Л.Н. Беляева

ЛЕКСИКОГРАФИЧЕСКИЙ ПОТЕНЦИАЛ ПАРАЛЛЕЛЬНОГО КОРПУСА ТЕКСТОВ

В своей статье «От Розеттского камня до информационного общества: обзор методов обработки параллельных текстов»[1], Дж. Веронис рассматривает параллельный корпус текстов как совокупность документов, переведенных на два или более языков, кроме того, следует подчеркнуть такую особенность корпуса параллельных текстов, как наличие метатекстовой, межтекстовой и морфологической разметки, а также включение в такой корпус полных текстов, а не (даже) представительных фрагментов. Такую совокупность текстов можно использовать для автоматизации лексикографических работ в области создания переводных словарей, а также для контрастивных исследований в области лексикологии и фразеологии. Использование параллельных текстов в двуязычной и многоязычной лексикографии дает возможность

- обогащения набора переводов, включаемых в словарную статью, за счет анализа значений устойчивых словосочетаний, используемых в исходных текстах;
- уточнения употребительности и значений конкретных слов и словосочетаний в текстах определенной предметной области для введения в соответствующие терминологические словари частотных и устойчивых конструкций;
- верификации значений лексических единиц, уже зафиксированных в двуязычных и многоязычных словарях, особенно в том, что касается идиом и терминологических выражений;

[1] *Véronis J.* From the Rosetta Stone to the Information Society: A Survey of Parallel Text Processing // J. Véronis (ed.). Parallel Text Processing. Kluwer, 2000. P. 1–25.

- выделения устойчивых словосочетаний и идиоматических выражений, которые целесообразно вводить в автоматические словари систем машинного перевода и глоссарии.

Таким образом, на основе анализа полнотекстовых баз параллельных текстов можно выделять устойчивые пары слов типа «исходная лексическая единица – перевод», кроме того, такое выделение можно автоматизировать.

Естественно, для выделения пар лексических единиц из разных текстов и, следовательно, для использования параллельного корпуса в лексикографических целях необходимо его предварительное выравнивание.

Выравнивание текстов по предложениям представляет собой сложную задачу, часто с множественными решениями, возникающими в результате:

- неоднозначности решения самой задачи сегментации текста на предложения (многозначности точки как знака препинания, особенностей описания прямой речи в текстах художественной литературы и публицистики, отсутствии фиксации конца предложения в случае заголовка и т.д.);
- несовпадения деления входного и выходного текстов на предложения, такое несовпадение возникает при ручном переводе текстов и выражается в следующих шести вариантах несовпадения границ предложений.

Выравнивание текстов осуществляется на основе предположения о существовании только шести возможных соответствий между переводными моделями:

1) одно предложение переводится одним предложением;
2) два предложения переводятся одним предложением;
3) одно предложение переводится двумя или несколькими предложениями;

4) два предложения переводятся двумя предложениями, но их внутренние границы не совпадают;
5) предложение исходного текста не переводится;
6) предложение в тексте перевода не имеет эквивалента в оригинале и вводится переводчиком.

При автоматизации процедуры выравнивания на основе совпадения параграфов текста выделяются пары, соответствующие этим моделям. Следовательно, автоматическое выравнивание параллельных текстов по предложениям всегда требует ручного редактирования, в рамках которого и устанавливаются границы поиска возможных соответствий слов, словосочетаний, синтаксических конструкций.

Одним из наиболее распространенных вариантов использования параллельных текстов в лексикографии является построение конкордансов, как в учебных целях, так и для собственно лексикографических исследований. Так, например, многоязычный параллельный конкорданс, разрабатываемый в Бирмингеме, создается как международный проект, в котором участвуют 6 университетов из 6 стран Европы[1]. В рамках этого проекта, называемого Collins Birmingham University International Language Database, создаются параллельные конкордансы для датского, английского, французского, немецкого, греческого и итальянского языков на базе корпуса текстов, включающего произведения художественной литературы и технические тексты.

Поскольку целью создания этого конкорданса является именно обучение языкам, то в каждом языке выбирается классическое произведение художественной литературы, и ему сопоставляются переводы. Отбор переводов и оценка их адекватности и/или эквивалентности осуществляется самими авторами проекта.

---

[1] *King P.* Trialling a Multilingual Parallel Concordancer // Second International Conference on Current Trends in Studies of Translation and Interpreting. Abstracts. Hungary, 1996. P. 49–50.

Важно подчеркнуть, что качество корпуса параллельных текстов и его применимость для лексикографии связаны именно с принципами отбора переводов, если мы определяем такой корпус как совокупность документов, переведенных на два или более языков.

В этом определении наиболее интересным представляется операция перевода как определяющая **параллельность** текстов. Дело в том, что использование корпусов параллельных текстов для получения лексикографической информации определяется не только корректностью выравнивания текстов на уровне меньшем, чем предложение, но и эквивалентностью самих выравниваемых текстов.

С этой точки зрения можно выделить различные соотношения между оригиналом и переводом в параллельном корпусе текстов. Рассмотрим их последовательно.

- *Оригиналу соответствует аутентичный перевод (перевод официального документа, имеющий одинаковую силу с оригиналом).*

В данном случае эквивалентность текстов подтверждается либо юридически (для международных правовых документов, контрактов, рекомендаций и постановлений международных организаций и т.д.), либо формально. В последнем случае мы имеем дело с документами государств, в которых есть несколько государственных языков. Классическим примером такого корпуса является выровненный по предложениям Корпус Hansard – отчеты о дебатах в канадской Палате общин за три года, которые включают 21,6 млн английских и 24,1 млн французских словоупотреблений[1]. В принципе, такого рода параллельные тексты достаточно ограничены по тематике и их явно недостаточно для решения различных

[1] *Langlois L.* Bilingual Concordances: A New Tool for Bilingual Lexicographers // Expanding MT Horizons. Proceedings of the Second Conference of the Association for Machine Translation in the Americas. Montreal, Quebec, Canada, 1996. P. 34–42.

лексикографических задач. В то же время подобные тексты могут использоваться для создания или уточнения нормативных словарей соответствующих областей знаний.

- *Оригиналу соответствует авторский перевод (перевод текста, осуществляемый человеком-переводчиком по заказу или на основе собственной инициативы).*

Эта ситуация является, пожалуй, самой распространенной, но требует рассмотрения адекватности переводов. В случае конкордансов системы COBUILD, авторы берут на себя ответственность за выбор перевода, который «назначается» адекватным. В случае использования в параллельном корпусе художественной литературы (во всех возможных вариантах определения этого термина) и набора переводов, выполненных разными авторами, появляется возможность лексикографического исследования вариантов перевода реалий, ксенонимов, просторечных элементов в разных переводах и, следовательно, возможность создания словарей соответствующих типов.

Другим случаем авторского перевода является перевод публицистической, научной и научно-технической литературы. В этом случае вопрос об адекватности перевода тоже является важным, поскольку такого рода корпус может использоваться для создания специализированных словарей. В то же время, при использовании подобных переводов следует оценивать либо адекватность самих переводов, либо источника, из которых они извлекаются. Вероятно, в этом случае целесообразно использовать материалы, которые прошли редактирование и опубликованы в серьезных изданиях.

- *Оригиналу соответствует машинный перевод (перевод, выполненный конкретной системой машинного перевода).*

Рассмотрим эту ситуацию более подробно. Такие массивы текстов могут использоваться для создания независимой от языка онтологии. В этом случае онтология представляет собой базу знаний, хранящую информацию о понятиях, существующих в мире

или предметной области, их свойствах, и о том, как они связаны друг с другом.

Онтология отличается от тезауруса тем, что содержит только независящую от языка информацию и множество семантических отношений, кроме того, она содержит таксономические отношения. Тем самым задача построения онтологии представляет собой задачу создания некоторой модели мира, необходимой для смысловой переработки текста. Онтология должна задавать понятия для представления значений слова в лексиконе и, соответственно

- должна хранить селективные ограничения понятий;
- должна иметь возможности для использования в любой прикладной программе, предназначенной для использования в системах переработки текстов на естественном языке и в любой области;
- должна быть независима от языка;
- должна быть удобна для работы самых разных пользователей.

Как правило, для описания онтологий используется расширяемый язык разметки (Extensible Markup Language – XML).

В рамках одного из вариантов построения подобной онтологии используются системы машинного перевода с корейского языка на японский и с японского на корейский, а также созданный в Японии тезаурус Kadokawa[1], в котором задано используется 1110 специальных семантических категорий и четырехуровневая таксономическая иерархия, кроме того, дополнительно выделено 30 типов семантических отношений. Каждое понятие, вводимое в исходную онтологию, имеет код по тезаурусу Kadokawa, имя на корейском языке, имя на английском языке, отметку времени введения понятия и определение понятия.

[1] *Li H.F., Heo N.W., Moon K.H., Lee J.H., Lee G.B.* Lexical Transfer Ambiguity Resolution Using Automatically-Extracted Concept Co-occurrence Information // International Journal of Computer Processing of Oriental Languages. World Scientific Publishing Co, 2000. № 13(1). P. 53–68.

Хотя коды понятий могут быть однозначно опознаны по кодам понятий Kadokawa, их имена на корейском и английском языках вводятся для читабельности и удобства разработчика онтологии.

Для создания онтологии на базе тезауруса Kadokawa используются двуязычные автоматические словари COBALT-J/K и COBALT-K/J, с помощью которых происходит разметка текстов оригинала и машинного перевода. При этом

- имя и глагольные слова аннотируются кодами понятий из тезауруса Kadokawa,
- для описания падежных рамок используется проект SELK, состоящий из различных подсловарей, при этом каждый подсловарь соответствует словарю определенной категории слов, типа словаря существительных, словаря глаголов и т.д.

Системой SELK называется электронный лексикон корейского языка Sejong[1].

На основе этих данных при создании онтологии методом машинного перевода был создан параллельный корпус патентных материалов в предметной области производства стали. Объем массива составляют 250 тыс. предложений на японском и корейском языке с соответствующей синтаксической и семантической разметкой, на основе которого и выделяются варианты переводов сложных конструкций и правила снятия синтаксической и семантической многозначности.

Подобную онтологию можно рассматривать

- как источник знаний о внешнем мире;
- как надежную словарную информацию;

[1] *Hong C.S., Pak M.G.* Developing a Large Scale Computational Lexical Database of Contemporary Korean: SELK // Proceedings of the 19th International Conference on Computer Processing of Oriental Languages. Seoul, Korea, 2001. P. 20–26.

- как контекстную информацию, необходимую для создания моделей снятия многозначности.

Рассмотренный метод построения онтологии требует тщательной ручной обработки, т.е. отображения перехода от синтаксических отношений до семантических отношений в соответствии со специфическими правилами и интуицией человека.

Использование машинных переводов для создания корпуса параллельных текстов имеет значение для автоматизированной лексикографии в той ее части, которая относится к созданию лексикографического обеспечения систем машинного перевода, поскольку позволяет при условии сохранения соответствия по предложениям оценить адекватность и постоянство перевода терминов и, следовательно, модифицировать словари. Кроме того, такой параллельный корпус позволяет уточнить неопознанные слова и конструкции.

В то же время следует понимать, что подобный параллельный корпус является рабочим инструментом отладки конкретной системы, обучающей выборкой, и при модификации системы потребует замены.

Крайним случаем отношений между оригиналом и переводом является ситуация, когда

- *Оригиналу соответствует не перевод, а сопоставимый по содержанию текст на другом языке.*

Эти так называемые псевдопараллельные тексты могут использоваться для составления словарей новых терминов, однако корпусы текстов, создаваемые из подобных источников, должны особым образом размечаться и использоваться.

Для решения вопроса о возможности использования корпуса текстов в конкретных целях отношение оригинал – перевод является критическим, более того, оно должно квалифицированно оцениваться тем, кто собирается использовать уже имеющийся корпус или создавать новый.

Поэтому чрезвычайно важным и для задачи создания параллельных корпусов текстов, для корпусной лингвистики в целом является аспект квалификации разработчиков и пользователей создаваемых корпусов. Действительно, общий переход на подготовку любых текстов в электронном формате, с одной стороны, и доступность персональных компьютеров и сети Интернет, с другой, определили развитие корпусной лингвистики, т.е. лингвистического направления, основанного на исследовании практически бесконечных массивов текстов. Возможность хранения, пополнения и исследования таких корпусов текстов определило появление нового метода в практике письменного перевода – систем с переводческой памятью.

Эти массивы дают совершенно новые возможности как для исследований в рамках одного языка (одного функционального стиля, одной тематики, произведений конкретного автора и т.д.), так и для контрастивного исследования параллельных или условно параллельных текстов на разных языках.

Совершенно по-новому выглядит на этом фоне автоматизированная лексикография, когда можно выбирать реально используемую лексику и описывать перевод или объем значения на огромном массиве, и делать это не вручную, а с помощью компьютера. В этом направлении особое место занимают задачи анализа уже имеющихся или создаваемых корпусов текстов, и создания новых. Особой задачей является разработка методов и средств извлечения лексических знаний (распознавание лексической единицы в тексте, формирование триады типа слово – словосочетание – перевод), использование параллельных текстов для разработки методов и средств представления знаний, создания специализированных словарей, создания многоязычных конкордансов.

Наличие большого количества полнотекстовых баз данных, а также выровненных корпусов параллельных текстов привело к возникновению новой дополнительной подсистемы в рамках операций, выполняемых лингвистическим автоматом (ЛА), – под-

системы автоматизированного создания словарей различного типа. При этом задачи подготовки частотных, алфавитных, обратных словников, конкордансов и т.п. являются вполне рутинными и традиционными. Новым аспектом является использование ЛА для создания и ведения двуязычных лексиконов (резидентных словарей или просто словарей на машинных носителях), а также создание и/или пополнение автоматизированных словарей для систем МП.

Решение этих задач в большой степени определяется методами выравнивания двуязычных параллельных текстов. Следует добавить, что использование статистических методов изучения контекста в параллельном корпусе дает возможность снять или уменьшить многозначность слова и подобрать его эквивалент в языке перевода[1].

[1] *Dan Melamed I.* Empirical Methods for Exploiting Parallel Texts. MIT Press, Cambridge (MA), 2001.