При поддержке РГНФ - грант № 12-04-12054

# ИНФОРМАЦИОННО-ПОИСКОВАЯ СИСТЕМА КОНТЕКСТНОГО ПОИСКА ПО ФРАЗЕОЛОГИИ ДОСТОЕВСКОГО

Результатом выполнения проекта «Создание системы контекстного поиска по фразеологии Достоевского» является создание электронного лингвистического ресурса

Целью проекта является создание нового информационно-поискового модуля, позволяющего извлекать, анализировать и обобщать данные о фразеологизмах у Ф.М. Достоевского. Этот модуль пользовательский интерфейс ввода-вывода поисковых запросов и результатов их обработки. Расположенный на сайте пользовательский интерфейс будет корреспондироваться с серверной частью базы данных, записи которой будут включать следующие виды лексикографической информации: словарная форма идиомы, в том числе ее стандартные и контекстно-зависимые варианты, пример употребления идиомы, источник примера. Информационно-поисковая система «Система контекстного поиска по фразеологии Достоевского» http://lexrus.ru/default.aspx?p=3173 является новой разработкой в области компьютерной лингвистики. Её новизна обусловлена, прежде всего, особым подходом к решению задач языкознания и обучения русскому языку, основанным на междисциплинарном взаимодействии гуманитарных и естественнонаучных дисциплин, и жёсткой ориентацией на определённый сектор лингвистической семиосферы. Последнее (целевая направленность ресурса) позволяет с максимальной эффективностью использовать академические разработки лингвистики.

Информационно-поисковая система «Система контекстного поиска по фразеологии Достоевского» (ИПС «Система контекстного поиска по фразеологии Достоевского») включает в себя специальные программы обработки и хранения баз данных, компьютерных картотек, программ обработки текста, позволяет в режиме реального времени динамически формировать словарные статьи и текстовые материалы на основе поисковых запросов пользователей.

Система, прежде всего, ориентирована на взаимодействие с пользователями Интернета, интересующимися вопросами русского языка и лингвистики. Модульный характер реализации системы обеспечивает, в рамках единого пользовательского интерфейса, хранение и обработку различных форматов данных, что позволяет не ограничивать возможности разработчиков по расширению и модернизации как отдельных частей ресурса, так и системы в целом. Благодаря этому, система обладает потенциалом развития.

Важно отметить факторы доступности, надёжности, удобства использования системы, ориентированной на пользователя Интернета. В то же время в ней реализованы сложные, с точки зрения пользователя, и скрытые от последнего, но необходимые алгоритмы хранения и обработки оцифрованной информации лингвистического содержания, которые позволяют использовать информационный потенциал не только в системах «человек – машина», но и «машина – машина». Последнее предоставляет возможность практически неограниченного наращивания информационной мощности системы не только с учётом существующих технологий обработки информации на ЭВМ, но и с учётом перспектив развития отраслей компьютерной лингвистики, информатики и цифровых технологий в целом.

Принципиальной позицией создателей данной разработки является реализация понятного и не отягощённого избыточной информацией интерфейса взаимодействия пользователя с системой, но в тоже время позволяющего реализовывать как простые, так и очень сложные запросы к базам данных.

Совокупность предоставляемых возможностей не имеет аналогов в сфере ресурсов Интернета, как по комплексности, так и по алгоритмизации решения задач.

Краткий обзор предоставляемых возможностей системы: использование собственного программного обеспечения позволило сделать систему автономной и независимой от сторонних производителей и

разработчиков, что обеспечивает надёжность работы всех модулей системы как единого целого; принцип многоуровневого администрирования системы позволяет обеспечивать работу администраторов, реализующих разные задачи в зависимости от их профессиональной принадлежности и компетенции в тех или иных вопросах, что гарантирует профессиональное администрирование на всех содержательных уровнях системы; существенным обстоятельством является то, что доступ ко всем перечисленным возможностям реализован через Интернет, посредством обычного Web-браузера

ИПС «Система контекстного поиска по фразеологии Достоевского» снабжена справочной информацией, представленной как в кратком изложении, достаточном для задания простых запросов к базам данных, так и в подробной форме, позволяющей на наглядных примерах быстро освоить профессиональное пользование системой.

Таким образом, ИПС «Система контекстного поиска по фразеологии Достоевского», прежде всего, ориентирована на достижение максимально эффективной и комфортной для пользователя машинной обработки данных по схеме «запрос – результат запроса» в системе «человек – машина».

В ходе выполнения работ в рамках проекта, были выработаны, критерии выделения идиом как особого лексического класса единиц. Критерии выявления употребляемых идиом и других видов фразеологизмов в текстах Достоевского. Выявлены характерные признаки, способствующие проводить классификацию фразеологизмов, позволяющую операционально отделять идиомы в точном смысле от других неидиоматичных, но устойчивых сочетаний слов, а также от слабоидиоматичных выражений.

Разработка программного обеспечения web-сайта базируется на технологии .NET компании Microsoft. В качестве среды выполнения используется Microsoft SQL Server 2008, MS ASP NET 4.0 и Net Framework 4.0.

Содержательная часть инженерно-технических работ ИПС «Система контекстного поиска по фразеологии Достоевского»: создание структуры базы данных и перенос в неё имеющихся исходных текстов; разработка модуля поиска, связывающего поисковые форму и встроенный поисковый механизм MS SQL Server; разработка интерфейса поисковых форм, позволяющего делать разнотипные запросы к информационной базе данных; разработка механизма предобработки поисковых запросов; встраивание ИПС «Система контекстного поиска по фразеологии Достоевского» в дизайн lexrus.ru.

Информатизация современного социума, развитие соответствующих технологий, количественное увеличение объемов коммуникаций потребовали новых подходов к принципам поиска лингвистической информации в Интернете. Некоторые специалисты считают, что в современном Интернете размещено почти два миллиарда документов. Чтобы в гигантском потоке информации найти материалы лингвистического характера необходима особая система поиска, так как существующие системы Яндекс, Рамблер, Google, обеспечивающие 88% всего поиска русскоязычного Интернета, не ставят своей задачей актуализацию лингвистических аспектов найденных слов. Недостатками существующих поисковых систем в контексте специфики лингвистической информации являются: наличие «лишней» информации, когда на запрос пользователя, желающего уточнить значение слова или его грамматические особенности (образование падежных форм, множественного числа существительных, определенных глагольных форм и т. д.), поисковая система выдает огромное количество адресов, среди которых не всегда есть те, которые представят лингвистическую информацию.

Преимущества ИПС «Система контекстного поиска по фразеологии Достоевского» состоят в том, что в ней реализована возможность предъявления пользователю лингвистического материала в системе, компоненты которой отражают три аспекта функционирования языка: язык

— как совокупность текстов, язык — как структура (в конечном счете тоже сводимая к совокупности, но уже единиц и правил), и язык — как способность создавать тексты. Указанный подход к предъявлению лингвистических материалов отражает общее направление развития языкознания, в котором на рубеже 80-х годов XX века произошел «переход от структурной модели языка к коммуникативно-функциональной». Теория связывает системно-структурные и функциональные свойства языковых единиц, что находит отражение в лингвистических исследованиях.

В ходе выполнения работ в рамках проекта, были адаптированы разработанные ранее базы данных к решению задачи описания фразеологии Достоевского, уточнены содержание различных ее полей и произведено наполнение базы данных с учетом возможных, в дальнейшем, унификаций и редактирования данных.

Лингвистические материалы ИПС «Система контекстного поиска по фразеологии Достоевского» помогут и ученым, и рядовым носителям языка решить возникающие проблемы. Необходимость таких материалов в мировой информационной сети продиктована потребностями развития компьютерных технологий и информатизации общества в целом. ИПС «Система контекстного поиска по фразеологии Достоевского» призвана обеспечить оптимальные условия для предоставления информационных услуг отечественным и зарубежным пользователям. Гибкая организация информации, её интегрированное представление, открытая архитектура системы являются ключевыми моментами в создании ИПС «Система контекстного поиска по фразеологии Достоевского».

При работе над проектом использовались методы корпусной и компьютерной лингвистики. В частности, будут использованы имеющиеся структуры лексикографических баз данных, созданные в отделе экспериментальной лексикографии Института русского языка РАН, а также корпус текстов, содержащихся в полном собрании произведений Ф.М. Достоевского.