

АВТОМАТИЧЕСКАЯ ГЕНЕРАЦИЯ ТРАНСКРИПЦИИ ДЛЯ СОЗДАНИЯ УЧЕБНЫХ МАТЕРИАЛОВ ПО ИНОСТРАННОМУ ЯЗЫКУ НА ОСНОВЕ ПОРОЖДАЮЩИХ ГРАММАТИК ХОМСКОГО

Бачурина Е.П.

Научный руководитель канд. техн. наук Личаргин Д.В.

Сибирский федеральный университет

Проблемной областью исследования является проблема повышения эффективности языковой подготовки в техническом вузе, разработка средств генерации учебных материалов в целях индивидуализации траекторий обучения. Разработка электронных обучающих ресурсов требует от преподавателя увеличения затрат времени на создание учебного материала, что затрудняет реализацию индивидуальной траектории обучения студентов, либо вообще определяет её отсутствие. При этом студенты, изучающие иностранные языки, также испытывают трудности с изучением правил чтения текстов на иностранных языках. В процессе создания современных систем электронного обучения требуется интеграция знаний из разнородных источников и решение задач тиражирования знаний.

Поэтому актуальна задача разработки ресурса, который может помочь студентам и лицам, изучающим иностранные языки, овладеть навыками чтения текста на иностранном языке за минимальное время, на основе опоры в виде практической и/или международной транскрипции.

В работе поставлены задачи:

- разработать правила порождающих грамматик для генерации транскрипции важнейших европейских языков;
- разработать функции, реализующие генерацию транскрипции на основе порождающих грамматик;
- разработать визуальный интерфейс на основе данных функций, что позволяет пользователю сгенерировать транскрипцию к любому тексту на важнейших европейских языках.

Программа «Генератор транскрипций» должна представлять программу с ограниченным набором функций (позволяет пользователю сгенерировать транскрипцию к любому тексту на важнейших европейских языках). Далее в неё должны быть добавлены функции перевода и пояснения значений иностранных слов по методу Ильи Франка, а также добавлены функции настройки, представляя её в форме генератора учебных материалов по иностранному языку.

Метод Ильи Франка дает возможность лицам, изучающим иностранный язык, читать на нем книги без словаря, быстро привыкая к системе языка и набирая лексический запас (при чтении, например, 2 часа в день – около 1000 слов в месяц). При ежедневном чтении таким методом можно в течение года научиться читать свободно. Запоминание слов и речевых оборотов происходит при этом не за счет заучивания, а за счет постоянной повторяемости слов в тексте.

В России выпущено около 200 книг иностранных авторов художественной литературы по методу Ильи Франка (переводы с 33 языков). Текст разбит на небольшие отрывки. Сначала идет адаптированный отрывок – текст с вкрапленным в него дословным русским переводом и небольшим лексико-грамматическим комментарием. Затем следует тот же текст, но уже неадаптированный, без подсказок.

Например: *I must have slept for more than nine hours* (я, вероятно, проспал более девяти часов; to sleep) *because when I woke up* (потому что, когда я

проснулся; to wake up) *it was daylight* (было совсем светло; *daylight* – дневной свет; день, светлое время суток). *I tried to get up* (я попробовал встать), *but I couldn't move* (но не мог двигаться / пошевелиться). *I was lying on my back* (я лежал на спине). *My arms and legs were tightly fastened to the ground on each side* (мои руки и ноги были крепко прикреплены / привязаны с каждой стороны = с обеих сторон к земле). *My long, thick hair was tied down in the same way* (мои длинные и густые волосы были точно так же / таким же образом / способом привязаны; *thick* – толстый; густой, частый; *way* – дорога; путь; метод, способ).

Метод Ильи Франка показал свою эффективность в изучении иностранного языка, однако существенным недостатком этого метода является высокая трудозатратность создания учебных материалов вручную. Этот недостаток может быть устранен предлагаемой программой, которая будет генерировать данный материал автоматически по заданному входному тексту.

На сегодняшнем этапе реализован первый этап построения генератора материалов, а именно, «Генератор транскрипции». Ниже приведен общий вид программы. Программа выполнена на языке Delphi.

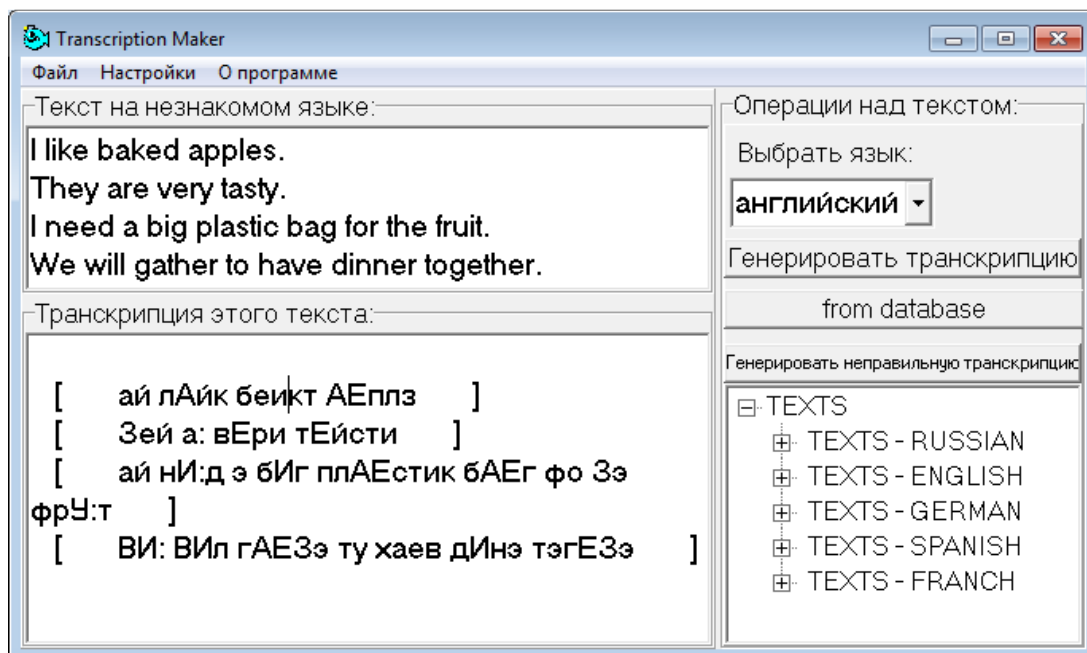


Рис. 1 – Генерация транскрипции английского языка

Реализуемое программное средство «Генератор учебных материалов» должно обладать следующими характеристиками:

- Принципиальная возможность составления материала с привлечением транскрипции;
- Простота работы по составлению материала;
- Возможность использовать практическую транскрипцию тем, кто не знаком с международной транскрипцией, детям, пожилым людям, неуспевающим ученикам.

При разработке программного средства преследуются следующие критерии качества:

1. Корректность. Минимальное количество ошибок в порождаемой транскрипции с точки зрения правил фонетики языка – более 97-99% корректных транскрипций, необходимо разработать модуль обработки исключений для некоторых языков.

2. Надежность. Не происходит обрушения программы, выдачи некорректных данных или ухода в бесконечный «мертвый» цикл, или бесконечный перебор результатов при любых исходных данных.

3. Эргономичность – удобство для пользователя, программа предельно проста и интуитивно понятна, что показало тестирование с привлечением нескольких участников.

4. Переносимость. Программа должна работать во всех разумно современных версиях Windows.

5. Расширяемость. Алгоритм порождающей грамматики и базы данных правил порождающей грамматики должны быть отделены от конкретных данных, должна быть возможность добавлять правила для различных языков.

6. Устойчивость. Алгоритм не должен давать сбоев. В случае неверных символов или слишком большого текста, лишняя информация просто игнорируется. Защита от бесконечного цикла в случае случайного употребления правил порождающей грамматики вида: $A \rightarrow A$. Даже употребление мертвого цикла в наборе правил вида: $A \rightarrow B, B \rightarrow C, C \rightarrow A$ не приведет к сбою в программе. Алгоритм универсален и легко обрабатывает любую строку символов на основе произвольной базы данных правил подходящего синтаксиса.

7. Повторное использование кода возможно в рамках выпуска дальнейших релизов программы и других программ по компьютерной лингвистике и другим областям (например, криптографии, стеганографии).

8. Универсальность. Возможность добавлять не только код html, но и xml под любую необходимую реализацию с применением тех же алгоритмов порождающих грамматик Хомского в отношении набора тегов.

База данных транскрипции слов насчитывает на сегодня более 10 000 слов английского языка из числа наиболее частотных слов языка.

Ниже приведена блок-схема алгоритма генерации транскрипции (рис. 2).

Отметим, что алгоритм генерации транскрипции для испанского, немецкого и французского языка более прост и может быть основан только на правилах чтения. Приведем пример транскрибирования текста с английского языка на основе рассмотренного выше алгоритма.

«Real-time and embedded systems operate in constrained environments in which computer memory and processing power are limited. They often need to provide their services within strict time deadlines to their users and to the surrounding world. It is these memory, speed and timing constraints that dictate the use of real-time operating systems in embedded software. The kernel is the part of an operating system that provides the most basic services to application software running on a processor».

[рИэл тАйм энд имбедид сИстемз оперЕйт ин констрейнд инвАйрэнментс | ин Вич кэмпйУ:тэ мемори энд прОусэсиН пАуэ а: лимитид || Зей Офн нИ:д ту прэвАйд ЗЕэ сЭ:висиз ВизИн стрИкт тАйм дедлайнз ту ЗЕэ йУ:зэз || энд ту Зэ сэрАундиН ВЭ:лд || ит из ЗИ:з мемори спИ:д энд тАймиН кэнстрЕйтс || Зает диктеит Зэ йУ:з ов рИэл тАйм оперЕйтиН сИстемз ин имБЕдид сОфтВеэ || Зэ кЭ:нел из Зэ пА:т ов эн оперЕйтиН сИстем | Зает прэвАйдз Зэ мОуст бЕисик сЭ:висиз ту эпликЕйшн сОфтВеэ | рАниН он э прОусесэ].



Рис. 2. Блок-схема алгоритма генерации транскрипции

При помощи специального макроса программы Microsoft Word мы можем преобразовать полученный текст практической транскрипции в нотацию международной транскрипции, как показано ниже.

[r'iəl t'aim ənd ɪmbedɪd s'istemz oper'eit ɪn konstr'eɪnd ɪnv'airənmənts | ɪn w'ɪtʃ kəm'pju:tə memori ənd pr'ousəsɪŋ pauə ə: 'lɪmɪtɪd || ðeɪ 'ɒfn n'i:d tu prəv'aɪd ð'eə s'ə:vɪsɪz wɪz'ɪn str'ɪkt t'aim d'edlɑɪnz tu ð'eə ju:zəz || ənd tu ðə sɜr'aʊndɪŋ w'ə:ld || ɪt ɪz ð'i:z m'eməri sp'i:d ənd t'aimɪŋ kənstr'eɪnts || ðæt dɪkt'eɪt ðə ju:z ov r'iəl t'aim oper'eɪtɪŋ s'istemz ɪn ɪmbedɪd s'ɒftweə bəsɪk kə:nəl s'ə:vɪsɪz || ðə k'ə:nəl ɪz ðə p'a:t ov ən oper'eɪtɪŋ s'istem | ðæt prəv'aɪdz ðə mɛust beɪsɪk s'ə:vɪsɪz tu əplɪk'eɪfn s'ɒftweə | r'ʌnɪŋ ɒn ə pr'ousəsə].

Необходимо отметить, что такого рода программа может практически постоянно находиться в процессе дальнейшего усовершенствования, например, в целях расстановки правильного ударения при конверсии. Следующим этапом работы над программой может быть генерация фонетического разбиения и интонационной маркировки анализируемых фраз на английском языке.

Выводы. В работе выполнен анализ проблемы генерации транскрипции. Предложен алгоритм транскрибирования известных из базы данных и не знакомых компьютерной системе слов. Подчеркивается важность продолжения исследований по повышению качества генерируемых по методу Ильи Франка учебных материалов к урокам иностранного языка, в частности, английского.