АВТОМАТИЧЕСКИЙ АНАЛИЗ СТИХА В СИСТЕМЕ STARLNG[1]

AUTOMATED ANALYSIS OF VERSE WITH STARLING SOFTWARE PACKAGE

А.В. Козьмин *(akozm@mail.ru)*

Центр типологии и семиотики фольклора РГГУ

Доклад посвящен использованию интегрированной информационной системы STARLING для автоматического анализа стиха. Описываются программные модули и алгоритмы, реализованные в системе.

Стиховедение требует выполнения огромного объема рутинных операций. Вероятно, именно поэтому в литературоведческой среде оно считается трудной областью, хотя работа именно в этой сфере приносит очень весомые и, главное, хорошо обоснованные результаты.

Эти операции хорошо формализуются, что, кажется, должно было бы привести к созданию программного инструментария для их выполнения. Однако до сегодняшнего дня нет программ, которые могли бы использоваться как «рабочее место стиховеда». Идеальной основой и прототипом для такого инструмента является интегрированная информационная система STARLING [Старостин 1994].

Сначала сформулируем круг наиболее массовых и рутинных задач, стоящих перед стиховедом, приступающим к обработке нового материала: Это прежде всего определение метрики и ритмики. Естественно, ямб от хорея отличить легко, ударения расставляются тоже без труда (если не брать сложные случаи, но и там, как правило, ясны по крайней мере альтернативы). Но если требуется обработать тысячи или десятки тысяч строк, задача становится весьма утомительной. То же самое относится и к анализу рифмы. Более маргинальные проблемы связаны с анализом фонетических эффектов, что также требует большой рутинной работы.

И автоматическое описание метрики и ритмики, и анализ фонетических эффектов требует использования прежде всего средств морфологического анализа, реализованных в системе STARLING [Крылов, Старостин 2003]. Результаты работы анализатора в удобном для пользователя виде также целесообразно представлять как базу данных в формате, поддерживаемым системой, поскольку она предоставляет богатые возможности для анализа уже полученных данных

Компьютерные инструменты являются средством решения массовых задач. Поэтому их разработка должна вестись по следующему принципу: от легко формализуемого и массового к трудно формализуемым исключениям. Первые версии программ должны обрабатывать наиболее массовые случаи. Поэтому все дальнейшие определения и, соответственно, алгоритмы принципиально ориентированы на массовый материал, хорошо разработанный в научной традиции.

Более того, предполагается, что определения метров являются в значительной степени вопросом соглашения. Например, считать ли ямбом строки, такие, что все ударения неодносложных слов приходятся на четные слоги, за исключением первой стопы (ямб с переакцентуацией)? Думается, что это чисто терминологическая, а не содержательная проблема.

Разработка должна вестись таким образом, чтобы алгоритмы легко могли быть улучшены. Поэтому работа программ разбивается на этапы, каждый из которых должен улучшать результаты, полученные на предыдущем шаге.

Основной единицей анализа в текущей версии программ является строка. Существующие определения метров принципиально требуют обращения к метрическому контексту, так что отдельная строка может быть интерпретирована как относящаяся к различным метрам одновременно. Снятие такой омонимии требует учета последовательности строк, к которой принадлежит данная. В данной версии такой анализ не производится,

[1] Работа поддержана Российским фондом фундаментальных исследований, грант 05-06-80236-а. Она является частью проекта «Автоматизированный лингвостиховедческий анализ русских поэтических текстов», которым руководил Сергей Анатольевич Старостин. С ним обсуждались некоторые решения, изложенные в докладе, ему же принадлежит часть кода. Разумеется, за все ошибки, неясности и неточности целиком отвечает автор доклада.

однако альтернативные варианты анализа сохраняются, так что в дальнейшем возможно применение алгоритмов, работающих с последовательностью строк.

Однако определенным образом метрический контекст все же учитывается. Алгоритм определения метров работает так: на вход подается строка, которая с самого начала считается ямбической до тех пор пока не обнаружатся противоречащие этому характеристики. Если это происходит, предполагается, что строка хореическая и т. д. Известно, что наибольшее число стихотворных строк русской поэзии (точнее, литературной поэзии Нового времени) ямбические. Поскольку это так, то произвольная строка, поступившая на вход программы, с наибольшей вероятностью относится к ямбу, с меньшей - к хорею, с еще меньшей – к одному из трехсложных метров и т. д. Этот порядок вероятностей и определяет последовательность работы алгоритма.

Формальное определение силлабо-тонических метров

Введем более или менее формальные определения (они неоригинальны и в конечном счете восходят к А.Н. Колмогорову). Метром называется последовательность слабых и сильных позиций (мест). Сильные позиции называются иктами. Если слабое место обозначить 0, а икт - 1, то метры бывают следующие:

010101 ... - ямб. Если пронумеровать позиции, то икты в ямбе будут иметь четные номера

101010 ... - хорей. Икты с нечетными номерами.

100100100 ... - дактиль.

010010010 ... амфибрахий.

001001001 ... анапест.

Если ставить в соответствие позициям слоги реальной строки, то срока определяется как определенный метр по Правилу 1:

Правило 1. Ударения МОГУТ падать только на икты соответствующей метрической схемы, если только эти ударения не подпадают под Исключение 1.

Исключение 1. Ударения могут приходиться на слабое место, если безударный слог ТОГО ЖЕ слова не попадает на икт.

Менее формально это можно описать следующим образом. "Евгений Онегин" - четырехстопный ямб. Но далеко не в каждой строке есть 4 ударения. "Когда не в шутку занемог" - все ударения приходятся на четные слоги, но не все четные слоги несут ударения. Об этом говорит правило 1.

На слабых местах могут стоять ударения, подчиняясь Исключению 1. "Мой" в строке "Мой дядя самых честных правил" несет ударение (если хочется, можно заменить его на слово «Я», которое не попадает даже в этом случае своим безударным слогом на сильную позицию - у него просто нет безударного слога. Поэтому в ямбе и хорее односложные слова могут стоять В ЛЮБОЙ ПОЗИЦИИ.В трехсложных размерах (дактиле, амфибрахии и анапесте), на слабых местах могут оказаться ударения двусложных слов, если безударные слоги этих слов не попадают на икты.

Основной алгоритм поэтому базируется на определении ударений, которые НАРУШАЮТ схему метра. Если нарушений нет, значит, это и есть ямб, хорей и т. д.Наконец, если строка не подпадает под определение одного из силлабо-тонических размеров, однако расстояние между ударными слогами любого сорта не превышает двух безударных слогов, назовем это дольником.

Следует помнить, что формальные определения устроены таким образом, что одна и та же строка может получать несколько интерпретаций. Например, строку «Сердце, тронутое холодком» можно считать хореем или анапестом (во втором случае первое ударение будет сверхсхемным). Это не недостаток формального определения, а отражение особенностей сложившейся терминологической системы в области описания метрики.

Реализованные программные модули

В настоящее время созданы следующие программные модули:

- модуль определения метра и размера;
- подмодуль определения рифмовки;
- модуль преобразования орфографической записи в фонетическую транскрипцию;
- модуль выявления анаграмматических эффектов.

Модуль определения метра и ритма

Шаги алгоритма:

1. разбить текст на строки;

2. для каждой строки выдать ее акцентуированную версию. Если одно или больше слов этой строки может быть проакцентуировано двумя способами, выдать две (четыре, восемь и т.д.) копии строки с вариантами акцентуации.

3. преобразовать строку в последовательность, состоящую из символов, обозначающие: **(1)** безударные

слоги, **(2)** ударные слоги односложных слов, **(3)** ударные слоги, занимающие первую позицию в двусложном слове, **(4)** ударные слоги, занимающие вторую позицию в двусложном слове, **(5)** ударные слоги слов, которые длиннее двух слогов.

4. определить, принадлежит ли последовательность тому или иному метру. Шаги алгоритма следующие:

(1) проверить: на нечетных позициях есть только символы **1** или **2.** Если да - это ямб. Перейти к шагу 5. Иначе

(2) проверить: на четных позициях есть только символы **1** или **2.** Если да – это хорей. Перейти к шагу 5. Иначе

(3) проверить: на позиции номер 2, 5, 8… присутствуют только символы **1, 2** или **3**, на позиции номер 3, 6, 9 присутствуют только символы **1, 2,** или **4.** Если да – это дактиль. Иначе

(4) аналогично для анапеста

(5) аналогично для амфибрахия. Иначе

(6) проверить, есть ли последовательности вида 111. Если нет – это дольник.

5. Для последовательности проверить наличие ударных слогов на позициях, определяемых размером. Указать номера иктов, несущих ударения.

В результате работы модуля на выходе получается таблица со следующими полями:

1. Исходная строка: «Мой дядя самых честных правил»

2. Проакцентуированная строка. Если вариантов акцентуации несколько, они выводятся все:

«Мо'й дя'дя са'мых честны'х прави'л»
«Мо'й дя'дя са'мых честны'х пра'вил»
«Мо'й дя'дя са'мых че'стных прави'л»
«Мо'й дя'дя са'мых че'стных пра'вил»

3. Последовательность, представляющая цепочку слогов, которая поступает на вход при определении метра и ритма (слоги заменены номерами соответствующих типов). Для акцентуированного варианта «Мо'й дя'дя са'мых че'стных пра'вил» - 231313131 (расшифровку символов см. выше).

4. Указание метра. Для «Мо'й дя'дя са'мых че'стных пра'вил» - ямб. Действительно, на нечетных местах последовательности 231313131 находятся символы 2,1,1,1,1. Ср. выше определение ямба: «на нечетных позициях есть только символы **1** или **2».**

5. Указание числа стоп. Для «Мо'й дя'дя са'мых че'стных пра'вил» - 4.

6. Последовательность номеров иктов, несущих ударения.. Для «Мо'й дя'дя са'мых че'стных пра'вил» - все икты несут ударения.

Предусмотрен режим работы с предуказанным метром (например, при анализе «Евгения Онегина» можно заранее указать, что это ямб)[2]. В результате сильно сокращается число избыточных вариантов и результаты анализа более «чистые».

Основные проблемы, связанные с разработкой модуля

Основная проблема, связанная с работой этого модуля – отбрасывание неверных вариантов акцентуации. Некоторые неверные результаты могут отбрасываться при помощи синтаксического компонента STARLING'а. Например, морфологический анализатор интерпретирует графическое слова «такая» в том числе как деепричастие от редкого глагола «такать» и выдает акцентуированную форму «та'кая». Синтаксический анализатор может отбрасывать такие варианты как не поддающиеся правильной синтаксической интерпретации. В настоящее время, однако, эта возможность пока не реализована. Собственно стиховедческий путь снятия неоднозначности заключается в использовании метрического контекста. Например, если большая часть строк произведения – ямбические, то разумно в случаях нескольких акцентуированных вариантов выбирать такой, который укладывается в ямб.

Вторая проблема связана с морфологией. Модуль работает с морфологическим анализом, использующим словарь Зализняка. Соответственно, не учитываются архаичные акцентуационные и морфологические нормы, отсутствует обработка имен собственных. Необходима доработка словаря Зализняка для корректной работы модуля.

Подмодуль определения рифмовки

Рифма рассматривается как пара словоформ, имеющих в своем составе схожие символы в определенных позициях. При этом схожесть символов и позиции, в которых они должны находится, определяется по соглашению. В текущей версии для признания пары рифмой требуется совпадение ударной гласной, заударного согласного, равносложность.

Шаги алгоритма:

1. составить массив конечных слов для всех строк
2. для каждой пары слов из массива проверить и записать как компонент вектора совпадение (1) ударного гласного, (2) заударного согласного, (3) позицию ударного гласного относительно конца слова.
3. Указать в качестве рифмующихся пар такие пары, у которых векторы имеют нужную последовательность компонент.

Основные проблемы, связанные с разработкой модуля

Необходимо разработать правила выявления рифм в зависимости от эпохи создания анализируемого текста. В текущей версии не учитываются составные рифмы.

Модуль преобразования орфографической записи в фонетическую транскрипцию

Модуль работает с базой правил, имеющих вид A->B, где A и B есть последовательности символов, а стрелка означает «заменить на» в строке. Список правил упорядочен. Предварительно для безударных гласных указывается (с использованием процедуры акцентуации) номер предударного или заударного слога.

Основные проблемы, связанные с разработкой модуля

Необходимо внести дифференцированные в зависимости от времени создания текста правила перевода орфографической записи в фонетическую (учет различных произносительных норм). В текущей версии фактически не учитывается собственно морфологическая информация, что приводит к ошибкам

Модуль выявления анаграмматических эффектов

Модуль работает с базой, в которой записаны множества символов (в том числе одноэлементные). Выбирается два файла – один тестируемый, другой «фоновый». Для обоих файлов вычисляется процентное отношение символов каждого множества к объему текста. Таким образом, можно видеть отклонения частот в тестируемом файле от фонового. Если в качестве фонового использовать стилистически нейтральный текст большого объема, то полученные частоты могут рассматриваться как приближенные к «языковым» частотам. Тексты, подаваемые на вход, могут быть как в орфографической, так и в фонетической записи, что позволяет исследовать эффекты обоих уровней – графического и фонетического.Использование многоэлементных множеств служит для того, чтобы выявить нетривиальные фонетические эффекты. Например, в качестве такого множества может быть использован набор шипящих согласных или узких гласных, или заднеязычных согласных и т. д.

Основные проблемы, связанные с разработкой модуля

Возможно, следует встроить стандартные статистические процедуры, верифицирующие надежность учета выявляемых отклонений частот.

Список литературы

Крылов С.А., Старостин С.А. *Актуальные задачи морфологического анализа и синтеза в интегрированной информационной среде STARLING // Международная конференция "Диалог": Компьютерная лингвистика и интеллектуальные технологии. Архив. 2003 (http://www.dialog- 21.ru/Archive/2003/Krylov.htm)*

Старостин С.А. *Рабочая среда для лингвиста // Гуманитарные науки и новые информационные технологии. М.: 1994. Вып.2. С.7-22.*

[2] Этой идеей я обязан В.А. Плунгяну.