ВОПРОСЫ ЯЗЫКОЗНАНИЯ

№ 5 2006

© 2006 г. А. В. КОРОЧКОВ

# О КОЛИЧЕСТВЕННОЙ ОЦЕНКЕ АДЕКВАТНОСТИ ЛИНГВИСТИЧЕСКИХ ПРАВИЛ (НА МАТЕРИАЛЕ ПРАВИЛ ЧТЕНИЯ ДЛЯ АНГЛИЙСКОГО ЯЗЫКА)

В статье рассматривается проблема объективной количественной оценки для систем правил, разрабатываемых лингвистами. В результате формализации и программирования правил чтения построена система автоматического транскрибирования английских слов, адекватность которой была оценена посредством сопоставления с несколькими эталонными корпусами транскрипции. Были получены следующие новые данные:

1) количественная оценка адекватности существующих правил чтения;

2) количественная оценка степени влияния морфемной и частеречной разметки на адекватность транскрибирования;

3) расширенные списки слов-исключений.

## 1. ВВЕДЕНИЕ

Рассмотрение языка как деятельности, т.е. системы или процедуры преобразований языковой информации одного типа в другой, хотя и имеет некоторую традицию, приобрело особую актуальность с момента появления такого универсального средства реализации различных деятельностных[1] моделей, как компьютер, что, в частности, проявилось в создании такой отрасли знаний о языке, которую принято называть "Обработка естественного языка (ОЕЯ)"[2]. Одним из типов преобразователей языковой информации является процедура, задающая переход от графемного представления слова к фонемному. Этот тип преобразования составляет основное содержание лингвистической части перехода от письменного представления информации к звуковому.

В области ОЕЯ подобного рода преобразование осуществляется в системах порождения речи "Текст → Речь", которые либо реализуются как отдельные самостоятельные системы преобразования, либо включаются в качестве компонента объемлющей системы ОЕЯ. Большая часть существующих систем подобного рода в лингвистическом аспекте теоретически основывается на работах Р.Л. Венезки [теория буквенно-звуковых (фонемных) соответствий] [Venezky 1970], Н. Хомского и М. Халле (генеративная фонология – абстрактные базовые формы и правила постановки словесных ударений) [Chomsky, Halle 1968][3].

В то же время параллельно с задачами построения автоматических систем порождения речи существует достаточно традиционная область обучения иностранному языку (в интересующем нас случае – английскому), и одним из неотъемлемых компонентов данного процесса является обучение умению читать, т.е. озвучивать письменный текст. Как правило, этот процесс предполагает использование некоторого набора правил чтения, которые теоретически не только не основываются на идеях генеративной фонологии, но и не являются даже в достаточной степени формализованными.

---

[1] В другой терминологии – когнитивных моделей.

[2] В англоязычных источниках – Natural Language Processing (NLP).

[3] См., например, работу Д. Клатта [Klatt 1987].

Если для правил первого типа существуют некоторые количественные оценки их адекватности[4], то для правил второго типа, которые широко используются в обучении (почти каждый учебник по английскому языку содержит некоторый набор правил чтения), существует лишь достаточно приблизительная качественная оценка их адекватности, что не позволяет научно-обоснованно решать вопросы степени применимости правил чтения при обучении чтению англоязычного текста. Отсутствие достаточно достоверной количественной оценки адекватности используемых в обучении (далее – существующих) правил чтения не является случайным. Полностью ручные (неавтоматические) методы получения этой оценки не могут быть признаны достоверными вследствие: 1) необходимости единообразного применения большого количества правил при анализе даже отдельного слова и 2) достаточно большого количества слов, которые при этом должны быть подвергнуты соответствующему анализу.

Как следствие, необходимая количественная оценка может быть получена только после применения некоторой автоматической процедуры, способной на единообразной основе обрабатывать большие объемы информации за достаточно короткие сроки, что предполагает осуществление компьютерного моделирования правил второго типа или, в более общей формулировке, компьютерного моделирования графемно-фонемного перехода (ГФП), базирующегося на существующих правилах чтения.

Далее в работе описываются задачи, возникающие при таком моделировании. Собственно данные будут приведены после краткого рассмотрения этапов моделирования, включающих как создание формальных сводов правил и на их основе автоматической системы получения фонематической транскрипции, так и создание машиночитаемых эталонов и оценочных процедур для получения необходимых количественных данных.

## 2. МОДЕЛИРОВАНИЕ ПРАВИЛ ЧТЕНИЯ

### 2.1. Источники правил

Хотя правила чтения широко известны и доступны, они изначально предназначены для их применения человеком, а не компьютером. Поэтому получение компьютерной модели этих правил предполагает прохождение их через несколько этапов предварительной обработки [Корочков 2002; Korochkov 2003].

Первоначально необходимо определиться с источниками правил чтения. Эти источники могут быть разбиты на две основные категории: общая учебная литература и специальные работы по графике английского языка или правилам чтения. Специальных работ по традиционным (в смысле используемых свойств лингвистических объектов) правилам чтения немного. Это работы В.И. Балинской [Балинская 1964], И.Г. Мкртчян [Мкртчян 1977] и Т.А. Абрамкиной [Абрамкина 1972].

Учебной литературы общей направленности, включающей описание правил чтения, несравнимо больше. Даже подсчитать точное ее количество не представляется возможным, т.к. почти каждое учебное издание по английскому языку может включать соответствующий раздел. Но можно априори (и это представляется разумным) предположить, что специальные работы охватывают значительно бóльшее количество рассматриваемых явлений, чем работы общего характера. Так, примерное число правил буквенно-звуковых (графемно-фонемных) соответствий в работе В.И. Балинской [Балинская 1964] составляет 415 (суммарное число звуковых значений графем по таблицам с 12 по 16 работы В.И. Балинской [Балинская 1964: 337–342]), а в учебнике английского языка для факультетов иностранных языков Т.И. Матюшкиной-Герке и др. [Матюшкина-Герке и др. 1974] – около 160.

---

[4] Таковыми можно считать приводимые в описании некоторых систем результаты их работы, см., например [Andersen et al. 1996; Coker et al. 1990].

Подобное распределение материала предполагает следующий порядок использования источников. Для последующей работы учитываются все правила из специальных работ [Абрамкина 1972; Балинская 1964; Мкртчян 1977], а из работ общего характера отбирается несколько работ, представляющих как разный временной срез, так и разную целевую направленность аудитории. Правила из этих работ могут быть приняты во внимание после рассмотрения правил из специальных работ и учитываться, если не встречались среди ранее рассмотренных и не противоречат им. В качестве источников второго типа были использованы работы [Андрианова и др. 1988; Матюшкина-Герке и др. 1974; Шаншиева 1991].

На следующем этапе из отобранных источников создаются исходные наборы (своды) правил. В рамках графемного подхода [Корочков 2003а][5], к которому может быть отнесено моделирование правил чтения, все правила разделяются на несколько категорий, т.е. создается несколько сводов правил. Это связано с тем, что собственно вычисление посимвольных соответствий и, прежде всего, для гласных элементов, возможно лишь при привлечении дополнительной информации о типе слога, что соответственно, также требует предварительного разбиения слова на слоги. Исходя из этого, для получения транскрипции входное слово в графемном подходе последовательно подвергается процессам слогоделения, постановки ударения (ударений) в изолированном слове (акцентуация), определения других характеристик (типов) слогов (открытость / закрытость и др.) и посимвольного преобразования. Для проведения преобразования может привлекаться и другая информация, например, о принадлежности слова к определенной части речи и морфемном составе анализируемого слова.

В источниках могут встречаться разнообъемные правила, когда в соответствующей паре правил одно является более общим, а другое – частным случаем первого. Встречаются также правила, которые прямо противоречат друг другу. Исходя из этого правила в полученных сводах затем подвергаются дополнительной обработке, в результате которой из сводов устраняются противоречащие правила и оставляется лишь один из вариантов разнообъемных правил. После этого оставшиеся правила алгоритмизуются (т.е. задается последовательность их применения) и формализуются. На последнем этапе моделирования эти правила (алгоритмы) реализуются на каком-либо языке программирования, что приводит к созданию работающей на компьютере модели соответствующих языковых процессов.

### 2.2. Автоматическая система фонематического транскрибирования

Описанные выше задачи решались в рамках проекта создания автоматической системы фонематического транскрибирования слов английского языка – KorLang Transcriber, начатого автором в 1999 г. и продолжающегося до сих пор [Korochkov 2003][6]. В настоящее время реализованы две версии, названные 1999 и 2002 по времени их первой реализации. Исходные правила первой версии в основном построены на работе И.Г. Мкртчян [Мкртчян 1977] с добавлением авторских правил. Исходные правила этой версии не имели промежуточного формального описания. С точки зрения программной отладки системы эта версия не является полностью завершенной.

Исходные правила второй версии (2002) построены на нескольких источниках (см. выше), но, прежде всего, на работах В.И. Балинской [Балинская 1964] и И.Г. Мкртчян [Мкртчян 1977], так как доля правил из других источников оказалась относительно невелика. Правила данной версии являются реализацией системы формальных правил,

---

[5] Подход, в основе которого лежит посимвольное преобразование в отличие, например, от пословного.

[6] Информация о проекте также доступна в сети Интернет по адресу: http://www.geocities.com/korochkov/kortextt_r.htm.

которая может быть отнесена к формальному классу конечных преобразователей (трансдьюсеров) [Kaplan, Kay 1994].

### 2.3. Формальная система правил

Для промежуточного этапа формализации (конечный – запись на языке программирования) правил слогоделения, акцентуации, определения типа слога и посимвольного преобразования используется система *подстановок* или, в другой терминологии, – *переписываний*, называемая *правилами*, которая задается определенными входными и выходными алфавитами и конечными совокупностями правил (формул) *вывода*. Совокупности правил вывода задаются как упорядоченные множества. Используется два типа правил вывода: правила *замены* описания некоторого лингвистического объекта на другое и правила *применимости* либо правил замены, либо других правил применимости. Операция замены в соответствующих формулах обозначается символом "→", а операция применимости – символом "⇒".

Общая форма двухуровнего правила применимости следующая:

$$\lambda\cdots\omega \Rightarrow (\alpha\cdots\beta\phi\alpha\cdots\beta \rightarrow \alpha\cdots\beta\varphi\alpha\cdots\beta),$$

где операция, обозначенная многоточием, соединяет части разрывного контекста, а алфавитные символы обозначают объекты, являющиеся операндами операций. Правило замены является частным случаем правила применимости, в котором не задано (опущено) первое условие (контекст). В части правила с операцией замены также может быть опущена любая контекстная часть.

Каждое правило обозначается типом и номером в этом типе. Тип задает множество правил. Порядок выборки элемента из множества определяется номером правила. Применимость правила с меньшим номером проверяется до применимости правила с большим номером.

Выбор правила для преобразования происходит следующим образом. Для каждого текущего символа (в более общем случае – объекта) анализируемого слова просматривается соответствующее множество правил, начиная с первого правила и далее по порядку номеров правил. При выборе правила сравнивается на совпадение левая часть правила (до знака преобразования) и часть представления слова такой же длины, что и левая часть правила, начинающаяся с текущего объекта (которому соответствует некоторый символ исходного слова). После выбора правила и преобразования по этому правилу происходит смена текущего элемента. Порядок смены текущего элемента и прекращение выборки правил зависят от множества, которому принадлежат применяемые правила.

Для примера рассмотрим более подробно первый из четырех подэтапов ГФП – слогоделение. Процесс слогоделения проходит в следующем порядке. Входная строка, полученная в результате либо автоматического, либо смоделированного "вручную" через возможные выходные данные морфологического анализа, инициализируется, сегментируется, категоризируется и маркируется. Инициализация предполагает обрамление исходной строки объектами, задающими границы строки и начальное значение для подсистемы маркирования слогов. При категоризации, совмещенной с графемной сегментацией, каждый выделенный графемный объект (символ или их последовательность) проинициализированной строки получает атрибут (признак) принадлежности к категории гласных или согласных элементов. Маркирование заключается в присвоении каждому объекту атрибута, обозначающего номер слога. В формальном своде правил имеется одно правило инициализации, 48 правил сегментации-категоризации и 18 правил маркирования. Ниже приводится множество правил маркирования в сокращенной форме отображения[7], при которой указываются только существенные для данного правила

[7] С полной формой записи и подробным описанием этапа можно ознакомиться в работе [Корочков 2003б]. Статья доступна также в электронном виде с сайта журнала "Web journal of formal, computational and cognitive linguistics" по адресу: http://fccl.ksu.ru/conf2003/cogmod/s20.rar.

свойства задействованных объектов. Гласные обозначены символом $V$, а согласные – символом $C$. Символ подчеркивания обозначает границу словоформы. Переменная $\eta$ содержит номер слога, остальные символы являются литералами[8].

Упорядоченное множество правил маркирования подэтапа слогоделения ($M$):

$$V_\eta V \rightarrow V_\eta V_{++\eta}, \quad \text{(M1)}$$

$$V_\eta \mathbf{x} V \rightarrow V_\eta \mathbf{x}_\eta V_{++\eta}, \quad \text{(M2)}$$

$$V_\eta C V \rightarrow V_\eta C_{\eta+1} V_{++\eta}, \quad \text{(M3)}$$

$$V_\eta \mathbf{ll} V \rightarrow V_\eta \mathbf{l}_\eta \mathbf{l}_{\eta+1} V_{++\eta}, \quad \text{(M4)}$$

$$V_\eta \mathbf{x}\mathbf{l} V \rightarrow V_\eta \mathbf{x}_\eta \mathbf{l}_{\eta+1} V_{++\eta}, \quad \text{(M5)}$$

$$V_\eta C\mathbf{l} V \rightarrow V_\eta C_{\eta+1} \mathbf{l}_{\eta+1} V_{++\eta}, \quad \text{(M6)}$$

$$V_\eta \mathbf{rr} V \rightarrow V_\eta \mathbf{r}_\eta \mathbf{r}_{\eta+1} V_{++\eta}, \quad \text{(M7)}$$

$$V_\eta \mathbf{x}\mathbf{r} V \rightarrow V_\eta \mathbf{x}_\eta \mathbf{r}_{\eta+1} V_{++\eta}, \quad \text{(M8)}$$

$$V_\eta C\mathbf{r} V \rightarrow V_\eta C_{\eta+1} \mathbf{r}_{\eta+1} V_{++\eta}, \quad \text{(M9)}$$

$$V_\eta C C_{\_} \rightarrow V_\eta C_\eta C_{\eta-}, \quad \text{(M10)}$$

$$V_\eta C C \rightarrow V_\eta C_\eta C_{++\eta}, \quad \text{(M11)}$$

$$C_\eta C \rightarrow C_\eta C_\eta, \quad \text{(M12)}$$

$$C_\eta V \rightarrow C_\eta V_\eta, \quad \text{(M13)}$$

$${}_{-\eta}V \rightarrow {}_{-\eta}V_{++\eta}, \quad \text{где} \quad \eta = 0, \quad \text{(M14)}$$

$${}_{-\eta}C \rightarrow {}_{-\eta}C_{++\eta}, \quad \text{где} \quad \eta = 0, \quad \text{(M15)}$$

$$V_\eta C_{\_} \rightarrow V_\eta C_{\eta-}, \quad \text{(M16)}$$

$$V_\eta C C C_{\_} \rightarrow V_\eta C_\eta C_\eta C_{\eta-}, \quad \text{(M17)}$$

$$V_\eta C C C C_{\_} \rightarrow V_\eta C_\eta C_\eta C_\eta C_{\eta-}. \quad \text{(M18)}$$

После слогоделения его результат подвергается процессу акцентуации. Этот процесс включает следующие последовательные этапы: частеречной инициализации, выделения частей слова, подлежащих отдельному анализу, специального морфологического анализа, предварительной инициализации, собственно акцентуации и терминации. В результате анализа формализуемого исходного свода правил акцентуации выделено (формализовано) 2 правила частеречной инициализации, 46 правил выделения частей слова, подлежащих отдельному анализу, 34 правила специального морфологического анализа, 2 правила предварительной инициализации, 41 правило собственно акцентуации (включающие 32 правила постановки первичного ударения и 9 правил постановки вторичного ударения) и 5 правил терминации.

Следующая процедура (определения характеристик слогообразующего элемента – типа слога) состоит из таких этапов, как: определение 1) модифицированности (наличия

[8] Самоопределенными величинами, т.е. обозначающими самих себя.

буквы *r* справа от гласного элемента), 2) первичной открытости/закрытости, 3) вторичной закрытости (слогов, исходно являющихся открытыми) и 4) степени полноты озвучивания гласного элемента. Она использует 3 правила определения модифицированности слога, 2 правила определения первичной открытости/закрытости слога, 12 правил определения вторичной закрытости слога и 31 правило определения степени полноты озвучивания гласного элемента слога.

Формализованный свод правил для этапа посимвольного преобразования включает 44 формализованных правила для буквы *A*, 3 правила для буквы *B*, 14 правил – для буквы *C*, 11 правил – для буквы *D*, 69 правил – для буквы *E*, 1 правило – для буквы *F*, 10 правил – для буквы *G*, 2 правила – для буквы *H*, 41 правило – для буквы *I*, 1 правило – для буквы *J*, 3 правила – для буквы *K*, 9 правил – для буквы *L*, 2 правила – для буквы *M*, 19 правил – для буквы *N*, 55 правил – для буквы *O*, 6 правил – для буквы *P*, 3 правила – для буквы *Q*, 8 правил – для буквы *R*, 35 правил – для буквы *S*, 17 правил – для буквы *T*, 52 правила – для буквы *U*, 1 правило – для буквы *V*, 6 правил – для буквы *W*, 15 правил – для буквы *X*, 9 правил – для буквы *Y*, 3 правила – для буквы *Z* и 24 правила для недопущения удвоения согласных звуков. Всего выделено 463[9] формальных правила этой категории.

Ниже приводится пример описания слова *atmospheric* после одного из этапов анализа. В примере описание отдельного объекта, построенное по позиционному принципу, обрамляется круглыми скобками, число в описании объекта обозначает номер слога. Другие обозначения: *Vb* обозначает глагол, *Adj* – прилагательное, *Pf* – префикс, *Rt* – корень, *St1* – первичную простую ударность, *St1c* – первичную сложную ударность, *St2* – вторичное ударение, *Us* – безударность. *Op* обозначает открытость слога, *Cl* – закрытость слога. *Md* обозначает модифицированность гласной, *Um* – немодифицированность, *Fl* – полное озвучивание гласной, *Rd* – редуцированное. Отсутствие звучания отдельного элемента обозначено символом ∅. Символ ~ обозначает морфемную границу, остальные символы являются литералами.

После этапа определения типа слога слово *atmospheric* может быть представлено следующим образом:

*atmospheric*: (_, 0) ~ (**a**, *V*, 1, *Adj*, *Rt*, *St2*, *Um*, *Cl*, *Fl*) (**t**, *C*, *1*, *Adj*) (**m**, *C*, 2, *Adj*) (**o**, *V*, 2, *Adj*, *Rt*, *Us*, *Um*, *Cl*, *Rd*) (**s**, *C*, 2, *Adj*) (**p**, *C*, 3, *Adj*) (**h**, *C*, 3, *Adj*) (**e**, *V*, 3, *Adj*, *Rt*, *St1c*, *Um*, *Cl*, *Fl*) (**r**, *C*, 4, *Adj*) ~ ~ (**i**, *V*, 4, *Adj*, *Rt*, *Us*, *Um*, *Cl*, *Rd*) (**c**, *C*, 4, *Adj*) ~ (_, 0).

То же в краткой форме записи: $a_1^{Um,\,Cl.Fl}\, t_1 m_2 o_2^{Um,\,Cl.Rd}\, s_2' p_3 h_3 e_3^{Um,\,Cl,\,Fl}\, r_4 i_4^{Um,\,Cl,\,Rd}\, c_4$.

Результат преобразования после завершающего этапа может быть представлен в краткой форме следующим образом: /ˌætməsˈferɪk/.

### 2.4. Реализация

Для реализации системы транскрибирования (формальной системы правил) была использована собственная система программирования ассоциативных (семантических) сетей *KorNet* [Корочков 1999][10], разработанная в 1993 году.

### 2.5. Словарные источники

Решение задачи количественной характеризации адекватности правил чтения помимо создания автоматической системы, построенной на базе свода таких правил, предполагает наличие некоторого эталонного машиночитаемого корпуса[11] транскрипций

---

[9] В это число не включены правила с полным контекстом (который описывает отдельную словоформу), предназначенные для обработки слов-исключений.

[10] Информация о системе также доступна в сети Интернета по адресам: http://www.geocities.com/korochkov/kornet.htm или http://korochkov.math.mrsu.ru/kornet.htm.

[11] Непривычное использование термина "корпус" призвано показать нерелевантность порядка вхождения элементов в это образование.

слов. Основой для построения такого корпуса может являться практически любой словарь, в котором приводятся транскрипции слов.

Первоначально для проведения машинного эксперимента был выбран раздел буквы "А" словаря А.С. Хорнби [Хорнби 1984]. Информация этого раздела (заглавие словарной статьи и варианты ее транскрибирования) была переведена в машиночитаемую форму. Создано 4 варианта представления этого раздела. В первом варианте подлежащее анализу слово полностью совпадает с заглавием исходной словарной статьи. Во втором – дополнительно содержит информацию о части речи этой словоформы и/или ее основы. В третьем варианте словоформа может быть разбита на морфемы. В четвертом варианте может одновременно предоставляться как частеречная информация, так и морфемная. После удаления некоторых словосочетаний этот источник содержал 1210 единиц.

Затем был добавлен второй источник. Таким источником послужил свободно распространяемый в Интернете машиночитаемый вариант англо-русского словаря В.К. Мюллера [Мюллер 1961], который был адаптирован для решения поставленных выше задач.

Адаптация включала следующее:

– выделение заглавий словарных статей и соответствующих им транскрипций из текста словарной статьи (для одного заглавия возможно несколько вариантов транскрибирования);

– приведение этой информации к формату, применяемому на входе подсистемы оценки результата преобразований;

– исключение части сложных (многокорневых) слов;

– исключение словосочетаний;

– замена конечной фонемной последовательности /⁅ən/ на /⁅n/ для синхронизации с соответствующим фонемным описанием в словаре А.С. Хорнби;

– замена фонемной последовательности /ɪkəl/ на /ɪkl/ с той же целью;

– замена конечной фонемной последовательности /⁅əl/ на /⁅l/;

– замена конечной фонемной последовательности /ɪtɪ/ на /ətɪ/;

– исправление фактических ошибок в описаниях.

Адаптированный вариант содержит 35068 единиц (входов), исходный (неадаптированный) вариант – 46120 словарных статей.

Помимо рассмотренных выше предполагалось использовать и третий источник – свободно распространяемый в Интернете машиночитаемый словарь университета Карнеги Меллона (США) – The Carnegie Mellon pronouncing dictionary[12]. Но этот источник не был использован по следующим причинам:

– словарь построен на основе американского варианта озвучивания и, соответственно, транскрибирования графем английского языка;

– в результирующей (эталонной) транскрипции обозначается не ударный слог, как это принято в большинстве существующих словарей и, соответственно, на выходе созданной автоматической системы транскрибирования, а ударность/безударность гласной. Проведение автоматического приведения к первой форме требует применения автоматического транскриптора для определения места расположения знака ударения, что нарушает принцип эталонности. Ручная правка невозможна из-за временных затрат (словарь содержит около 129.5 тыс. статей).

## 2.6. Оценочные процедуры

Еще одним компонентом, необходимым для решения поставленной задачи, являются оценочные процедуры для получения таких характеристик как:

---

[12] The CMU Pronouncing dictionary; адрес в Интернете (URL): http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

– словесная адекватность системы ($\alpha_W$), вычисляемая по формуле:

$$\alpha_W = W^C * 100 / W^T,$$

где $W^C$ обозначает количество правильно транскрибированных слов, а $W^T$ – общее количество проанализированных слов;

– и фонемная адекватность[13] системы ($\alpha_P$), вычисляемая по аналогичной формуле:

$$\alpha_P = P^C * 100 / P^T,$$

где $P^C$ обозначает полученное число корректных фонем, а $P^T$ – общее количество фонем в эталонном корпусе.

Вычисление первой из двух характеристик является достаточно тривиальной задачей, что нельзя сказать о вычислении фонемной адекватности. В этом случае возникает проблема соотнесенности[14] сравниваемых значений фонем (проблема выбора сравниваемых позиций в результирующей и эталонной строках). Эта проблема является следствием того, что в результирующей строке анализа помимо прямых несоответствий (замен) фонем по сравнению с эталоном могут быть вставки и пропуски различной длины в различных совместных сочетаниях. Отсюда текущее несоответствие может быть результатом либо замены, либо вставки или опущения (пропуска) одной или нескольких фонем, либо сочетаний одной или нескольких из обозначенных причин. Для решения этой проблемы была предложена и использована[15] специальная 83-шаговая процедура (алгоритм).

## 3. РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТА

В результате эксперимента были получены данные следующих типов:

– количественные данные о степени адекватности смоделированных правил чтения;

– количественные данные о степени влияния наличия морфемной и частеречной разметки анализируемого материала на адекватность ГФП;

– расширенные списки слов, которые могут быть отнесены к исключениям.

### 3.1. Адекватность

Упомянутые выше количественные данные о степени адекватности смоделированных правил чтения приведены в табл. 1.

Обозначения, принятые в таблице: *Хорнби 1* обозначает неразмеченный вариант машиночитаемого раздела словаря А.С. Хорнби, *Хорнби 2* – вариант, содержащий морфемную разметку, *Хорнби 3* – вариант, содержащий частеречную разметку, *Хорнби 4* – вариант, содержащий и морфемную, и частеречную разметку, *Хорнби 1_2* – это вариант *Хорнби 1* с подключенным словарем приводимых в существующих правилах исключений, начинающихся на букву *А*. Такое же соотношение между *Хорнби 4_2* и *Хорнби 4*. Адаптация словаря В.К. Мюллера, обозначенная как *Мюллер*, имеет только один вариант, который не содержит ни морфемной, ни частеречной разметки.

---

[13] На фонемном уровне возможно также вычисление с учетом числа полученных несоответствующих фонем, которое могло бы быть названо фонемной точностью или корректностью. В данной работе это значение не вычислялось.

[14] В англоязычных исследованиях (например [Lawrence, Kaye 1986; Wightman, Talkin 1997]) используется термин "выравнивание" (alignment).

[15] Предполагается отдельная публикация с описанием этой процедуры.

*Таблица 1*

| Версия правил | Вход | $\alpha_W$ | $\alpha_P$ |
|---|---|---|---|
| 1999 | Хорнби 1 | 20 | 69 |
| 1999 | Мюллер | 28 | 76 |
| 2002 | Хорнби 1 | 28 | 80 |
| 2002 | Хорнби 2 | 33 | 82 |
| 2002 | Хорнби 3 | 43 | 84 |
| 2002 | Хорнби 4 | 54 | 88 |
| 2002 | Хорнби 1_2 | 29 | 80 |
| 2002 | Хорнби 4_2 | 55 | 88 |
| 2002 | Мюллер | 39 | 83 |

### 3.2. Влияние дополнительной разметки

Из апробации на основе словаря А.С. Хорнби, для которого было создано несколько машиночитаемых вариантов представления, различающихся наличием или отсутствием дополнительной частеречной и/или морфемной разметки, были получены следующие данные, касающиеся степени влияния на адекватность системы наличия этой дополнительной информации.

Разница между анализом с учетом частеречной принадлежности (используется в правилах специального морфологического анализа, постановки ударений, определения полноты озвучивания гласной слога и некоторых правилах посимвольных соответствий) и без него составляет в абсолютном выражении для словесной адекватности 15%, а для фонемной – 4%. Увеличение адекватности относительно неразмеченного варианта (далее – относительное изменение), адекватность которого может быть взята в этом случае за базовую и приравнена к 100%, составляет для словесной разновидности 53.57%, а для фонемной – 5%.

Абсолютная разница между анализом с учетом морфемного деления (используется в большинстве правил) и без него составляет для словесной адекватности 5%, а для фонемной – 2%. Относительное изменение словесной адекватности составило 17.85%, фонемной – 2.5%.

Разница между анализом на основе неразмеченного варианта словаря А.С. Хорнби и анализом на основе частеречно и морфемно размеченного варианта составляет в абсолютном выражении для словесной адекватности 26%, а для фонемной – 8%. Относительное изменение словесной адекватности – 92.85%, фонемной – 10%.

Подключение стандартного словаря слов-исключений к анализу на основе неразмеченного варианта словаря А.С. Хорнби привело к увеличению в абсолютном выражении словесной адекватности на 1%, т.е. составила 29%, фонемная же адекватность осталась практически неизменной. Относительное изменение словесной адекватности составило 3.57%.

Подключение стандартного словаря слов-исключений к анализу на основе полностью размеченного варианта словаря А.С. Хорнби привело к увеличению в абсолютном выражении словесной адекватности на 1%, т.е. составила 55%, фонемная же адекватность осталась также неизменной. Относительное изменение словесной адекватности – 1.85%.

Эти результаты могут быть сведены в табл. 2, в которой абсолютное увеличение (разница) обозначено через индекс *Abs*, а относительное – через *Rel*:

*Таблица 2*

| Вход 1 | Вход 2 | $\alpha_W^{Abs}$ | $\alpha_P^{Abs}$ | $\alpha_W^{Rel}$ | $\alpha_P^{Rel}$ |
|---|---|---|---|---|---|
| Хорнби 1 | Хорнби 3 | 15 | 4 | 53.57 | 5 |
| Хорнби 1 | Хорнби 2 | 5 | 2 | 17.85 | 2.5 |
| Хорнби 1 | Хорнби 4 | 26 | 8 | 92.85 | 10 |
| Хорнби 1 | Хорнби 1_2 | 1 | <1 | 3.57 | –[16] |
| Хорнби 4 | Хорнби 4_2 | 1 | <1 | 1.85 | – |

Полученные данные позволяют сделать следующие выводы.

1. Упорядоченные (алгоритмизированные) правила чтения позволяют правильно озвучивать (читать) значительную часть входного материала, особенно с точки зрения фонемного состава результата (более 88% адекватности). Сообразно этому можно считать вполне целесообразным их применение при обучении чтению.
2. Учет частеречной принадлежности при преобразовании позволяет повысить его словесную адекватность в 1.5 раза, что является достаточно большой величиной и косвенно говорит о соответствующей степени использования частеречной характеристики в рассматриваемых правилах и объеме подпадающих под эти правила входных данных.
3. Учет морфемного деления позволяет повысить словесную адекватность преобразования менее чем в 1.2 раза.
4. Одновременный учет и частеречной принадлежности, и морфемного деления позволяет повысить словесную адекватность преобразования почти в 2 раза, что в конечном итоге говорит о важности учета этих характеристик при преобразовании (чтении).
5. Фонемная адекватность мало зависит от использования дополнительной информации о частеречной принадлежности и/или морфемном делении, что может говорить о том, что доля в результате несоответствующих эталону фонем относительно невелика, что косвенно также подтверждается общей степенью фонемной адекватности (88%).
6. Стандартные списки слов-исключений не являются исчерпывающими.

При последующем использовании этих выводов необходимо учитывать их некоторую гипотетичность, связанную с использованием для апробации в нескольких вариантах представления только относительно небольшой части словаря А.С. Хорнби.

Также некоторой гипотетичностью будет обладать перенос полученных данных на словарь В.К. Мюллера. Эта экстраполяция могла бы дать следующие абсолютные результаты, просчитанные от базовых вариантов в 39 и 83%. Полностью размеченный вариант словаря В.К. Мюллера мог бы дать при использовании правил версии 2002 словесную адекватность в 75.2%, а фонемную – в 91.3%. Расхождение с результатами, полученными на основе машиночитаемой версии части словаря А.С. Хорнби, может объясняться различиями в фонемном описании словарного материала, что действительно было обнаружено в процессе работы с рассматриваемыми словарями.

### 3.3. Перспективы развития проекта

Полученные в результате экспериментов с системой автоматического транскрибирования *KorLang Transcriber* количественные оценки адекватности правил чтения нельзя

[16] Значения, которые не просчитывались, в таблицах данной работы обозначены тире.

считать окончательными вследствие существования возможности усовершенствования как собственно формальных наборов правил в сторону их большего соответствия исходным наборам правил, так и исходных сводов правил. Кроме того, существуют еще следующие три направления возможного развития с целью получения более точных количественных результатов:

– усовершенствование программной реализации (отладка) формальных наборов правил;
– усовершенствование оценочной процедуры, в частности, того ее компонента, который отвечает за решение проблемы соотнесенности (выравнивания) сравниваемых значений;
– усовершенствование эталонного материала как в аспекте устранения неточностей в описаниях и разночтений в разных источниках (синхронизация источников), так и в аспекте увеличения числа и объема словарных источников. Под увеличением объема источников имеется в виду возможность расширения машиночитаемой версии словаря А.С. Хорнби до его полного покрытия.

### *3.3.1. Ошибочные ситуации*

Часть отмеченной выше работы по уточнению результатов машинной обработки была проделана после завершения соответствующих машинных экспериментов. При ручном рассмотрении ошибочных результатов по источнику *Хорнби 4* все ошибки были распределены по следующим классам:

– ошибки в постановке ударения;
– ошибки в определении типа слога (открытый или закрытый);
– ошибки в определении статуса гласной в безударном положении (степень редукции или ее отсутствие);
– ошибки слогоделения;
– ошибки определения или учета морфемной структуры слова;
– ошибки посимвольного преобразования;
– ошибки в эталонных описаниях.

Кроме того, часть ошибочно проанализированных слов была отнесена к группе иноязычных слов, сохранивших в основном свое исходное озвучивание и, как следствие этого, полностью выпадающих из системы существующих правил чтения для английского языка (например, *attaché, avalanche, avoirdupois*).

В каждой из выделенных категорий (классов) все ошибки далее были распределены по нескольким типам (типовым ситуациям). Ошибки первой подгруппы были отнесены к одной из 52 типовых ситуаций, таких, например, как:

– нестандартная (т.е. не вытекающая из существующих правил) постановка вторичного ударения на предударном слоге (например, *au'tumnal*, где слог со вторичным ударением подчеркнут);
– неучет в исходных правилах возможного соположения глагольного суффикса *-ise/-ize* и последующего безударного зияния[17] как в словах *acclimatization* или *authorisation*;
– и другие.

Для второй подгруппы были выявлены 14 типовых ситуаций, например, таких как:

– нестандартное незакрытие слога перед суффиксом *-ic* (*acetic, analgesic* и др.);
– нестандартное закрытие открытого слога (*Adam, atom, allege* и др.);
– неучет в стандартных правилах влияния наличия безударного зияния на действие вторичного ударения (*alienation, amiability, affiliation* и др.);
– и другие.

---

[17] Соположения гласных, озвучиваемых в отличие от диграфа по отдельности [Мкртчян 1977].

Ошибки третьей подгруппы были распределены по 43 типовым ситуациям, таким, например, как:

– нестандартное отсутствие редукции гласной межударного слога (*affectation, analgesic*[18] и др.);
– неучет глагольного суффикса *-yse/-yze* (*analyse, analyze*);
– нестандартное отсутствие полной редукции конечного *-e* (*acme, adobe* и др.);
– и другие.

Для ошибок, зависимых от определения морфемной структуры слова, было обнаружено 24 типовых случая (ситуации), среди которых, например, такие как:

– невыделение суффикса *-hood* для проведения отдельного анализа оставшейся части слова, содержащего такой суффикс (*adulthood*);
– неучет полнозначного префикса *a-* (*aseptic, asexual* и др.);
– неправильное выделение суффикса *-or/-er* для проведения отдельного анализа оставшейся части неотглагольного существительного (*ambassador*);
– и другие.

Для ошибок посимвольного преобразования было выявлено 52 типовых случая (ситуации), среди которых, например, такие как:

– неправильный анализ *ch* как /t\|/ вместо /k/ (*ache, alchemy*);
– нестандартное озвучивание *d* как /dʒ/ (*adulation*);
– неучет озвучивания *-gue* как /g/ (analogue);
– и другие.

После выделения типовых ошибочных ситуаций каждая из них была отнесена к одной из трех категорий. К первой были отнесены ошибочные ситуации, которые могут быть исправлены без изменения исходных (существующих) правил, например, ситуации, связанные с ошибками в эталонных данных или неправильными формализациями исходных правил, ко второй – ситуации, для устранения которых требуется либо каким-то образом изменить исходные (стандартные) правила, либо добавить в исходные наборы правил новые обобщения (правила) и, при этом, необходимые правила могут быть сформулированы. К третьей категории были отнесены все остальные случаи, т.е. ошибочные ситуации, для устранения которых соответствующие правила в настоящее время не могут быть сформулированы[19].

### *3.3.2. Уточнение адекватности*

Проведенная категоризация позволила получить новые (в какой-то степени гипотетичные) данные о степени адекватности существующих правил чтения. Для этого было определено, какие слова были бы проанализированы правильно, если бы были устранены ошибочные ситуации 1) только первой категории, 2) только второй категории и 3) первой и второй категории в совокупности. При этом устранение ошибочной ситуации не всегда может приводить к полностью правильному анализу соответствующего слова, т.е. к увеличению $\alpha_W$. Это является следствием возможности сочетания при анализе слова действия множества различных ошибочных ситуаций. В этом случае вместо изменения $\alpha_W$ в расчетах должно изменяться значение $\alpha_P$, что при "ручной" обработке представило определенную проблему, связанную с тем, что вручную было сложно отследить точное воздействие устранения только одной из множества ошибочных ситуаций, влияющих на возможный результат анализа с точки зрения его фонемного состава.

В этом случае нет прямой зависимости: "одна ошибочная ситуация – одна фонемная ошибка". Поэтому далее при рассмотрении фонемной адекватности будут приведены результаты для следующих трех[20] возможных случаев, когда устранение ошибочных си-

---

[18] Подчеркнут нередуцируемый элемент.
[19] Что не исключает возможности их формулирования в будущем.
[20] Возможно, наиболее вероятных.

туаций приводит к увеличению фонемной адекватности в среднем 1) на 1 единицу на каждое ошибочно проанализированное слово, 2) на 1.5 единицы на слово и 3) на две единицы на слово.

В результате такой ручной работы были получены следующие уточнения достигнутых ранее результатов. При учете исправления ошибочных ситуаций только первого типа появляется 138 новых правильно проанализированных слов, что дает новую степень адекватности $\alpha_W = 66.36\%$. Учет исправлений только второго типа дает прибавку в 67 слов и, соответственно, совокупная прибавка по двум категориям составляет 205 слов, что приводит к степени адекватности на словном (словесном) уровне ($\alpha_W$), равной 71.9%.

На фонемном уровне было выявлено 45 случаев фонемных ошибок первого типа, устранение которых не приводило к последующему правильному анализу слова, что говорит о том, что при анализе соответствующего слова были совершены и другие ошибки. Таких фонемных ошибок второго типа было выявлено 17.

При виртуальном исправлении ошибок первого типа фонемная адекватность при учете трех возможных соотношений между количеством устраненных ошибочных ситуаций и количеством исправленных фонем может увеличиться до 90.84%, 91.72% или 92.6%. При совокупном устранении ошибочных ситуаций первого и второго типа фонемная адекватность может увеличиться до 91.92%, 93.22% или 94.53%.

Более подробная информация приведена в таблицах 3 и 4.

*Таблица 3*

| Тип ошибки | Приращение (слов) | $\alpha_w$ (%) |
|---|---|---|
| I | 138 | 66.36 |
| II | 67 | – |
| I + II | 205 | 71.9 |

*Таблица 4*

| Тип ошибки | Приращение (фонем на слово) | Общее приращение (фонем) | $\alpha_P$ (%) |
|---|---|---|---|
| I | 1 | 45 + 138 = 183 | 90.84 |
| I | 1.5 | 45 + 138*1.5 = 252 | 91.72 |
| I | 2 | 45 + 138*2 = 321 | 92.6 |
| II | 1 | 17 + 67 = 84 | – |
| II | 1.5 | 17 + 67*1.5 = 117.5 | – |
| II | 2 | 17 + 67*2 = 151 | – |
| I + II | 1 | 183 + 84 = 267 | 91.92 |
| I + II | 1.5 | 252 + 117.5 = 369.5 | 93.22 |
| I + II | 2 | 321 + 151 = 472 | 94.53 |

## 4. ЗАКЛЮЧЕНИЕ

Окончательная интерпретация полученных результатов с точки зрения степени применимости правил чтения скорее всего будет прерогативой специалистов по методике обучения английскому языку. Но нельзя не заметить, что даже по предварительным результатам степень адекватности применяемых правил чтения представляется достаточ-

но высокой, чтобы говорить о целесообразности их использования в комбинации с заучиванием отдельных слов, не подходящих ни под одно правило.

С общелингвистической точки зрения полученные данные могут быть использованы в типологическом аспекте для описания степени регулярности графемно-фонемных соответствий в различных языках. Кроме того, полученная информация о количестве применяемых правил (около 500, см. выше) позволяет более точно охарактеризовать степень сложности графемно-фонемного перехода для английского языка. Эта числовая оценка, в отличие от распространенного мнения "Написано Манчестер – читай Ливерпуль", может быть использована при типологическом сравнении языков по критерию сложности графемно-фонемных переходов.

## СПИСОК ЛИТЕРАТУРЫ


Абрамкина 1972 – *Т.А. Абрамкина.* Обучение произношению и технике чтения на английском языке. М., 1972.

Андрианова и др. 1988 – *Л.Н. Андрианова, Н.Ю. Багрова, Э.В. Ершова.* Английский язык для заочных технических вузов: Учебник. М., 1988.

Балинская 1964 – *В.И. Балинская.* Графика современного английского языка. М., 1964.

Корочков 1999 – *А.В. Корочков.* KorNet – модель данных и система программирования семантической сети для платформы Win32/PC // Методы и средства управления технологическими процессами: Сб. трудов III Междунар. научн. конференции. Саранск, 1999.

Корочков 2002 – *А.В. Корочков.* Компьютерное моделирование графемно-фонемного преобразования в английском языке (постановка задачи) // Социальные и гуманитарные исследования: традиции и реальности: Межвузовский. сб. науч. трудов. Саранск, 2002. Вып. 2.

Корочков 2003а – *А.В. Корочков.* Компьютерное моделирование графемно-фонемного преобразования в английском языке (выбор подхода) // Вестник Мордовского ун-та. 2003. № 1–2.

Корочков 2003б – *А.В. Корочков.* Компьютерное моделирование графемно-фонемного преобразования в английском языке (слогоделение) // Обработка текста и когнитивные технологии. Вып.8. Междун. конференция "Когнитивное моделирование в лингвистике": Сб. докладов. Варна, 2003.

Матюшкина-Герке и др. 1974 – *Т.И. Матюшкина-Герке и др.* Учебник английского языка для 1 курса институтов и факультетов иностранных языков. Учебник. М., 1974.

Мкртчян 1977 – *И.Г. Мкртчян.* Learn to read English words. (Пособие по технике чтения на английском языке). М., 1977.

Мюллер 1961 – *В.К. Мюллер.* Англо-русский словарь. М., 1961.

Хорнби 1984 – *А.С. Хорнби.* Учебный словарь современного английского языка при участии К. Руз. М., 1984.

Шаншиева 1991 – *С.А. Шаншиева.* Английский язык для математиков (интенсивный курс для начинающих). Учебник. М., 1991.

Andersen et al. 1996 – *O. Andersen* et al. Comparison of two tree-structured approaches for grapheme-to-phoneme conversion // Proceedings of ICSLP-96. V. 3. Philadelphia, 1996.

Chomsky, Halle 1968 – *N. Chomsky, M. Halle.* The sound pattern of English. New York, 1968.

Coker et al. 1990 – *C. Coker, K. Church, M. Liberman.* Morphology and rhyming: Two powerful alternatives to letter-to-sound rules for speech synthesis // Proceedings of the First ESCA workshop on speech synthesis / G. Bailly, C. Benoit (eds.). Autrans (France), 1990.

Kaplan, Kay 1994 – *R. Kaplan, M. Kay.* Regular models of phonological rule systems // Computational linguistics. 20.3. 1994.

Klatt 1987 – *D.H. Klatt.* Review of text-to-speech conversion for English // Journal of the Acoustical society of America. 82 (3). September. 1987.

Korochkov 2003 – *A.V. Korochkov.* Computer modeling of grapheme to phoneme conversion for English // Proceeding of International workshop SPECOM'2003 (Speech and Computer). M., 2003.

Lawrence, Kaye 1986 – *S.G.C. Lawrence, G. Kaye.* Alignment of phonemes with their corresponding orthography // Computer speech and language. 1. 1986.

Venezky 1970 – *R.L. Venezky.* The structure of English orthography. 's-Gravenhague, 1970.

Wightman, Talkin 1997 – *C.W. Wightman, D.T. Talkin.* The aligner: text-to-speech alignment using Markov models // Progress in speech synthesis / Jan P. H. van Santen et al. (eds.). New York, 1997.