УДК 004.35

Р.Х.Хакимов
Нижневартовск, Россия

СТАНДАРТИЗАЦИЯ ГРАФИЧЕСКИХ ПОДСИСТЕМ ЯЗЫКОВ — ВЫГОДЫ И ПОТЕРИ

Аннотация. Рассмотрены вопросы оптимизации ввода информации в компьютер с клавиатуры за счет стандартизации графических подсистем языков. Путем расчетов на основе статистических данных о частоте встречаемости русских букв определяются величины удлинения текстов на русском языке при переходе с традиционной кириллицы на Inalif и транслит. На основе анализа и сравнений делаются выводы о выгодах и потерях от такого перехода и стандартизации графических подсистем языков народов Российской Федерации.

Ключевые слова: графическая подсистема языка; раскладка клавиатуры; транслитерация; Inalif; поддержка татарской раскладки клавиатуры; кодирование символов; алфавит языка; частота встречаемости символов; время реакции выбора.

Сведения об авторе: Хакимов Роберт Хаккиевич, кандидат технических наук, доцент кафедры информатики и методики преподавания информатики.
Место работы: Нижневартовский государственный гуманитарный университет.

R.K.Khakimov
Nizhnevartovsk, Russia

STANDARDIZATION OF GRAPHICS SUBSYSTEMS OF LANGUAGES — GAINS AND LOSSES

Abstract. The article is concerned with optimization of keyboard entry due to the standardization of graphics subsystems of languages. By means of calculations based on the statistical data on frequencies of Russian letters, the author determines the elongations of Russian texts during the transition from the traditional Cyrillic alphabet to the Inalif and translit. On the basis of the analysis and comparison, the conclusions have been derived with regard to gains and losses of such transition and standardization of graphical subsystems of languages spoken in the Russian Federation.

Key words: graphics subsystem of a language; keyboard layout; transliteration; Inalif; Tatar keyboard layout support; character coding; alphabet of a language; frequency of letters in text; choice reaction time.

About the author: Robert Khakkievich Khakimov, Candidate of Engineering, Associate Professor of the Department of Informatics and its Teaching Methodology.
Place of employment: Nizhnevartovsk State University of Humanities.

Контактная информация: 628611, г. Нижневартовск, ул. Дзержинского, д. 11; тел. (3466) 454403.
E-mail: 02_10@yandex.ru

В настоящее время в мире в разных языках используется множество графических систем для письма (для краткости будем их называть алфавитами). Вот наиболее распространенные из них: латиница, кириллица, иероглифическое письмо, арабская письменность, санскрит.

Кроме этого в алфавитах народов, перешедших на латиницу, для лучшей передачи специфических звуков этих языков для модификации латинских букв используют различные диакритические знаки.

Примеры этих знаков: ˆ, ˇ, ˚, ˜, ˙.

Примеры букв с диакритическими знаками: ï, À, Á, Â, Ã, Ä, Å.

Нестандартизованность графических подсистем языков имеет два неудобства:

1. На клавиши устройств ввода в компьютер приходится кроме символов латиницы наносить символы национальных алфавитов. В России для некоторых применений на клавиши приходится наносить даже символы двух национальных алфавитов: русского и, например, татарского (см. рис. 1).
2. Для кодирования букв национальных алфавитов требуются кодовые комбинации в кодовой таблице.

Рис. 1. Клавиатура с совмещенной латинско-русско-татарской раскладкой

Как видно из рисунка, на клавиши с «редко используемыми» в татарском языке русскими буквами Ё, Ц, Щ, Ъ, Ж, Ь нанесены специфические буквы татарского алфавита Һ, Ө, Ә, Ү, Ң, Ж. Так что для работы должны быть активизированы три раскладки клавиатуры. Естественно, переключаться между тремя раскладками сложнее, чем между двумя. Более того, даже при вводе только татарских текстов приходится пользоваться двумя раскладками, потому что в татарском алфавите есть и буквы Ё, Ц, Щ, Ъ, Ж, Ь, которые на татарской раскладке отсутствуют. Так что хотя операционная система Windows, начиная с версии XP, и поддерживает татарскую раскладку клавиатуры, проблема решена недостаточно хорошо.

Из-за первого неудобства (удвоения или утроения количества символов на клавишах), согласно закону Хика—Хаймана о времени реакции выбора, увеличивается среднее время ввода символов [3].

Из-за второго неудобства (для букв национальных алфавитов требуются кодовые комбинации в кодовой таблице) при восьмиразрядной кодировке (количество возможных кодовых комбинаций — 256) для каждого языка приходится использовать собственную страницу в кодовой таблице (для русского языка кодовая страницаWindows-1251). В свою очередь использование отдельных кодовых страниц для разных языков усложняет межнациональный обмен программными продуктами.

Для устранения второго неудобства используются два пути.

Первый — внедряется двухбайтовый код Unicode. В нем можно закодировать 65 576 различных символов. Этого достаточно для кодирования символов всех языков мира, в том числе и китайских иероглифов. Однако при переходе с восьмиразрядной кодировки на Unicode длина любой информации, которая хранится в памяти ЭВМ или передается по сети, увеличивается в два раза. Это влечет существенные экономические потери.

Второй путь — переход от национальных алфавитов на стандартную латиницу. Стран, которые пошли по этому пути, уже много: Вьетнам, Турция, Узбекистан, Южная Корея и т.д. Однако во всех этих языках для лучшей передачи звуков в алфавите оставили буквы с диакритическими знаками. То есть проблема решена не полностью.

Принципиально новый путь был предложен Координационным советом по созданию татарского алфавита для применения в Интернете [5] — использовать в алфавите Inalif (In — Internet, Alif — Alifba = алфавит) только символы, отображенные на стандартной латинской клавиатуре ЭВМ. Это решает обе названные выше проблемы — с клавиатурой и кодированием. Проблема решается за счет использования в качестве оператора смягчения символа апостроф «’» и использования для обозначения некоторых звуков буквосочетаний, принятых в мировой практике (в транслите, в частности). Например: Ч-CH, Ш-SH, Х-KH и т.д.

Однако использование вместо одной буквы буквосочетания и частое применение апострофа приводит к удлинению текстов.

Определение величины этого удлинения для русского языка составляет цель данного исследования.

По аналогии с татарским языком Inalif можно использовать для более точной передачи русских звуков при использовании латиницы (см. табл. 1).

Таблица 1

Частоты появления букв в русских текстах

№	Буква кириллицы	Обозначение буквы в Inalif	Вероятность	Удлинение, доля	Удлинение, процент
	пробел		0,175		
1	О	O	0,090		
2	Е	E	0,072		
3	А	A	0,062		
4	И	I	0,062		
5	Т	T, T’	0,053	0,053/2=0,0265	2,65
6	Н	N, N’	0,053	0,053/2=0,0265	2,65
7	С	S	0,045		
8	Р	R	0,040		
9	В	V	0,038		
10	Л	L, L’	0,035	0,035/2=0,0175	1,75
11	К	K	0,028		
12	М	M	0,026		
13	Д	D, D’	0,025	0,025/2=0,0125	1,25
14	П	P	0,023		
15	У	U	0,021		
16	Я	JA	0,018	0,018	1,8
17	Ы	Y	0,016		
18	З	Z	0,016		
19, 20	Ь, Ъ	‘	0,014		
21	Б	B	0,014		
22	Г	G	0,013		
23	Ч	CH	0,012	0,012	1,2
24	Й	J	0,010		
25	Х	KH	0,009	0,009	0,9
26	Ж	ZH	0,007	0,007	0,7
27	Ю	JU	0,006	0,006	0,6
28	Ш	SH	0,006	0,006	0,6
29	Ц	TS	0,004	0,004	0,4
30	Щ	SCH	0,003	0,003*2=0,006	0,6
31	Э	EH	0,002	0,002	0,2
32	Ф	F	0,002		
Итого					15,3%

Для смягчения согласных в качестве оператора можно использовать тот же апостроф. В кириллице твердый и мягкий варианты произношения согласных звуков (д, л, н, т) в письме не различаются. Хотя для иностранцев такое различение твердого и мягкого произношения было бы полезно. Например, в слове «Николай» звук «н» можно произносить

не только мягко, но и твердо, как часто делают иностранцы. Уточненное написание слова «Николай» — Н’иколай. В слове «критический» звук «т» также можно произносить и мягко (как положено), и твердо (как часто делают иностранцы). Уточненное написание слова «критический» — крит’ический.

В других случаях звуки «н» и «т» можно произносить только твердо, например, «Наина», «такси».

Тот же апостроф мог бы заменить «ь» и «ъ», например, al’bom, khod’ba, pod’’ezd.

Для передачи букв «я», «ю», «ё» также можно было бы использовать апостроф, который будет смягчать предыдущие согласные, например, t’anet, t’uk.

Но в транслите [1] буквы «я», «ю», «ё», «ж», «ц», «щ» передаются диграфами: Я — JA, Ю — JU, Ё — JO, ZH, TS, SCH и менять это правило не имеет смысла.

Итак, приступаем к исследованию. Используя частотный словарь русских букв [4. С. 238], определим, на сколько процентов удлинятся русские тексты при записи их на Inalif. Напомним еще раз, удлинение происходит по двум причинам:

- некоторые буквы заменяются двух-, трехбуквенными сочетаниями (как в транслите);
- некоторые буквы используются для обозначения двух вариантов произношения (твердый и мягкий).

Из-за отсутствия экспериментальных данных примем, что частоты мягких и твердых звуков в русском языке одинаковы.

Получается, что при использовании принципов Inalif русские тексты удлиняются на 15%. Это плата за перечисленные удобства и большую точность передачи звуков в письме.

Аналогично можно подсчитать процент удлинения русских текстов при использовании транслита (там твердое и мягкое произношение согласных не различается). Получаем удлинение на 8,4%. Это плата за использование вместо некоторых букв кириллицы двух-, трехбуквенных сочетаний букв латинского алфавита.

А удобства, напомним еще раз, такие:

- На клавишах клавиатуры компьютера было бы по одному символу (только латинские). По закону Хика—Хаймана это повысило бы скорость ввода примерно на 7%. Данный вывод подтверждается и предварительными экспериментальными исследованиями [2. С. 8].
- Не приходилось бы переключать раскладку клавиатуры при вводе смешанных (англо/национальных) текстов.
- Для кодирования символов достаточно было бы восьмиразрядной кодировки в соответствии с первой страницей кода ASCII.

В случае признания показанных выгод бóльшими, чем недостатки, было бы возможно пользоваться стандартной латинской клавиатурой (рис. 2).

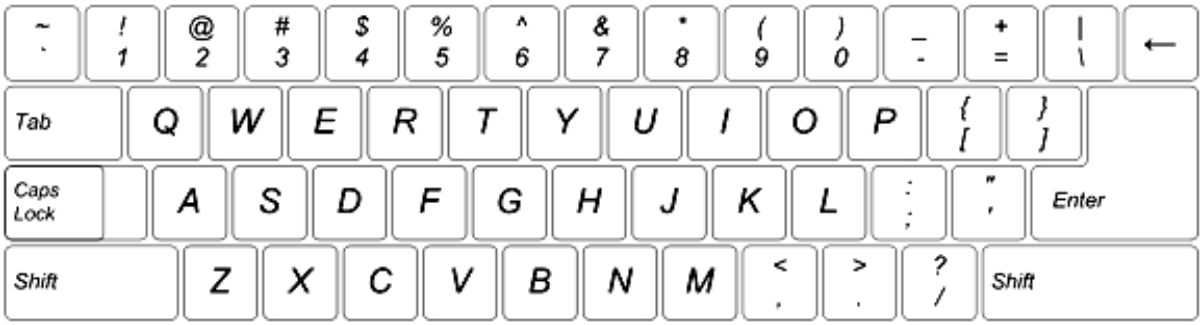

Рис. 2. Клавиатура с латинской раскладкой

Конечно, сразу же возникнут вполне обоснованные возражения против стандартизации раскладки клавиатуры за счет перехода от кириллицы к латинице.

Действительно, такой переход не может быть применен для поэтических, художественных, историографических текстов.

Но для деловых и технических текстов решение выглядит приемлемым. Об этом говорит уже то, что транслит, то есть написание русских текстов латинскими буквами (в частности), находит широкое применение, пишутся программы для транслитерации. Но в этой области пока нет согласованности — ведомственной, международной. Подтверждением является наличие многих транслитерационных алфавитов для русского языка (ГОСТ РФ, МВД РФ, ОВИР, ГИБДД, Библиотека Конгресса США и т.д.).

Отсутствие единообразия порождает потребность в стандартизации графических подсистем языков и повод для дискуссий и исследований. Обсуждаемая проблема еще более, чем для русского языка, актуальна для народов Российской Федерации, для языков которых на клавишах начертано по три буквы.

ЛИТЕРАТУРА

1. Транслитерация ввода с клавиатуры. URL: http://transliteration.ru/hand_transliteration
2. Хакимов Р.Х. Информационные технологии в автоматизации научных исследований. Учебно-методическое пособие. Нижневартовск, 2007.
3. Хика—Хаймана, закон.URL: http://vslovare.ru
4. Яглом А.М., Яглом И.М.. Вероятность и информация. М., 1973.
5. «Inalif» Алфавит татарского языка для использования в сети Интернет. URL: www.tatarlar.ru