ОБЩАЯ ТЕОРИЯ ПЕРЕВОДА

Цзинь Ифан,
аспирант Высшей школы перевода (факультета) МГУ
имени М.В. Ломоносова;
e-mail: 1554583267@qq.com

КОРПУСА МЕЖЪЯЗЫКОВЫХ БОЛЬШИХ ДАННЫХ И ПЕРЕВОД

Корпуса больших данных на сегодняшний день являются универсальным средством обработки, анализа и поиска необходимой информации. Если раньше для проведения научных исследований требовалось проделать большой объём работы: подобрать материал, проанализировать и обработать большие текстовые объёмы с целью выявления некоторых языковых закономерностей, то в настоящее время благодаря методам компьютерного анализа текста значительно сокращается время на работу и обработку объёма полученных данных. Кроме того, компьютерный анализ текста помогает избежать неточностей и ошибок в подсчётах, способствует установлению языковых закономерностей, основанных не на теоретических, а на эмпирических данных. Благодаря использованию технических средств происходит формирование новых межъязыковых корпусов, что представляет собой сложную задачу, часто с множественными проблемами.

Ключевые слова: перевод, переводческие технологии, межъязыковые большие данные.

Важнейшая роль в современном мире, где повсеместно используются сети и электронные устройства, отводится большим данным и методам их анализа. Происходит непрерывное объединение возможностей больших данных, аналитики и машинного/глубинного обучения. В сущности, само понятие «большие данные» подразумевает работу с огромными потоками информации, которая регулярно обновляется и поступает из разных источников с целью увеличения эффективности её функционирования, создания новых продуктов и повышения конкурентоспособности самого ресурса. Иными словами, большие данные — это нематериальный ресурс, который объединяет технику и технологии, чтобы анализировать огромные массивы данных.

Примером структурированного представления больших данных являются лингвистические корпуса. В принципе, любой набор,

в структуру которого включено более одного текста, может приравниваться к корпусу. Однако уникальность корпусной лингвистики заключается в использовании современных компьютерных технологий, в сборе и структурировании большого объёма языковых данных, в методах, применяемых для их обработки, а также предполагаемых стратегий их применения во всех видах деятельности, связанных с языком, в частности, в процессе перевода, и исследованиях результатов перевода и их оценке.

Межъязыковые большие данные представляют собой информацию, сформированную по некоторым выборкам из области реализации языковой системы, содержащей феномены, которым требуется лингвистическое описание и межъязыковое представление

«Представление о больших данных как о качественно новом состоянии глобальной информационной системы заставляет задуматься над тем, к каким изменениям переводческой деятельности и образовательных моделей подготовки переводчиков может привести этот феномен», — пишет Н.К. Гарбовский [Гарбовский, 2018: 38].

«Круг вопросов, касающихся корреляции «больших данных» и перевода, — продолжает исследователь, — сводится на первый взгляд к двум объективно существующим сторонам этого взаимодействия: во-первых, речь идёт об обработке информации, воспринимаемой переводчиком для принятия решения и достижения успешной межъязыковой коммуникации и, во-вторых, это генерирование в результате переводческой деятельности новой информации, пополняющей так называемые межъязыковые большие данные» [там же].

Во втором случае «проявляются связи между актантами двуединой подсистемы, включающей в себя текст оригинала и текст перевода как преобразованную субстанцию» [см.: Гарбовский, Костикова, 2018].

Таким образом, в результате переводческой деятельности происходит постоянное обогащение межъязыковых больших данных, представленных как в параллельных лингвистических корпусах, структурированных определённым образом на основании избранных критериев, так и неструктурируемых, поступающих «из множества сайтов, социальных сетей, блогов, переводческих форумов, общения с коллегами, корпоративных архивов документов, средств массовой коммуникации и пр.» [Гарбовский, 2018: 38].

Разумеется, в этом многогранном информационном процессе особое значение имеют структурированные данные, сформированные как в национальные, так и в параллельные межъязыковые лингвистические корпуса.

Лингвистический корпус представляется собранием текстов, объединённых логическим замыслом или идеей, которые воплощаются в правилах организации текстов, алгоритмах и программах их анализа, идеологической и методологической базах [см.: Рыков, 2002: 29].

В трактовке зарубежных исследователей корпус представлен как собрание текстов, отобранных в соответствии с чётко обозначенными языковыми критериями и для использования в качестве определённой модели языка [McEnery, Wilson, 2001: 15]. Т.е. с их точки зрения — это набор данных, представленный в электронном виде и имеющий определённую композицию.

В корпусе содержится большая коллекция специально подготовленных (в соответствии с поставленной исследовательской задачей), имеющих определённую разметку и структуру, представленных в унифицированном виде образцов текстов, охватывающих различные разновидности языков. Как правило, корпус снабжают специальным поисковым интерфейсом, благодаря которому пользователи могут осуществлять необходимые фрагменты текста по заданным параметрам.

Практика разработки и использования корпусов текстов показала, что создание универсального массива данных представляет собой довольно длительный и трудоёмкий процесс. Его можно сформировать, лишь исходя из задачи и целей, поставленных в исследовании, которое предполагается проводить с помощью данных корпусов, и, соответственно, определить его тип, правила отбора информации, способ и степень её обработки. Поэтому на сегодняшний день создано множество корпусов, предназначенных для различных типов исследований, в основу классификации которых положены разнообразные характеристики [Захаров, Богданова, 2013: 16]. При этом практически во всех классификациях присутствует деление корпусов на одноязычные и многоязычные, а сами данные создаются и размещаются в электронном виде. Подобного рода корпуса активно используются как для лингвистических исследований, так и практической переводческой деятельности, а также для решения каких-либо прикладных задач. Национальные корпуса создаются лингвистами (специалистами по корпусной лингвистике, представляющей быстро развивающуюся область современного языкознания) для проведения научных изысканий и с целью обучения языку. Большинство стран уже имеют свои корпуса данных, которые отличаются по степени наполненности информацией и уровнем её научной обработки. Общепризнанным образцом среди созданных на сегодняшний

день является Британский национальный корпус (BNC) — сборник из 100 миллионов слов из письменного и устного языка из широкого круга источников, предназначенных для широкого изучения британского английского языка. Это одноязычный корпус, так как он использует данные на современном британском английском, а не на других языках, используемых в Великобритании. Однако небританские слова английского и иностранного языка встречаются в корпусе.

В открытом американском национальном корпусе (OANC) содержится электронная коллекция американского английского языка, включая тексты всех жанров и стенограммы устных данных, выпущенных с 1990 года. Все данные и аннотации полностью открыты и не ограничены для пользования. Существует также несколько корпусов испанского языка, итальянский, украинский, французский, греческий и польский корпуса.

В России функционируют сразу несколько корпусов, подчинённых разным исследовательским задачам.

Национальный корпус русского языка (НКРЯ) — информационно-справочная система, основанная на собрании в электронной форме русских текстов разного периода и разных жанров, — существует в Интернете с 2003 года. Чуть более двенадцати лет тому назад корпус включал в себя тексты общим объёмом около 140 миллионов словоупотреблений [см.: Ягунова, 2007: 77]. Сегодня корпус современного русского языка уже включает более 600 млн словоупотреблений и охватывает период от XVIII до XXI века «в разных социолингвистических вариантах — литературном, разговорном, просторечном, отчасти диалектном» [http://www.ruscorpora.ru/corpora-intro.html]. В этот корпус включены непереводные (оригинальные) произведения художественной литературы, высокой культурной значимости, а также тексты, интересные с точки зрения употреблений русского языка в литературе разных речевых жанров: научная и научно-популярная литература, публицистика, публичные выступления, документы и пр.

Национальный корпус русского языка включает целый ряд подкорпусов (основной, синтаксический, корпус современных СМИ, а также параллельные корпуса, «в которых можно найти все переводы для определённого слова или словосочетания на русский язык или с русского языка» [Там же]: англо-русский и русско-английский, немецко-русский и русско-немецкий, французско-русский и русско-французский, испанско-русский и русско-испанский, итальянско-русский и русско-итальянский, и некоторые другие языковые комбинации.

Параллельные корпуса представляют собой особый тип корпуса, в котором текст на русском языке сопоставляется с его переводом на другой язык или, наоборот, тексту на иностранном языке сопоставлен его перевод на русский язык. Между предложениями оригинального и переводного текстов с помощью специальной процедуры установлено соответствие, называемое выравниванием.

Выравненный параллельный корпус представляет собой важный инструмент для научных исследований в области теории и методологии перевода [см.: Добровольский и др., 2003–2005: 263–296].

В Национальном корпусе русского литературного языка (НКРЛЯ) представлены отобранные определённым образом тексты (с опорой на филологическую экспертизу) на русском языке, необходимые для осуществления поиска в них лексических, грамматических, стилистических единиц и явлений, интересующих пользователя. В состав данного ресурса входят Тюбингенский корпус — первый морфологически аннотированный корпус, появившийся в интернете в открытом доступе Opencorpore — включает распространяемые тексты, размечаемые силами волонтёров, для учебных целей — Хельсинкский аннотированный корпус (ХАНКО), система баз данных Интегрум, содержащая электронный архив СМИ России, материалы законодательного характера, справочники, аналитические исследования и обзоры; биржевые и фондовые новости, полные тексты произведений русской классики, часть которых содержит параллельный перевод на английский язык и озвученные фрагменты в исполнении профессиональных актёров [Копотев, 2014: 230].

Сказанное позволяет сделать вывод, что тип корпуса и структура обусловлены его предназначением и представляют собой мощные информационные ресурсы, которые могут быть использованы в проведении различных лингвистических направлений. Например, как источник данных для лексикографии, поскольку позволяет пересмотреть текстовые материалы гораздо быстрее за счёт процесса автоматизации, помогающего сформировать массив согласно течению времени.

Вместе с тем возникает дилемма: когда новости и информация, собранные со всех уголков мира, передаются в режиме перекрёстного языка, в режиме реального времени и больших данных, предприятия в процессе глобализации сталкиваются с двумя проблемами: как нарушить языковые барьеры и получать данные в режиме реального времени по всему миру для принятия оперативных решений. В данном случае речь идёт о создании корпуса

межъязыковых больших данных, который будет обладать многомиллионными формами словоупотреблений на разных языках. Теоретическое значение разработки данного корпуса заключается в том, что появится возможность проводить на большом объёме текстов сравнительный анализ лексических и грамматических средств выражений на примере двух или более языков, например, русского и китайского. Между данными языковыми системами всего существовала большая разница, в первую связанная с морфологией. В отличие от русского языка с категориями рода, числа и падежа у имён существительных, времени, вида и наклонения, и словообразованием различного типа, китайский язык характеризуется фактически полным отсутствием морфологии [Тао, Захаров, 2015: 18]. Поэтому целью разработки русско-китайского корпуса должно стать создание программно-лингвистической платформы, необходимой для исследований в области перевода с русского языка на китайский, с китайского языка на русский и для обучения русскому языку.

Примечательно, что сегодня корпус русского языка в комбинации с английским включает 1 608 376 предложений, 24 681 277 слов, а в комбинации с китайским — 15 735 предложений, 279 478 слов. Этот факт говорит о необходимости более интенсивного обогащения данными параллельных корпусов в комбинации русского и китайского языков.

В Китае межъязыковые большие данные стали новым фактором производства, и в настоящее время примером создания подобного рода корпуса можно назвать платформу YeeSight, это аналитическая платформа с большими данными. Она интегрирует, обеспечивает и анализирует кросс-языковые данные в массовом масштабе, позволяет пользователям плавно взаимодействовать с данными и принимать более правильные решения. Это интеллектуальный движок, объединяющий машинный перевод и семантический поиск.

Ещё один пример — YeeCloud — профессиональная языковая сервисная платформа, выпущенная компанией China Translation Language Technology Co., Ltd. Она обеспечивает перевод информации, предоставляя ответ в режиме реального времени. Если данный корпус разработан для удовлетворения потребностей пользователей для повышения точности и эффективности перевода с помощью человеческого и машинного перевода, открытого API, многоязычных инструментов и других технологий, то YeeSight построен на миллионах источников данных на более чем 60 языках от независимых доменных сайтов в более чем 200 странах мира, с более чем 30 миллионами ежедневных обновлений новостных данных, более 500 миллионов данных в социальных сетях и общедоступных текстов

на базе Интернета и мультимедийных данных на основе Интернета. Благодаря массивным и точным кросс-языковым данным YeeSight предлагает различные пользовательские решения для больших данных. Что касается предприятий, YeeSight даёт представление о пути распространения новостей, тенденциях продуктов и технологий, целевых групп и рынков, а также маркетинговых стратегий конкурентов. Она также предупреждает о потенциальных рисках на рынке для предприятий, чтобы оперативно излагать или корректировать свои стратегии развития и добиваться устойчивого роста прибыли.

Сейчас YeeSight мониторит сотни миллионов источников информации, включая новостные порталы и социальные сети по всему миру. Алгоритм искусственного интеллекта отражает общую статистику: объем охваченных данных, их распределение по каналам (новости, соцсети, форумы), эмоциональную окраску сообщений (негативные, позитивные, нейтральные), распределение по странам и ключевые слова.

По мнению вице-президента компании-разработки платформы GTCOM Чжан Сяодань, искусственный интеллект в сочетании с языками и большими данными должны стать новыми двигателями инноваций и развития, поэтому необходимо развивать межъязыковое взаимодействие в области больших данных, используя технологии искусственного интеллекта.

Сказанное, однако, требует решения ряда проблем, которые, несомненно, возникнут при создании корпуса межъязыковых больших данных. Например, проблемы перевода в процессе создания корпуса межъязыковых больших данных. Переводческое исследование на базе корпуса межъязыковых больших данных сталкивается с двумя основными техническими проблемами: 1) построение базы языкового материала, особенно корпуса межъязыковых больших данных, затруднено, а количество и тип базы языкового материала в стране и за рубежом не могут эффективно удовлетворять исследовательские потребности; 2) контекст, обеспечиваемый базой языкового материала, недостаточен, а репрезентативность корпуса не идеальна. Корпуса межъязыковых больших данных является незаменимой исследовательской платформой для исследования перевода на базе языкового материала. Его строительство технически сложно и зачастую требует больших трудовых, материальных и финансовых ресурсов. В результате немногие лица или организации могут самостоятельно разрабатывать корпус межъязыковых больших данных. Кроме того, промежуток времени и количество языковых пар в корпусе, установленных для изучения перевода к настоящему времени, как правило, ограничено, что затрудняет ис-

следователям понимание норм перевода в определённый исторический период или соответствие между другими языковыми парами. Например, жёлтый цвет (黄色) на китайском языке представляет не только цвет, но в древнем Китае жёлтый представляет императора, а в современном Китае жёлтый представляет нездоровую культуру. Очевидно, что выводы исследований, полученные с использованием этих параллельных корпусов, не являются достаточно универсальными, а нормы перевода или языковые соответствия, основанные на корпусе, не являются достаточно объективными и научными. Как справедливо отмечает Н.К. Гарбовский, важно знать, каким «инструментарием» владеет переводчик в процессе работы над текстом, какие приёмы и методы использует, чтобы не допустить возникновения противоречивых ситуаций [Гарбовский, 2007: 8]. Ведь перевод должен передавать не только то, что выражено подлинником, но и то, как это выражено в нём, что может оказаться довольно затруднительным, поскольку разные языки отражают действительность по-разному [Рецкер, 1974: 7]. Следовательно, двумя важнейшими проблемами при создании корпуса межъязыковых больших данных могут стать вопросы, связанные с переводимостью и инвариантностью, поскольку в процессе перевод речь идёт о субъективном выборе средств конкретного переводчика [Malmkjaer, 1998: 536].

Однако база языкового материала в основном состоит из литературного языкового материала. В политической, экономической, юридической и коммерческой сферах люди собирают меньше корпусных данных. Например, в политической сфере российско-китайская база данных в основном отражается в корпусе ВСНП (Всекитайское собрание народных представителей) и НПКСК (Народный политический консультативный совет Китая).

Корпусе ВСНП и НПКСК (часть 1)	
反腐倡廉	борьба с коррупцией, разложением и утверждение неподкупности
社会保障	социальное обеспечение
教育改革	реформы образования
住房制度	жилищная система
就业收入	занятость и доходы
环境保护	защита окружающей среды

国防军改	реформа национальной обороны и армии
乡村振兴	возрождение деревни
大国外交	дипломатия крупной державы с китайской спецификой
网络强国战略	стратегия интернет-державы
共建网络空间命运共同体	совместное создание сообщества единой судьбы в киберпространстве
双创（大众创业 万众创新）	широкая предпринимательская инициатива и массовая инновационная деятельность
工匠精神	дух мастера
精准扶贫、精准脱贫	точечная помощь малоимущим и целенаправленная ликвидация бедность
社会主义核心价值观	концепция социалистических базовых ценностей
命运共同体	Сообщество единой судьбы
生态补偿制度	Система экологической компенсации

Ещё одной проблемой может стать автоматический анализ естественного языка, поскольку он не безошибочен и многозначен и, как правило, предлагает несколько вариантов анализа для одной лексической единицы. Подобного рода инвариантность может вызвать ряд проблем касательно трактовки слов и выражений. Например, когда вы вводите китайские слова 胡同 (старая, небольшая улочка) в систему Yeesight, результатом перевода будет проспект. Снятие неоднозначности должно стать одной из важнейших и сложнейших задач перевода. Для этого следует использовать автоматические и ручные способы, ориентироваться на новые принципы разработки систем, которые бы минимизировали вмешательство человека. В отличие от ручной корректировки, автоматическое разрешение использует информацию более высокого уровня и компонует материл с применением статистических методов.

Создание корпуса межъязыковых больших данных — это масштабная, многонаправленная, инициативная работа, сложный процесс, отнимающий много времени и подверженный ошибкам. Он требует эффективной системы обработки данных, как с технической точки зрения, так и касательно их хранения и анализа. И чем больше массив данных, тем больше возникает проблем с его переводом, адаптацией и управлением.

Список литературы

Гарбовский Н.К. Перевод: cognition и communicatio в эпоху «больших данных» // Когнитивные исследования языка. Вып. XXXIV. Cognition и communicatio в современном глобальном мире. Материалы VIII Международного конгресса по когнитивной лингвистике. 10–12 октября 2018. Москва — Тамбов. 2018.

Гарбовский Н.К. Теория перевода: Учебник / Н.К. Гарбовский. М., 2007. 544 с.

Гарбовский Н.К., Костикова О.И. Перевод и общество // Вестник Моск. ун-та. Сер 22. Теория перевода. № 1, М., 2018.

Добровольский Д.О., Кретов А.А., Шаров С.А. Корпус параллельных текстов: архитектура и возможности использования // Национальный корпус русского языка: 2003–2005. М.: Индрик, 2005. С. 263–296.

Захаров В.П., Богданова С.Ю. Корпусная лингвистика / В.П. Захаров, С.Ю. Богданова. 2-е изд., перераб. и дополн. СПб.: СПбГУ, 2013.

Копотев М. Введение в корпусную лингвистику: Учебное пособие для студентов филологических и лингвистических специальностей университетов / М. Копотев. М., 2014. 230 с.

Рецкер Я.И. Теория перевода и переводческая практика / Я.И. Рецкер. М., 1974.

Рыков В.В. Корпус текстов как реализация объектно-ориентированной парадигмы / В.В. Рыков // Труды Международного семинара Диалог 2002. М.: Наука, 2002.

Тао Ю., Захаров В.П. Разработка и использование параллельного корпуса русского и китайского языков / Ю. Тао, В.П. Захаров // НТИ. Сер. 2. ИНФОРМ. ПРОЦЕССЫ И СИСТЕМЫ, 2015. № 4. С. 18–29.

Ягунова Е.В. Использование Национального корпуса русского языка для решения задач моделирования речевой деятельности / Е.В. Ягунова // Национальный корпус русского языка и проблемы гуманитарного образования. Материалы международной научной конференции. Москва, 19–20 апреля 2007. М., 2007. С. 77–79.

Malmkjaer K. Love thy Neighbour: Will Parallel Corpora Endear Linguists to Translators? K. Malmkjaer. Meta: journal des traducteurs. Meta: Translators' Journal, 1998. Vol. 43, No. 4, pp. 534–541.

McEnery T., Wilson A. Corpus Linguistics. T. McEnery, A. Wilson. 2nd ed. Edinburgh: Edinburgh University Press, 2001.

Jin Yifang,
Postgraduate Student at the Higher School of Translation and Interpretation, Lomonosov Moscow State University, Russia;
email: 1554583267@qq .com

CORPUS OF INTERLINGUAL BIG DATA AND TRANSLATION

The article develops the idea that How do different countries develop their own corpus and the development of interlingua big data. Nowadays computer technology leads the development of big data and in this process various problems that may exist in creating a big data corpus, For example, the phenomenon of non-correspondence between different languages due to cultural differences. This article focuses on the relationship between translation theory and big data corpus, and the possible impact of the development of big data corpus on translation theory.

Key words: translation, translation technology, interlanguage big data.

References

Dobrovol'skij D.O., Kretov A.A., Sharov S.A. Korpus parallel'nyh tekstov: arhitektura i vozmozhnosti ispol'zovanija [Corpus of parallel texts: architecture and use]. Nacional'nyj korpus russkogo jazyka: 2003–2005. Moscow: Indrik, 2005, pp. 263–296 (In Russian).

Garbovskij N.K. Perevod: cognition i communicatio v jepohu "bol'shih dannyh" [Translation: cognition and communication in the era of "big data"]. Kognitivnye issledovanija jazyka. Vyp. XXXIV. Sognition i communicatio v sovremennom globalnom mire. Materialy VIII Mezhdunarodnogo kongressa po kognitivnoj lingvistike. October 10–12.2018. Moscow — Tambov. 2018 (In Russian).

Garbovskij N.K. Teorija perevoda: Uchebnik [Theory of Translation]. Moscow, 2007. 544 p. (In Russian).

Garbovskij N.K., Kostikova O.I. Perevod i obshhestvo [Translation and Society] // *Vestnik Mosk. un-ta. Ser 22. Teorija perevoda*. No. 1. Moscow, 2018 (In Russian).

Jagunova E.V. Ispol'zovanie Nacional'nogo korpusa russkogo jazyka dlja reshenija zadach modelirovanija rechevoj dejatel'nosti [Using the National Corpus of the Russian language to solve problems of speech activity modeling]. Materialy mezhdunarodnoj nauchnoj konferencii. Moscow, 19–20 April 2007. Moscow, 2007, pp. 77–79 (In Russian).

Kopotev M. Vvedenie v korpusnuju lingvistiku: Uchebnoe posobie dlja studentov filologicheskih i lingvisticheskih special'nostej universitetov [Introduction to corpus linguistics]. Moscow, 2014. 230 p. (In Russian).

Recker Ja.I. Teorija perevoda i perevodcheskaja praktika [Translation Theory and Translation Practice]. Moscow, 1974 (In Russian).

Malmkjaer K. Love thy Neighbour: Will Parallel Corpora Endear Linguists to Translators? Meta: Translators' Journal.1998. Vol. 43, No. 4, pp. 534–541.

McEnery T., Wilson A. Corpus Linguistics. 2nd ed. Edinburgh: Edinburgh University Press, 2001.

Rykov V.V. Korpus tekstov kak realizacija ob'ektno-orientirovannoj paradigm [Corpus of texts as the implementation of an object-oriented paradigm]. Trudy Mezhdunarodnogo seminara Dialog, 2002. Moscow: Nauka, 2002 (In Russian).

Tao Ju., Zaharov V.P. Razrabotka i ispol'zovanie parallel'nogo korpusa russkogo i kitajskogo jazykov [Development and use of a parallel corps of Russian and Chinese languages]. *NTI. Ser. 2. INFORM. PROCESSY I SISTEMY,* 2015. No. 4, pp. 18–29 (In Russian).

Zaharov V.P., Bogdanova S.Ju. Korpusnaja lingvistika [Corpus linguistics]. 2-e izd., pererab. i dopoln. St. Petersburg: St. Petersburg GU, 2013 (In Russian).