МИНИСТЕРСТВО НА ОБРАЗОВАНИЕТО И НАУКАТА НА РЕПУБЛИКА БЪЛГАРИЯ
КИРИЛО-МЕТОДИЕВСКИ НАУЧЕН ЦЕНТЪР ПРИ БЪЛГАРСКА АКАДЕМИЯ НА НАУКИТЕ
ИЖЕВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ИМ. М. Т. КАЛАШНИКОВА
НАУЧНОЕ СООБЩЕСТВО “ПИСЬМЕННОЕ НАСЛЕДИЕ”
DIGITAL MEDIEVALIST SCHOLARLY COMMUNITY
ФОНДАЦИЯ „УСТОЙЧИВО РАЗВИТИЕ НА БЪЛГАРИЯ “

Писменото наследство и информационните технологии

El’Manuscript–2014

Материали от V международна научна конференция
Варна, 15–20 септември 2014 г.

София · Ижевск
2014

Сборникът е издаден с финансовата подкрепа на Министерството на образованието и науката на Република България по процедура за подкрепа на международни научни форуми.

Отговорни редактори: проф. дфн В. А. Баранов
доц. д-р В. Желязкова
д-р А. М. Лаврентьев

Редактори: Нели Ганчева, Веселка Желязкова (български текст)
О. В. Зуга, В. А. Баранов (руски текст)
Кевин Хокинс (Kevin Hawkins) (английски текст)

Писменото наследство и информационните технологии [Текст] : материали от V международна науч. конф. (Варна, 15–20 септември 2014 г.) / отг. ред. В. А. Баранов, В. Желязкова, А. М. Лаврентьев. — София ; Ижевск, 2014. — 448 с.

Сборникът съдържа материали от конференция, посветена на разработването и създаването на съвременни средства за съхраняване, описване, обработка, анализ и публикуване на ръкописни и старопечатни книжовни паметници и исторически извори, а също и на въпросите за подготвянето на електронни ресурси в областта на хуманитаристиката и тяхното използване в научните изследвания и преподаването.

ISBN 978–954–9787–25–2

Компьютерный фонд древнерусской скорописи

Ю. Н. Филиппович, А. Ю. Филиппович, И. А. Зеленцов

Системы распознавания рукописных текстов, древнерусская скоропись XVII века, технология распознавания, структурные фреймовые модели

An Electronic Corpus of Old Russian Cursive

Yuriy Philippovich, Anna Philippovich, Ivan Zelentsov

The paper presents the project to create an electronic corpus of Old Russian cursive. A model for and algorithms of Old Russian cursive handwriting recognition are being developed. Key aspects of recognition technology are a structural approach, duplex recognition circuit (letter/word), recognition controlled by hypotheses, vagueness in the description, the use of knowledge and participation of the expert. A mockup of the recognition system is being developed. The main components of the system are a tracer, word and letter recognizers, a knowledge base containing information about letters and words and a training module.

О проекте

Данный проект посвящен созданию электронного фонда древнерусской скорописи и разработке информационной технологии распознавания скорописных древнерусских текстов и документов. Его цель — сохранение культурно-исторического наследия России [Филиппович А. 2012].

Скоропись в русских документах появляется с XIV века и к XVII веку становится основным почерком деловых текстов. Скоропись — это почерк, рассчитанный на существенное ускорение процесса письма. Ее отличают более свободные взмахи, росчерки, большое разнообразие графических вариантов отдельных букв.

Сложность создания электронного фонда древнерусской скорописи связана с ограниченным кругом людей, способных к чтению скорописных текстов, трудоёмкостью ручного перевода рукописей в электронное представление и отсутствием современных средств распознавания, работающих с древнерусской скорописью.

Материалы проекта — скорописные книги XVII в.

1. Книга отводная Карельского села Онежского Крестного монастыря приказчика старца Тихона старцу Иринарху (РГАДА, фонд 1195. Оп. 1, ед. хр. 34, л. 1–12. 3 октября 169 г. = 1660 г.).
2. Книги отводные Онежского Крестного монастыря казначея старца Иринарха Каменева новому казначею старцу Игнатию (РГАДА, фонд 1195. Оп. 1, ед. хр. 86, л. 1–26. 1 апреля 173 г. = 1665 г.)

3. Книга записная властелинским указам Онежского Крестного монастыря (РГАДА, фонд 1195. Оп. 1, ед. хр. 387, л. 1–12 об. 1 января 195 г. = 1687 г.).
4. Отводная книга Белого двора Онежского Крестного монастыря (РГАДА, фонд 1195. Оп. 1, ед. хр. 412, л. 1–16 об. 6 апреля 196 г. = 1688 г.).

Основные задачи проекта:

1) разработка концепции и архитектуры компьютерного фонда древнерусской скорописи;
2) разработка информационной технологии автоматизированного перевода (распознавания) скорописных древнерусских текстов и документов из растровых изображений в электронную форму;
3) выявление особенностей древнерусской скорописи;
4) исследование и разработка методов, алгоритмов, процесса распознавания; проектирование и реализация информационной технологии;
5) создание автоматизированной системы распознавания древнерусской скорописи (АСРДС).

В рамках данного проекта переведены в электронный вид более 300 скорописных листов. Создана графическая база данных, включающая изображения слов, отдельных букв и страниц текста [Графический словарь-справочник].

Технология распознавания

Для решения задачи распознавания скорописи в проекте используется структурный метод. В его основе лежит представление об изображении буквы как о наборе структурных элементов, расположенных друг относительно друга определённым способом.

В процессе распознавания программа использует базу знаний о начертаниях букв. Она формируется экспертом на этапе обучения программы. Обучение системы заключается в наполнении базы знаний системы сведениями о символах алфавита, встречающихся в рукописях, и способах их начертания. База состоит из двух связанных частей: базы фреймов букв и словника.

Фреймы букв отражают знания эксперта об элементах, формирующих изображение символов, и их пространственных взаимоотношениях. Словник состоит из слов, составляющих лексикон программы. Слова представлены в виде фреймовых структур, ссылающихся на фреймы букв из соответствующей части базы знаний и отражающих порядок следования букв в каждом конкретном слове. Заполнение словарной части базы знаний осуществляется автоматическим преобразованием текстового словника рукописей.

В рабочем режиме перед обученной программой стоит задача оффлайн-распознавания. В общем виде задача решается следующим образом. Графический анализатор пытается выделить в изображении элементы букв всех типов, встречающихся в базе знаний. При нахождении очередного элемента определяются его отношения к уже найденным элементам. Эта информация передаётся

вышестоящему распознавателю букв. Последний производит поиск в базе знаний вхождений элементов найденных типов со схожими взаимоотношениями в одну из букв-эталонов. Как только подходящее вхождение обнаружено, распознаватель выдвигает гипотезу о наблюдаемой в текущей точке изображения букве. Далее он получает из буквы-эталона информацию об элементах, которые должны присутствовать в букве (если гипотеза верна), но ещё не обнаружены сканером. Эта информация служит для ориентирования сканера на поиск элемента определённого типа в определённой области изображения относительно одного из найденных элементов [Зеленцов 2010].

Для облегчения процесса распознавания в программу также входит распознаватель слов, опирающийся на словник. Он служит источником гипотез для распознавателя букв, помогая последнему скорее находить правильные гипотезы. Подробнее описание технологии распознавания представлено в [Зеленцов, Филиппович Ю 2011а; б; в]

В результате реализации проекта разработана теоретическая модель и алгоритмы распознавания скорописных текстов, которые позволят в дальнейшем ввести значительную часть скорописных рукописей со схожими графическими характеристиками и качеством.

Литература

Графический словарь-справочник — *Графический словарь-справочник по скорописи* [Электронное издание]. Режим доступа: http://it-claim.ru/Projects/Skoropis/SkoropisBooks.htm, свободный.

Зеленцов 2010 — *Зеленцов И. А.* Выдвижение и проверка гипотез в системе распознавания древнерусской скорописи // Информационные технологии и письменное наследие: материалы междунар. науч. конф. Уфа; Ижевск, 2010. С. 99–101.

Зеленцов, Филиппович Ю. 2011а — *Зеленцов И. А., Филиппович Ю. Н.* Распознавание букв и слов древнерусской скорописи XVII в. [Электронный ресурс] // Наука и образование: электронное научно-техническое издание. М., 2011. № 12. Режим доступ: http://technomag.edu.ru/doc/296965.html, свободный (дата обращения: 01.04.2014).

Зеленцов, Филиппович Ю. 2011б — *Зеленцов И. А., Филиппович Ю. Н.* Распознавание образов на основе структурных фреймовых описаний в скорописных текстах XVII в. [Электронный ресурс] // Наука и образование: электронное научно-техническое издание. М., 2011. № 12. Режим доступа: http://technomag.edu.ru/doc/296744.html, свободный (дата обращения: 01.04.2014).

Зеленцов, Филиппович Ю. 2011в — *Филиппович Ю. Н., Зеленцов И. А.* Распознавание скорописи XVII века // Проблемы полиграфии и издательского дела. М., 2011. № 3, С. 87–97.

Филиппович А. 2012 — *Филиппович А. Ю.* Информационные технологии сохранения культурно-исторического наследия России // Международный Форум славянской веб-культуры. Российско-болгарский круглый стол (Пловдив, 26–28 сентября 2011 г.). Москва: МГУП им. Ивана Фёдорова, 2012. С. 38–44.