**С. Я. Вишневский**
*(ГрГУ им. Я. Купалы, Гродно)*

**АННОТИРОВАНИЕ ЕСТЕСТВЕННО-ЯЗЫКОВОГО ТЕКСТА**

Одним из эффективных средств для быстрой обработки больших объемов текстовой информации является аннотирование текстов или составление их рефератов. Таким образом, пользователь может проанализировать краткое содержание текста и принять решение о необходимости дальнеешего его изучения.

Для решения данной задачи используются два основных метода: статистический и лингвистический. Статистический метод основан на присвоении предложениям весов и построении аннотации из предложений с наибольшим весом. На вес влияют частота употребления слов, длина предложения, позиция в тексте, употреблении дат, мест, имен. В лингвистическом методе выявляются зависимости между частями текста и предложений, выбираются наиболее значимые части и строится связный реферат в отличие от статистического, где результатом является набор элементов текста.

Была поставлена задача построения аннотации текста и анализ особенностей полученных аннотаций в зависимости от стилистики исходного текста.

В качестве решения был выбран подход включающий в себя оба классических метода (статистический, лингвистический).

Обработка исходного текста включает следующие этапы: построение синтаксического дерева, определение морфо-характеристик от слов, удаление шумов (исключение из дерева предлогов, местоимений, частиц, союзов), статистический анализ дерева, визуализация.

Синтаксическое дерево строится разбиением текста на параграфы, предложения, слова. Для определения морфо-характеристик используются продукты в открытом доступе команды разработчиков проекта «Диалинг»(www.aot.ru). Шумы удаляются при обходе построенного дерева, затем используя основные характеристики статистического метода выбираются наиболее весомые части. Полученная аннотация сохраняется в XML файл и визуализируется в браузере при помощи XSLT трансформаций.

При анализе результатов замечено, что наиболее полные аннотации получаются при анализе художественных текстов. В аннотациях по публицистическим и научным текстам выявлено смешивание смысловых частей из-за большого количества сложно-зависимых предложений. Для рещения данной проблемы необходимо разработать алгоритм анализа таких предложений, который бы выделял независимые части таких предложений до обработки статистическим алгоритмом.

Средством реализации выбран язык программирования C++, который позволяет разрабатывать эфективные, кроссплатформенные системы работающие на вычислительном кластере.