Л.Н. Беляева

КОРПУС ПАРАЛЛЕЛЬНЫХ ТЕКСТОВ КАК БАЗА ДАННЫХ ПРИКЛАДНОЙ ЛЕКСИКОГРАФИИ

Развитие информационных технологий (ИТ), составляющих сегодня часть профессиональной подготовки специалистов-филологов, определяет необходимость разработки компьютерных средств обучения и подготовки специалистов, различных систем автоматической переработки текста, специальных средств интерактивной электронной издательской деятельности, компьютерной лексикографии и, в целом, совершенствования технологий в области анализа и обработки естественного языка.

В современном мире в условиях открытой и многоязычной научной коммуникации и развития средств непрерывного и открытого обучения возникает целый ряд задач, решение которых связано с качеством и практической применимостью различных информационных технологий, связанных с анализом текстов на естественном языке и звучащей речи. К таким задачам в самом общем виде относятся:

- автоматический поиск, извлечение и обогащение информации и знаний, получаемых из различных мультимедийных многоязычных источников и источников, связанных с коммуникацией различных участников;
- межъязыковое или многоязычное извлечение, презентация и распространение информации;
- автоматическое обнаружение и «отслеживание» новой фактографической информации из неструктурированных мультимедийных данных;
- использование источников знаний для того, чтобы облегчить разметку знаний и доступ к ним (в качестве таких структурированных источников знаний могут выступать одно- и многоязычные лексиконы, толковые и энциклопедические словари, тезаурусы, энциклопедии и т.д.);

- поддержка вопросно-ответного взаимодействия человека и компьютера, а также людей между собой с помощью компьютера как посредника для извлечения знаний из источников различной природы, структуры и состава;
- поддержание дистанционного обучения в системах открытого образования, включая автоматизированное тестирование уровня знаний, разработку электронных учебников и диалоговых обучающих систем;
- создание интеллектуальных средств поддержки автоматизированного ведения библиографической работы, анализа и понимания документов для того, чтобы обеспечить возможности доступа к информации различных экспертов или групп экспертов;
- моделирование знаний, потребностей и намерений пользователей на основе анализа их запросов к различным системам, созданных ими продуктов и взаимодействия с компьютером;
- обеспечение возможности устного диалога с компьютером и поддержки анализа и порождения звучащей речи.

Все это определяет необходимость создания и использования (в том числе, обучения использованию) специализированных систем обработки многоязычной информации, в частности, систем компьютерной поддержки обучения в условиях традиционного и открытого образования, а также систем автоматической переработки текстов (АПТ), предназначенных для специалиста в конкретной области знаний.

Особым направлением прикладной лингвистики сегодня является прикладная лексикография, в задачу которой входит создание и ведение автоматизированных и автоматических словарей и баз данных, жестко проблемно и предметно ориентированных. Полноценность и адекватность спектра лексикографических систем в значительной степени определяют уровень и достоверность извлечения информации и знаний из текстов различного состава, структуры и назначения. Однако анализ современного набора

переводных словарей, издаваемых в нашей стране и/или включенных в различные автоматизированные словарные системы, позволяет фиксировать их отставание от современного развития науки и техники, несоответствие основным направлениям развития отраслей знаний. Это связано не только с естественным отставанием лексикографии, связанным с необходимостью переработки больших массивов современной информации, а прежде всего с традиционным подходом к созданию словарей с опорой на уже опубликованные источники, а уже затем на результат анализа текстов[1]. Использование ИТ при таком преимущественно компилятивном подходе ничего не меняет по сути, но только уменьшает трудоемкость работы при сравнении и объединении словарных источников, а также при редактировании сформированного массива. Следовательно, оперативное создание словарей, отражающих картину лексического состава отдельных отраслей знаний, фиксирующих нормативную (рекомендуемую) и реально встречающуюся лексику, представляет собой особую задачу.

Аналогичная задача решается и при создании автоматических словарей (АС) систем, предназначенных для анализа и переработки текстов. Подобные словари не только жестко ориентированы на конкретные предметные области и типы текстов, но и в большой степени определяют качество работы систем переработки информации, что определяет особый подход к отбору лексических единиц: при создании АС предварительно осуществляется анализ достаточно больших корпусов текстов по заданной тематике. В ряде систем для формирования АС используются уже имеющиеся архивы текстов или уже созданные словари информационных систем.

---

[1] См., например, использование 9 словарей, опубликованных в США и Великобритании в период с 1973 по 1994 гг., при создании Большого англо-русского медицинского словаря (М., 2005).

В самом общем виде отбор лексических единиц (слов и словосочетаний) в АС должен осуществляться на основе:

– статистического критерия, определяющего обязательность введения в АС всех единиц, обеспечивающих 85-процентное распознавание текстовых словоформ;
– критерия синтаксической независимости, определяющего необходимость введения только тех единиц, структура которых не зависит от структуры предложения и ближайшего контекста;
– критерия релевантности, определяющего обязательность введения в АС терминологических единиц из конкретной предметной области независимо от частоты их появления в обучающей выборке текстов.

Следует особо подчеркнуть, что с расширением функций информационных систем и включением в них систем машинного перевода и, в частности, систем переводческой памяти архивы систем, сгруппированные по определенным тематикам, являются оптимальным источником для отбора лексики в словарь. Дело в том, что ориентация на конкретную предметную область является важной характеристикой любого автоматического словаря, так как позволяет на уровне лексического анализа снимать многозначность лексических единиц и стандартизировать перевод терминологии. В то же время современное развитие науки свидетельствует о возникновении новых направлений на стыках традиционных отраслей знаний и, следовательно, настоятельную потребность создания узкоспециализированных словарей, как автоматизированных, так и «бумажных».

Современный подход к созданию словарей предполагает формирование и использование параллельного корпуса современных текстов, который может рассматриваться как база данных для решения не только исследовательских задач, но и практических лексикографических задач. Корпусы письменных текстов, как правило, включают сами тексты, а также разметку текстов с

точки зрения формата и предложений по результатам парсинга, позволяющими установить принадлежность лексических единиц к конкретным частям речи. Эти тексты могут служить для создания конкордансов, словарей слов и словосочетаний в случае одноязычного корпуса, а также для создания многоязычных лексиконов и многоязычных конкордансов в случае корпуса параллельных массивов.

Создание базы полнотекстовых данных предполагает обеспечение хранения, модификации и поиска текстов произведений художественной и научной литературы на разных языках с формированием массивов параллельных и псевдопараллельных текстов. Эта база может непосредственно использоваться в учебном процессе для анализа конкретных лингвистических и литературоведческих фактов, проведения сравнительного стилистического анализа, изучения особенностей авторского стиля и т.д. Кроме того, подобная база является важным источником сведений для создания словарей разного состава и назначения.

Следует учитывать, что лексикографическая работа даже с использованием возможностей ИТ в значительной степени остается работой творческой и не может быть полностью автоматизирована. В то же время, существуют возможности подготовки массивов текстов для лексикографического анализа, одной из таких возможностей является формирование особых корпусов текстов, включающих параллельное представление исходных текстов, их машинных переводов и отредактированных переводов, согласованных с экспертами в конкретной области знаний (см. табл. 1). Важно отметить, что качество и потенциал такого корпуса в большой степени зависит от сотрудничества с экспертами при отборе исходного материала и редактировании переводов.

На основе корпуса псевдопараллельных текстов может быть, в частности, проведен анализ номинации экстралингвистических объектов и особенностей терминообразования (структуры именных терминологических сочетаний) в условиях разных родных

языков. Особенности номинации одних и тех же референтов в текстах носителей разных языков отражаются в вариантах структур именных терминологических сочетаний в этих языках. При этом следует иметь в виду, что построение лексического спектра глобального английского языка дает возможность проанализировать не только особенности использования лексических единиц (ЛЕ) – слов и словосочетаний в текстах, написанных носителями разных языков, но и выявить особенности когниции, свойственные «усредненному» носителю конкретного языка и, как следствие, конкретной культуры.

Таблица 1. Фрагмент структуры выровненного по предложениям корпуса текстов с использованием результатов машинного перевода

Исходный текст	**Машинный перевод**	**Отредактированный перевод**
Student-centred learning produces a focus on the **teaching-learning-assessment** relationships and the fundamental links between the design, **delivery**, assessment and **measurement of learning**.	**Ориентированное на обучающегося обучение** производит фокус на отношениях **teaching-learning-assessment** и фундаментальных связях между проектом, **поставкой**, контролем знаний и **измерением обучения**.	При **личностно-ориентированном обучении** основное внимание уделяется отношению **преподавание-обучение-оценка и** фундаментальным связям между проектом, **подачей материала**, контролем знаний и **измерением качества обучения**.

В таблице жирным шрифтом выделены словосочетания, фиксация которых целесообразна в словарях соответствующей предметной области.

Рассмотрение полнотекстового корпуса параллельных текстов в качестве лексикографической базы предполагает необходимость ее дополнения корпусом машинных переводов текстов, что позволяет явным образом выделить те лексические единицы, которые должны быть введены в словарь или модифицированы с точки зрения набора переводных эквивалентов.

Использование корпуса параллельных текстов в двуязычной лексикографии позволяет не только максимально автоматизировать отбор терминологических словосочетаний, но также служит для

- обогащения набора словарных статей за счет выбора свободных словосочетаний, используемых в исходных текстах, что чрезвычайно важно для тех, кто переводит на язык, не являющийся родным;
- уточнения употребительности конкретных словосочетаний в текстах определенной предметной области;
- верификации значений лексических единиц, зафиксированных в двуязычных словарях, особенно в том, что касается идиом и терминологических выражений;
- выделения устойчивых словосочетаний и идиом, которые целесообразно вводить в словарь конкретной отрасли знаний.

На основе полнотекстовых баз параллельных выровненных текстов возможно выделение устойчивых пар слов типа «исходное слово – перевод», однако применение статистических процедур, как правило, допускает соответствие слов, но не словосочетаний. Это достаточно жесткое ограничение для выбора потенциальных компонентов словаря может быть уменьшено, если параллельные тексты предварительно проходят процедуру парсинга или подвергаются ручной разметке, что позволяет соотносить не отдельные слова, а фрагменты предложения.

Корпус текстов может использоваться для выявления структуры предметной области и соответствующей терминосистемы. Рассмотрим это на примере словосочетаний с элементом *higher education*, являющимся самым частотным в корпусе текстов Болонского процесса. Для выявления структуры лексикографического описания – терминосистемы поля «*higher education*» получим частотный словарь лексем, составляющих эти словосочетания, исключив из него служебную лексику. Поскольку базой терминосистемы являются имена существительные, то оставляем

в списке только их. Следующим этапом является выявление синонимов и объединение их в блоки. В результате получаем словник ключевых единиц, который можно использовать как базу структурирования терминосистемы (см. фрагмент в табл. 2).

Таблица 2. Фрагмент словника ключевых единиц

Ключевая единица (наиболее частотные ЛЕ)	**Частота**	**Синонимы**	**Суммарная частота**
institution	11	institute	2
		schools	1
		university	1
program(mes)	7	cycle	3
		guidelines	1
		schemes	1
minister	5	ministry	1
		agency	1
		authority	1
		bodies	1
		community	2
		council	2
		department	1

Используя выделенные ключевые единицы как основу для объединения, получаем следующую структуру терминосистемы (см. табл. 3). Исследование словосочетаний, выявленных и организованных на основе предложенной процедуры, позволяет установить те из них, которые следует включать в словари и глоссарии, а также определить потенциально возможные (см., например, *higher education law, higher education program*) словосочетания.

Таблица 3. Фрагмент структуры терминосистемы «*higher education*» в текстах Болонского процесса

Словосочетания	**Длина**	**Частота**
Программы высшего образования		
higher education program	3	0
higher education programme of study	5	1
higher education programmes of learning	5	1
recognised higher education programme of study	6	1
accredited vocational higher education programs	5	2
classic higher education programs	4	1
shorter higher education programs	4	1
interdisciplinary higher education study programs	5	1
Системы высшего образования		
higher education system	3	10
higher education systems	3	15
higher educational system	3	2
European higher education system	4	4
global higher education system	4	1
Italian higher education system	4	1
long-established and powerful higher education systems	7	1
mass higher education system	4	2
national higher education system	4	2
national higher education systems	4	3
own higher education system	4	1
respective higher education systems	4	1