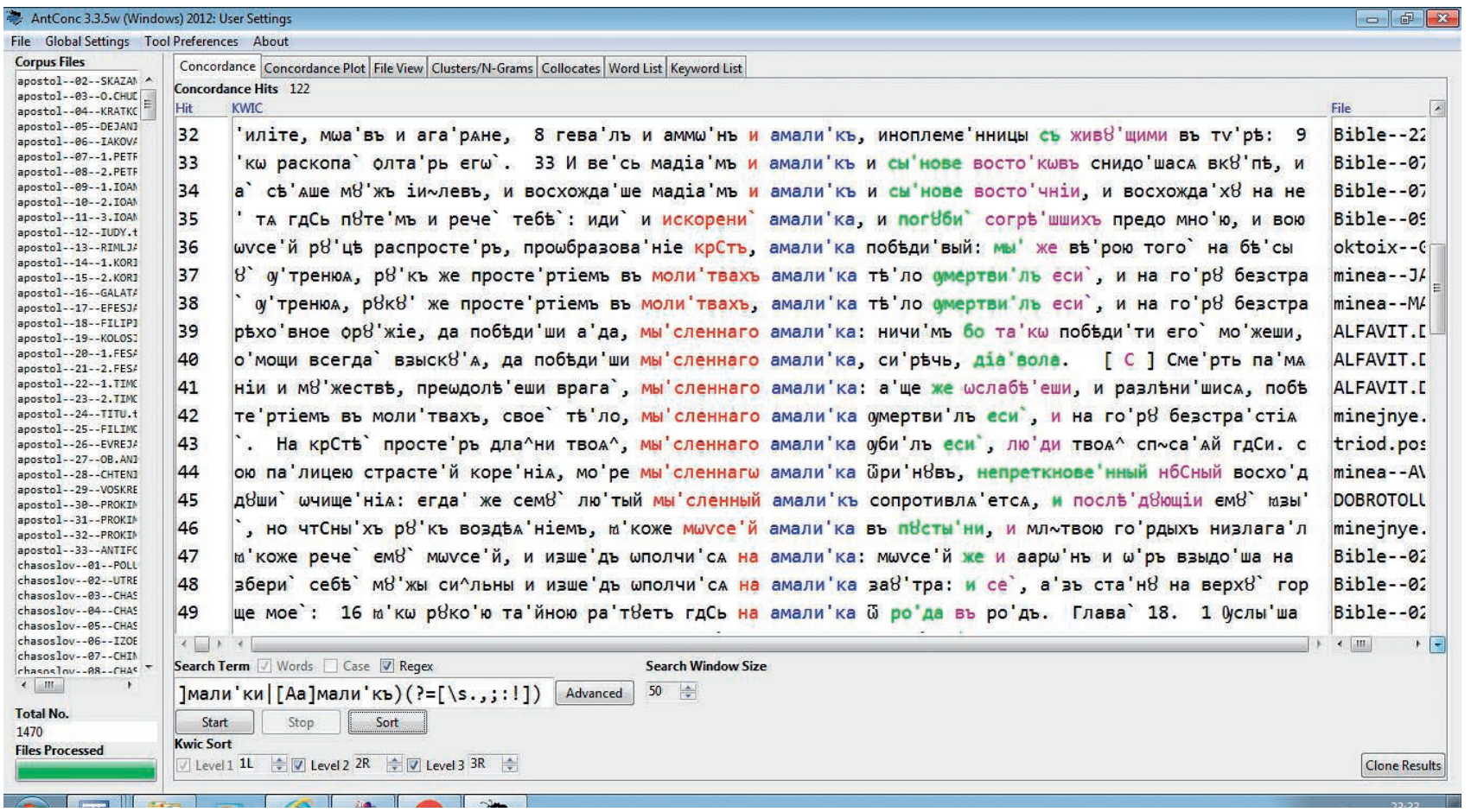
__Современные компьютерные технологии (на экране сортировка вокруг имени Амалик) позволяют просматривать огромные массивы материала и вычленять не только отдельные слова, но и устойчивые сочетания



__Цифровые гуманитарные науки (Digital Humanities) — быстро развивающаяся и перспективная научная дисциплина на пересечении компьютерных и гуманитарных наук. История дисциплины началась сравнительно недавно, в 1949 году, но до сих пор не существует ее стандартного определения.

Сохранение и исследование культурного наследия человечества невозможно без собирания, систематизации и хранения письменных документов. Перевод документов в электронную форму и создание онлайновых баз данных со свободным доступом предоставляют возможность проведения исследований для более широкой научной аудитории и существенно ускоряют работу ученых.

Цель междисциплинарного проекта SlaVaComp (COMPutergestuetzte Untersuchung von VAriabilitaet im KirchenSLAvischen) Федерального министерства образования и научных исследований Германии — создание новых возможностей для исследований географической и хронологической дифференциации церковнославянского языка в X–XVI веках. Исходный набор данных составляют 15 церковнославянско-греческих словарей, соответствующих различным регионам и временным периодам. Проект является одной из первых работ по переводу печатных словарей в электронный формат TEI XML для использования в славистике. Решаемая задача относится к интеллектуальному анализу текстов.

На этапе предварительной обработки с помощью специального программного обеспечения SlaVaComp-Konvertierer кодовая точка каждого символа в уже имеющихся файлах словарей в бинарном формате Microsoft Word (.doc) преобразуется в соответствующий эквивалент в системе Юникод. Затем документы переписываются в текстовые файлы в формате Юникод (.txt) с малым количеством элементов форматирования.

Распознавание образов при анализе текста включает в себя не только определение текстовых элементов, но и идентификацию правильных позиций размеченных элементов в записи в файле TEI XML. Словарные статьи имеют общую структуру, но могут существенно различаться в деталях. Реализация последовательной разметки XML, начиная с первого элемента и далее, может быть очень сложной. Мы предлагаем другой подход, в котором разметка выполняется по блокам в соответствии с «картой словарной статьи». Идея появилась у одного из авторов на основе опыта работы в области вычислительной гидродинамики. Одним из подходов к решению широкого круга задач о течениях жидкости в областях со сложной геометрией является взаимно-однозначное отображение в новые независимые координаты, в которых область имеет более простую форму (например, единичный квадрат). Основные уравнения становятся более сложными, но их конечно-разностные аппроксимации могут быть сделаны проще и с большим порядком точности. Аналогичный подход был нами успешно применен в области цифровых гуманитарных наук.

Тип каждого элемента словарной статьи записывается в массив — «карту». Элементы карты (типы текстовых элементов) определяются с помощью набора критериев, который может быть расширен в случае появления новых типов или специальных случаев. Между массивом с элементами карты и массивом с текстовыми элементами есть взаимно-однозначное соответствие. Мы работаем с картой в «пространстве типов элементов», проводя объединение, разделение, сдвиг и другие необходимые операции.

Элементы «карты нулевого уровня» называются «коробочками нулевого уровня», поскольку текстовые элементы как бы накрываются коробочками с названиями типов. Каждая часть словарной статьи делится на блоки так, чтобы каждый блок содержал или 1) лемму с ее информацией; или 2) грамматическую информацию варианта и первый (и возможно, единственный) графический вариант; или 3) другой (не первый) графический вариант. Карта обрабатывается по блокам, а каждый блок — поэлементно.

Согласно карте нулевого уровня проводится разметка и запись для временного хранения основных элементов (лемма, библиографическая информация и др.). Могут появиться пустые «коробочки нулевого уровня», которые удаляются с карты, и тогда массив с текстовыми элементами корректируется согласно взаимно-однозначному соответствию. Затем «коробочки нулевого уровня» с грамматической информацией накрываются сверху «коробочками первого уровня». Тэги корректируются и составляется «карта первого уровня». Далее «коробочка с леммой» или «коробочка с графическим вариантом» («коробочки второго уровня») накрывают наборы «коробочек» нулевого и первого уровней. Выполняется разметка, создается «карта второго уровня», где каждый блок или начинается с леммы, или содержит графический вариант, обновляется массив с текстом. Далее собираются «коробочки» с графическими вариантами, представляющие варианты и соответствующую грамматическую информацию («коробочки с вариантами» или «коробочки третьего уровня»). Размеченные лемма и вариант(-ы) собираются вместе и записываются в файл XML.

Формы слов и лемм с одинаковыми значениями могут иметь различные графические представления. С помощью лемматизации — присвоения нормализованной формы, называемой гиперлеммой, каждой церковнославянской или греческой словоформе — можно соотнести эти формы с правильной единственной главной формой. Ранее проверка по словарям и исправления могли быть сделаны только лингвистом вручную. В нашей работе после успешного перевода нескольких словарей в формат TEI XML был составлен электронный лексикон, содержащий гиперлеммы и ассоциированные с ними формы слов и позволяющий проводить автоматическую лемматизацию через XML-базу данных. С каждым обработанным словарем новые формы слов добавляются в базу. Лемматизация выполняется с помощью и автоматической проверки форм по электронному лексикону, и ручной проверки лингвистом. Для ручной лемматизации нами было найдено оригинальное решение: при первом запуске программы одновременно с разметкой TEI XML создается специальный текстовый файл для лингвиста, содержащий формы слов для лемматизации, а при последующих запусках предварительно считывается файл с уже сделанными ручными исправлениями и исправленные формы слов отмечаются в файле TEI XML.

Благодаря универсальной модульной структуре разработанная система интеллектуального анализа текста может быть использована не только для изучения церковнославянского языка, но и других языков. Кроме того, в нашей оригинальной системе происходит эффективное взаимодействие человеческих и вычислительных ресурсов.

НИНА ШОКИНА, кандидат физико-математических наук;
Dr. СЮЗАННЕ МОКЕН,
Фрайбургский университет, Фрайбург-им-Брайсгау, Германия

СИСТЕМА ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ТЕКСТА ДЛЯ СОЗДАНИЯ ЭЛЕКТРОННЫХ СЛОВАРЕЙ В ПРИМЕНЕНИИ К ИССЛЕДОВАНИЮ ЦЕРКОВНОСЛАВЯНСКОГО ЯЗЫКА