ	Ђ	Ћ	Һ	Җ	Ҙ	Љ	8x
	Ї	Қ	Ҡ	Ҝ	Ӕ	Ң	Ҥ	Ѕ	
9x	Ө	Ҫ	Ў	Ү	Ұ	Ҳ	Ҵ	Ҹ	9x
	Ҷ	Є	Ә	Њ	Ё	№	¤	§	
Ax	ґ	ғ	ђ	ћ	һ	җ	ҙ	љ	Ax
	ї	қ	ҡ	ҝ	ӕ	ң	ҥ	ѕ	
Bx	ө	ҫ	ў	ү	ұ	ҳ	ҵ	ҹ	Bx
	ҷ	є	ә	њ	ё	„	«	»	

a) T2A encoding

	x0/x8	x1/x9	x2/xA	x3/xB	x4/xC	x5/xD	x6/xE	x7/xF	
8x	Ӻ	Ғ	Ӷ	Ҕ	Һ	Җ	δ	Ӡ	8x
	Љ	Қ	Ӆ	Ӄ	Ԓ	Ң	Ҥ	Ӈ	
9x	Ө	Ҫ	Ў	Ү	Ӿ	Ҳ	Ӽ	Ӌ	9x
	Ҷ	Њ	Ә	Ԑ	Ё	№	¤	§	
Ax	ӻ	ғ	ӷ	ҕ	һ	җ	δ	ӡ	Ax
	љ	қ	ӆ	ӄ	ԓ	ң	ҥ	ӈ	
Bx	ө	ҫ	ў	ү	ӿ	ҳ	ӽ	ӌ	Bx
	ҷ	њ	ә	ԑ	ё	„	«	»	

b) T2B encoding

	x0/x8	x1/x9	x2/xA	x3/xB	x4/xC	x5/xD	x6/xE	x7/xF	
8x	Ԥ	Ҵ	Ҭ	Ҕ	Һ	Р̦	Ҏ	Ӡ	8x
	Ӎ	Қ	Ӆ	Ҟ	Ԓ	Ң	М̡	Ӈ	
9x	Ө	Ҽ	Ҿ	Ѣ	Ѵ	Ҳ	Ҵ	Ѹ	9x
	Ҷ	Ӊ	Ә	Ж	Ё	№	¤	§	
Ax	ԥ	ҵ	ҭ	ҕ	һ	р̦	ҏ	ӡ	Ax
	ӎ	қ	ӆ	ҟ	ԓ	ң	м̡	ӈ	
Bx	ө	ҽ	ҿ	ѣ	ѵ	ҳ	ҵ	ѹ	Bx
	ҷ	ӊ	ә	ж	ё	„	«	»	

c) T2C encoding

Figure 3: The national Cyrillic letters in T2A/T2B/T2C

Code	T2A	T2B	T2C
`"80/"a0`	ghe-upturned	ghe-with-descender hcrossed	pe-with-tail
`"81/"a1`	ghe-hcrossed	ghe-hcrossed	te+tse
`"82/"a2`	dje (T+h with tail)	ghe-with-descender	te-with-descender
`"83/"a3`	tshe (T+h)	ghe-with-tail	ghe-with-tail
`"84/"a4`	h-special	h-special	h-special
`"85/"a5`	zhe-with-descender	zhe-with-descender	er-with-descender
`"86/"a6`	ze-with-descender	delta-phonetical	er-gravecrossed
`"87/"a7`	lje	zet	zet
`"88/"a8`	i-with-umlaut	lje	em-with-descender
`"89/"a9`	ka-with-descender	ka-with-descender	ka-with-descender
`"8a/"aa`	ka-with-left-poker	el-with-descender	el-with-descender
`"8b/"ab`	ka-vcrossed	ka-with-tail	ka-hcrossed
`"8c/"ac`	a+e	el-with-tail	el-with-tail
`"8d/"ad`	en-with-descender	en-with-descender	en-with-descender
`"8e/"ae`	en+ghe	en+ghe	em-with-tail
`"8f/"af`	S	en-with-tail	en-with-tail
`"90/"b0`	o-barred	o-barred	o-barred (fita)
`"91/"b1`	es-with-descender	es-acutecrosseds	abkhazian-ch
`"92/"b2`	u-with-cyrbreve	u-with-cyrbreve	abkhazian-ch with descender
`"93/"b3`	Y-special	Y-special	yat (semisoft sign)
`"94/"b4`	Y-special-hcrossed	ha-hcrossed	izhitsa
`"95/"b5`	ha-with-descender	ha-with-descender	ha-with-descender
`"96/"b6`	tse-macedonian	ha-with-tail	tse-macedonian
`"97/"b7`	che-vcrossed	che-with-left-descender	abkhazian-ha
`"98/"b8`	che-with-descender	che-with-descender	che-with-descender
`"99/"b9`	e-ukrainian	nje	en-with-right-tail
`"9a/"ba`	shwa	shwa	shwa
`"9b/"bb`	nje	epsilon	big yus

Table 1: The national Cyrillic letters in encodings T2A/T2B/T2C

to achieve the coverage of X2: the problem is that most characters in the T$\langle n\rangle$ tables are the same. And to support each such T$\langle n\rangle$ encoding it is necessary to have a separate font class like the EC fonts.

To keep such enormous numbers of fonts is too large a price for people who use Cyrillics only occasionally. On the other hand, if all the Cyrillic glyphs are put into just one table without the Latin letters in 32−127, but in a way that satisfies the `\lccode`–`\uccode` requirements, one table and one font class is enough provided the user obeys some elementary rules of safety. This "economy mode" is implemented by X2.

There is a similar situation for Old Slavonic characters and some other encodings which are only occasionally used by normal users. To resolve this problem, "glyph containers" like X2 could again be helpful. The "glyph container" encodings X$\langle n\rangle$ should be an intermediate case between T$\langle n\rangle$ and the "free style" X$\langle n\rangle$: such encodings do not have ASCII in 32−127, but they do have the standard, compatible `\lccode`–`\uccode` values.

Currently the LaTeX Team only supports the T$\langle n\rangle$ encodings and TS encodings, while the support of an X$\langle n\rangle$ encoding is entirely the responsibility of the designer of the encoding.[5] For the "glyph container" X2 such support should include listing some simple rules which should be followed by Users so as to avoid strange and undesirable effects. The X2 encoding does not contain Latin ASCII letters but only digits, punctuation and mathematical symbols, etc., therefore the rules should guarantee that text containing Latin ASCII letters is never used with the X2 encoding:

[5] It seems that there *should be* support for such "glyph container" encodings by the LaTeX Team as well (such support should include the registration procedure for glyph containers and maintenance of the official list of exceptions where the glyph container encodings produce undesirable results).

1. If you use ASCII Latin letters in the text part of your document, some Latin encoding (i.e., OT1 or T1) must be active for this piece of your text;
2. The following commands should be used only outside the range where the X2 encoding is active (or should be preceded by the explicit specification of some "Latin" encoding) because they may implicitly include the Latin text in your document:

   \part, \chapter, \section, \subsection, \subsubsection, \paragraph, \subparagraph, \caption, \tableofcontents, \listoftables, \listoffigures, \ref, \pageref, \cite, \item, \labelenum*xx*, \labelitem*xx*, \makelabel, \numberline, \thechapter, \thesection, \thesubsection, \date, \today.

   Exactly the same requirement is needed for all user-defined macros (or those loaded from external packages) that have ASCII Latin text in their body but without explicit specification of the "Latin" encodings OT1 or T1 for this text;
3. When you deal with floating objects (or moving arguments), you should not rely on the assumption that the Cyrillic letters are used by LaTeX when the floating material is inserted into the document. For example, if Cyrillic letters are used inside some command defining the floating object, the encoding X2 should be activated explicitly in front of Cyrillic letters even if X2 is active at the point where the command is issued. Among such commands are:
   (a) the floating environments:

   \begin{table}−\end{table}, \begin{figure}−\end{figure}, \begin{table*}−\end{table*}, \begin{figure*}−\end{figure*},

   (b) the commands that define floating text explicitly:

   \author, \title, \date, \address, \name, \signature, \telephone, \footnote, \footnotetext, \thanks, \marginpar, \markboth, \markright, \bibitem, \topfigrule, \botfigrule, \dblfigrule, \footnoterule,

   (c) the commands that define headers, footers, margin remarks, etc., implicitly:

   \part, \chapter, \section, \subsection, \subsubsection, \paragraph, \subparagraph, \caption,

   (d) the commands that write, explicitly or implicitly, text to external files, which may be loaded outside the X2 encoding:

   \addtocontents, \addcontentsline, \glossary, \index, \part, \chapter, \section, \subsection, \subsubsection, \paragraph, \subparagraph, \caption.

   Similarly, if such a floating command includes Latin letters and the resulting object may appear inside the range where X2 is active, some Latin encoding (i.e., OT1 or T1) should be activated explicitly before the Latin-encoded text.
4. Just the same requirement holds for *all* commands which can occasionally insert Cyrillic or Latin text where the Latin encodings OT1/T1 or the "Cyrillic glyph container" encoding X2 are active. For example, you should be careful with the definition and re-definition of the following commands:
   (a) commands which create automatically generated text used by other commands:

   \labelenum*xx*, \labelitem*xx*, \makelabel, \numberline, \thechapter, \thesection, \thesubsection, \today,

   (b) commands which are used in international LaTeX to define language-specific names:

   \abstractname, \appendixname, \alsoname, \ccname, \chaptername, \contentsname, \enclname, \headtoname, \figurename, \indexname, \listfigurename, \listtablename, \notesname, \pagename, \partname, \prefacename, \seename, \tablename,

   (c) commands which implicitly define headers, footers, margin remarks, etc., and/or implicitly write something into external files:

   \part, \chapter, \section, \subsection, \subsubsection, \paragraph, \subparagraph, \caption, \addtocontents, \addcontentsline, \glossary, \index,

   (d) commands which create floating text and floating environments:

   \author, \title, \date, \address, \name, \signature, \telephone, \footnote, \footnotetext, \thanks, \marginpar, \markboth, \markright, \bibitem, \topfigrule, \botfigrule, \dblfigrule, \footnoterule, \begin{table}−\end{table}, \begin{figure}−\end{figure},

`\begin{table*}`–`\end{table*}`,
`\begin{figure*}`–`\end{figure*}`,

(e) macros and user-defined commands which may be expanded unintentionally inside or outside X2:

`\def`, `\newcommand`, `\newcommand*`,
`\renewcommand`, `\renewcommand*`,
`\providecommand`, `\providecommand*`,
`\newenvironment`, `\newenvironment*`,
`\renewenvironment`,
`\renewenvironment*`,
`\newtheorem`, `\ProvideTextCommand`,
`\ProvideTextCommandDefault`,
`\AtBeginDocument`, `\AtEndDocument`,
`\AtEndClass`, `\AtEndPackage`,
`\DeclareRobustCommand`,
`\DeclareTextCommand`,
`\DeclareTextCommandDefault`.

7 The weak points of X2 and T2∗

The X2 and T2∗ encodings do not contain accented letters, and (for some languages) this throws the user back on the `\accent` primitive which prevents construction of correct hyphenation tables and destroys kerning pairs. The encodings (especially X2) are also overloaded (to some extent) with rare glyphs, which arise from the attempt to collect *all* Cyrillic glyphs in one table.

There are the cross-modifiers (horizontal stroke "-", vertical stroke "|", diagonal strokes "⟍" and "⟋") which are included in X2 but are absent in T2A/T2B/T2C. Although there is a great chance that these glyphs will be included in TS2 (see section 8), their status at this stage of the project is undefined. Similarly, there are the title forms[6] for the letters Љ/љ and Њ/њ which were included in previous (intermediate) versions of X2 but are now excluded for some reason.

Another disadvantage of minor importance is that there are two glyphs (Ѣ/ѣ and Ѳ/ѳ) which correspond to logically different letters: Ѣ/ѣ stands for Saam *semisoft sign* and for old Russian *yat*, and Ѳ/ѳ stands for *o-barred* and old Russian *fita*. Although graphically these symbols are similar, they are different logically.

This situation can be accepted taking into account the status of X2 as a glyph table rather than a table for direct text coding, and the status of T2∗ as the *modern Cyrillic* encodings. In structured markup, the ambiguity would be addressed by assigning *two* symbolic names for each glyph (say, `\yat`/`\semisft` and `\fita`/`\obarred`) and only using the semantically correct one to code texts.

Some preliminary information about exotic glyphs and pure phonetic symbols has been provided by linguists studying some minor writing systems. These letters and symbols are not currently included in X2 and T2∗ at all. The reason for not including the glyphs at this stage is that the writing systems are very unstable and are subject to change from publication to publication. There is no justification for including such symbols in the version of X2 and T2∗ proposed as a *standard* until the situation becomes stable.

It seems that all the specific Cyrillic glyphs used in modern Cyrillic alphabets are included in X2 and in one of the T2∗, but there is also a chance that some minor writing system is omitted. There is also a chance that some linguists suggest a new alphabet for some minor language using their own glyphs not available in X2. Until this happens we can consider X2 and T2A/T2B/T2C as comprehensive glyph collections for modern Cyrillic texts (although not very comfortable and not specifically adjusted for intensive Cyrillic writing).

8 Some remarks about the TS2 encoding

The TS2 is expected to be the collection of accents and special symbols which are necessary for Cyrillic typography, but which are not included into the encodings X2 and T2A/T2B/T2C for some reasons (i.e., TS2 is the encoding supplementary to X2 and T2⟨*n*⟩ as TS1 is supplementary to T1).

For typographical reasons, 'wide' versions of some accents — macron, tilde, breve, etc. — are desirable. These versions would be used for extra wide letters: as compared with the Latin alphabet, Cyrillic has a far higher proportion of wide letters. Such wide versions of the accents are good candidates for a TS2 encoding. Similarly, the lowercase/uppercase variants of cedilla, ogonek and the accents absent in T1 and TS1 may make useful additions to TS2.

The letters Љ/љ and Њ/њ used in some Cyrillic languages are actually ligatures "Л+Ь" and "Н+Ь". As well as the uppercase and lowercase forms there is also a *title* form for these letters: the combination of the uppercase form for "Л" or "Н" and the bowl for the lowercase "ь". This form is used for titles where the first letter is capital while the other letters are ordinary (a similar effect occurs for 'IJ' used in Dutch). Such title letters

[6] The title form is a combination of the uppercase "Л" or "Н" and the bowl from the *lowercase* "ь". These glyphs are used for first letters in titles, etc., where the first letter is capital and other letters are in lowercase mode. For example, there is the title form "Ij" for the ligature "IJ"/"ij" used in Dutch. (Note, that the title form "Ij" is absent in T1 and TS1 encodings.)

should be placed in TS2 and shared by the X2 and T2∗ encodings.

To construct some exotic letters from pieces, special modifiers are necessary: horizontal stroke “-”, vertical stroke “|”, diagonal strokes “ˎ” and “ˏ”. The diagonal strokes are used only for letters Ç/ç (Enetz) and Ҏ/ҏ (Saam, or Lappish). Vertical strokes are used only for letters Ҝ/ҝ and Ҹ/ҹ which have become obsolete since modern Azerbaijan writing is based on the Latin alphabet. Horizontal strokes are used in several Cyrillic letters (Ғ/ғ, Ҟ/ҟ, Ұ/ұ, etc.). There are serious reasons for keeping these modifiers in TS2: there are still minor languages for which alphabets based on Cyrillic could be proposed. The availability of these modifiers in TS2 would support such developments without the necessity to include more glyphs in the X2 and *T2∗* encodings.

It is still a question how, and whether, the TS2 encoding should be realized. Taking into account that there are only a few glyphs really necessary for it and that there are several positions in TS1 reserved for future extension of this encoding, it may be a good decision just to combine these two encodings.

9 Acknowledgments

There are many people who have contributed to this project, and it is difficult to list all of them in this section. Among the people who contributed the essential components are Mikhail Grinchuk, Vladimir Volovich, Werner Lemberg, Frank Mittelbach, Jörg Knappen, Michel Goossens, Andrew Slepuhin, but the list is not restricted to these names only.

We are especially thankful to Vladimir Volovich and Werner Lemberg for their work on macro support for the X2 and T2∗ encodings and to the members of the mailing list *CyrTEX-T2* who discussed enthusiastically the X2 and *T2∗* problems. (To subscribe to this list send email to `Majordomo@vvv.vsu.ru` with the command: `subscribe cyrtex-t2` *your-e-mail-address*.)

We are grateful to Robin Fairbairns for his time spent polishing the text of our papers submitted to the *EuroTEX-98* conference where the preliminary results of this project (namely, the X2 encoding) was presented for the first time, and to Michel Goossens for his efforts to organize the support for Russian participants at *EuroTEX-98*.

Finally, although most work on this project was done on a voluntary basis (as it is traditional for TEX community), it is worth mentioning that part of the research was supported by a grant from the Dutch Organization for Scientific Research (NWO grant No 07 – 30 – 007).

References

[1] A. Berdnikov, O. Lapko, M. Kolodin, A. Janishevsky, A. Burykin. The encoding paradigm in LATEX 2ε and the projected X2 encoding for Cyrillic texts. Proceedings of *EuroTEX-98*, Saint-Malo, 1998.

[2] A. Berdnikov, O. Lapko, M. Kolodin, A. Janishevsky, A. Burykin. Alphabets necessary for various Cyrillic writing systems (towards X2 and T2 encodings). Proceedings of *EuroTEX-98*, Saint-Malo, 1998.

[3] A. Berdnikov, O. Grineva. Some problems with accents in TEX: letters with multiple accents and accents varying for uppercase/lowercase letters. Proceedings of *EuroTEX-98*, Saint-Malo, 1998.

[4] K. Píška, *Cyrillic Alphabets*, in: Proceedings of TUG'96, eds. M. Burbank and C. Thiele, pp. 1 – 7, JINR, Dubna; *TUGboat* **17**(2), pp. 92 – 98.

[5] O. Lapko, *Full Cyrillic: How many languages*, in: Proceedings of TUG '96, eds. M. Burbank and C. Thiele, pp. 164 – 170, JINR, Dubna; *TUGboat* **17**(2), pp. 174 – 180.

[6] K. Kenneth. The languages of the world. London, Henley.

[7] M. Ruhlen. A Guide to the languages of the world. Stanford University, 1975.

[8] C. F. Voegelin, F. M. Voegelin. Classification and Index of the World Languages. Academic Press, 1977.

[9] WWW page by Karel Píška: `http://www-hep.fzu.cz/~piska/`

[10] Ethnologue Database: `ftp://ftp.std.com/obi/Ethnologue/eth.Z`

◇ A. Berdnikov, O. Lapko,
M. Kolodin, A. Janishevsky,
A. Burykin
Institute of Analytical
Instrumentation
Rizskii pr. 26, 198103
St. Petersburg, Russia
`berd@ianin.spb.su`