УДК 80:004

Е. С. Астапкина (astapkina@gmail.com),
канд. филол. наук, преподаватель кафедры теоретического и славянского языкознания

Белорусский государственный университет г. Минск, Республика Беларусь

Т. В. Астапкина (astapkina.tat@gmail.com),
ст. преподаватель кафедры информационно-вычислительных систем

Белорусский торгово-экономический университет потребительской кооперации г. Гомель, Республика Беларусь

ПЕРСПЕКТИВЫ МЕЖДИСЦИПЛИНАРНОГО ПОДХОДА В СФЕРЕ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ (НА ПРИМЕРЕ ИСПОЛЬЗОВАНИЯ ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ NOOJ)

Статья посвящена перспективам использования современного программного обеспечения NooJ в интересах филологического знания. В работе дан обзор потребностей филологии, которые могут быть удовлетворены при использовании информационных технологий. Обращается внимание на необходимость подготовки специалистов в области компьютерной филологии.

This paper considers the advantage of using environment development NooJ in the field of philology. A brief overview the philological requirements, that can be satisfied with the help of computer technologies, is given. The importance of education and training of specialists in the field of "Computer philology" is paid attention to.

Ни одну научную дисциплину, в том числе и гуманитарную, нельзя рассматривать обособленно, изолированно, а только во взаимосвязи со смежными ей отраслями. Это верно и для филологии – происходит ее постепенная компьютеризация и математизация. Так, на стыке информационных технологий и лингвистики возникла компьютерная (вычислительная или математическая) лингвистика.

Без компьютерных информационных технологий невозможно представить ни одну прикладную область языкознания: корпусную лингвистику, машинный перевод, создание электронных словарей и т. д. Компьютерные программы облегчают научно-исследовательскую работу: обработка и анализ текстов, поиск и учет текстовой информации. Современные информационные технологии упрощают оперирование текстами, автоматизируют работу исследователя, что в конечном счете сказывается на ее продуктивности.

На чем основана такая междисциплинарная связь информационных технологий и филологии? С одной стороны, филологию можно рассматривать как «совокупность наук, изучающих культуру народа, выраженную в языке и литературном творчестве» [1], т. е. специфика филологии – в ее ориентировке на понимание текстов. Информационные технологии же, с другой стороны, представляют процессы, связанные с использованием компьютеров и программного обеспечения для создания, хранения, обработки, ограничения к передаче и получению информации. Филология и информационные технологии заинтересованы в кооперации и взаимопомощи.

На сегодняшний день подавляющее большинство возможностей, которые предоставляет современная вычислительная техника и информационные технологии, используются достаточно эффективно в области филологии[1]. Однако для создания специализированных программ, которые разрабатывали бы специальные алгоритмы, необходимо сотрудничество программистов и филологов. Так, например, средства разметки прозаических или стихотворных текстов разрабатываются не только филологами, но и «компьютерщиками». Отметим также и необходимость повышения информационной культуры специалистов-филологов, под которой в данном случае понимаются «знания и навыки эффективного пользования информацией, предполагающие разностороннее умение поиска нужной информации и ее использования, от работы с библиотечным каталогом, компьютерной грамотности до просмотра информации в сети "Интернет"» [3].

Так, растущая роль компьютерных технологий в образовательном процессе, их задействованность практически в каждой отрасли лингвистики обусловливает и соответствующие образовательные инновации. В новый образовательный стандарт Республики Беларусь (ОСРБ 1-21 05 02-2009) было включено направление «Компьютерное обеспечение» и дополнительная квалификация

[1] Например, используются строковые функции при работе с лингвистической картотекой в MS Access [2].

– «Специалист по компьютерной филологии». Среди требований к профессиональным компетенциям отмечены такие, как использование в работе современных компьютерных методов сбора, обработки и хранения информации, работа с современными информационными ресурсами и экспертными компьютерными системами, работа с прикладными системами обработки информации и др. [4]. Такие изменения в системе образования позволяют говорить о новых возможностях для выпускников на рынке труда. Это также служит повышению привлекательности и престижа филологических специальностей. Конечно, новое направление требует доработки и более детального разъяснения как требований к профессиональным компетенциям специалиста по компьютерной филологии, так и этапов, их целей, задач и содержания.

Важным является то, что специальные программы создавались и продолжают создаваться для филологов и активно используются для автоматизации работы исследователя. Так, например, на языке JAVA разработаны пакеты Cogito, LingPipe, LinguaStream и др., на языке C++ – пакеты Distinguo, FreeLing и др. В зависимости от целей исследования такие программы могут быть «простыми, включающими несколько программных надстроек, или комплексными, включающими программные, аппаратные, технологические, организационные и иные компоненты» [5]. В статье «Филология и современные информационные технологии (К постановке проблемы)» авторы дают перечень основных общих, информационных и функциональных требований к информационным системам, которые рассчитаны на филологическую работу с текстами. Обратим внимание на функциональные требования, которые предполагают наличие развитых функциональных возможностей: «навигация (возможность получать сведения о составе и структуре информационной базы и иметь доступ к ее компонентам); лексический поиск (т. е. поиск по свободной лексике естественных языков с учетом их морфологии); атрибутный поиск (т. е. поиск информационных объектов по значениям их формальных характеристик); комбинированные поиски; сортировка результатов поиска по заданному параметру; просмотр содержания информационного объекта – последовательный (страница за страницей), выборочный (переход на заданную страницу), параллельный (одновременный просмотр нескольких документов); поддержка гипертекстовых связей; экспорт информации с указанием источника» [5].

Таким требованиям, по нашему мнению, соответствует созданное профессором Максом Сильберцтайном программное обеспечение NooJ, основанное на INTEX и предназначенное для обработки естественного языка [6]. Данное программное обеспечение находится в постоянной разработке, ежедневно обновляется и доступно для скачивания по адресу: http://www.nooj4nlp.net/pages/nooj.html. NooJ основан на платформе .NET, что является ответом Microsoft на JAVA/J2EE и работает в операционных системах Windows 95-98-ME, Windows NT-2000, Windows XP и Windows VISTA, хотя некоторые ее функции (а именно поддержка UNICODE и XML) доступны только в ОС Windows 2000, Windows XP и Windows VISTA.

NooJ является бесплатным программным обеспечением, которое применяется для формализации различных типов текстовых явлений (орфографии, словообразовательной и словоизменительной морфологии, структурного и трансформационного синтаксиса). Оно интегрирует широкий спектр вычислительных технологий – от автоматического вывода конечных состояний до расширенных или рекурсивных и трансляционных взаимосвязей.

Набор инструментов позволяет построить, протестировать, отладить, сохранить и накопить большой набор языковых ресурсов. С помощью данной программы можно описать словообразовательную и словоизменительную морфологию, вариативность в написании или терминологии, лексику (простые слова, сложные слова, устойчивые выражения), синтаксис (грамматику фраз и предложений) и семантику (извлечение сущностей и трансформационный анализ).

На сегодняшний день NooJ может обрабатывать текстовые массивы разной алфавитной традиции на десятках языков: для скачивания доступны модули франко-акадийского диалекта, арабского, армянского, белорусского, болгарского, каталанского, хорватского, французского, английского, немецкого, греческого, иврита, венгерского, итальянского, латинского, польского, португальского, русского, сербского, испанского, турецкого, вьетнамского языков. Каждый модуль представлен различными ресурсами, например, модуль русского языка, разработанный Винсентом Бенетом, включает в себя пять художественных текстов и один учебник[2], главный словарь (содержит более 90 тыс. слов: 20 тыс. прилагательных, 45 тыс. существительных и

[2] Пользователь может загружать любые тексты, которые имеют формат .NOT.

25 тыс. глаголов), словарь собственных имен и субстантивированных прилагательных, словарь соответствующих парадигм, несколько грамматик для устранения неоднозначности дат, предлогов, дательного и родительного падежей, глаголов движения и др.

Важной характеристикой является то, что NooJ может обрабатывать как отдельные тексты, так и корпусы из сотен текстовых файлов.

От соответствующих задач исследования будут зависеть и системные требования программы. Для работы в программе NooJ с небольшими текстами (менее 1 Мб) требования к системе минимальные: 512 Мб оперативной памяти, 1 Гб свободного места на жестком диске. Однако если в задачи исследования входит анализ крупных текстовых массивов (сотни или тысячи текстовых файлов) или составление словарей с десятками тысяч слов, то конфигурация должна быть иной: Pentium 4,2 Гб или больше оперативной памяти. Если же запланировано составление грамматической разметки (сотни графов), то необходим хороший монитор – не менее 19 дюймов, с разрешением 1600×1024 16 бит и минимальной частотой обновления 80 Гц.

К уникальным характеристикам программы относят возможность распознавания и работы с более чем сотней форматов текстовых файлов, включая HTML, PDF, MS Office, все варианты Unicode, ASCII и др. NooJ может импортировать и экспортировать информацию о грамматической разметке в документы XML. Система грамматической разметки может быть применена на любом уровне, сохраняя текст без изменения. Это позволяет лингвистам формализировать различные явления независимо друг от друга.

Программа достаточно проста в использовании и отвечает требованиям, приведенным выше. Так, при загрузке текста выводится информация о его размере – количестве глав, символов, словоформ. Приводятся статистические данные о встречаемости букв, слов, словосочетаний. Просмотр текста осуществляется как постранично, так и при переходе на произвольную (заданную) страницу или часть текста. Возможна параллельная работа с двумя текстами (например, оригиналом и переводом).

По тексту осуществляется лексический поиск – по слову, словосочетанию, двум и более словам или словоформам (рисунок 1). При этом можно установить следующие параметры сортировки результатов поиска: увеличить или уменьшить длину контекста до и после искомого слова, упорядочить по левому или правому контексту или по искомым словам (если их несколько), установить ограничение на количество вхождений в выдаче.

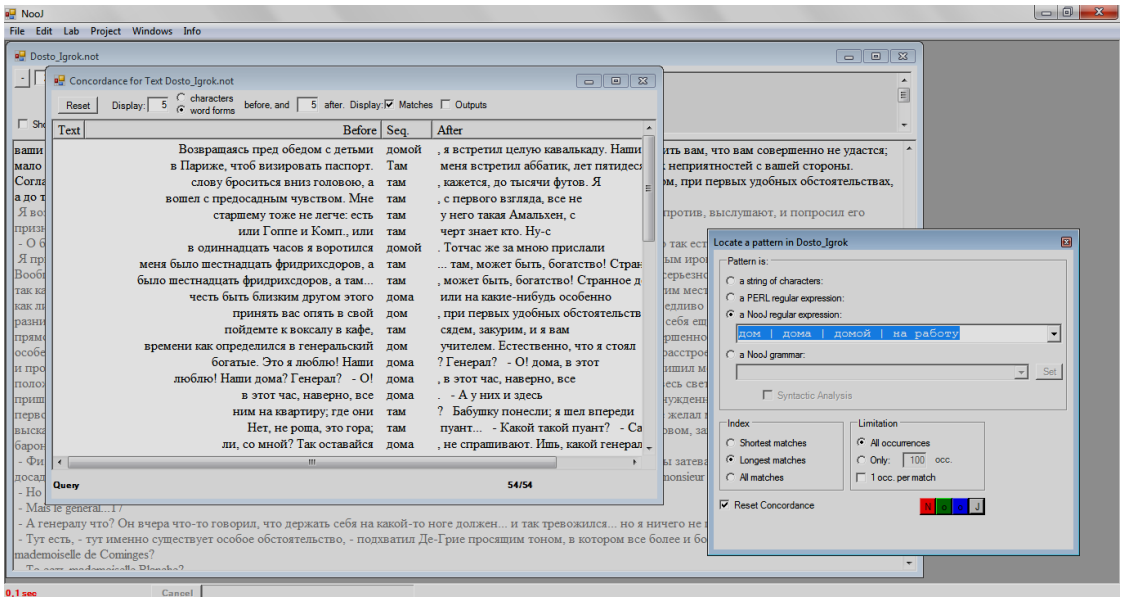

Рисунок 1 – **Лексический поиск по тексту в программе NooJ**

Далее такой конкорданс можно обработать статистически: NooJ построит график частотности соответствующих слов или словоформ, их встречаемости по главам текста, а также рассчитает численные показатели встречаемости слов/словоформ по главам (рисунок 2).

Также возможен и атрибутный поиск, т. е. поиск объектов по формальным характеристикам – части речи, роду и т. д. Комбинированные поиски выполняются при запросе, состоящем из, например, определенного глагола, например, *идти,* и формального указателя, например, *prep*

(*preposition*) – предлог. Результатом такого поиска будет конкорданс из контекстов, в которых содержится сочетание искомого глагола с любыми предлогами, найденными в тексте.

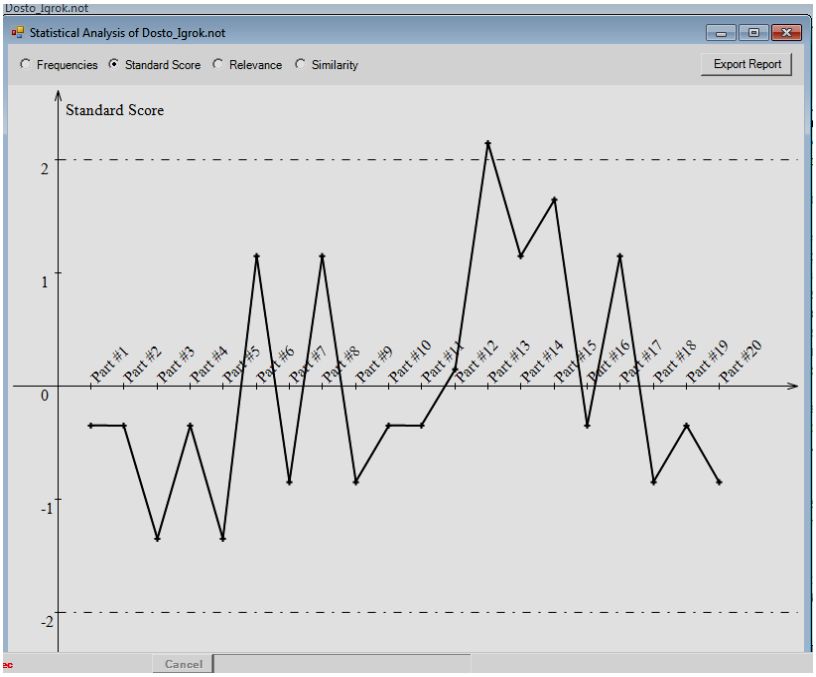

Рисунок 2 – **Статистический анализ частотности слов/словоформ в тексте**

В статье описаны только основные функции работы с текстом или корпусом текстов. Возможности NooJ зависят отчасти от степени разработанности модуля определенного языка, от наличия доступных словарей и уровня грамматической разметки.

Простота в использовании и функциональность NooJ объясняют активное использование программы в лингвистических исследованиях [7; 8]. Однако его популярность в исследованиях русскоязычных филологов не так высока. Следует еще раз обратить внимание на простоту интерфейса программы и легкость в ее использовании и отметить, что программное обеспечение несомненно поможет лингвистам автоматизировать однотипную, трудоемкую часть исследовательской работы, а также облегчит работу педагогам при обучении иностранному языку.

Список литературы

1. **Толковый** словарь русского языка / под ред. проф. Д. Н. Ушакова. – М., 1940. – Т. IV. – С. 1081.
2. **Станкевич, А. Ю.** Использование строковых функций при работе с лингвистической картотекой MS Access / А. Ю. Станкевич // Компьютерная лингвистика: научное направление и учебная дисциплина : сб. науч. ст. / ГГУ им. Ф. Скорины, научно-методический центр русистики ; отв. ред. В. И. Коваль. – Гомель : ГГУ им. Ф. Скорины, 2012. – С. 108–113.
3. **Райзберг, Б. А.** Современный экономический словарь / Б. А. Райзберг, Л. Ш. Лозовский, Е. Б. Стародубцева. – 2-е изд., испр. – М. : ИНФРА-М, 1999. – 479 с.
4. **Образовательный** стандарт Республики Беларусь по специальности 1-21 05 02 «Русская филология (по направлениям)» [Электронный ресурс]. – 2009. – Режим доступа : edubelarus. info/uploads/standart/2pokolenie/1–21%205%2002.pds. – Дата доступа : 15.12.2012.
5. **Вигурский, К. В.** Филология и современные информационные технологии (К постановке проблемы) / К. В. Вигурский, И. А. Пильщиков [Электронный ресурс]. – Режим доступа : http://rd.feb-web.ru/vigursky-03.html. – Дата доступа : 06.01.2013.
6. **NooJ Community** [Электронный ресурс]. – Режим доступа : http://www.nooj4nlp.net/pages/nooj.html. – Дата доступа : 18.12.2012.
7. **Ehmann** Bea et. al. NooJ as a Tool for Psychological Content Analysis of Small Group Communication in Isolated, Confined and Extreme (ICE) Environment [Electronic resourse]. – Mode of access : http://gavriilidou.gr/nooj2010/abstracts.pdf#page=29. – Date of access : 05.01.2013.

8. **Jelena Jovanović**, Uroš Milošević. NooJ linguistic tool – porting on mono platform [Electronic resourse]. – Mode of access : http://www.e-drustvo.org/icist/2012/html/pdf/499.pdf. – Date of access : 05.01.2013.