



**Ордена Ленина
ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ
имени М.В. Келдыша
Российской академии наук**

**А.В. Бондаренко, В.А.
Галактионов, А.М. Мусатов,
С.В. Ёлкин, Э.С. Клышинский,
О.Ю. Слѣзкина**

**Метод фонетической
транскрипции с
использованием единого
промежуточного фонетического
представления**

Препринт №

Москва 2003

Аннотация

В данной работе приведены основные проблемы, возникающие при передаче фамильно-именных групп с одного языка на другой.. Дается обоснование выбора практической транскрипции в качестве метода такой передачи, вводится необходимость создания единой фонетической таблицы для многоязыковой транскрипции. В работе также приведены математические основы процесса передачи слов одного языка в другой, на базе которых может быть построена программа многоязыковой машинной транскрипции.

The paper is devoted to main problems of multilanguage name groups' translation. The paper consists substantiation of selecting of practical transcription as a method of translation and substantiation of the need of unified phonetic table creation. The mathematical bases of transcription words from one language into another are designed in the paper. These bases were used as a ground for a program of multilanguage machine transcription.

Содержание

Содержание	3
1. Введение.....	4
2. Анализ существующих систем.....	4
4. Создание единой фонетической таблицы.....	11
5. Математическое описание машинной транскрипции.....	17
6. Заключение.....	26
Список литературы.....	27

1. Введение

Необходимость передачи с одного языка на другой имен и фамилий людей, географических названий, а также названий различных организаций, пароходов и кинотеатров и мн. др. появилась уже довольно давно. И с увеличением количества информации в современном мире потребность в подобной передаче все больше увеличивается. Использование персонала, производящего транскрипцию вызывает целый ряд объективных трудностей. Причиной может являться, например, человеческий фактор, с которым связаны ошибки при написании, переписывании и других процессах передачи информации и переноса из одного документа в другой; слабое знание персоналом специфики транскрипции с конкретных языков; смешивание различных языков и диалектов. В качестве другой причины следует указать наличие различных лингвистических школ, а также традиций написания имен. Все это приводит к неоднозначности транскрипции и, как следствие, к снижению узнаваемости и идентифицируемости имени.

Все вышеизложенное делает актуальным создание машинной системы передачи имен с одного языка на другой. Однако создание качественной вычислительной процедуры немислимо без соответствующей теоретической проработки вопроса и, возможно, создания математической модели процесса. Именно этим вопросам посвящена данная работа.

2. Анализ существующих систем

Прежде чем приступить непосредственно к описанию машинной транскрипции, кратко охарактеризуем уже существующие системы, которые используются для передачи имен собственных с одного языка на другой: транслитераторы, транскрипторы и машинные переводчики, отражающие 3 основных метода перевода фамильно-именных групп:

- *Перевод*, при котором некоторому часто встречающемуся имени ставится в соответствие его эквивалент, устоявшийся в данном языке (языке перевода) в данный период времени.
- *Транскрипция* (точнее *практическая транскрипция*), когда имени собственному одного языка ставится в соответствие слово другого языка, наиболее точно отражающее его звучание в родном языке¹.
- *Транслитерация* – побуквенная передача имен собственных, записанных с помощью одной графической системы, средствами другой графической системы².

Нами было проведено тестирование различных программных средств, доступных для пользователей сети интернет или существующих на внутреннем рынке. Для тестирования были составлены выборки имен собственных, причем в их состав входили не только имена, характерные для исходного языка, но и заимствованные в него с других языков. Таким образом, моделировалась ситуация, характерная для практических задач. Результаты тестирования 10 систем записывались в виде таблицы, в которой сравнение производилось по количеству ошибок разных видов.

Среди протестированных программ было семь транслитераторов, никак не учитывающих особенности и правила чтения в том или ином языке, один машинный переводчик и два транскриптора. Для выборки в 100 имен собственных количество ошибок колебалось от 70 до 150, а мера качества перевода, зависящая от серьезности ошибки (менее серьезной ошибке соответствует меньшая мера), колебалась от 3,3 до 10,5.

Отсутствие ориентации на определенный язык при транслитерации естественно приводило к множеству ошибок. Так, например, José Cristóbal (Хосе Кристобаль) одним из транслитераторов (HS Transliter) был передан

¹ Когда же один и тот же звук можно передать различными буквами/буквосочетаниями, выбирается тот вариант, который максимально отображает *графическую* форму слова.

² Базируясь на каком-либо алфавите, транслитерация допускает условное употребление букв, введение дополнительных и диакритических знаков.

как Жосе Цристорбал, а другим (Cifrica) Джосе Сристорбал. В том и в другом случае не учитывались правила чтения испанского языка (такие как: «с» перед согласной читается как «к», «j» - как «х», а не как «ж» или «дж» и т.д.). Транслитерация не может обеспечить качественного перевода фамильно-именной группы на определенный язык, при котором носители этого языка произносили бы ее максимально близко к ее звучанию на родном языке. Таким образом, этот метод передачи имен собственных, широко используемый пользователями компьютерных сетей, оказывается непригодным для качественной передачи произвольных имен собственных.

Две протестированные программы при передаче имен использовали метод транскрипции, однако недостатков в их работе было не намного меньше. Первая из них (Hieroglyph) довольно хорошо передала всего лишь около трети предложенных ей фамильно-именных групп, остальные же оставила без транскрипции. Вторая (Translit to Cyrillic) перевела все фамилии, однако процент ошибок был очень велик.

И последняя из протестированных нами программ – машинный переводчик Промт – выдала самый лучший результат, однако в переводе, помимо некоторого количества оставленных без перевода имен и неверно переведенных букв и буквосочетаний, были зафиксированы также ошибки, когда слово именно «переводилось», то есть вместо звукового соответствия имело место смысловое. Например, Corse Matin (Корс Мартин) было переведено как Корсика Утро, что неприемлемо при работе с фамильно-именными группами.

На основе этого исследования были сделаны два вывода: во-первых, проблема адекватной машинной передачи имен собственных с одного языка на другой не имеет удовлетворительного решения; во-вторых, для решения этой проблемы необходимо использовать метод практической транскрипции. Поэтому было решено для машинной передачи имен собственных с одного языка на другой использовать именно этот метод, так

как два других имеют значительные минусы, а именно: метод перевода возможен лишь при создании бесконечной базы имен, а метод транслитерации не позволяет ориентироваться на конкретный язык перевода.

3. Основные проблемы

При создании экспериментальной системы машинной транскрипции мы столкнулись с целым рядом проблем:

1. В некоторых странах существует несколько национальных систем транскрипции и транслитерации с национального языка на латиницу, которые зачастую конкурируют. Примером таких систем могут служить пиньин и Уэйда в китайском; ромадзи, кунрэй ("официальная") и система Хёпберна в японском; ГОСТ 16876-71, ISO 9, Библиотеки конгресса Соединенных Штатов, АН СССР, Yellow pages в русском и т.д.

2. В других странах системы транскрипции на кириллицу либо еще совсем не разработаны, либо разработаны, но вызывают много вопросов (т.е. даны лишь основные соответствия, а правильная передача многих буквосочетаний остается не ясной). Примером могут служить арабский и турецкий языки.

Даже для английского языка, являющегося сейчас одним из наиболее употребительных мировых языков, правила практической транскрипции, существующие на настоящий момент, опираются на фонетическую транскрипцию. Это вызвано тем, что историческое развитие английской орфографии привело к ее значительному расхождению с произношением. Из-за этого часто оказывается невозможным определить, какой из возможных вариантов чтения слова оказывается правильным. Так, например, английские сочетания **ou**, **ow** могут соответствовать дифтонгу [ou] и тогда передаются через **ou**: *Barrow* ['bærou] → *Бэппоу*, *Boulder*

['bouldə] → *Боулдер*, но также могут выражать (обычно в ударном слоге) и дифтонг [au], в этом случае они передаются соответственно через **ау**: *Founder* → *Фаундер*. Однако нельзя с уверенностью сказать, когда они читаются так, а когда иначе. В подобных случаях при транскрипции возникают два потенциально возможных варианта имени, выбрать одно из которых не представляется возможным.

3. Помимо сложностей с определением фонетического образа слова, возникает целый ряд затруднений с обозначением звуков, отсутствующих в данном языке. Это приводит к тому, что при создании систем транскрипции приходится ставить в приблизительное соответствие звукам одного языка звуки другого. При этом, зачастую, теряется важная фонетическая информация, такая как длительность, палатализованность, высота тона и др. Разные звуки обозначаются одинаково: звуки [t], [θ] и [ð] передаются буквой «т», [u] и [w] в большинстве случаев передаются русской буквой «у» и др. Из-за этого возникают различные варианты транскрипции: до сих пор, например, не решен окончательно вопрос о том, передавать ли звук [æ]³ как **а**, **е** или **э** - у каждого из этих вариантов находятся свои плюсы и минусы.

4. Еще одна проблема возникает вследствие отсутствия взаимнооднозначного соответствия при транскрипции слова с исходного языка (L1) на язык перевода (L2) и обратно. То есть, если взять слово языка L1, протранскрибировать его в язык L2, а затем уже по правилам языка L2 протранскрибировать его обратно в L1, то полученное слово в значительном количестве случаев будет отличаться от исходного.

³ Буква **а** в закрытом ударном слоге произносится как нечто среднее между русскими буквами **а** и **э**.

В соответствии с международными требованиями машиночитаемые документы оформляются латинскими буквами, в связи с чем при транскрипции берется не само слово языка-оригинала со всеми его специфическими буквами и диакритиками, а оно же, записанное латиницей, а это уже потеря информации.

5. Другая проблема возникает при транскрипции с одного языка (например, английского) имен собственных, исконно принадлежащих другому языку. То есть, если попытаться протранскрибировать «с английского» имя и фамилию мексиканца, например, *Jose Enrique Martinez* (*Хосе Энрике Мартинес*), то по правилам английского языка получится *Джоус (или Джоуз) Энрайк Мартинез*.

6. И последней из проблем, о которых хотелось бы упомянуть в данной статье, является «борьба» между правилами транскрипции, принятыми в настоящее время, и исторической традицией при переводе иностранных имен.⁴

Приведем теперь для наглядности пример обработки одной из самых сложных английских букв W. С передачей звука [w], обозначаемой на письме при помощи английской буквы w, возникают некоторые сложности, так как соответствующая фонема отсутствует в русском языке. Здесь нельзя не учитывать сложившуюся во многих случаях традицию передачи w через русское в; сравните, например, фамилии Вашингтон, Веллингтон, Вильсон, а также образованные от них или аналогичных им фамилий географические названия. Еще в XVI—XVII веках передавали Greenwich через Гренович. Следует принять во внимание также, что передача типа Queen → Куин

⁴ В этом списке были перечислены лишь основные проблемы, встающие при осуществлении практической транскрипции при оформлении машиночитаемых документов. Такие же частные случаи как проблемы неблагозвучности и встречающийся в основном в художественной литературе перевод имен по смыслу, здесь не рассматриваются.

вводит в русскую транскрипцию лишний по сравнению с английской исходной формой слог.

С другой стороны, при передаче необходимо дифференцировать английские *v* и *w*. В последнее время наблюдается сильная тенденция передавать *w* → *y* даже вопреки традиции (*William* → Уильям, вместо Вильям; фамилия *Darwin* → Дарвин, во город *Darwin* следует передавать Даруин). «Гласная» передача *w* → *y* принята во всех советских картографических источниках. По-видимому, передача *w* → *y* является все же несколько более близкой к английскому оригиналу, чем передача *w* → *v*.

Исходя из сказанного, а также учитывая трудности произношения *y* в некоторых положениях, рекомендуем следующее правило: английское *w* → *y*. Исключение составляет *w* между гласными, не образующее дифтонга, которое передается как *v*. Также в начале или середине слова *w* → *v*, если иначе получается сочетание двух русских *y*:

Gatewood → Гейтвуд

Haywood → Хейвуд

Woolwich ['wu:lɪdʒ] → Вулидж (населенный пункт)

Attwood ['ætwu:d] → Атвуд (фамилия)

Что касается произносимого английского *w* (в частности *wɪ*, в начале слова), оно не передается:

Blisworth ['blɪzwo:ð] → Блисуэрт (населенный пункт)

Lerwick ['lerwɪk] → Леруик (населенный пункт)

Wright [raɪt] → Райт (фамилия).

Во многих случаях не удастся точно установить, является ли *w* произносимым или «немым». Тогда следует условно передавать его через русское *y*. Доводом в пользу такой передачи является и то, что в ряде английских географических названий после 30-х годов под влиянием

правописания восстановлено и узаконено в произношении ранее немое w. Примерами таких названий служат: Butterwick, Ganwick, Colwick, Fordwick, Kinwarton, Renwick, Southwark, Southwick, Tideswell, Urswick, Walwick. Здесь передача будет соответственно:

Uutterwick → Баттеруик,

Ganwick → Гануик,

Colwick → Колуик,

Fordwick → Фордуик,

Kinwarton → Кинуартон,

Renwick → Ренуик и т. д.

Впрочем, иногда в аналогичном положении можно встретить o → o, например Wordsworth ['wo:dzwoθ] → Вордсворт (поэт, конца XVIII—начала XIX века). Такая передача допустима только как традиционная; фамилия того же написания, принадлежащая другому лицу, будет транскрибироваться через Уэрдсуэрт.

Затруднения доставляет транскриптору также передача w в именах, английское происхождение и чтение которых только предполагается, но не доказано. Если имя или название имеет валлийское происхождение, w всегда передается через у. Например, валлийские Islwyn → Ислуин, Gwynn → Гуинн.

Во многих случаях возможно по различным справочным изданиям типа «Who is who» определить языковую или национальную принадлежность транскрибируемого имени и затем передавать его сообразно правилам транскрипции с соответствующего языка. В случае недоказанности английского происхождения имени w → в, как в прочих германских языках.

4. Создание единой фонетической таблицы

Традиционно практическая транскрипция осуществляется при помощи отдельного алгоритма (базы правил) транскрипции с каждого из множества языка L1 на каждый язык из множества L2. Создание подобных правил обязывает лингвиста знать оба языка (как L1, так и L2), либо же требует совместной работы двух лингвистов. Поскольку на начальном этапе внедрение этих правил в код программы требует совместной работы как лингвистов, так и программистов, это может являться затруднительным, особенно при большом количестве языков. В связи с этим было принято решение разработать экспериментальную систему машинной транскрипции, работа которой основывается на использовании единой фонетической таблицы. Процесс транскрипции в этом случае можно представить следующим образом:

Слово X исходного языка L1

↓ каждому элементу слова L1 подбирается соответствующий элемент общей фонетической таблицы

(≈звук)

Промежуточное фонетическое представление слова X

↓ каждому элементу промежуточного фонетического представления подбирается соответствующая буква или буквосочетание, выражающие этот звук в языке L2.

Слово X языка транскрипции L2

Создание единой фонетической таблицы для всех языков позволило намного сократить количество правил транскрипции, работу по их написанию, не ухудшив при этом качество транскрипции.

Если в существующих системах перевод осуществлялся напрямую с исходного языка L_1 на язык перевода L_2 (что, как отмечалось, требует написания правил транскрипции для каждой такой пары языков):

$L_1 \rightarrow L_2$	$L_2 \rightarrow L_1$...	$L_N \rightarrow L_1$	
$L_1 \rightarrow L_3$	$L_2 \rightarrow L_3$...		всего $N*(N-1)$ групп
правил				
...	...			где N – общее количество
языков				

$L_1 \rightarrow L_N$	$L_2 \rightarrow L_N$	$L_N \rightarrow L(N-1)$
-----------------------	-----------------------	--------------------------

Использование же единой фонетической таблицы позволяет писать для каждого языка лишь правила с исходного языка в некоторое фонетическое представление (ФП) и обратно. Таким образом, группы правил для всех языков выглядят следующим образом:

$L_1 \rightarrow \text{ФП}$	$\text{ФП} \rightarrow L_1$	
$L_2 \rightarrow \text{ФП}$	$\text{ФП} \rightarrow L_2$	
$L_3 \rightarrow \text{ФП}$	$\text{ФП} \rightarrow L_3$	Всего $2N$ групп правил.
...		
$L_N \rightarrow \text{ФП}$	$\text{ФП} \rightarrow L_N$	

Транскрипция при этом осуществляется в два этапа: на первом этапе фамильно-именная группа переводится с языка транскрипции в промежуточное фонетическое представление в соответствии с таблицей, о которой подробнее речь пойдет чуть ниже, а на втором этапе из ФП – в написание на нужном языке. Сама транскрипция осуществляется за счет работы программного «движка», который остается неизменным при присоединении к нему баз транскрипции различных языков. В связи с этим совместная работа программистов и лингвистов требуется лишь на начальных этапах – при создании и отладке программного «движка» транскрипции.

Важной задачей при таком подходе является само создание фонетической таблицы, то есть отбор звуков таким образом, чтобы в таблице присутствовали все звуки исследуемых языков, и в то же время ни один звук не был представлен двумя символами. Использование уже имеющихся таблиц вызывает объективные трудности. Представляется невозможным просто представить эту таблицу как пересечение множеств звуков разных языков, во-первых, из-за того, что одни и те же звуки в фонетических системах разных языков обозначаются по-разному (или различные звуки – одинаково), а во-вторых, вследствие того, что в каждом конкретном случае приходится принимать решение, должны ли два похожих, но все же различающихся звука обозначаться в ФП двумя разными символами или одним символом (возможно с разными параметрами). В качестве примеров этих трех случаев можно привести:

1. звуки [п] и [п̣], соответствующие разным символам фонетической таблицы,
2. английское «л» и немецкое «ль», обозначаемые одинаково как «l», но имеющие разные значения параметра «мягкость/твердость»;
3. французское (дорсо-увулярное) «r» японское (или русское) «р» (апико-альвиолярное), которые в ФП обозначаются одним и тем же символом r.

Прежде чем приступить к созданию таблицы нами был тщательно проанализирован материал различных языков. Ориентируясь при транскрибировании в основном на фонетическую форму слова, необходимо было одновременно учитывать и орфографический момент, с тем чтобы, не препятствуя правильному чтению, по возможности сохранить при передаче слова близость к его графической форме. Так, например, возник вопрос, следует ли английское [θ] (на письме “th”) и испанское [ç] (на письме “c”), похожее на него по звучанию, обозначать одним и тем же символом или нет. Тут вступают в противоречие принципы фонетического и графического

подобия. В данном конкретном случае вопрос был решен в пользу их различия (передачи испанского «с» в английском буквой «s») из-за того, что в американских диалектах испанского языка эта буква читается как [s], что сближает ее с графическим написанием в английском.

Помимо этого вставал также вопрос о необходимости учета традиции передачи имен или же при транскрипции стоит опираться лишь на фонетический облик слова. Многие фамилии и имена были транскрибированы достаточно давно и в отношении строго определенных людей, оставивших свой след в истории. Однако истории известны примеры, когда людей, принадлежащих к одной семье, транскрибировали в разные периоды времени различным образом. Даже транскрипция имени одного человека может сильно изменяться со временем⁵. Поэтому все говорит в пользу того, чтобы имена современных однофамильцев знаменитых исторических личностей передавать по общим правилам. То есть, *Hamlet, Prince of Denmark* — останется *Гамлетом, принцем Датским*, ибо именно в таком виде он давно уже вошел в русскую культуру и всем знаком. Но его современные тезки будут по-русски Хэмлетами, так как русское орфоэпическое «г» — звук взрывной, а не фрикативный (как английское «h»).

Ниже приведены основные элементы, при помощи которых строится промежуточное фонетическое представление (аналоги простых звуков, а также дифтонгов и аффрикат).

∫ («ш»)	а
---------	---

⁵ Эволюция написания под влиянием фонетической тенденции ясно прослеживается на передаче фамилии английского политического деятеля XVIII века R. Walpole. В энциклопедическом словаре Брокгауза и Ефрона (изд. 1891 г.) он значится как Вальполь, в 6 томе БСЭ (изд. 1951 г.) дается транскрипция Вальпол, а в 44 томе БСЭ изд. 1956 г. и позже - чисто фонетический вариант: Уолпол.

æ	г
b	г' (непроизносимое «р» из англ)
B~w (рус в)	s
b ^x («б» придыхательное)	s' (арабск. «с»)
cj ^x («ч» придыхательное)	S~ θ (рус с)
d	t
ð	T' (арабск. «т»)
ð' (арабск)	u
d' (арабск. «д»)	ü («ю» без йотации)
D~ ð (рус д)	v («в»)
dj ^x («дж» придыхательное)	w
DZ	W'
d ^x («д» придыхательное)	Ya (русск. «я»)
e	yo (русск. «ё»)
f	yu (русск. «ю»)
g	z
Ġ (турецк. «г» мягкое)	z' (арабск. «з»)
Ġ'	Z~ θ (рус с)
g ^x («б» придыхательное)	θ
h	Θ' (арабск.)
h' (арабск. «х»)	Ĝ (смычное)
h'' (арабск. «х»)	ĥ (англ. «нг»)
i	E (je) (на рус всегда e)
ı («ы»)	zh («ж»)
j	dzh («дж»)
k	З'
k' (арабск. «к»)	n' (испанск. «нь»)
ks (аффриката «кс»)	ts («ц»)
k ^x («к» придыхательное)	цз
l	cj («ч»)
m	чж
m ^x («м» придыхательное)	чз
n	э
n ^x («н» придыхательное)	э (яп на рус всегда Э)
o	э: (франц)
ö («ё» без йотации)	' (пунк)
p	< (начало слова)
p ^x («п» придыхательное)	> (конец слова)

Каждая фонема имеет при себе целый ряд параметров, таких как гласный/согласный, мягкий/твердый и др.

Эти параметры приписываются фонемам на первом этапе превращения слов языка L1 в элементы промежуточного фонетического представления, например:

L1 – испанский, L2 – русский

$l \Rightarrow l + \text{мягкость}^6$

На втором этапе преобразований для «л мягкого» находится буквосочетание, передающее этот звук в русском языке. Например:

$l + \text{мягкость, в конце слова} \Rightarrow \langle \text{ль} \rangle$, т.е. на конце слова мягкое «л» превращается в русском языке в «ль».

Еще один пример.

L1 – английский, L2 – русский

На первом этапе (английский – промежуточное ФП) суффикс *-tion* представляется следующим образом:

$t, i, o, n, \text{ в конце слова} \Rightarrow \int \text{согл.}, e \text{ гласн.}, n \text{ согл.}$

На втором этапе это сочетание будет передано на русский язык уже при помощи 3 правил:

$\int \Rightarrow \text{ш}$

$e \text{ после согласной} \Rightarrow e$

$n \Rightarrow \text{н,}$

Таким образом, аффикс *-tion* будет передан на русский при помощи буквосочетания «шен».

В данном случае использовалось правило передачи звука «ε» именно после согласного, так как в противном случае (после гласной или в начале слова) он выражался бы в русском языке при помощи буквы «е».

5. Математическое описание машинной транскрипции

⁶ Для облегчения понимания в статье вместо цифровых кодов, соответствующих элементам таблицы, используются аналогичные символы транскрипции.

Изложим проблему машинной транскрипции с использованием языка математики.

Здесь мы принимаем, что сама буква, а не только обозначаемый ею звук, обладает некоторыми параметрами (например, гласность/согласность, ряд и так далее). Это необходимо для того, чтобы выяснить, какой звук обозначает данный символ в определенном месте слова и какой набор параметров будет соответствовать данному звуку. В противном случае подобная операция представляется затруднительной или трудоемкой.

Определим параметр как пару $P = \langle N, V \rangle$, где N – имя параметра, а V – его значение. Параметр будет отображать некоторые характеристики буквы, важные для транскрипции, или позволяющие классифицировать буквы по группам. Например: <”ряд“, ”передний“>, <”тип“, ”гласная“>, <”ударение“, ”безударная“>. Два параметра равны, если совпадают их имена и значения.

Также дадим определение буквы, удобное для дальнейшего изложения. Буква состоит из графемы, однозначно идентифицирующей данную букву, и набора параметров, либо изначально присущих данной букве, либо отражающих положение буквы в слове. В связи с этим определим букву как пару $S = \langle C, \{P\} \rangle$, где C – фиксированный символ (графема), обозначающий данную букву, а P – набор ее параметров. При этом будем считать, что различные написания одной и той же буквы (например, строчное и прописное или начальное, срединное, конечное и изолированное) имеют одно и то же обозначение, однако могут обладать (в зависимости от применения) различными значениями определенных параметров. Набор параметров определяется критичностью различения таких написаний при транскрипции и особенностями языка.

Примером буквы может служить пара $\langle 'A', \{ \langle \text{“тип”}, \text{“гласн”} \rangle, \langle \text{“написание”}, \text{“прописн”} \rangle, \langle \text{“ряд”}, \text{“задний”} \rangle \} \rangle$, где $'A'$ – графема, идентифицирующая данную букву, а множество, заключенное в фигурные

скобки – множество параметров данной буквы. Здесь и в дальнейшем выделим с помощью апострофов графемы, относящиеся к символам некоторого языка. Служебные графемы, предназначенные для обеспечения процесса транскрипции, будут обозначаться несколькими символами и не будут заключаться в апострофы.

Определим следующие операторы сравнения букв.

Оператор $=$ производит сравнение как графем букв, так и их наборов параметров. Две буквы S_1 и S_2 равны в смысле оператора $=$ ($S_1=S_2$), если равны их графемы и множество параметров S_2 является подмножеством параметров S_1 .

Оператор \approx производит сравнение только наборов параметров букв. Две буквы S_1 и S_2 равны в смысле оператора \approx ($S_1\approx S_2$), если множество параметров S_2 является подмножеством параметров S_1 .

В целом транскрипция будет состоять из двух частей – перевода с языка оригинала на язык-посредник и перевода с языка-посредника на язык транскрипции. Плюсом такого подхода является сокращение количества наборов правил транскрипции в случае работы со многими языками. Как это было показано выше, при отсутствии языка-посредника приходилось бы создавать базы для транскрипции с каждого языка на все остальные, что составило бы $N_L*(N_L-1)$ баз, где N_L – количество языков, с которыми производится работа. При транскрипции через язык-посредник это количество составит лишь $2* N_L$, так как потребуются базы лишь для транскрипции на язык-посредник и с него.

Однако подобный подход налагает дополнительные требования на язык-посредник. Алфавит языка-посредника должен содержать звуки всех языков, с которых производится транскрипция. Кроме алфавита для языка-посредника должен определяться набор параметров, которыми могут обладать буквы этого языка. Для того, чтобы корректно произвести транскрипцию, правила транскрипции с языка-посредника должны

охватывать все буквы алфавита этого языка, что несколько увеличивает объем правил. Одновременно с этим за счет проведения дополнительных работ скорость транскрипции падает.

Также имеется необходимость определить алфавит каждого языка с тем, чтобы сопоставить любому символу, встречающемуся в данном языке, букву из этого алфавита (графему и набор параметров).

В целом, процесс транскрипции разобьем на пять этапов:

1. преобразование написания слова на языке оригинала во внутреннее представление;
2. выделение слогов, расстановка переносов и ударений;
3. перевод внутреннего представления слова в промежуточное фонетическое написание;
4. перевод промежуточного фонетического написания слова во внутреннее представление слова на языке транскрипции;
5. преобразование внутреннего представления слова на языке транскрипции в написание слова на языке транскрипции.

Опишем каждый из этих этапов подробнее

1. Преобразование написания слова на языке оригинала во внутреннее представление состоит в преобразовании слова языка, записанного как множество символов $W=\langle G \rangle$, во множество букв $W'=\langle S \rangle$. Здесь G – символ (знак), а в случае машинной транскрипции - информационный код знака в одной из компьютерных кодировок (ASCII, ANSI или иной другой). Для такого преобразования вводится множество правил, называемых правилами алфавита, сопоставляющих символу (информационному коду знака) G букву S . $\mathcal{R}_a=\{R_a\}$, где \mathcal{R}_a – база правил алфавита, а $R_a=\langle G,S \rangle$ – правило.

Примерами правил алфавита могут служить следующие множества.

$\langle 'A', \langle 'A' \rangle, \{ \langle \text{“тип”}, \text{“гласн”} \rangle, \langle \text{“написание”}, \text{“прописн”} \rangle, \langle \text{“ряд”}, \text{“задний”} \rangle \} \rangle$

$\langle 'a', \langle 'A', \{ \langle \textit{тип}, \textit{“гласн”}, \langle \textit{”написание”}, \textit{”строчн”}, \langle \textit{”ряд”}, \textit{”задний”} \rangle \} \rangle \rangle$

$\langle 'B', \langle 'B', \{ \langle \textit{тип}, \textit{”согласн”}, \langle \textit{”написание”}, \textit{”прописн”}, \langle \textit{”звонкость”}, \textit{”звонкая”} \rangle \} \rangle \rangle$

$\langle 'b', \langle 'B', \{ \langle \textit{тип}, \textit{”согласн”}, \langle \textit{”написание”}, \textit{”строчн”}, \langle \textit{”звонкость”}, \textit{”звонкая”} \rangle \} \rangle \rangle$

Курсивом здесь выделена часть, относящаяся к букве (S), а полужирным шрифтом – параметры буквы.

Для всех графем входного слова последовательно находятся такие правила, что графема входного слова совпадает с графемой из найденного правила. Внутреннее представление слова получается путем последовательной конкатенации букв, входящих в полученные правила. Кроме того, в начало и конец слова добавляются специальные буквы, обозначающие начало и конец слова. Все графемы, для которых не было найдено соответствия в правилах алфавита, считаются знаками препинания и передаются дальше без изменений с соответствующей пометкой. Перед началом группы знаков препинаний ставится буква конца слова, после нее – начала слова. Подобный подход позволяет вычленить не только знаки препинания, но и символы из других алфавитов, которые не должны транскрибироваться в рамках данного языка.

Таким образом $W \Rightarrow W' = \prod_{m=1}^N S_m$, причем

a) $S_1 = \langle \textit{BEG}, \{ \} \rangle$,

b) $S_N = \langle \textit{END}, \{ \} \rangle$, здесь BEG и END – графемы, обозначающие начало и конец слова,

c) $S_m = S$, если $\exists R_a = \langle G, S \rangle \in \mathfrak{R}_a : G = G_j$, здесь $j = 1..M$, где M – общее количество графем во входном слове,

d) $S_m = \langle G_j, \{ \} \rangle$, если не $\exists R_a = \langle G, S \rangle \in \mathfrak{R}_a : G = G_j$,

e) $S_m = \langle \text{BEG}, \{\} \rangle$, если S_{m-1} получено по правилу d), а S_{m+1} получено по правилу c),

f) $S_m = \langle \text{END}, \{\} \rangle$, если S_{m-1} получено по правилу c), а S_{m+1} получено по правилу d),

Здесь $m \in (1, N)$, где N – общее количество букв в выходном слове (во внутреннем формате).

2. Выделение слогов и расстановка переносов производятся для того, чтобы определить закрытые/открытые слоги и ударные/безударные буквы. Любая буква, находящаяся в конце слога, приобретает дополнительный параметр «буква в слоге» со значением «открытая». Для остальных букв значение этого параметра – «закрытая».

Выделение слогов производится по следующему алгоритму. Для алфавита каждого языка может быть задан набор слогообразующих букв. В качестве части слога, присоединяемой к слогообразующей букве, берется половина букв между двумя слогообразующими. При нечетном количестве букв, средняя передается следующему слогу. Исключение делается для приставок, суффиксов и окончаний, разделение на слоги которых фиксировано. Они присоединяются к остальной части слова как отдельный слог или несколько самостоятельно выделенных слогов. Написание и деление на слоги таких приставок, суффиксов и окончаний задается отдельной базой правил.

Расстановка ударений, как и выделение слогов, не является обязательной. Их необходимо производить для языков, в которых буквы читаются различным образом в зависимости от того, в какой позиции находится данная буква – в ударной или безударной, в конце слога или нет.

3. Задачей перевода внутреннего представления слова в промежуточное фонетическое написание является приведение слов различных языков к единой записи в рамках алфавита языка-посредника. На

вход данного этапа поступает последовательность букв языка. Выходом этапа является набор фонем, входящих в состав языка-посредника.

Под строкой (словом) здесь будем понимать упорядоченное множество букв. Подстрокой слова будет являться подмножество последовательно идущих букв данного слова. Обозначим через W_i^i подстроку слова W длиной l , начинающуюся с буквы в позиции i . В дальнейшем верхний индекс подстроки будет обозначать позицию, с которой начинается данная подстрока в слове, а нижний индекс будет обозначать длину подстроки. Символом $*$ будем обозначать произвольное значение позиции.

Под правилом перевода будем понимать пару $R_t = \langle W_{l_1}^*, \bar{W}_{l_2} \rangle$, где $W_{l_1}^*$ - строка-образец, а \bar{W}_{l_2} - строка-результат. Правило R применимо к подстроке $W_{l_1}^i$, если строка-образец сравнима с $W_{l_1}^i$. Под сравнимостью понимается нахождение равенства букв из $W_{l_1}^*$ и $W_{l_1}^i$ в одних и тех же позициях подстрок. При этом здесь две буквы S_1 и S_2 равны, если $S_1 = S_2$ или $S_1 \approx S_2$. Подробный алгоритм определения применимости правила к строке приведен ниже.

Под переводом подстроки $W_{l_1}^i$ будем понимать функцию $\bar{W}_{l_2} = F^i(W_{l_1}^i)$, такую, что $\exists R_t = \langle W_{l_1}^*, \bar{W}_{l_2} \rangle \in \mathfrak{R}_t$ применимое к $W_{l_1}^i$. Здесь $\mathfrak{R}_t = \{R_t\}$ - база правил перевода.

Задача перевода в промежуточное фонетическое написание в этом случае может быть представлена следующим образом.

Пусть имеем на входе на данный этап некоторое слово $W = \langle S_1, S_2, \dots, S_a \rangle$ и набор правил перевода \mathfrak{R}_t . Перевод внутреннего представления в промежуточное фонетическое написание в этом случае будет заключаться в нахождении и применении упорядоченного подмножества правил $\mathfrak{R} = \langle W_{l_1}^*, \bar{W}_{l_2} \rangle$, таких что:

- 1) $i = \langle i_1, i_2, \dots, i_n \rangle$, где n – число правил в подмножестве \mathfrak{R} ;
- 2) $l = \langle l_1, l_2, \dots, l_n \rangle$;

- 3) $\sum_{j=1}^n l_j = a$;
- 4) $i_l = 1$;
- 5) $i_{k+1} = i_k + l_k$ для $k < n$ и $i_n + l_n = a + 1$;
- 6) $\forall i, l \exists R_t = \langle W_l^*, \bar{W}_{l2} \rangle : \exists \bar{W}_{l2} = F^l(W_l^i)$.

Здесь множество i – это множество позиций, с которых применимы правила, а множество l – множество длин подстрок.

Результатом перевода будет являться конкатенация результатов последовательного применения правил перевода.

$$\bar{W} = \prod_{i,l} F^l(W_l^i)$$

Проверка применимости правила к строке производится следующим образом. Правила могут содержать в себе буквы со специально определенной графемой ЕМРТУ. Сравнение буквы правила и буквы строки производится при помощи оператора =, если графема буквы правила не равна ЕМРТУ, и при помощи оператора \approx в противном случае.

В начале перевода внутреннего представления слова в промежуточное фонетическое написание текущая позиция во входной строке устанавливается в 1. Далее, до тех пор, пока не будет достигнут конец слова, последовательно применяется следующий алгоритм.

Сохраняем текущую позицию. Далее пытаемся найти все правила, применимые для строки, начинающейся с текущей позиции. Если первые несколько последовательно идущих букв в правиле имеют графему, равную ЕМРТУ, то уменьшаем текущую позицию на количество таких букв. Если текущая позиция меньше 1, то считаем, что правило не применимо, восстанавливаем текущую позицию и переходим к следующему правилу.

Начиная с полученной текущей позиции последовательно сравниваем буквы строки и правила. Если хотя бы одна буква строки не равна соответствующей букве правила, то считаем, что правило не применимо,

восстанавливаем текущую позицию и переходим к следующему правилу. Если сравнение всех букв прошло успешно, то считаем, что правило применимо. В этом случае помещаем сохраненную текущую позицию в множество i . Во множество l помещаем количество букв в правиле за вычетом последовательно идущих букв в начале и в конце правила, имеющих графему равную ЕМРТУ. Далее восстанавливается сохраненная текущая позиция и алгоритм переходит к следующему правилу.

По окончании перебора всех правил текущая позиция увеличивается на величину, сохраненную в множестве l . В случае если к одной и той же позиции в слове применимо несколько правил, то для каждого правила заводятся свои множества i и l .

4. Этап перевода промежуточного фонетического написания слова во внутреннее представление слова на языке транскрипции аналогичен этапу 3, но имеет противоположные задачи. Он служит для того, чтобы сформировать последовательность букв, отражающих полученное звучание слова в языке транскрипции. Работа этапа осуществляется по тем же принципам, что и этапа 3. Здесь правила являются не столь многозначными, как на этапе 3, так как имеется возможность задать одно определенное правило для передачи данного набора звуков при наличии альтернативы.

5. Преобразование внутреннего представления слова на языке транскрипции в написание слова на языке транскрипции является обратным относительно этапа 1. Здесь могут использоваться те же самые правила, что и на этапе 1, так как в большинстве случаев должно существовать взаимнооднозначное соответствие между графемой и буквой с данным набором параметров. Буквы с графемами ВЕГ и END удаляются, знаки препинания передаются соответствующими символами.

Предложенный метод позволяет формально подойти к проблеме машинной транскрипции в многоязыковых системах. Это позволит строго сформулировать требования к языку-посреднику и языкам, участвующим в

транскрипции, исследовать их особенности и свойства. Формализация процесса транскрипции позволяет проще перейти к решению задачи машинной транскрипции.

6. Заключение

Как это было показано выше, методом, наиболее подходящим для передачи фамильно-именных групп одного языка в другой, является практическая транскрипция. Она позволяет сохранить «узнаваемость» имени одного языка в другом.

Использование людей позволяет опереться на огромный опыт специалистов, однако транскрипция, проводимая лицом слабо разбирающимся в прикладных вопросах передачи слов с различных языков, наталкивается на ряд трудностей. Это так называемые ошибки человеческого фактора. Кроме того, на практике существует целый ряд правил, противоречащих друг другу и используемых различными лингвистическими школами и традициями. Это приводит к тому, что получить однозначную передачу слова с одного языка на другой в ряде случаев не возможно. Изложенные проблемы приводят к задаче формализации процесса машинной транскрипции и набора правил, по которым она проводится.

Применение промежуточной фонетической таблицы позволяет проводить транскрипцию сразу между несколькими языками. Создание такой таблицы и ее использование приводит к уменьшению работы, которую необходимо проделать при создании правил транскрипции. В работе приводится вариант такой таблицы, разработанной в ходе исследования основных европейских и азиатских языков.

В данной работе процесс транскрипции был разбит на пять этапов, после чего на основе теории множеств была предложена математическая модель каждого из этапов. Создание такой модели позволяет перейти к практической

реализации программного комплекса, имеющего возможность производить транскрипцию слов между несколькими языками. Формализация самого процесса транскрипции позволяет перейти к созданию рекомендаций по написанию правил транскрипции.

Список литературы

1. Клышинский Э.С., Слезкина О.Ю. К проблеме математического описания многоязычной математической транскрипции // Сб. Трудов научно-практического семинара «Новые информационные технологии-6», МГИЭМ, Москва, 2003;
2. Реформатский А.А. Введение в языкознание. Гл. 3. Фонетика // М.: Аспект Пресс, 1996;
2. Трубецкой Н.С. Основы фонологии // М.: НЛ, 1960.