O'zbekiston Respublikasi Vazirlar Mahkamasi huzuridagi O'zbek tilini rivojlantirish jamg'armasi

O'zbekiston Respublikasi Raqamli ta'lim texnologiyalar vazirligi

O'zbekiston Respublikasi Oliy ta'lim, fan va innovatsiyalar vazirligi

Sankt-Peterburg davlat universiteti

Istanbul texnika universiteti

Toshkent davlat o'zbek tili va adabiyoti universiteti

Samarqand davlat universiteti

Toshkent axborot texnologiyalari universiteti Samarqand fiiliali

MUHAMMAD AL-XORAZMIY NOMIDAGI TOSHKENT AXBOROT TEXNOLOGIYALARIUNIVERSITETI SAMARQAND FILIALI

“O‘ZBEK TILINING MILLIY KORPUSI: MUAMMOLAR VA VAZIFALAR”

MAVZUSIDAGI XALQARO
ILMIY-AMALIY ANJUMANI MA’RUZALAR TO‘PLAMI
2023 yil 25 mart

PROCEEDINGS
of the International Scientific and Practical Conference

“THE NATIONAL CORPUS OF THE UZBEK LANGUAGE: PROBLEMS AND TASKS”

March 25, 2023

ПРИМЕНЕНИЕ ПАРАЛЛЕЛЬНЫХ КОРПУСОВ В СОЗДАНИИ МНОГОЯЗЫЧНЫХ ПЕРЕВОДНЫХ СЛОВАРЕЙ: ПРОБЛЕМЫ И ПЕРСПЕКТИВЫ
APPLICATION OF PARALLEL CORPORA IN CREATING MULTILINGUAL TRANSLATION DICTIONARIES: ISSUES AND PROSPECTS

Авезов Сухроб Собирович
*Бухарский государственный университет, Бухара, Узбекистан
1990senigama@gmail.com

Аннотация: В данной статье рассматриваются трудности, связанные с созданием лексикографической компоненты автоматизированного рабочего места лексикографа и переводчика, а также общие проблемы, возникающие в области прикладной лексикографии, которая основана на использовании параллельных корпусов текстов, специализированных по тематике и предметной области.

Abstract. This article examines challenges related to the development of the lexicographic component of an automated workstation for lexicographers and translators, as well as the broader issues that arise in applied lexicography, which relies on the utilization of specialized parallel corpora of texts within particular subjects and subject areas.

Ключевые слова: корпус параллельных текстов, лексикографическая база, машинный перевод, терминология, сравнительный анализ, автоматизация, перевод.

Keywords: corpus of parallel texts, lexicographic base, machine translation, terminology, comparative analysis, automation, translation.

В настоящее время проблема получения точного и оперативного перевода научно-технической литературы стала особенно актуальной. Однако, несмотря на длительные дискуссии, научное сообщество в целом и переводчики в частности пока не готовы обрабатывать огромные объемы информации на разных языках ни технически, ни психологически. В то же время давно известно, что использование компьютеров при обработке текстов и потоков документов на разных языках может помочь специалистам искать, переводить и использовать необходимую информацию более эффективно[1].

Для эффективного извлечения информации и создания переводных словарей необходим комплекс лингвистических и программных инструментов, которые поддерживают работу переводчика и терминолога. Такой комплекс может быть представлен в виде специально организованного автоматизированного рабочего места (АРМ). Ключевыми элементами АРМ являются резидентные словари, тезаурусы, системы проверки орфографии и доступа к информации через различные сети передачи данных. Грамотное использование такого комплекса инструментов должно стать обязательным не только для переводчиков, но и для специалистов в различных областях знаний.

Анализ существующих переводных словарей, в том числе тех, которые включены в различные автоматизированные словарные системы, показывает, что они не отвечают современному уровню развития науки и техники и не отражают основные направления развития отраслей знаний. Такая ситуация обусловлена не только естественным отставанием лексикографии, связанным с необходимостью обработки больших объемов современной информации, но и традиционным подходом к созданию словарей, который основывается на уже опубликованных источниках и результате анализа текстов.

Профессиональная лексикография, ориентированная на конкретную тематику, требует применения технологического комплекса для создания и поддержки словарей и словарных баз данных, которые позволяют объединить выполнение нескольких задач, таких как: выделение терминов из текстов, получение статистических данных о частоте использования терминов в обрабатываемых текстах, просмотр конкорданса и терминосочетаний с заданными параметрами контекстного окна, автоматическое пополнение словарей, а также построение онтологии.

Для поддержания актуальности переводных словарей, которые использует переводчик, необходимо постоянно отслеживать новые термины. Для этого можно использовать методы лингвистического и лингвостатистического анализа, а также метрики, которые позволяют выделять терминологические словосочетания из текста. Эти метрики могут оценивать различные характеристики словосочетаний, такие как устойчивость сочетаемости лексических единиц, терминологичность словосочетаний и характерные особенности для конкретного корпуса текстов или терминологии языка для специальных целей.

Создание обширных терминологических ресурсов, как правило, возлагается на специалистов-терминологов, однако данный процесс требует значительных ручных трудозатрат. Это не соответствует современным требованиям, поскольку терминологам необходимо быстро и стандартизированно реагировать на новые или еще не зарегистрированные термины для обеспечения требований по обработке информации. Однако различия в исходных текстах, уровнях специализации, целях и профилях конечных пользователей и уровнях автоматизации, объясняют отсутствие универсальных методов для извлечения терминов из текстов.

При интеграции АРМ в образовательную среду, ориентированную на языковое обучение, важно учитывать возможности, которые она предоставляет различным категориям пользователей: самостоятельной работы, экспертам и менеджерам общего лингвистического ресурса. Также следует учитывать, что этот лингвистический ресурс может формироваться на основе словарей, создаваемых в индивидуальных АРМ.

В АРМ для работы терминолога и лексикографа следует предусмотреть возможность использования заранее выбранной системы управления контентом, которая обеспечивает сохранение данных. Выбор такой системы, которая является ключевой составляющей высокотехнологичной образовательной среды, должен осуществляться менеджерами системы. Кроме того, АРМ должен иметь доступ к онлайновым инструментам для работы с терминологией. Каждый индивидуальный АРМ также должен иметь свои собственные ресурсы, которые используются совместно с онлайновыми.

Современный подход к созданию лексикографических ресурсов включает использование корпуса реальных текстов, который может быть использован как база данных для решения как исследовательских, так и практических задач. Корпусы письменных текстов содержат сами тексты и их разметку в соответствии с форматом и предложениями по результатам парсинга, что позволяет определить принадлежность лексических единиц к конкретным частям речи. Эти тексты могут быть использованы для создания конкордансов, словарей слов и словосочетаний в случае одноязычного корпуса, а также для создания многоязычных лексиконов и конкордансов в случае параллельных массивов [2].

При разработке переводного словаря с использованием корпуса текстов необходимо определить принципы отбора образцов для создания исследовательского параллельного корпуса текстов, а также достаточный объем выборки. Требования к разметке текстов в корпусе должны быть установлены, и базовая лингвистическая информация должна быть получена. Для работы с корпусом требуются различные информационные технологии, включая системы машинного перевода, переводческие памяти, средства выравнивания параллельных текстов, редакторы разметки (тегирования), средства формального извлечения терминов из текста и т.д.

В рамках автоматизации лексикографических процессов, средства информационных технологий могут быть направлены на выбор и обработку терминов и понятий, работу с многоязычными или одноязычными ресурсами, а также на работу с конкретной языковой парой. Для этого рекомендуется интегрировать подобные средства в автоматизированное

рабочее место лексикографа, которое позволит осуществлять запись, обработку, сохранение и использование различных лингвистических и лексикографических данных.

При создании исследовательского корпуса текстов для лексикографических задач процесс формирования выборочной совокупности включает следующие этапы:

а. определение подобластей предметной области, связанных со структурой языка для специальных целей, на основе выбранной классификации знаний;

б. отбор текстов различных жанров, сопоставимого объема и соответствующих определенным критериям;

в. разметка текстов, включающая фиксацию коллокаций и других лингвистических характеристик[3].

В корпусной лингвистике существует различие между корпусами параллельных текстов и корпусами псевдопараллельных текстов. Они отличаются принципами отбора текстов. При создании псевдопараллельных текстов отбор может быть осуществлен на основе достаточно ясных критериев. Однако, при создании корпуса параллельных текстов для последующего терминологического анализа и извлечения пар типа термин -перевод, качество выбранных переводов является важным условием успешности и адекватности созданного ресурса. В случае создания корпуса параллельных текстов для других целей, качество перевода не является столь важным.

Для создания корпуса параллельных текстов, оценка качества перевода является важным вопросом, и может быть осуществлена с использованием различных критериев, включая последовательность использования номинаций в тексте перевода, соблюдение норм языка перевода, сохранение логической структуры исходного текста, а также экспертную оценку.

Для формирования корпуса и извлечения терминологии из текстов необходимо учитывать специфику языка научных текстов, который является разновидностью функционального языка и обладает большой автономностью в рамках специальной предметной области. В отличие от многих искусственных систем обработки и передачи информации, язык является открытой, динамической и неравновесной метасистемой. При подборе текстов для корпуса необходимо учитывать эту специфику и выбирать тексты соответствующих жанров и тематик. Качество извлекаемой терминологии напрямую зависит от качества выбранных текстов и переводов, поэтому особое внимание следует уделить этому этапу работы.

Для передачи информации в тексте используются различные языковые конструкции, такие как имена существительные и именные словосочетания. При этом научный текст содержит универсальную часть смысла, которая может быть извлечена при совпадении тезаурусов автора и получателя, и которая в основном определяется информацией об объектах, описываемых в тексте. Следовательно, текст можно рассматривать как результат решения задачи передачи информации и источник ее извлечения [4].

Следовательно, информационная составляющая специального текста, т.е. информация, которую получатель может извлечь из текста, включает в себя в основном денотативный компонент, связанный с номинацией описываемых объектов.

В контексте перевода научных текстов, возможность извлечения информации для получателя зависит от точности передачи терминов, то есть имен объектов. Качество восприятия текста на лексическом уровне зависит от того, насколько насыщен текст именными единицами, а также от степени компрессии или развернутости номинации объектов. В случае, если текст содержит слишком много длинных или чрезмерно свернутых лексических комплексов, он может быть непонятен или даже невозможен для восприятия. Учет сложности таких лексических образований и их преобразование в более простые комплексы помогает создавать адекватный текст перевода. В реальных текстах также могут встречаться корпоративные номинации, отличительные для конкретной организации,

значения которых не фиксируются словарями и представляют особую проблему для перевода. Сопоставительный анализ структуры именных групп разных языков позволяет выявлять различия в принципах номинирования сложных объектов и степени отражения особенностей этих объектов при расчлененной (многокомпонентной) номинации.

Необходимо учитывать, что при выравнивании текстов в параллельном корпусе по предложениям мы получаем лишь приблизительные соответствия между лексическими единицами исходного языка и их переводом. Даже при сохранении выравнивания по частям речи между термином исходного языка и его аналогом в языке перевода возникают расхождения. Расхождения могут быть категориальными, когда в переводе используются единицы других частей речи, результатом могут быть также расхождения в номинации, когда коллокация переводится универбом - знаменательным словом, и структурные расхождения, когда в переводе изменяется синтаксическая структура предложения.

В связи с этим следует отметить, что даже при использовании автоматических систем выравнивания текстов в параллельных корпусах и методов поиска соответствий в псевдопараллельных корпусах результаты необходимо тщательно проверять специалистами в области терминологии и/или перевода. Тем не менее, проведение такой проверки и ручной выбор терминов из предложенного списка могут существенно сократить затраты труда на отбор и перевод новых терминов, включая универбы и терминологические словосочетания.

Использование параллельного корпуса текстов позволяет не только автоматизировать отбор терминологических словосочетаний, но также может быть использовано для:

1. расширения словарного запаса путем включения свободных словосочетаний, используемых в исходных текстах, что особенно важно для переводчиков, работающих на язык, отличный от родного;
2. уточнения употребительности конкретных словосочетаний в текстах определенной предметной области;
3. проверки значений лексических единиц, занесенных в двуязычные словари, особенно в отношении идиом и терминологических выражений;
4. выделения устойчивых словосочетаний и идиом, которые целесообразно включать в словарь конкретной области знаний.

Использование корпуса параллельных и/или псевдопараллельных текстов в качестве лексикографической базы предполагает необходимость расширения этой базы путем добавления в нее корпуса машинных переводов текстов. Такое дополнение позволяет выявить те лексические единицы, которые необходимо включить в словарь, или перевод которых требует модификации.

Список использованной литературы

1. Нигматова, Л. (2022, June). ПАРАЛЛЕЛЬНЫЕ КОРПУСЫ, КАК ЭФФЕКТИВНЫЙ МЕТОД ОБУЧЕНИЯ ИНОСТРАННОГО ЯЗЫКА. In *«УЗБЕКСКИЕ НАЦИОНАЛЬНЫЕ ОБРАЗОВАТЕЛЬНЫЕ ЗДАНИЯ ТЕОРЕТИЧЕСКОЕ И ПРАКТИЧЕСКОЕ СОЗДАНИЕ ВОПРОСЫ" Международная научно-практическая конференция* (Vol. 1, No. 1).

2. Нигматова, Л. Х. (2020). Некоторые проблемы узбекской лексикографии и разработки словаря. *Ташкент Журнал Вопросы филологии*.

3. Avezov, S. S. (2022, December). MACHINE TRANSLATION TO ALIGN PARALLEL TEXTS. In *International Scientific and Current Research Conferences* (pp. 64-66).

4. Авезов, С. (2022, June). О корпусной лингвистике, трудностях перевода и принципах организации параллельных корпусов текстов. In *«узбекские национальные образовательные здания теоретическое и практическое создание вопросы" Международная научно-практическая конференция* (Vol. 1, No. 1).

MUNDARIJA

YALPI YIG‘ILISH MA’RUZALARI

1-SHO‘BA. O‘ZBEK TILINING MILLIY KORPUSI: NATIJALAR, MUAMMOLAR, VAZIFALAR