

УДК 414:410.51:51:155.001.57

## АЛГОРИТМ АВТОМАТИЧЕСКОГО ТРАНСКРИБИРОВАНИЯ ТЕКСТА

В.Н.Ёлкина, Л.С.Юдина

Широко проводимые в настоящее время лингвостатистические исследования требуют фронтального статистического анализа больших массивов текстов как печатных, так и в фонемной записи. Перевод больших массивов печатных текстов в фонетическую транскрипцию оставался до сих пор областью кропотливого ручного труда. Кроме того, в процессе транскрибирования вручную следует заранее предполагать наличие как случайных ошибок, так и ошибок, обусловленных отсутствием унифицированной транскрипции и, следовательно, субъективизмом лингвиста.

В результате работы предлагаемого алгоритма вводимый в ЭВМ письменный язык автоматически преобразуется в фонетическую транскрипцию.

Преимущество автоматического транскрибирования в том, что оно

- позволяет обрабатывать материал практически неограниченного объема с минимальными затратами ручного труда;
- позволяет принимать однозначные решения относительно фонем в аналогичных фонетических ситуациях.

Особое значение разработка алгоритма транскрибирования имеет для управления фонемными синтезаторами речи.

Алгоритм автоматического транскрибирования текста основан на формализации принципов, по которым строится описание русского литературного произношения: установить все возможные фонетические позиции; использовать сформулированные правила позиционных изменений фонем; провести строгое соответствие между буквами и представляющими их в соответствующих позициях звуками.

Алгоритм реализован на ЭВМ БЭСМ-6.

По сравнению с опубликованным нами ранее алгоритмом, реализованным на ЭВМ М-20 [1], данный алгоритм существенно изменен и дополнен.

Получаемая в результате работы алгоритма фонетическая транскрипция является общепризнанной и наиболее распространенной. Транскрипция рассчитана на полный стиль произношения в соответствии с основными орфоэпическими правилами современного русского литературного языка [2,3,4].

**Методика работы.** При определении состава фонем мы исходили из того, что в современном русском языке 42 фонемы — 6 гласных (а, о, у, э, и, ы) и 36 согласных (бб', вв', гг', дд', ж, зз', кк', лл', мм', нн', пп', рр', сс', тт', фф', хх', ц, ч, ш, ж, ь, ъ). Кроме того, ввиду важности ударения как одного из основных параметров при автоматическом распознавании речи особое внимание обращено на различение ударных и безударных гласных, поэтому в наш инвентарь фонем, помимо основных 42 фонем, включены безударные (А, ь, и<sup>э</sup>, ь, и, н, у).

Рамкой анализа служило фонетическое слово, то есть слово с проклитиками и энклитиками. Таким образом, в данной работе принято, что все взаимовлияния звуков заканчиваются внутри слова. Исключение составляет переход [н] в [н] в начале слова после твердого согласного предыдущего знаменательного слова (например, "конец [н]гры") и после предлогов-проклитиков (к Ире — к [н]ре). Озвончения через пробел не производилось.

Специфика слов-проклитиков, которые то имеют самостоятельное ударение, то не имеют его, учитывалась на этапе кодирования исходного текста в соответствии с мелодикой фразы. Особая проверка производится в алгоритме на односложные союзы с безударными [о], [а] : (то, что, сквозь, но). Ввиду способности гласной в них сильно редуцироваться в безударной позиции, они при транскрибировании получают самостоятельное ударение.

Кроме перечисленных выше, в алгоритме предусмотрены следующие фонетические изменения.

**В области гласных.** Изменения гласных в зависимости от положения по отношению к ударению:

— **иканье** — произношение безударных ([о, а] в I предудар-

ной позиции как [А] (тр[А]ва), а во II предударной и в заударных — как редуцированного [э] (н[э]рубил);

— **иканье** — произношение гласных [е], [я] в I предударной позиции как звука [и<sup>э</sup>] (в[и<sup>э</sup>]сна).

Безударные а ≠ и ≠ я ≠ в качестве отдельных слов записываются как [А] и [jА] соответственно, безударное [а] в слове ≠ как ≠ транскрибируется как [А].

Явления начала слова: гласные [о] и [а] в абсолютном начале слова транскрибируются как [А]; начальный безударный гласный [э] в иностранных словах транскрибируется как [и<sup>э</sup>]: ([и<sup>э</sup>]таж, [и<sup>э</sup>]кзамен); начальные безударные [е], [я] транскрибируются как [jи<sup>э</sup>]: ([jи<sup>э</sup>]пония).

После [к, п, ц] в I предударной позиции [е] изменяются в [ы]: ([к]на, [п]лка), в остальных позициях в [э]: ([ш]лама).

Безударные сочетания гласных, кроме сочетаний [-уэ], [-юэ], транскрибируются без [j]: (син[иэ]).

Алгоритм составлен таким образом, что слова, не имеющие самостоятельного ударения, примыкают, как правило, к последующему слову. В целях получения правильной транскрипции слова с переносом ударения со знаменательного слова на предлог (например, на воду, не был и пр.), а также слова, которые пишутся через черточку (все-таки, где-то), в исходном тексте должны кодироваться как одно слово.

**В области согласных.**

Оглушение звонких согласных на конце слова.

Ассимиляция по звонкости-глухости.

Взрывной [г] в положении перед глухими и на конце слова изменяется в [к], а в сочетании [гк] [г — х] : (смя[к]чение); (ле[х]ко).

Ассимиляция зубных [з], [с] перед [к, ш, ч, щ].

Полное уподобление сочетаний: [тч — ч', жч — ж', тс — ц].

Упрощение сочетаний: [стн — сн, здн — зн, лнц — нц, рдц — рц]

Смягчение согласных перед переднерядными гласными.

Как известно, в ассимилятивном (позиционном) смягчении согласных в настоящее время допускается большая вариативность. Однако отмечено, что в современном русском языке наметилась четкая тенденция к утрате ассимилятивного смягчения согласных [5].

Если полвека назад консонантизм русского языка характеризовался последовательным ассимилятивным смягчением, то в настоящее время часть бывших позиционно мягких согласных или отвердела, или находится на пути к отвердению.

В нашем алгоритме предусмотрено смягчение:

- [с, з] перед [т', д', л', н'] ;
- [н] перед [т', д', ч', ш'] ;
- двойных согласных, если вторая из них мягкая.

Особая проверка в алгоритме производится на звонкий целевой [в], занимающий в русской фонетической системе особое положение, - перед ним не происходит чередования согласных по глухости-звонкости, как перед всеми другими звонкими шумными согласными.

Подробно фонетические изменения звуков представлены на блок-схеме алгоритма.

Как известно, построение текста определяется сложным переплетением качественно различных законов. Различные языковые явления не одинаково просто поддаются формализации, то есть полному и в то же время краткому описанию. Функционирование языковых единиц зависит от столь большого числа факторов, что учесть их все практически очень трудно, и поэтому в отношении языковых элементов не всегда можно сформулировать строгие, однозначные правила.

Это положение действительно не только на уровне значащих единиц языка. Более того, на фонетическом уровне отражаются ограничения, действующие на других уровнях.

Выражается это прежде всего в том, что в некоторых группах слов варьирование одних и тех же фонем по позициям неодинаково.

Так, из фонетической системы русского литературного языка исключаются произносительные особенности терминов и заимствованных слов, образующих особую фонетическую подсистему. Их произношение требует для своего описания принципиально иных правил [6]. Как правило, фонема [о] в I предударной позиции представлена как [А]. Однако в некоторых словах она реализуется как [о] : (б[о]а, рож[о]ко).

В позиции перед переднерядным [е] в некоторых случаях встречается твердый согласный: (темп, антенна).

Неоднозначно реализуется также [г] в [в] в сочетаниях [-ого], [-его] : (мно[бгъ], но люб[бвъ]).

Сочетание [чт] также по-разному передается в словах: (чтобы - [шт]обы и по [чт]а).

При транскрибировании вручную в ряде случаев оказывается необходимым учитывать семантику, морфологическую структуру слова и пр. Например, как термин слово "конечно" затранскрибируется как [клн'эчнъ] (в выражении "пространство конечно"), а как вводное слово оно должно быть затранскрибировано как [клн'эшнъ].

Сочетание [тс] реализуется неодинаково: (по [тс]удимый) - приставка + корень, но в окончаниях: (волнуе [цъ]). Это перечисление можно было бы продолжить.

В нашем алгоритме предпочтение отдавалось тому варианту транскрипции сочетания букв, при котором будет правильно за-транскрибировано большинство слов, содержащих это сочетание. Так, сочетание [чч] транскрибируется как [чн], поскольку мы получим менее грубую ошибку при неточном транскрибировании вводного слова "конечно", чем, исказив слова "точно", "съемочный", "многостаночный", термин "конечно" и др.

Для слов со специфическим произнесением, часто встречающихся в текстах, с целью их правильного транскрибирования алгоритмом предусмотрена дополнительная проверка (см. блок-схему).

Ниже приводится образец автоматически затранскрибированного текста, выданного печатающим устройством ЭВМ БЭСМ-6.

Поскольку в наборе символов печатающего устройства нет некоторых знаков, принятых в фонетической транскрипции, для их обозначения использованы имеющиеся символы.

Таким образом, при чтении машинной выдачи следует помнить, что [j] заменено символом [й]; [и<sup>э</sup>] - [и']; [ц] обозначено как [цц]; [ш'] обозначается как [щ]; знак ударения (') помещен не над ударной гласной, а справа от неё; знаком (^), помещенным также справа от согласного, обозначена мягкость этого согласного; знаками [ъ] и [ь] обозначены редуцированные гласные; [А] обозначает [а, о] в I предударной позиции и в абсолютном начале слова. Например: [АддАхн'от'ь] - отдохнете.

Если полвека назад консонантизм русского языка характеризовался последовательным ассимилятивным смягчением, то в настоящее время часть бывших позиционно мягких согласных или отвердела, или находится на пути к отвердению.

В нашем алгоритме предусмотрено смягчение:

- [с, з] перед [т', д', л', н'] ;
- [н] перед [т', д', ч', ш'] ;
- двойных согласных, если вторая из них мягкая.

Особая проверка в алгоритме производится на звонкий целевой [в], занимающий в русской фонетической системе особое положение, - перед ним не происходит чередования согласных по глухости-звонкости, как перед всеми другими звонкими шумными согласными.

Подробно фонетические изменения звуков представлены на блок-схеме алгоритма.

Как известно, построение текста определяется сложным переплетением качественно различных законов. Различные языковые явления не одинаково просто поддаются формализации, то есть полному и в то же время краткому описанию. Функционирование языковых единиц зависит от столь большого числа факторов, что учесть их все практически очень трудно, и поэтому в отношении языковых элементов не всегда можно сформулировать строгие, однозначные правила.

Это положение действительно не только на уровне значащих единиц языка. Более того, на фонетическом уровне отражаются ограничения, действующие на других уровнях.

Выражается это прежде всего в том, что в некоторых группах слов варьирование одних и тех же фонем по позициям неодинаково.

Так, из фонетической системы русского литературного языка исключаются произносительные особенности терминов и заимствованных слов, образующих особую фонетическую подсистему. Их произношение требует для своего описания принципиально иных правил [6]. Как правило, фонема [о] в I предударной позиции представлена как [А]. Однако в некоторых словах она реализуется как [о] : (б[о]а, рок[о]ко).

В позиции перед переднерядным [е] в некоторых случаях встречается твердый согласный: (темп, антенна).

Неоднозначно реализуется также [г] в [в] в сочетаниях [-ого], [-его] : (мно[бгъ], но люб[бвъ]).

Сочетание [чт] также по-разному передается в словах: (чтобы - [шт]обы и по [чт]а).

При транскрибировании вручную в ряде случаев оказывается необходимым учитывать семантику, морфологическую структуру слова и пр. Например, как термин слово "конечно" затранскрибируется как [клн'эчнъ] (в выражении "пространство конечно"), а как вводное слово оно должно быть затранскрибировано как [клн'эшнъ].

Сочетание [тс] реализуется неодинаково: (по [тс]удимый) - приставка + корень, но в окончаниях: (волнуе [цъ]). Это перечисление можно было бы продолжить.

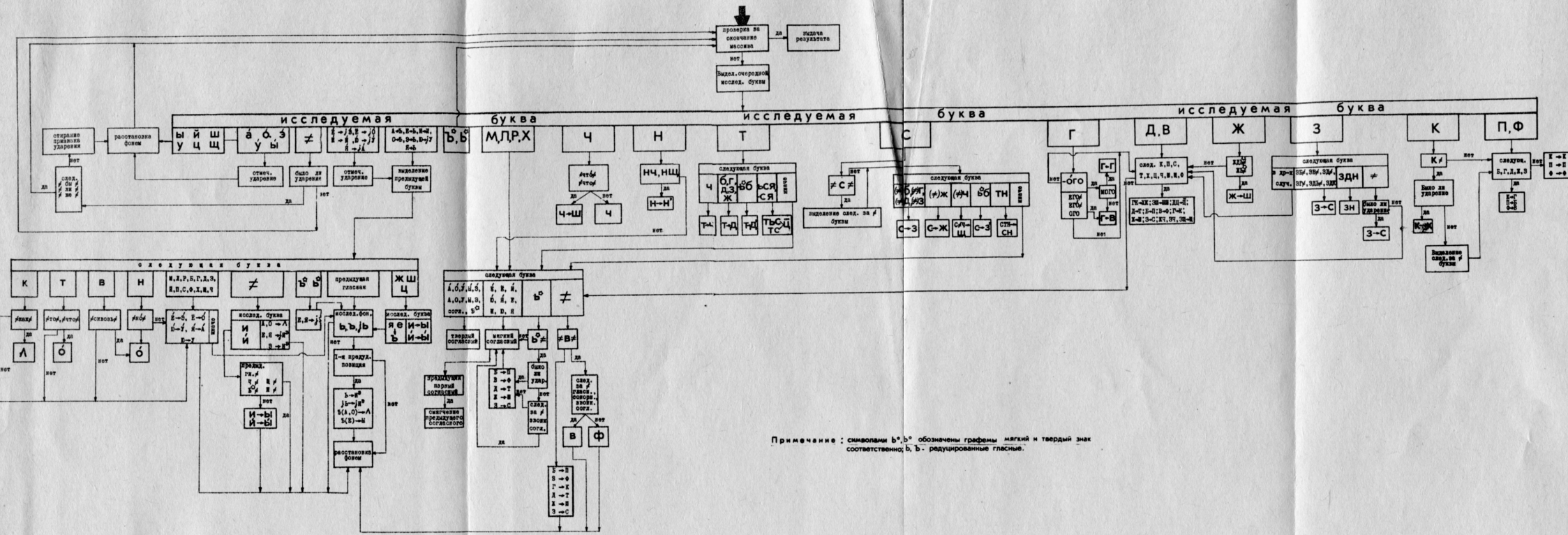
В нашем алгоритме предпочтение отдавалось тому варианту транскрипции сочетания букв, при котором будет правильно за-транскрибировано большинство слов, содержащих это сочетание. Так, сочетание [чч] транскрибируется как [чн], поскольку мы получим менее грубую ошибку при неточном транскрибировании вводного слова "конечно", чем, исказив слова "точно", "съемочный", "многостаночный", термин "конечно" и др.

Для слов со специфическим произношением, часто встречающихся в текстах, с целью их правильного транскрибирования алгоритмом предусмотрена дополнительная проверка (см. блок-схему).

Ниже приводится образец автоматически затранскрибированного текста, выданного печатающим устройством ЭВМ БЭСМ-6.

Поскольку в наборе символов печатающего устройства нет некоторых знаков, принятых в фонетической транскрипции, для их обозначения использованы имеющиеся символы.

Таким образом, при чтении машинной выдачи следует помнить, что [j] заменено символом [й]; [и<sup>э</sup>] - [и']; [ц] обозначено как [цц]; [ш'] обозначается как [щ]; знак ударения (') помещен не над ударной гласной, а справа от неё; знаком (^), помещенным также справа от согласного, обозначена мягкость этого согласного; знаками [ъ] и [ь] обозначены редуцированные гласные; [А] обозначает [а, о] в I предударной позиции и в абсолютном начале слова. Например: [АддАхн'от'ь] - отдохнете.



Примечание : символами б\*, б° обозначены графемы мягкий и твердый знак соответственно; б, в - редуцированные гласные.

Блок-схема алгоритма транскрибирования

# ТЕКСТ

— Это и есть наш корабль, — сказал Осоргин-старший. — Мы тут посоветовались и дали ему хорошее имя: "Гром и молния". Вот эта нижняя часть — "Гром", а планер — "Молния". Значит, в совокупности — "Гром и Молния". Если, конечно, вы не возражаете. Как заказчик.

Гром и молния, подумал я, гром и молния, пятнадцать человек на сундук мертвеца, а также сто тысяч чертей. Похоже, это сооружение не сдвинется с места. Корабли без двигателя. Овальная платформа, выкрашенная провозительной желтой краской. На платформе — обыкновенный планер. Малиновый планер на желтом диске. И все.

Я ответил машинально:

— Не возражаю. Отчего же мне возражать?

Горит мой эксперимент, вот о чем я думал. Горит самым натуральным образом.

— Очень удачное название, — подтвердил вежливый Каплинский. — Звучное. В таком...э...морском стиле.

Осоргин-старший одобрительно взглянул на него.

— Вы тоже со студии? — спросил он.

Я быстро ответил за Каплинского:

— Да, конечно. Михаил Семенович тоже работает для этого фильма.

Похоже, это полный крах. А ведь они внушали такое доверие, этот Осоргин-старший с его прекрасной адмирал-макаровской породой и Осоргин-младший с такими интеллектуальными манерами.

— А вы все худеете, — благожелательно сказал Осоргин-старший. — Ну ничего, здесь отдохнете. Здесь у нас хорошо, спокойно. Вам бы с дороги искупаться. А потом соответственно закушать. Видите палатку? Там мы вас обоих и устроим. Поутру, если трасса будет свободна, махнем на тот бережок. — Он вдруг рассмеялся. — Ребята думали, вы прибудете со всем хозяйством, ну, с аппаратами и это... с кинозвездами. А вы вдвоем... Без кинозвезд, вот что огорчительно... Так вы купайтесь.

Увязая в бедем песке, мы бредем к заливику, и Каплинский восторженно взмахивает руками.

— А ведь здесь и в самом деле хорошо, — говорит он. — Про сто здорово, что вы меня сюда вытащили! Пять лет не был на Черном море.

— Это Каспийское море, Михаил Семенович, — терпеливо поясняю я. — Каспийское. Понт Хазарский, как говорили в старину. Сняв очки, Каплинский удивленно смотрит на волны.

— Никогда здесь не был, не приходилось, — говорит он. — Э, да все равно! Понт как понт. Давайте окунемся, а? Меня, кажется, опять немного искрит...

Сумасшедший дом, такой небольшой, но хорошо организованный сумасшедший дом. Каплинского то и дело искрит. Все-таки удачно, что я не оставил его в Москве.

Купаться мне совсем не хочется. Наскоро окунувшись, я вытираюсь на берег и валюсь в раскаленный песок.

(Г.Альтов. Создан для бури.  
Научно-фантастич. рассказы. М., 1970)

# Литература

1. ЕЛКИНА В.Н., ЮДИНА Л.С., ХАЙРЕТДИНОВА А.Г. Статистика двух- и трехфонемных сочетаний русской речи. — "Вычислительные системы", Новосибирск, 1969, вып. 37, с. 48-74.

2. ГВОЗДЕВ А.Н. Современный русский литературный язык. ч. I, изд. 2, 1961.

3. ЗИНДЕР Л.Р. Общая фонетика. Изд. ЛГУ, 1960.

4. ГАЛКИНА-ФЕДУРК Е.М., ГОРШКОВА К.В., ШАНСКИЙ Н.М. Современный русский язык. ч. I, М., 1957.

5. БАН А. Палатализация в группах согласных в современном русском языке. Автореферат кандид. дисс. М., 1963.

6. ГЛОВИНСКАЯ М.Я. Фонологическая подсистема редких слов в современном русском литературном языке. Автореферат канд. дисс. М., 1967.

Поступила в ред.-изд.отд.

12 апреля 1973 года

