Ю.Б. Смольская , С.И. Дубинин *

Принципы составления параллельного немецко-русского корпуса текстов

Smolskaja J., Dubinin S.

Prinzipien der Zusammenstellung eines deutsch-russischen Parallelkorpus

Im Aufsatz werden Fragen der Zusammenstellung eines modernen deutsch-russischen Korpus behandelt: Kriterien und Probleme der Textauswahl, Urheberrechte und elektronische Textverarbeitung. Dabei werden auch grundlegende Aspekte der Korpuslinguistik im Allgemeinen erwähnt. In Russland, das nach der Zahl der Deutschlernenden auf Platz eins in der Welt steht, ist die kontrastive deutsch-russische Korpuslinguistik, obwohl es ein typisch auslandsgermanistisches Thema ist, auf Grundlagen der modernsten EDV-Technologien trotzdem kaum ausgearbeitet. [17]

Уже к 1960-м гг. в лингвистике (т.н. "докомпьютерный период") наметился определенный кризис традиционных технологий сбора языкового материала в виде картотек, обусловленный недоступностью их данных широкому кругу исследователей, громоздкостью, неудобством обработки и хранения [1. С.112]. С началом компьютерной эры кризис механических способов сбора и обработки языковых данных лишь обострился. Электронная обработка и автоматизация информации последовательно и уверенно выходит на первый план.

* Ю.Б. Смольская, С.И. Дубинин, Самарский государственный университет

[17] Neben der größeren Projektarbeit mit den deutschen Kollegen von der Moskauer Lomonossow-Universität (Leiter *Prof. W.Gladrow* [6]) sind nur noch mehr oder weniger sporadische oder kooperierte Arbeitsgruppen an den Universitäten Voronesch und Halle (*Prof. I.Sternin*), an der Universität Twer (*Prof. B.Balin*), an der Pädagogischen Universitäten Kaluga und Tula (*Prof. A.Selenezkij*), an der Universität Uljanowsk (Prof. A.Fefilov), an der Moskauer Linguistischen Universität (*Prof. R.Potapowa*) zu erwähnen. Diese sind aber weniger auf die Schaffung von korpusbezogenen elektronischen Datenbanken aus Paralleltexten orientiert, sondern tendieren zur traditionellen typologisch-vergleichenden Sprachanalyse auf Grund der illustrativen Korpora, oder haben aktuell aufgehört zu existieren.

Одним из способов представления и анализа феноменов языковой системы стало составление корпусов текстов. В идеале корпус текстов может быть моделью функционирования языка в какой-либо дискурсивной сфере [1. С.135].

Первый опыт корпусной лингвистики на базе новых электронных технологий в институте немецкого языкознания университета г.Вюрцбурга состоялся в рамках крупного проекта 1984 – 1992 гг. "Wissensorganisierende und wissensvermittelnde Literatur des Mittelalters / DFG-SFB 226" (руков. проф. Н.Р.Вольф). В рамках данного проекта был составлен не только корпус объемных оригинальных текстов / фрагментов XIV века, но и были заложены основы электронной обработки филологических данных.

В ФРГ корпусная лингвистика давно и успешно развивается как в направлении систематизации разговорных, так и письменных источников [2. с.64-69]. На данный момент наиболее известными являются проекты института немецкого языка в Мангейме (IDS), где составляются не только текстовые корпусы, но и разрабатываются программы поиска грамматических и лексических явлений в корпусе и статистического анализа найденных данных, в частности COSMAS I, II, LIMAS (см.www.ids.de).

С сентября 2002 года разрабатывается совместный проект кафедры немецкой филологии СамГУ и института немецкого языкознания университета г. Вюрцбурга по созданию немецко-русского корпуса, в котором нашло бы отражение современное состояние как немецкого, так и русского языков.

В конце февраля 2003 г. в Вюрцбурге состоялась международная конференция по теме „Korpuslinguistik Deutsch: synchron, diachron, kontrastiv", на которой был представлен проект создания немецко-русского корпуса [3]. На конференции были также представлены проекты других билингвальных корпусов, в которых немецкий язык является исходным. Целью конференции было создание координационного центра по составлению текстовых корпусов (www.uni-wuerzburg.de/germanistik/spr/klde2003).

В проект в настоящее время уже включены девять крупных, функционально значимых и контактирующих европейских языков различных семей, а также типологической и ареальной принадлежности:

- индоевропейская семья

западный ареал / германская группа:

западногерманские языки

немецко-английский корпус (Великобритания, университет г.Эксетер; координатор Д.Льюис)

немецко-нидерландский корпус (ФРГ, университет г.Вюрцбурга; координатор Р.Леклерк)

скандинавские языки:

немецко-шведский корпус (Швеция, университет г.Умео; координатор А.Стедье)

романская группа

иберо-романская подгруппа

немецко-португальский корпус PORTDE (Португалия, университет г.Брага; координатор И.Диаш)

славянская группа

западно-славянская подгруппа

немецко-польский корпус (Польша, университет г.Ополе; координаторы К.Лазатович, Д.Пелка)

немецко-чешский корпус (Чехия, университеты г.Острава и г.Опава; координаторы Л.Ванкова, И Кратохвилова)

- неиндоевропейская семья

уральские языки / финно-угорская семья

северная группа прибалтийско-финских языков

немецко-финский корпус FINDE (Финляндия, университеты г.Хельсинки и г.Ювяскюля; координаторы Л.Колехмайнен и А.Янтти, университет г.Вюрцбурга - П.Шталь).

Интересен опыт германистов университета г.Осло (Норвегия), работающих с 1994 года над созданием мульти-лингвального корпуса переводных текстов (Oslo Multilingual Corpus / OMC) под руководством С.Йоханссона (http://www.hf.uio.no/german/sprik/english/corpus.shtml). Первоначально двуязычный англо-норвежский корпус из фрагментов художественных / нехудожественных текстов (до 15 тыс. слов) в оригинальной кодировке К.Хофланда (http://www.hf.uio.no/iba/projekt) включает дополнительно языки немецкий и французский [5].

Отсутствие надежной международной координации по составлению текстовых корпусов приводит к дублированию, несовместимости программ и чересполосице форматов, затрудняет конвертирование и обмен данными. Выработка единых критериев составления корпусов [1. С.136] является важной задачей в развитии корпусной лингвистики, имеющийся опыт пока не велик [7, 8, 10].

Русско-немецкий текстовый корпус, получивший "аграмматическое" название "DER-Korpus" планируется составить из двух частей (модулей):

1. переводной параллельный субкорпус, состоящий из оригинальных немецких / русских текстов (беллетристика последних десятилетий и специальные научные тексты) и их русских / немецких переводов соответственно;

2. контрастивный субкорпус из текстов СМИ (газеты и периодика, в первую очередь информативного типа и по одинаковой тематике), в том числе интернет-версии.

Предлагаемая модель "DER-Korpus" была уже апробирована в "вюрцбургской ассоциации". Так, контрастивный корпус германистов из финского университета г.Ювяскюля (руков. А.Янтти) состоит из переводных параллельных текстов современной художественной литературы (немецкие авторы: Г.Грасс, К.Хайн, Б.Штраус; финские авторы: П.Хаавикко, А.Идстрем и А.Туури), дополненных текстами международных новостей [9].

Большие трудности представляет отбор художественных текстов для включения их в корпус. Литература последних десятилетий, как немецкая, так и русская не слишком активно переводится на русский и немецкий языки.

Отсутствие единых критериев отбора текстов также затрудняет работу.

МакЕнери [4] выделяет в критериях отбора текстов в первую очередь репрезентативность выбранного текста для языка, но критериев ее определения не дает, таким образом, определение репрезентативности является субъективным критерием, и в большой степени зависит от личных взглядов исследователей на литературу.

Трудность также представляет поиск текстов специальной научной литературы, т.к. основная масса научной литературы, как на немецкий, так и на русский языки переводится с английского.

На следующем этапе отбора литературы составители „DER-Korpus" столкнулись с проблемой поиска переводов выбранных текстов. Для корпуса важен не только сам перевод как таковой, но и время его создания, насколько актуален этот перевод по отношению ко времени написания оригинала. Качество перевода при этом тоже играет важную роль. Хотя именно качество и полноту перевода можно установить лишь в процессе электронной обработки текста при синоптизации оригинала и перевода, когда можно точно выявить все пропущенные и/или добавленные в переводе фрагменты текста.

Так как немецко-русский корпус предполагается как параллельный корпус, то необходимо также согласовывать и

жанровую специфику немецких и русских оригинальных текстов и их временную соотнесенность между собой, что создает дополнительные трудности при отборе текстов в корпус.

Все критерии отбора текстов для составления корпуса, таким образом, являются субъективными и в большой степени зависят от пристрастий и взглядов лингвистов, работающих над проектом.

При составлении немецко-русского корпуса возникли в первую очередь внелингвистические проблемы такие, как урегулирование авторских прав, получение разрешения на работу с текстами, на их электронное тиражирование, хотя бы в рамках рабочей группы проекта.

Форматом электронной кодировки текстов немецко-русского корпуса был выбран ТУСТЕП (TUSTEP-Tübinger System von Textverarbeitung-Programmen, см. www.itug.de). На основе этой системы в Вюрцбурге разработаны и успешно применяются ряд программ, облегчающих машинную обработку текстов. При этом также соблюдается установка на создание единого формата билингвальных корпусов, что должно в дальнейшем способствовать созданию переводных корпусов, в которых сравнивались бы переводы на разные языки произведений немецких авторов, а также составлению мультилингвальных корпусов по образцу проекта лингвистов университета Осло. Реализация проекта позволит преодолеть и наметившийся отрыв отечественных и зарубежных исследователей в данной области [11, с. 9-11], в частности, в германистике.

Библиографический список

1. Баранов А.Н. Введение в прикладную лингвистику. М., 2001.

2. Баранов А.Н., Добровольский Д.О. Немецкая корпусная лингвистика // Вестник МГУ. Серия 15: лингвистика и межкультурная коммуникация. 1998. № 1.

3. Vadijaew S., Smolskaja J., Dubinin S. Probleme und Schwerpunkte eines modernen deutsch-russischen Textkorpus // Tübingen, 2004 (im Druck).

4. MacEnery, Tony Corpus linguistics / Tony McEnery and Andrew Wilson. Edinburgh : Edinburgh University Press, 1996.

5. Multilingual Corpora in Teaching and Research / Ed. Botley. Amsterdam, 2000.

6. Gladrow W. Russisch im Spiegel des Deutschen. Leipzig, 1986.

7. Корпусная лингвистика и лингвистические базы данных. СПб., 2002.

8. Тихонова К.А. Контрастивное исследование баз данных (на материале немецких и русских неологизмов XX в.). Автореф. канд. дисс. М.: МГЛУ, 2002.

9. Wolf N.R. Plädoyer für eine Korpuslinguistik // http://kds.german.or.kr/wwwb/data/board9/Prof. Wolf Korpuslinguistik.doc.

10. Гак В.К. О контрастивной лингвистике // Новое в зарубежной лингвистике / Ред. В.К.Гак. М., 1989. Вып. XXV: контрастивная лингвистика.

11. Выгурский К.В., Пильщиков И.А. Филология и современые информационные технологии // Известия РАН. Серия литературы и языка, 2003. №2.