

ИСПОЛЬЗОВАНИЕ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ В ЛИНГВИСТИКЕ

В. В. Петрашко

*Белорусский государственный университет, Минск, Республика Беларусь,
bsu@bsu.by*

В статье выделены и проанализированы основные информационные технологии в лингвистике.

Ключевые слова: информационные технологии; машинный перевод; автоматический ввод текста; автоматическое аннотирование и реферирование текстов; автоматический поиск информации; автоматизированные словари.

В современном мире информационные технологии являются неотъемлемой частью любой сферы профессиональной деятельности, в том числе лингвистики. Если когда-то использование компьютеров и соответствующих программ в лингвистических исследованиях, переводе и в обучении языку не являлось обязательным, то сегодня будущим преподавателям иностранных языков, переводчикам и лингвистам необходимы знания, связанные с использованием информационных технологий в сфере своей профессиональной деятельности.

К наиважнейшим задачам, с которыми связаны информационные технологии в лингвистике, относятся: автоматический или машинный перевод; автоматический ввод текста; автоматическое аннотирование и реферирование текстов; автоматический поиск информации; автоматизированные словари различных типов; специальные системы генерации речи; компьютерное обучение иностранному языку.

Наряду с рассмотренными выше способами использования информационных технологий в лингвистике существуют и другие области пересечения лингвистики и информатики: извлечение знаний из текста, автоматическое индексирование и рубрицирование документов, гипертекстовые технологии в лингвистике и многое другое [2].

Решение этих задач невозможно при отсутствии качественных технологий. Поэтому разработки в области прикладной лингвистики очень актуальны. Особенность использования информационных технологий в лингвистике состоит в том, что продукты, созданные таким путём, следует постоянно дополнять и перерабатывать, актуализировать. Это в первую очередь относится к словарям, энциклопедиям и автоматическим переводчикам.

В настоящее время один из основных источников профессиональной информации для лингвиста считается сеть Интернет, где есть разнообразные ресурсы в виде массивов (баз) информации: электронные энциклопедии (feb-web.ru), сети словесных ассоциаций (wordassociations.net), электронные библиотеки (eLIBRARY. RU, КиберЛенинка, Арбикон), базы данных университетов, авторские сайты, форумы, текстовые процессоры (MicrosoftWord, LotusWordPro, OpenOffice Writer), ментальные карты (ultra outliner) и др.

Достаточно существенную часть информационных технологий в лингвистике занимают образовательные и научные ресурсы. Они существуют в основном в виде интернет-библиотек: словари, энциклопедии, базы данных, архивы. В связи с мобильностью и доступностью предоставляемой информации они пользуются широким спросом в научном и образовательном сообществе. Основная их отличительная черта – абсолютная доступность электронного информационного ресурса: книги, учебники, пособия, сборники, исторические источники, статьи и авторские монографии. Подобные ресурсы упрощают и ускоряют работу как в обучении, так и в лингвистических исследованиях.

Однако многие проблемы остаются неразрешёнными, к примеру, требуются специальные программы по эффективному статистическому подсчёту и выборке лексических единиц, которая традиционно выполняется до сих пор вручную, но без чего невозможно провести ни одно исследование в области лингвистики.

Еще одной из главных областей, в которых применяются информационные технологии в лингвистических исследованиях, является корпусная лингвистика. Корпус – это большой набор текстов, собранный и проанализированный на компьютере. Корпусные данные могут быть представлены как национальными, так и международными корпусами, а также специально собранными корпусами, которые содержат тексты, собранные в рамках конкретного лингвистического исследования.

Корпусные исследования являются способом изучения языка на основе анализа большого объема релевантных текстовых данных. Использование корпусных данных позволяет получать более точную статистику по определенным вопросам, выполнять сравнительный анализ и изучать языковые тенденции в большом масштабе.

Существует несколько программных средств для работы с корпусами, таких как Sketch Engine, AntConc, WordSmith, TextSTAT и др. Они предоставляют инструменты для фильтрации и выборки нужной информации, анализа текста, построения статистики и многое другое. Например, эти программы позволяют искать определенное слово или выражение, находить классы слов по определенным критериям, рассчитывать

статистические параметры для текста, определять частотность использования слов и т.д.

Использование корпусных данных в лингвистике позволяет выявлять закономерности в языке, изучать специфику использования языка в разных контекстах (например, в разных жанрах или в разных культурах). Кроме этого, корпусные исследования могут помочь в создании словарей, грамматик и других лингвистических ресурсов, что является особенно полезным для разработки программных приложений, например, машинного перевода.

На сегодняшний день много вопросов вызывает возможность синтеза прагматической структуры текста из тех целей, которые стимулируют создание текста. Например, неясно, как структуру разбить на структуры предложений и как от них перейти к глубинным семантическим структурам. Поиск информации является часто возникающей задачей в наши дни.

Сейчас возникает необходимость в систематизации информации и облегчению её поиска. Эту проблему решают информационно-поисковые системы. Сущность любого поиска заключается в отборе информации по запросу в базе данных, где хранится информация. При этом не важно, в каком виде представлена информация. Для большей эффективности применяют специальные программы, производящие лингвистическую обработку текстов с учетом морфологии и синтаксиса, ибо не всегда простое использование названия документа является оптимальным выбором поискового образа [1].

В условиях все возрастающего количества текстов в окружающем человека мире возникает проблема поиска нужного документа и ознакомления с его содержанием. В решении проблемы помогают составление рефератов и аннотаций полнотекстовых документов. Они дают читателю представление о содержании исходных документов и позволяют оценить степень необходимости обращения к полным текстам каждой работы. Кроме того, рефераты и аннотации акцентируют внимание читателя на новых сведениях, т.е. позволяют за небольшой промежуток времени узнать много новой информации.

Наиболее простыми системами автоматического реферирования и аннотирования является функция AutoSummarize в MS Word, системы Intelligent Text Miner, Oracle Context и Inxight Summarizer (компонент поискового механизма AltaVista). Правда, возможности этих программ ограничены выбором оригинальных фрагментов из исходного документа и их соединением в короткий текст [2].

Все эти направления и многие другие развиваются с каждым днём благодаря современным компьютерным технологиям. На сегодняшний

момент существует множество систем и методов, которые помогают не только лингвистам, но и обычным пользователям при взаимодействии с компьютером.

Сейчас эти сервисы облегчают жизнь людям в их повседневной жизни, а уже завтра, возможно, будут облегчать существование машин с искусственным интеллектом в повседневном общении с людьми на естественном языке.

Библиографический список

1. *Шилихина, К. М.* Основы прикладной лингвистики: учеб. пособие / К. М. Шилихина. – Воронеж: Изд-во ВГУ, 2006. – 51 с.
2. *Щипицина, Л.Ю.* Информационные технологии в лингвистике: учеб. пособие / Л. Ю. Щипицина. – Москва: ФЛИНТА, 2013. – 314 с.