

Восприятие и порождение текстов электронной почты на русском языке в различных вариантах латинской кодировки

Архипов Александр

МГУ, ОТиПЛ

Введение

Электронная почта — быстро развивающееся средство коммуникации. По некоторым оценкам, в январе 1998 года в мире насчитывалось около 120 миллионов пользователей электронной почты в разных ее вариантах. Как у всякой сложной системы, у электронной почты есть свои слабые места, которые в большей или в меньшей степени снижают ее эффективность.

Одна из проблем, возникающих при ее использовании у носителей русского языка, заключается в том, что сообщения, написанные кириллицей, не всегда приходят к получателю в том виде, в котором были отправлены. Это происходит из-за того, что, в отличие от букв латинского алфавита, символы кириллицы по-разному кодируются разными программами. Вот и получается, что слово «Привет», записанное в отечественной системе КОИ-8, на другом компьютере может прочитаться как «đŎĖ×ĀŎ», «pTЙЧЕФ» или «№vЩзХд». Автору не однажды приходилось получать в таком виде весьма развернутые письма, чтение которых происходило с не меньшим азартом, чем решение олимпиадных лингвистических задач (см., например, ЛЗ, с. 9, задача №1; с. 62, задача №136).

По причинам, связанным с технической организацией сетей электронной почты, а также с различием в предпочтениях, отдаваемых в разных странах тому или иному аппаратному и программному обеспечению, вероятность возникновения подобного рода проблемы во много раз возрастает при электронной переписке с адресатом, находящимся за пределами России.

Из этой непростой ситуации существует несколько выходов. Первый и, очевидно, самый эффективный, — технический, состоит в полной русификации компьютера-получателя. Но, во-первых, при этом от получателя требуются определенные познания в области информатики, во-вторых, необходимо наличие под рукой необходимых программ. В третьих, трудность выполнения указанных выше требований значительно возрастает в случае, когда адресат читает и пишет свою почту, как часто бывает, на компьютере, открытом для публичного доступа — например, в библиотеке, интернет-кафе и т.п. В-четвертых, обнаруживается, что иногда, в связи со спецификой механизмов электронной почты, текст

письма претерпевает необратимые изменения уже на этапе пересылки, и прочитать его уже невозможно, независимо от оснащённости компьютера-получателя. (Существуют, однако, специальные программы, призванные «обратить» эти изменения и расшифровать исходный текст). Все перечисленные сложности ведут к сокращению числа тех, кто может воспользоваться этим способом решения проблемы.

Второй возможный выход — вести переписку на языке, использующем латинский алфавит, например на английском. Английский язык замечателен еще и тем, что практически не использует, за исключением некоторых заимствований, никаких букв, кроме 26 букв стандартного латинского алфавита (в отличие, скажем, от немецкого, французского, польского и чешского). А при переписке на французском возможно, например, пожертвовать надстрочными знаками без большого ущерба для смысла. Но, во-первых, не все люди, имеющие постоянных или временных корреспондентов за границей, владеют английским (или другим подходящим) языком, а во-вторых даже среди лиц, таким языком владеющих, далеко не всем удобно вести на нем более или менее регулярную переписку: процесс написания письма на иностранном языке почти всегда требует определенных творческих усилий, повышенного контроля за производимым текстом, что может также усугубляться (как, впрочем, и во всех остальных случаях электронной переписки) трудностями при общении с компьютером.

Наконец, третий выход, и он реализуется очень часто, состоит в том, чтобы записывать текст НА РУССКОМ ЯЗЫКЕ, но ЛАТИНСКИМИ БУКВАМИ. Гибрид, рожденный таким способом, получил название «рунглиш» (Runglish < Russian + English). Данный доклад посвящен проблеме орфографии Runglish, т.е. возможным правилам передачи русского текста латинскими буквами. Оказывается, вариантов подобных правил может быть очень много, и у каждого пользователя имеется свой персональный вариант. Более того, он может меняться с течением времени, и часто на протяжении одного и того же письма сменяют друг друга несколько вариантов орфографии.

Будем называть Кодировкой систему правил переписывания, позволяющую текст на русском языке записать латинскими буквами. Каждое правило имеет вид $u\xi v \rightarrow uuv$ (знак “ $\rightarrow$ ” читается «передается через»), где ξ — буква русского алфавита, u — последовательность букв латинского алфавита, u и v — контекст (правила могут быть контекстно-зависимыми). Последовательность латинских букв в большинстве случаев состоит из одной буквы; диграфы используются в большинстве Кодировок в основном для обозначения шипящих (ч, ш, щ, ж), а также некоторых других букв. (На практике самое длинное обозначение — из пяти латинских букв — зафиксировано для буквы щ: *chtch*).

Возможно также определить Кодировку как определенную систему транслитерации русского текста латинскими буквами. Однако здесь необходимо сделать некоторые оговорки.

О терминах «транскрипция» и «транслитерация»

В связи с нашей темой уместно вспомнить два термина — «транскрипция» и «транслитерация», а также тесно связанный с ними термин «практическая транскрипция».

Транскрипция — «способ однозначной фиксации на письме звуковых характеристик отрезков речи» [ЛЭС]. Основное требование к транскрипции — однозначное соответствие «знак — звук». Примером является система Международной фонетической транскрипции

(МФА), включающая более 100 основных символов и более 30 модификаторов, призванных дать наиболее детальную характеристику каждому звуку каждого языка мира.

Транслитерация — «побуквенная передача текстов и отдельных слов, записанных с помощью одной графической системы, средствами другой графической системы» [ЛЭС]. Основное требование к транслитерации — однозначное соответствие «знак ó знак». «Базируясь на каком-нибудь алфавите, транслитерация допускает условное употребление букв, введение дополнительных знаков и диакритических знаков» [там же]. Транслитерация призвана в первую очередь позволить восстановление исходного написания слова (текста), поэтому в общем случае транслитерированный текст не должен *произноситься*, что и позволяет использование дополнительных знаков, для которых в языке вообще нет правил чтения.

Практическая транскрипция (термин А.М. Сухотина (1935); введен в широкое употребление А.А. Реформатским (1947)) — «запись иноязычных слов средствами национального алфавита с учетом их произношения» [ЛЭС]. Основное требование к практической транскрипции — сохранение, по возможности, звукового облика исходного слова (текста), и, во вторую очередь, его написания — например, наличия двойных согласных, и т.д. В частности, в практической транскрипции одна и та же буква языка-источника в конечном тексте может передаваться разными буквами, в зависимости от фонетической позиции. Практическая транскрипция предполагает чтение по правилам того языка, в графике которого ведется запись. Отсюда вынужденный компромисс между адекватной передачей написания и адекватной передачей звучания, зависящий от соответствия друг другу фонетических систем двух языков. Так, во французский язык были заимствованы русские слова *пирожки* и *блины*, что дало фр. слова *pirojki* и *blini*; эти слова по правилам французского языка произносятся [piroʒki] и [blini]. Как видно, при заимствовании утрачено различие графем (и фонем) *и* и *ы*, а также оглушение *ж* перед глухим *к*, т.е. и звуковая, и графическая сторона слов передаются приблизительно.

В русской традиции практическую транскрипцию иноязычных слов средствами русской графики иногда называют транслитерацией. С другой стороны, «широкое» употребление термина «[практическая] транскрипция» может включать в себя как транскрипцию, так и транслитерацию, что создает определенную путаницу в терминах. Не желая усиливать эту путаницу, примем все же, вслед за *ГуС* слово «транскрипция» как сокращение для сочетания «практическая транскрипция», а транслитерацию будем именовать, как и прежде, транслитерацией.

О некоторых особенностях текстов электронной почты на Runglish

По словам *ГуС* практическая транскрипция, транслитерация и перевод «служат средствами передачи слова из какого-либо языка в заимствующий язык с использованием графики последнего. Различие заключается в средствах, используемых для этой передачи».

Первая особенность состоит в том, что, в большинстве случаев, на Runglish пишется весь текст целиком, а не отдельные слова, вкрапленные в русский текст.

Во-вторых, именно язык-источник (т.е. русский) является обычно родным языком как для отправителя, так и для получателя сообщения. Это позволяет автору снизить контроль за орфографией в надежде, что адресат сможет восстановить исходный текст даже при наличии опечаток, непоследовательном применении Кодировки и т.д. Это же обстоятельство

позволяет прибегать к не вполне детальным Кодировкам, порождающим неоднозначность при восстановлении русского текста (например, если в по правилам Кодировки *сà s, mà t, çà ts*, то только знание языка помогает установить, что строка «*otsenka*» является записью слова «оценка», а не *«отсенка».

К тому же, коммуникативная ситуация общения на Runglish чаще всего обладает выражено сниженным статусом по сравнению с «нормальной» перепиской по-русски, а тем более с перепиской на престижном иностранном языке, что авторы иногда отмечают эксплицитно, ср.:

«Sovsem zabyla, chto tiebie nado pisat' cherti-kak».

«Ujasno pisat' takim varvarskim sposobom, no nitchego ne podelaech».

(Из личной переписки)

Общение таким «варварским способом» во-первых, непрестижно (русский язык приходится «коверкать»), во-вторых, для начинающих требует некоторой практики. Низкий престиж ситуации общения тоже является одним из поводов к снижению контроля. С другой стороны, пока способ письма еще не вполне освоен, естественно ожидать некоторое количество «системных» сбоев (хотя самоконтроль в этот период может быть как раз высоким).

Как следствие двух названных особенностей, а именно сравнительно большой длины текста и сниженного самоконтроля автора (либо сбоев в период обучения), в текстах на Runglish, в отличие от обычной транслитерации или транскрипции, у большинства людей наблюдаются колебания: одни и те же буквы, одни и те же слова могут в пределах одного текста встретиться в нескольких разных написаниях.

И еще одно следствие: поскольку родным языком является чаще всего русский, применяемая данным человеком Кодировка может находиться под влиянием разных языков, которыми он владеет или которым он обучался. Напротив, когда адресант, например, немец, его Кодировка будет, скорее всего, основываться только на правилах немецкого языка. Русский же студент (в особенности — лингвист), изучающий французский, польский и английский, с большей вероятностью будет пользоваться попеременно то польской Кодировкой, то английской, то французской (или одной смешанной).

Как мы видим, способ передачи русского текста на Runglish не совпадает полностью ни с практической транскрипцией в традиционном понимании, ни с транслитерацией, а является некоторым новым (видимо, хорошо забытым), способом.

Задачи

Какие задачи целесообразно поставить в связи с понятием Кодировки в приложении к текстам электронной почты на Runglish ?

1. а) Предложить значимые параметры для классификации возможных Кодировок. Предложить оценку степени удобства произвольной Кодировки.
2. б) Предложить Кодировку или Кодировки, оптимальную (-ые) для электронной переписки на Runglish. Эти оптимальные Кодировки могут быть использованы не только в

переписке по электронной почте, но и на сайтах Интернета, на которых содержатся тексты на Runglish.

Какие бывают Кодировки ?

Кроме перечисления конкретных правил, составляющих ту или иную Кодировку, может оказаться полезным учитывать и некоторые более общие ее признаки.

Во-первых, необходимо помнить о том, что на протяжении одного текста Кодировка может изменяться, или, другими словами, могут сменять друг друга различные Кодировки, что должно быть отражено в описании текста (Кодировки).

Во-вторых, важно учитывать то, под влиянием какого (каких) иностранного (-ых) языка (-ов) создана та или иная Кодировка. В связи с этим возможно, не повторяя для каждого случая перечисление всех правил, говорить, например, об английском или французском стандарте передачи шипящих (который будет являться подмножеством данной Кодировки как системы правил). Могут также встречаться, как и во всяком другом типе текста, вкрапления отдельных слов или более крупных фрагментов текста на иностранных языках, напр. *jeszcze* (польс. «ещё») вместо, скажем, *eshho* (в соответствии с правилами данной Кодировки) для всех вхождений слова «ещё»; только в нашем случае эти вкрапления менее заметны, не сразу бросаются в глаза. Если, как в примере со словом «ещё», такие вкрапления регулярны, они, по-видимому, должны быть включены в Кодировку как отдельные правила (одно правило для каждой лексической единицы).

В-третьих, некоторые «нерегулярности» Кодировок могут быть адекватно описаны как элементы транскрипции (т.е. отражения звуковой стороны слова) в системе, ориентированной на орфографию. Действительно, основная направленность Runglish — передача именно написания, а не произношения. Это легко видеть по тому, что такие фонетические явления, как редукция гласных, ассимиляция согласных на Runglish не передаются (пишут не *“malako”*, а *“moloko”*, не *“z drugom”*, а *“s drugom”* etc.), а произносимые сочетания согласных передаются согласно русской орфографии. И все же некоторые элементы фонетичности присутствовать могут. Так, в некоторых Кодировках *chà ch*, *shà sh*, при этом в словах «что» и «чтобы» пишется не *ch*, а *sh*; в некоторых Кодировках сочетание *жж* в словах с корнем *-езж-* (*уезжать*, *приезжать* и т.д.) передается не как *з+ж*, а как *ж+ж* (напр., *zhzh*), что тоже является отражением произношения.

Наконец, еще один немаловажный параметр, который оказывает свое влияние на удобство как пишущего, так и читающего — русских букв с помощью графически сходных с ними знаков, включая прописные и заглавные латинские буквы, а также цифры (3, 4, 6). Примером могут служить такие представления, как *xà x*, *cà c*, *zà 3*, *chà 4*, *shà w*, и другие. Назовем правила, использующие этот прием, правилами типа «*zà 3*». Замечено, что разные правила этой серии имеют разную распространенность; по-видимому, правило *xà x* встречается у гораздо большего количества авторов, чем правило *shà w*. Есть также некоторые указания на то, что в исконный «ареал распространения» данного рода приемов входят прежде всего люди, более или менее регулярно практикующие *chat* — переписку по Интернет в реальном времени. Так или иначе, применение правил типа «*zà 3*» серьезно влияет на удобство письма и на скорость прочтения сообщения. В ожидании проведения более детального исследования проблемы орфографии Runglish вообще и данного явления в частности можно сделать предварительное утверждение о том, что для людей, специально не практиковавшихся в замене знаков по «графическому» принципу, чтение текста,

закодированного с помощью правил типа «зѧ З», существенно затрудняется. С другой стороны, затруднения человека, к такой замене привыкшего, при чтении Кодировки, не использующей правил типа «зѧ З», представляются гораздо менее значимыми. Если эти утверждения найдут экспериментальное подтверждение, можно будет обоснованно рекомендовать пользователям электронной почты избегать применения в переписке правил типа «зѧ З».

Примеры существующих Кодировок: две официальные Кодировки

В настоящее время в России существует несколько Кодировок (предназначенных для транслитерации), официально принятых к употреблению в различных областях деятельности. Например, одна из них используется Министерством иностранных дел РФ при оформлении загранпаспортов, другая — ГИБДД при оформлении водительских прав, и т.д. Что кажется нам особенно важным, эти Кодировки (обе нормативные !) друг с другом не совпадают. Неудивительно, что и в электронной переписке нет единодушия: Runglish объединяет многие десятки различных Кодировок. (По данным ЛЭС, в разных странах мира для передачи заимствований из русского языка используется более 20 Кодировок, ориентированных на правила различных языков Европы и мира).

Приведем две официально принятые Кодировки (таблица 1). Первая была разработана в 1951-56 гг. АН СССР (см. столбец 3 и условия), вторая — «Правила транслитерации букв кирилловского алфавита буквами латинского алфавита» (стандарт СЭВ 1362-78), утвержденные СЭВ в 1978 году, вариант для русского языка без использования диакритических знаков (см. столбец 2). [Кодировка СЭВ цит. по ГИС, с. 277 и сл.; Кодировка АН СССР цит. по ЛЭС.]

Наиболее существенное различие состоит в том, что Кодировка АН СССР содержит фонетически обусловленные варианты представления, в частности йотированных гласных букв (я, ю, ѐ, е); пилотное исследование (см. след. раздел) показало, что более половины опрошенных на практике проводят хотя бы для одного из этих гласных различие в позициях в начале слова / после гласной / после согласной. Следовательно, отражение подобных различий в Кодировке АН СССР является, по всей видимости, ее значительным преимуществом. При этом она, как и Кодировка СЭВ, позволяет однозначно восстановить русский текст, за одним исключением: в обеих Кодировках одинаковое представление получает буква ё в начале слова или после гласной и сочетание ъо. Недостатком Кодировки АН СССР, с точки зрения электронной почты, является использование надстрочных знаков. В Кодировке СЭВ, в отличие от Кодировки АН СССР, различаются представления букв е и э. Напротив, это различие большинством пользователей Runglish игнорируется: и здесь более адекватной оказывается Кодировка АН СССР.

	1	1 (условия)	2	Рус.	1	2	Рус.	1		1 (условия)	2
Рус.											
а	а		а	к	к	к	ш	š			sh
б	b		b	л	l	l	щ	šč			shh

в	v		v	м	m	м	ъ	опускается	”
г	g		g	н	n	н	ы	y	y
д	d		d	о	o	о	ь	’	в конце; ’ перед согл.
е	е	после согл.	е	п	p	р		j	перед <i>и</i>
	je	в начале; после гл., <i>ь, ъ</i>		р	r	г		опускается	перед гл., кроме <i>и</i>
ё	о	после <i>ч, ж, ш, щ</i>	jo	с	s	с	э	e	eh
	’о	после согл., кроме <i>ч, ж, ш, щ</i>		т	t	т	ю	’u	после согл. ju
	jo	в начале; после гл., <i>ь, ъ</i>		y	u	и		ju	в начале; после гл., <i>ь, ъ</i>
ж	ž		zh	ф	f	ф	я	’a	после согл. ja
з	z		z	х	ch	kh		ja	в начале; после гл., <i>ь, ъ</i>
и	i		i	ц	c	с			
й	j		j	ч	č	ch			

Таблица 1. Две официальные Кодировки

Эксперимент

В рамках поставленных задач в декабре 1999 — январе 2000 г. было проведено небольшое (пилотное) психолингвистическое исследование. Эксперимент включал в себя две части: (1) на порождение Кодировки, (2) на восприятие Кодировки.

Порождение

Испытуемому предлагался заранее приготовленный текст на русском языке, который он должен был набрать на клавиатуре (или, в случае затрудненного доступа к компьютеру, написать от руки) латинскими буквами. Его предупреждали, что русскими буквами пользоваться нельзя, зато можно пользоваться обычными знаками препинания, включая апостроф. Текст содержал не менее одного вхождения всех букв русского алфавита и всех важных для определения Кодировки буквосочетаний (напр., «Ь» на конце слова vs. «ь» перед гласной, и т. п.).

Обработка полученных записей русского текста заключалась, в основном, в извлечении из них существенных для описания Кодировки каждого испытуемого параметров (например, передача йотированных букв (я, ю, ё, е) в позиции в начале слова / после гласной / после согласной / после твердого или мягкого знака); передача буквы щ перед гласной / не перед гласной, и т. п.). Затем параметры, релевантные для большинства испытуемых, были занесены в общую базу данных, и была проанализирована частота встречаемости разных значений этих параметров.

Восприятие

Испытуемому предлагалось прочитать (про себя) несколько текстов (наборов предложений) на русском языке, записанных латинскими буквами. Экспериментатор засекал время прочтения каждого текста. Каждый текст был написан в одной определенной Кодировке, причем все тексты в разных Кодировках. Испытуемому заранее не сообщалось, что тексты написаны в разных Кодировках.

Тексты содержали примерно одинаковое количество слов (и знаков) и были

примерно одинаковой сложности (тематика, синтаксические конструкции, редкие слова и т.п.). Для каждой Кодировки суммировалось время, затраченное всеми испытуемыми на чтение соответствующего текста. В результате была выявлена наименее «удобная» Кодировка из использованных: среднее время чтения текста, записанного с ее помощью, было в 1,35 раза больше значения того же показателя для двух Кодировок, наиболее «удобных» для восприятия.

Выводы

Среди множества возможных Кодировок не существует такой, которая одна пользовалась бы статистическим предпочтением, с одной стороны, и была бы одобрена всеми безоговорочно хотя бы для чтения. Это происходит оттого, что есть такие параметры, у которых, условно говоря, одни пользователи принимают значение *a* и не принимают значение *b*, а другие пользователи принимают значение *b* и не принимают значение *a*. Пример: некоторые пользователи одобряют (и сами применяют) запись русской *x* с помощью латинской *x*, и категорически не одобряют ее запись латинскими *kh* (и, скорее всего, среди них есть такие, для которых эта последняя запись затрудняет прочтение транслитерированных текстов). В то же время, некоторые другие пользователи одобряют и сами применяют запись в виде *kh*, но категорически не одобряют запись в виде *x* (мотивируя это тем, что латинскую букву *x* они привыкли воспринимать как аналог русского *кс*, а не иначе).

Выбор личной Кодировки — во многом вопрос привычки, а не тщательного обдумывания и расчета. Во время проведения эксперимента обнаружилось, что некоторые испытуемые говорили о явном неудобстве того или иного способа передачи букв, но сами его использовали, мотивируя это привычкой и затруднением в поиске другого способа. Хотя перед этим им были предъявлены тексты в различных Кодировках (которые содержали более одного способа передачи этих букв). Поэтому, принимая во внимание сказанное выше, можно предположить, что хотя не существует Кодировки, одинаково удобной для всех в плане порождения текста, тем не менее есть надежда, что можно построить такую Кодировку (или, например, две Кодировки), которые будут удобны большинству пользователей для чтения.

Некоторые существенные признаки такой Кодировки можно считать выявленными. Это, в частности, фонетически обусловленные варианты передачи йотированных гласных, и нежелательность использования правил типа «за 3» (основывающихся на графическом сходстве знаков). Получены также предварительные статистические данные, позволяющие выделить одно или два оптимальных представления для букв, у которых их встречается несколько в разных Кодировках. (Это в первую очередь буквы *ч, ш, щ, ж, х, ц*, а также буквы *ы* и *й*).

Остается провести подробное и тщательное исследование, с привлечением как можно более широкого круга пользователей электронной почты, с тем чтобы установить «рейтинг удобства восприятия» для каждой Кодировки, отвечающей названным условиям. Это позволит выбрать одну или несколько лучших Кодировок, которые можно будет рекомендовать к применению не только пользователям электронной почты, но и тем, кто поддерживает в Интернете сайты с текстами на Runglish (а таких существует несколько десятков).

Другое направление исследования — изучение сбоев и колебаний в применении Кодировок, которые, будучи некоторой (новой) разновидностью речевых ошибок, являются новым источником психолингвистических знаний.

Библиография

ГиС — *Гиляревский Р.С., Старостин Б.А.* Иностранные имена и названия в русском тексте. Справочник. М.: Высшая школа, 1985.

ЛЗ — Лингвистические задачи. М.: Просвещение, 1983.

ЛЭС — Лингвистический энциклопедический словарь. М., 1990.