пример, в либретто оперы Дж. Верди «Дон Карлос» (действие третье, картина пятая) читаем: «Мадрид. Кабинет короля. "Когда король спит, не спит враг". Считая себя *преданным*, Филипп пытается получить благословение на решительные действия у Великого инквизитора…»

В данном случае нельзя не заметить нарушений в формальной организации текста. Кроме того, некоторую двусмысленность создает слово преданный (предан кому-то или кто-то предал его?)

Итак, исследованный материал показал, что в подавляющем большинстве случаев тексты оперных либретто, напечатанные в программках, представляют собой краткий (в ряде случаев – очень краткий) пересказ литературной основы оперы с довольно нерегулярно привлекаемыми экспрессивно-образными средствами. Нерегулярно проявляет себя и автор, который как рассказчик излагает свою точку зрения и свои оценки.

Представляется, что жанр оперного либретто требует дальнейшего исследования в расширенном поле текстов (оригинальных и переводных, классических и современных) в направлении уточнения роли различных языковых средств, регулярности и результативности их использования, а также более пристального внимания к текстовой структуре и, возможно, в других направлениях, так как рассмотренные в данной статье тексты лингвистами фактически не анализировались.

Список литературы

1. Гулая Т. Н. Оперные либретто как феномен интертекстуальности: на материале оперы Г.Г. Вдовина «Пасынок судьбы» : автореф. дис. … канд. культурологии / Т. Н. Гулая. – Саранск, 2006. – 22 с.

2. Пивоварова И. Л. Либретто отечественной оперы: аспекты интерпретации литературного первоисточника : автореф. дис. … канд. искусствоведения / И. Л. Пивоварова. – Магнитогорск, 2002. – 22 с.

3. Словарь иностранных слов. – 15-е изд., испр. – М. : Русский язык, 1988. – 608 с.

4. Райс К. Классификация текстов и методы перевода / К. Райс // Вопросы теории перевода в зарубежной лингвистике. – М., 1978. – С. 202–208.

ОСОБЕННОСТИ ФОРМИРОВАНИЯ СПЕЦИАЛЬНОГО КОРПУСА ТЕКСТОВ В ОБЛАСТИ РОБОТОТЕХНИКИ НА НЕМЕЦКОМ ЯЗЫКЕ

М.А. Большакова

Статья посвящена исследованию особенностей формирования специального корпуса текстов в области робототехники на примере немецкого языка. Описаны лингвистические особенности технических текстов с точки зрения когнитивного терминоведения.

The article is devoted to research features of the formation of a special corpus of texts in the field of robotics by the example of the German language. In article is described the linguistic features of technical texts from the perspective of cognitive terminology.

Ключевые слова: корпусная лингвистика, специальный корпус текстов, научный текст, терминология.

Key words: corpus linguistics, a special corpus of texts, scientific text, terminology.

Корпусная лингвистика (КЛ) является быстро развивающейся областью современного языкознания. Распространение этой дисциплины в значительной степени мотивировано развитием технических средств и информационных технологий за последнее время. Предмет изучения в этой области – составление моно- и многоязычных корпусов текстов различной тематики, с помощью которых осуществляются исследования определенных объектов языка и особенно совершенствование лингвистического обеспечения разнообразных информационных систем. Преимуществом использования корпусов является возможность автоматической обработки данных корпуса, в том числе наличие механизмов автоматического поиска необходимой информации. Обычно для ее сбора необходимо просматривать большое количество текстов и выписывать примеры вручную. Этот недостаток затрудняет обработку больших массивов материала. С появлением текстовых корпусов (КТ) на машинных носителях процесс сбора данных существенно упрощается, а качество значительно улучшается.

Основное понятие КЛ – это «корпус текстов», под которым понимается сформированная по определенным правилам выборка текстовых данных из проблемной области. Проблемная область определяется как «область реализации языковой системы, содержащая феномены, подлежащие лингвистическому описанию» [1, с. 24].

С точки зрения КЛ автором в статье рассматривается корпус текстов области робототехники на немецком языке. Основной целью исследования является построение модели отбора научно-технических текстов выбранной узкой предметной области из массивов, в том числе Интернет, использующей модуль автоматического анализа текста на естественном языке для полного представления знаний.

В соответствии с этой целью и гипотезой исследования поставлены следующие задачи:

1) определить лингвистические особенности технических текстов с точки зрения когнитивного терминоведения;

2) модифицировать процедуру анализа текстов, настроив на обработку выбранные ресурсы;

3) сформировать материал исследования: представительный корпус технических текстов предметной области робототехники;

4) отобрать в качестве объекта исследования набор ключевых понятий – многозначные общеупотребительные слова, существительные и глаголы, которые способны воплощать характерные для заданной предметной области понятия.

В связи с этим использовался специальный корпус текстов. Под специальным КТ будет пониматься сбалансированный корпус, как правило, небольшой по размеру (несколько тысяч словоупотреблений), подчиненный определенной исследовательской задаче и предназначенный для использования только в целях, соответствующих замыслу составителя [7, с. 114–123].

Первый компьютерный корпус создан в США (так называемый Брауновский КТ) около 40 лет назад, за это время сформировали и другие корпуса текстов. Они используются в различных исследованиях. Появилось много публикаций, описывающих результаты этих исследований, а также свойства корпуса текстов как нового типа словесного единства. Создана новая наука – корпусная лингвистика, получили названия разнообразные типы КТ – двуязычные, учебные и т.п.

Возникновение корпусных методов связано с бурным развитием компьютерных технологий во второй половине XX в. Возможность сканирования и распознавания текста (перевод в текстовый формат), появление баз данных и систем управления базами данных сделали возможным сбор, хранение и обработку огромных массивов текстовых данных. Не последнюю роль в развитии корпусной лингвистики сыграла популяризация сети Интернет, так как корпуса стали доступны широкому кругу пользователей, значительно расширились возможности их наполнения. С тех пор накоплен значительный опыт разработки и их применения. Обсуждению проблем корпусной лингвистики посвящена специализированная электронная рассылка Corpora List и периодические издания "International Journal of Corpus Linguistics", "Corpora", "Corpus Linguistics and Linguistic Theory", "ICAME Journal".

В России разработкой и исследованием корпусов занимаются специалисты Центра лингвистической документации при Независимом Московском университете, Московского государственного лингвистического университета, отдела экспериментальной лексикографии Института русского языка им. В.В. Виноградова РАН, Института языкознания РАН, Института проблем передачи информации РАН, Всероссийского института научной и технической информации РАН, Института лингвистических исследований РАН в Санкт-Петербурге и др.

Теоретические и практические проблемы КЛ обсуждаются на специализированных семинарах и в рамках научных конференций по прикладной и компьютерной лингвистике: ежегодная международная конференция по компьютерной лингвистике «Диалог», конференция «Мегалинг», конференция «Корпусная лингвистика» при кафедре математической лингвистики СПбГУ. Компьютерной лингвистике посвящен раздел форума на сайте конференции «Диалог».

Таким образом, современная практика составления и использования корпусов текстов наглядно демонстрирует разнообразие КТ, построенных по определенным принципам, основанным на том, что из доступного составителям множества текстов составляется корпус, отвечающий заданной специфической потребности его составителя (моделирование терминологии, отладка системы машинного перевода, обучение иностранному языку и т.п.) [10, p. 75–96]. При этом КТ, представляющие подборку текстов предметной области знаний и созданные исследователем для решения конкретных задач, можно назвать специальными, и использоваться они должны в целях, для которых спроектированы [1, с. 27–35]. При проведении исследований специального КТ следует учитывать следующие факторы:

- ✓ специальный корпус не всегда может быть объективным отражением внешней действительности;
- ✓ специальный корпус предназначен для использования только в целях, соответствующих замыслу составителя [2, с. 35–46].

Разнообразие существующих корпусов и осознание необходимости консолидации исследований в области КЛ привели в 1992 г. к созданию Европейской корпусной инициативы (ECI). Результатом ее работы являются около 40–50 корпусов текстов на европейских языках, каждый объемом от 12 тыс. до 5 млн слов. Целью этой организации является не только создание более представительных, универсальных КТ на как можно большем числе языков, но и создание корпусов параллельных текстов. При этом наибольший интерес для проведения различных лингвистических исследований представляют

корпуса текстов, которые создавались как параллельные. Примером такого корпуса является корпус Hansard – отчеты о дебатах в канадской Палате общин. Отметим, что Канада – страна, в которой два языка имеют статус государственных, что дает право утверждать, что данный корпус является истинно параллельным [3, с. 318–333]. Современная лингвистическая практика составления и использования корпусов текстов демонстрирует разнообразие подходов (см. табл.) не только к их составлению, но и к собственно корпусам [4, с. 27–29]. Это связано с тем, что при решении различных задач, таких, как моделирование терминологии, отладка системы машинного перевода, обучение иностранному языку [10, p. 75–96] или изучение определенных лексико-семантических, а также семантико-синтаксических категорий, из доступного множества текстов формируется корпус, соответствующий основной цели его составителя.

Таблица

Лингвистические подходы к формированию корпусов текстов

Глобалистический подход	Персоналистский подход	Иллюстративно-дидактический подход	Информационно-технологический подход	Конструктивно-синтезирующий подход
Национальные корпусы Жанровые корпусы	Корпусы выдающихся авторов	Корпусы учебных текстов Корпусы параллельных текстов	Корпусы как речевой материал для создания и тестирования информационных систем	Лингвистически представительские корпусы

Собственно КТ рассматривается как способ сбора и хранения данных, при этом специальные программы анализа и лингвистической обработки текстов, в частности, программы построения конкорданса, алфавитных и частотных словарей, используются в качестве инструментов обработки текстов [8, с. 52–53].

Накоплен значительный опыт построения КТ, однако информация о наличии корпусов ограничена. Это связано с тем, что практически все они создаются в рамках локальных проектов отдельными организациями, поэтому для решения исследовательских задач приходится создавать собственный корпус текстов. В нашем случае он представлен текстами на немецком языке, которые относятся к технической области, точнее, к области робототехники.

С точки зрения объекта исследования корпус текстов представляет собой сверхсложную многоуровневую динамическую систему, которая предоставляет возможность не только для совершения таких элементарных операций, как, например, вычисление частот отдельных букв, но и для проведения глубокого лингвистического анализа [9, с. 334–352].

В настоящем исследовании КТ предназначен для изучения структуры и соотношения немецкой терминосистемы в области робототехники.

При создании корпуса научных текстов будем исходить из того, что любой КТ обладает минимальными базовыми признаками, позволяющими назвать некоторое собрание текстов корпусом [2]. К таким признакам относятся следующие:

- ✓ расположение на магнитном носителе;
- ✓ наличие единой процедуры отбора;
- ✓ репрезентативность;
- ✓ единство разметки или представления корпуса;
- ✓ конечность размера корпуса [6, с. 72–73].

Расположение на магнитном носителе обеспечивает возможность автоматической, компьютерной обработки корпуса текстов и позволяет сформировать выборку лексического материала из корпуса, которая соответствует разнообразно сформулированным запросам, что является непременным условием работы с любым корпусом.

Единство процедуры отбора характеризует способ формирования выборки текстов для включения в конкретный корпус, при этом критерии построения корпуса и процедура отбора текстов для корпуса должны адекватно отражать задачу создания корпуса. В процессе процедуры отбора определяются тексты, которые должны составить базу корпуса [5, с. 7].

Основным объектом исследования в данной работе являются термины, их поведение в научных текстах, составленных на немецком языке, поэтому при создании КТ в него включены только те тексты, которые отвечают следующим требованиям.

Каждый текст должен быть представлен на немецком языке и соответствовать основным критериям текстуальности, то есть обладать информативностью, интенциональностью, воспринимаемостью, ситуативностью, интертекстуальностью, быть когезивным и когерентным.

В этой связи следует отметить, что в данной работе основное внимание уделяется только элементам, которые являются значимыми для исследования концепта текста, то есть собственно научно-технические тексты. Вся остальная информация и, следовательно, структурные элементы, которые содержат эту информацию, находятся за рамками нашего внимания, но входят (за исключением последнего элемента) в структуру корпуса, что расширяет возможности его дальнейшего использования.

Представленный в нашем исследовании КТ отражает особенности одного типа языка – языка научной прозы, относится к предметной области робототехники как одной из областей научного знания. Конечный размер корпуса определяется целями планируемого исследования или задачами использования корпуса, в соответствии с которыми корпус может быть как дополнен, так и сокращен. Постоянное дополнение и обновление корпуса способствует поддержанию его репрезентативности [11, p. 30].

Полученный корпус соответствует заданным характеристикам и отражает подъязык робототехники. На основе классификации корпусов, предложенной в проекте EAGLES, созданный корпус научно-технических текстов имеет следующие характеристики:

- ✓ по модусу, форме существования представляет собой письменные тексты и алфавитно-частотные словари к ним;
- ✓ по объему – небольшой подкорпус;
- ✓ по языковому покрытию относится к корпусу подъязыка робототехники; достаточно представителен по объему выборки и позволяет делать достоверные выводы с учетом поставленных перед ним задач;
- ✓ по типу текстов: исследуемые тексты написаны на немецком языке и являются как первичными, так и вторичными научными текстами;
- ✓ по жанру (регистру) представляет собой опубликованные или представленные в Интернете научные тексты по основным предметным областям робототехники;
- ✓ по варьированию языков является непереводным одноязычным корпусом;

✓ по общности, которая продуцирует корпус, охватывает не только носителей немецкого языка в разных его вариантах, но и носителей других языков, которые используют немецкий язык для профессиональной коммуникации;

✓ по маркированности: тексты являются простыми неаннотированными, форма текстов стандартизирована и унифицирована в соответствии с целью исследования;

✓ по степени открытости – открыт для постоянного пополнения новыми текстами подобного типа;

✓ по национальному варьированию: часть корпуса немецкого языка, представленная текстами, созданными как его носителями, так и теми, для кого родным является русский, французский, испанский или другой язык мира;

✓ по историческому варьированию – синхронный подкорпус;

✓ по возрасту тех, кто продуцирует тексты: все тексты, вошедшие в корпус, написаны взрослыми людьми;

✓ по доступности корпуса: в перспективе этот корпус предполагается сделать свободно доступным в Интернете для изучения и использования в научных целях.

Представленный корпус отвечает требованиям, которые предъявляются к специализированным исследовательским корпусам, и, соответственно, может быть базой для проведения лингвистических исследований.

Обработка корпуса текстов вручную подразумевает следующие действия:

✓ создание общей базы научных текстов;

✓ редактирование и форматирование каждого отдельного текста, вошедшего в корпус;

✓ создание общей базы заголовков научных текстов;

✓ фрагментация заголовочных комплексов на отдельные тематические разделы;

✓ лексико-семантический анализ лексического состава научных текстов.

Наше исследование проводится на материале терминологии робототехники, поэтому представляется целесообразным кратко остановиться на вопросе формирования данной терминосистемы. Актуальность рассмотрения этого вопроса связана с тем, что робототехническая терминология возникла сравнительно недавно и система терминов этой области знания еще окончательно не сложилась. При всей существенности терминологических отличий, в первую очередь, в стремлении к однозначности, терминологическую систему определяют общеязыковые особенности: общность семантики в ряду вариантов, значение инвариантной модели, функциональность в определенной профессиональной среде, которая обусловливает границы терминосистемы.

Термины подчиняются словообразовательным, грамматическим и фонетическим правилам данного языка, создаются терминологизацией слов общенародного языка, заимствования или калькирования иноязычных терминоэлементов. В современной науке существует стремление к семантической унификации систем терминов одной науки в разных языках (однозначное соответствие между терминами разных языков) и к использованию интернационализмов в терминологии.

Глубинный признак терминов позволяет отделить их от других единиц языка и расчленить множество терминов. Этим глубинным признаком терминов является обозначение ими общих понятий. Выделяются термины категорий, общенаучные и общетехнические термины, межотраслевые термины, специальные термины.

Одна из особенностей терминосистемы заключается в использовании в качестве терминов имени существительного из-за своей семантической емкости, например: “Beim Wagen sind Sechsgang-Schaltgetriebe, Zentralverriegelung, Parktronic, Kühlbox Standard” [12, S. 3].

Номинализированный стиль более удобен и легок для авторов: безличность придает ему относительную свободу от грамматического времени. Он более абстрактен, чем вербальный.

Таким образом, полученный корпус текстов использовался в дальнейшем в качестве экспериментальных данных для решения ряда практических задач. Результаты обработки специальных текстов позволили окончательно разработать алгоритм, определяющий принадлежность текста к определенной предметной области. Алгоритм автоматической обработки естественного языка использован при разработке приложений, связанных с автоматической обучающей системой, которая включает пакеты тестов немецкого языка для технических наук.

Список литературы

1. Баранов А. Н. Введение в прикладную лингвистику / А. Н. Баранов. – М. : Эдиториал УРСС, 2001. – 358 с.

2. Баранов А. Н. Проблема репрезентативности корпуса текстов / А. Н. Баранов // Диалог : тр. Междунар. конф. – Режим доступа: http://www.dialog-21.ru/archivearticle, свободный. – Заглавие с экрана. – Яз. рус.

3. Беляева Л. Н. Автоматизация в лексикографии / Л. Н. Беляева, А. С. Герд, И. И. Убин // Прикладное языкознание : учеб. – СПб. : Изд-во СПб. ун-та, 1996. – С. 318–333.

4. Бородина О. А. Текст-реферат-концепт в аспекте номинации / О. А. Бородина // Прикладная лингвистика без границ : мат-лы Междунар. конф. – СПб., 2004.

5. Виландеберк А. А. Принципы и методы гармонизации терминологии на основе корпуса специальных параллельных текстов (на материале документов ООН) : автореф. дис. … канд. филол. наук / А. А. Виландеберк. – СПб., 2005. – 22 с.

6. Захаров В. П. Чешский национальный корпус текстов: организация и способы использования / В. П. Захаров // Корпусная лингвистика и лингвистические базы данных : докл. науч. конф. (г. Санкт-Петербург, 5–7 марта 2002 г.). – СПб., 2002.

7. Камшилова О. Н. Разработка корпуса текстов петербургских школьников (learner corpus): задачи и перспективы / О. Н. Камшилова // Прикладная лингвистика в науке и образовании : мат-лы IV Междунар. науч.-практ. конф. (г. Санкт-Петербург, 27–28 марта 2008 г.). – СПб., 2008.

8. Лингвистические ресурсы автоматизированного рабочего места филолога : монография / Л. Н. Беляева, А. А. Виландеберк, Л. А. Девель, И. Н. Ларченков, С. В. Молчанова, В. Р. Нымм, Т. Н. Петрова-Маслакова. – СПб. : Инфо-да, 2004. – 181 с.

9. Френсис У. Н. Проблемы формирования и машинного представления большого корпуса текстов / У. Н. Френсис // Новое в зарубежной лингвистике. – М. : Прогресс, 1983. – Вып. 14: Проблемы и методы лексикографии. – С. 334–352.

10. Conrad S. Corpus Linguistics Approach for Discourse Analysis / S. Conrad // Annual Review of Applied Linguistics. – 2002. – Vol. 22. – P. 75–96.

11. Hunston S. Corpora in Applied Linguistics / S. Hunston. – Cambridge : Cambridge Univ. Press, 2002. – XI, 241 p.

12. Zeitung: Berliner Zeitung. – 2004. – № 26.