**Информационная технология распознавания**
**печатных источников второй половины XVIII – начала XIX вв.**

Анна Юрьевна Филиппович,
к.т.н., доцент кафедры Медиасистем и технологий МГУП им. Ивана Федорова

Для решения проблемы сохранения письменного культурно-исторического наследия России предлагается информационная технология, которая основана на выделении группы исторических источников с определенными свойствами (в числе которых способ печати, используемые средства оформления, шрифтовые гарнитуры и т.п.) и обработке их системами распознавания текстов.

Материалами для исследований выступают печатные источники вт. пол. XVIII – нач. XIX вв. В результате их исследования, были описаны их палеографические и лексические характеристики, сформированы шрифтовые эталоны, построены квантитативные модели текстов, получены параметры функций распределения частот слов.

Исследование эффективности распознавания текстов рассматриваемых источников показало, что большинство ошибок было связано с особенностями графем шрифтов и старинной лексикой. Использование шрифтовых эталонов, функции распознавания с обучением увеличивало точность распознавания, однако для более эффективной работы систем распознавания требуется учет лексического состава текста.

Для улучшения качества распознавания текстов необходимо сформировать в электронной форме лексическое ядро языка коллекции документов – основу проверочного словаря-спеллера. В качестве основы ядра используется расширенный словник Словаря Академии Российской 1789-1794 гг. (САР), в 2001-2005 гг. переизданного с использованием современных информационных технологий. Отсутствие четких правил написания вариантов делает невозможным сформировать расширенный словник заголовочных слов САР автоматически, поэтому была создана автоматизированная методика формирования расширенного словника.

В настоящее данные исследования поддерживаются грантом Президента РФ МК-3732.2010.9 «Разработка словарных компонентов интегрированной информационной технологии переиздания печатных источников XVIII – нач. XIX вв.».

Результаты исследований могут быть использованы для переиздания значительного массива конкретных источников XVIII – нач. XIX вв., разработки систем распознавания исторических текстов, решения практических задач электронного и полиграфического издания древних памятников.