27 января 2011 г., в четверг, в Институте русского языка им. В.В.Виноградова РАН (Волхонка, 18) состоится первое в текущем году заседание Центра текстологии и стиховедения. Начало – в 17:30 в Малом конференц-зале – комнате 22 на 2-м этаже. После заседания (примерной общей продолжительностью в два – два с половной часа) мы приглашаем всех гостей Центра на чаепитие.

Анатолий Сергеевич Старостин прочтёт доклад «О компьютерной среде идентификации параметров стихотворного текста», затем в ходе обсуждения доклада аудитория может предлагать докладчику конкретные стихотворные тексты для их обработки в разработанной докладчиком компьютерной среде TREETON (нацеленной на осуществление компьютерных задач в области обработки естественного языка). Вы можете подготовить предварительно тексты для такого тестирования стиховедческой подсреды в составе среды TREETON, для чего требуется предоставить их в компьютерном формате – желательно (но не обязательно) с расширением .doc или .txt (в едином файле для всех текстов или в директории с файлами, отведёнными каждый для одного текста; каждый текст желательно предварить «шапкой» – информацией о нём, включающей по минимому имя автора и название произведения). Такие подготовленные тексты можно принести на флэшке и предоставить докладчику перед заседанием или непосредственно в ходе обсуждения; можно также послать их докладчику на следующий электронно-почтовый адрес Anatoliy_St@abbyy.com.

Мы будем рады видеть Вас среди аудитории намечаемого заседания и предлагаем Вам принять участие в обсуждении доклада и в тестировании среды TREETON.

Аннотация доклада

В докладе представлены результаты нового этапа работ по автоматизации лингвостиховедческого анализа поэтических текстов. Описывается программный модуль определения метра стихотворного текста, являющийся частью системы Treeton — программно-информационной среды для решения различных проблем в области обработки естественного языка (natural language processing).

Описываемый программный модуль получает на вход стихотворный текст на русском языке, снабжает его информацией о возможных вариантах расстановки ударений (в соответствии с данными в грамматическом словаре А.А. Зализняка) и определяет один или несколько стихотворных метров, к которым может относиться данный текст.

Программа работает с описанием поддерживаемых стихотворных метров на специально разработанном формальном языке, которому будет посвящена первая половина доклада.

Это описание представляет собой набор метрических схем – шаблонов, организованных в группы. Каждому из распознаваемых системой метров соответствует один шаблон. Шаблон представляет собой последовательность обозначений слогов, часть из которых взята в круглые скобки, после которых стоит знак ‘?’ (фрагмент может отсутствовать) или ‘*’ (фрагмент может повторяться 0 или более раз). Для удобства изложения обозначения слогов в метрической схеме называются «местами». Выделяется четыре типа мест:

1) метрически ударные (‘—’);
2) метрически безударные (‘∪’);

3) обязательно ударные (‘—;′’);

4) обязательно безударные (‘∩’).

Обязательно ударные места используются для учета константной ударности последнего икта, а обязательно безударные — для учета специфики клаузул и анакрус.

В программе реализована процедура множественного сопоставления акцентуаций с шаблонами. Это означает, что любая акцентуация может сопоставиться с несколькими шаблонами. При этом:

а) метрически ударные и метрически безударные места могут сопоставляться как со слогами, несущими языковое ударение, так и с безударными слогами;

б) обязательно ударные места могут сопоставляться только со слогами, несущими языковое ударение;

в) обязательно безударные места могут сопоставляться только со слогами, не несущими языкового ударения

Содержащиеся в этих пунктах ограничения могут нарушаться в системе, но в этих случаях вариант сопоставления существенно «штрафуется» — оценивается как «плохой»

Благодаря тому, что в шаблонах предусмотрена опциональность, то есть некоторые места во время сопоставления можно пропускать, во многих случаях акцентуацию оказывается возможным сопоставить с одним и тем же шаблоном несколькими способами. На оценку конкретного варианта сопоставления акцентной формы с шаблоном влияет два фактора: количество нарушений запрета на переакцентуацию и количество нарушений обязательной ударности и безударности.

Оценка того, насколько данная акцентуация соответствует данному шаблону, вычисляется как наилучшая по всем возможным вариантам сопоставления. Общая оценка того, насколько данная строка соответствует данному шаблону, вычисляется путем усреднения оценок всех возможных акцентуаций строки.

После того, как про каждую строку становится известно, насколько хорошо она укладывается в тот или иной метр (то есть какова усредненная оценка ее сопоставлений с соответствующим шаблоном), программе остается вычленить на множестве всех строк текста те метры, которые имеют «хорошие» показатели – их процент должен быть выше некой экспериментально определяемой константы (наряду с шаблонами такие константы считаются частью формального описания). Если таких метров оказывается несколько, предпочтение отдается тому, у которого меньше вариативность (v) объема междуиктовых интервалов (i). Последнее правило есть по сути заложенное в программу представление об *иерархии* метров. На сегодняшнйи день эта иерархия такова:

I. *Классическая силлаботоника* ($v = 0$):
Ямб
Хорей
Дактиль
Амфибрахий
Анапест
II. *Дольники* ($v = 1$):
На трехсложной основе ($1 \le i \le 2$)
На двусложной основе ($0 \le i \le 1$)
III. *Тактовики* ($v = 2$)
Тип 1-2-3 ($1 \le i \le 3$)
Тип 0-1-2 ($0 \le i \le 2$)

Могут возникать ситуации, при которых один и тот же текст «хорошо» укладывается в несколько метров внутри одной группы. В таких случаях иногда удается выбрать правильный метр за счет анализа таких показателей, как средний уровень пропуска схемных ударений и средний уровень сверхсхемных ударений. Обсуждению этого вопроса посвящена финальная часть доклада.