Тао Юань

Создание и использование параллельного корпуса русского и китайского языков

В данной работе мы представляем сведения об общей структуре параллельного корпуса, даём характеристику отобранных текстов, рассматриваем процесс создания корпуса и его использования. Кроме того, мы описываем метаданные и поиск словоизменительных парадигм, в том числе падежных форм существительных, прилагательных и спрягаемых форм глаголов, также представляем автоматическую генерацию словника терминов. Создание корпуса тесно связано с его назначением, поэтому в работе представлены примеры лингвистического анализа на основе корпуса.

This article introduces the parallel corpora general structure, characterizes the selected texts and dwells on the corpora construction procedure. Besides, it describes the metadata and retrieval of inflectional paradigms including nominal and adjectival inclination forms and verbal conjugation forms. The article also pays special attention to automatic term base generation. Given that corpora design and its research purpose are closely related, this article also displays the samples of linguistic analysis on the basis of the corpora.

Ключевые слова: параллельный корпус русского и китайского языков; разработка корпуса; генерация словника терминов; поиск.

Keywords: parallel Russian-Chinese corpora; corpus construction; term-base generation; intersection retrieval.

1. Введение

Статья посвящена изучению процесса разработки параллельного корпуса русского и китайского языков и возможного направления исследований на его основе. Это первый параллельный и компаративный корпус русского и китайского языков, который предназначен для анализа переводов, поэтому разработка дизайна корпуса, формирование коллекции текстов, проведение предварительной обработки, создание поисковой системы и метаданных — всё это сформировало целую цепь сложных задач.

1.1. Существующие корпусы

Ещё в 1897 году немецким лингвистом Кэйдингом был составлен первый корпус текстов для сравнения частоты распределения букв в словах и определения их последовательности [4–5]. Однако результаты исследования нельзя было назвать многообещающими в силу того, что без применения технических средств ни один человек не смог бы проанализировать такое большое собрание текстов самостоятельно.

Появление компьютеров позволило частично решить эту проблему. Последующие разработки программного обеспечения для работы с корпусами текстов увенчались созданием программ-конкордансеров [6].

Центральное место в корпусной лингвистике занимает электронный корпус текстов. Один из вариантов определения этого ресурса был предложен С. Лавиоза, которая высказала мнение, что под электронным корпусом текстов следует понимать тексты и фрагменты текстов, отобранные по определённому принципу, размеченные и упорядоченные для поиска необходимой лингвистической информации [2: с. 31].

В данной статье мы ставим целью рассмотреть, как необходимо структурировать параллельный корпус, предназначенный для анализа переводов. Стоит отметить, что эта проблема уже рассматривалась Н.В. Владимовым [1: с. 24], однако для исследования он использовал Национальный корпус британского варианта английского языка — British National Corpus (BNC). В Китае [8] составлен параллельный корпус русского и китайского языков, но только для военных текстов. Обзор специализированных корпусов для русского языка даётся В.П. Захаровым [3; 7], но параллельных русско-китайских тематических корпусов и там не зарегистрировано.

Наш корпус отличается от существующих большим масштабом (5 млн словоупотреблений) и использованием универсальных принципов перевода и сопоставления двух языков. В нашем случае мы опираемся на специально составленные нами корпусы оригинальных текстов по (1) политике и международным отношениям; (2) лингвистике; (3) литературоведению; (4) переводоведению. Для составления конкорданса использовалась программа-конкордансер ParaConc: для согласования и создания конкордансов параллельных текстов (перевод) в ней параллельно могут быть проанализированы до 4-х языков.

2. Структурирование и создание корпуса

Цель разработки корпуса — создать платформу для обучения переводу и его исследования. Поэтому порядок создания должен быть следующим: постановка задачи → создание корпуса → выбор программного инструмента (e.g. Wordsmith) → статистический анализ корпусных данных → использование корпуса. В соответствии с указанной целью были приняты следующие проектные решения.

2.1. Структурирование корпуса: общая структура и отбор текстов

Наш корпус — специальный и гомогенный [9], в него были включены только тексты гуманитарной области. Причина выбора этого типа текстов состоит в следующем:

1) с точки зрения направленности практики перевода: в прошлом веке в Китае большинство переводных произведений — художественные тексты, а в этом веке быстро развивается и перевод научных текстов, что требует выявления специфики перевода в данной области;

2) с точки зрения стиля языка: научный стиль имеет ряд общих черт, что даёт возможность говорить о специфике стиля в целом. Научный стиль характеризуется логической последовательностью изложения, упорядоченной системой связи между частями высказывания, стремлением авторов к точности, сжатости, однозначности при сохранении насыщенности содержания. Стремлением к информационной насыщенности обусловливается отбор наиболее ёмких и компактных синтаксических конструкций.

2.2. Структура и объём корпуса

Корпус состоит из двух частей: параллельный корпус текстов на русском языке и их переводов на китайский язык (parallel corpora — PC) и сопоставимый корпус (comparable corpus — CC) из тематических текстов на китайском языке. На начальном этапе объём параллельного корпуса — 5 млн слов, а сопоставимого — 1 млн. Оба подкорпуса будут открытыми и динамическими. Корпус будет увеличиваться каждые два года, и техника разметки будет улучшаться.

Параллельный корпус (PC) включает в себя 14 монографий в области политики, международных отношений, лингвистики, литературоведения и переводоведения на русском языке и их переводы на китайский. В этих монографиях представлены тексты на современном русском языке, они переведены в последние тридцать лет. Сравнительный корпус (CC) включает в себя 10 монографий на китайском языке из тех же предметных областей. Совпадение стиля текстов параллельного и сравнительного корпусов обеспечивает сопоставимость особенностей переводного и оригинального китайского языка.

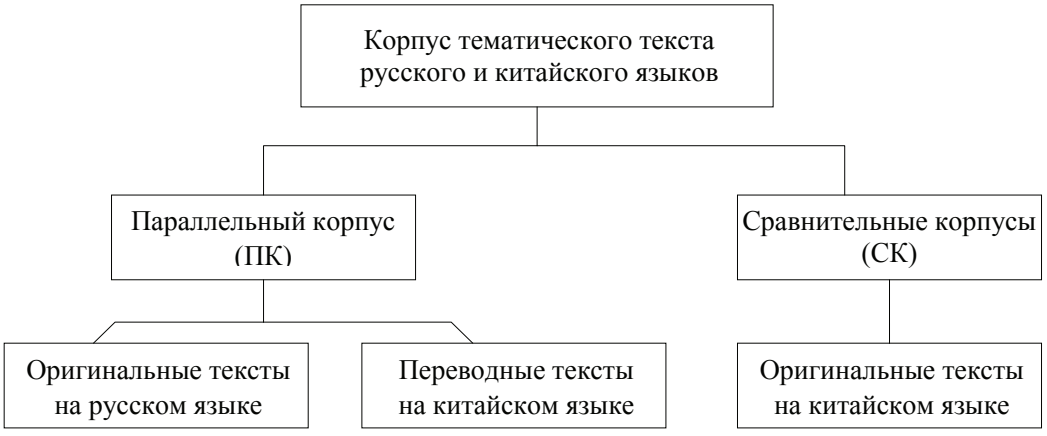

Рис. 1. Структура тематического корпуса русского и китайского языков

Переводной китайский язык сопоставляется не только с оригинальным русским языком в параллельном корпусе, но и с оригинальным китайским языком в сравнительном корпусе. Оригинал и перевод каждой монографии параллельного корпуса создают свой подкорпус, и оригинал каждой монографии сравнительного корпуса создаёт соответствующий подкорпус. Так, мы получаем PC1, PC2, PC3... и CC1, CC2, CC3... Таким образом, сопоставление

и количественный анализ идёт не только между корпусами, но и между парами подкорпусов. Для увеличения надёжности исследования мы обращаем внимание на баланс количества слов между подкорпусами.

Таблица 1

Объём тематических корпусов русского и китайского языков

	РС			СС	
	Кол-во слов (русский)	**Кол-во слов (китайский)**	**Кол-во текстов**	**Китайский (слова)**	**Кол-во текстов**
Политика и международные отношения	418 100	710 856	4	303 718	2
Лингвистика	568 738	855 326	3	335 546	2
Литературоведение	208 643	315 258	5	316 328	2
Переводоведение	401 223	598 816	2	287 914	2
Всего:	1 596 704	2 480 256	14	1 243 496	8

2.3. Метаданные

Метаданные тематического корпуса включают следующую информацию: язык (language), тип (type), автор (author), переводчик (translator), время издания (time) и название (title).

В метаданных мы указываем два языка — русский и китайский, четыре типа: политика и международные отношения, лингвистика, литературоведение и переводоведение, два вида времени издания — год публикации оригинала и перевода, название на двух языках — русском и китайском.

2.4. Процедуры обработки текстов при создании корпуса:

1) заполнение текстовых таблиц. Импорт данных таблицы в параллельный корпус;

2) выравнивание. На первом этапе выравнивание корпуса выполняется автоматически с использованием Paraconc (точность выравнивания 60–70 %).

На втором этапе ошибки выравнивания исправляются в ручном режиме (100 %).

1. Загрузка. Загрузка выравненных текстов выполняется автоматически на платформе Paraconc. На платформе нажмём file – export – export corpus files, и загрузка осуществится. Загруженный файл должен быть сохранён вместе с микрософтом Paraconc, иначе файл не открывается.

2. Поиск в корпусе. Морфологическая нормализация для русского языка не делается. Поиск осуществляется на основе языка регулярных выражений (regular expressions).

Приведём пример такого поиска. Введя регулярное выражение <[Кк]отор*>, мы получим все словоизменительные парадигмы лексемы *который*. [Кк] включает большие и маленькие буквы «к», а звездочка «*» включает в себя все возможные формы окончаний.

Так как целью создания нашего корпуса является исследование переводов, то основной режим поиска — поиск на платформе Paraconc. Результат поиска <[Кк]отор*> на платформе Paraconc.

Кроме того, мы создаём платформу поиска в Интернете. Там будет реализован поиск по лексическим единицам/словоформам с добавлением элементов метаданных (retrieval intersection). Например, мы ищем контексты для слова «стратегия» в текстах по политике, по конкретному автору или переводчику.

3. Генерация словника терминов. Материал корпуса — специальные тексты гуманитарного направления, и, естественно, в оригиналах и переводах много терминов, перевод которых является приоритетным в нашем исследовании, поэтому создание словника или базы данных терминов — одна из целей нашей работы.

На первом этапе мы в ручном режиме ищем термины из текстов на русском языке и их переводов в текстах на китайском языке и выравниваем их. Выровненные термины составляют новый текст (в формате txt).

На втором этапе мы импортируем выровненные тексты по строчке (с помощью инструмента (eclipse и java) в базу данных (database) и выполняем генерацию словника терминов по данным корпуса.

4. Разработка сайта корпуса. Мы разработали сайт корпуса, через веб-интерфейс которого реализуются поиск морфологических форм, поиск с добавлением элементов метаданных и поиск терминов.

3. Возможные виды лингвистического анализа на основе корпуса

Морфологическая разница между китайским и русским языками очевидна [12]: морфология русского языка богата, а в китайском языке морфологические формы фактически отсутствуют. Одна из особенностей синтаксиса китайского языка заключается в отсутствии подчинительного способа соединения слов в словосочетании и предложении, сравнимого с русским языком. Современные китайские учёные в связи с этим утверждают, что «семантика занимает центральное место» и «семантика определяет грамматику» [10; 13].

Мы предлагаем опыт составления корпуса этих двух языков, будучи уверенными, что китайский язык в переводах с русского обнаруживает как типологию соответствий, так и своеобразие своей природы.

Параллельный корпус переводов с русского языка на китайский позволяет решать различные лингвистические, переводческие и образовательные задачи. Приведём примеры некоторых таких задач.

1. Исследование универсальных принципов перевода на китайский язык на основе анализа перевода неопределённых местоимений и неопределённых наречий, например, перевод лексем *кто-то, кто-нибудь, кое-кто, что-то, что-нибудь, когда-нибудь, кое-когда, где-то, где-нибудь, кое-где* и т. д.

2. Исследование универсальных принципов перевода на китайский язык, основанных на анализе перевода пассивного залога, в том числе страдательных причастий и глаголов с постфиксом *-ся* и т. д.

3. Исследование норм перевода сложноподчинённых предложений с русского языка на китайский. В частности, при переводе сложноподчинённых

предложений с союзом *чтобы* существует четыре вида нормы — экспликация и импликация, упрощение-осложнение, нормализация и отчуждение, с использованием и без использования идиом.

4. Определение основной переводной единицы на основе корпуса. Анализ производится в аспекте прагматики. Мы выдвинули предположение, что высказывание является основной единицей перевода по классической теории о высказывании. На основе корпуса мы показали, что:

а) значение слова *узел* (node) неопределённо, значение его уточняется только в контексте;

б) соответствие перевода осуществляется только на уровне высказывания [11].

5. Практическое исследование перевода на основе корпуса:

а) перевода дискурсивных маркеров, в том числе таких, как *речь идёт о..., как указывалось, как отмечалось, согласно этому, в результате, следовательно, ввиду этого, в зависимости от этого, в связи с чем..., рассмотрим, перейдём к рассмотрению..., иначе говоря, из этого следует...*;

б) правил трансформации предложений с союзными словами и союзами *который, так как* и т. д.

4. Выводы

В данной работе мы описали цели, специфику и процесс создания параллельного корпуса русского и китайского языков.

Способы разработки параллельного корпуса русского и китайского языков несовершенны, и мы предлагаем внести определённые изменения и корректировки в данный корпус. В частности, намечено разработать дополнительные программы предварительной обработки текстов. Также планируется автоматическая лемматизация слов в текстах русскоязычной части корпуса. Кроме того, предстоит преодолеть немало трудностей в части поиска и генерации корпуса терминов и т. д. Решение этих задач позволит оптимизировать последующую работу по расширению параллельного корпуса и по его использованию.

В настоящее время в теории принятия решений существует два подхода: нормативный и дескриптивный. Создание данного корпуса позволяет анализировать перевод с русского языка на китайский в аспекте дескриптивного подхода. На платформе корпуса возможно проведение не только качественного, но и количественного исследования, не только осуществление оценки (плюсы и минусы) переводных текстов, но и исследование природы и универсальности переводного языка.

Библиографический список

Литература

1. *Владимов Н.В.* Корпусный подход к решению переводческих проблем: дис. ... канд. филол. наук: 10.02.19. М., 2005. 198 с.

2. Захаров В.П. Корпусная лингвистика. СПб.: Изд-во СПбГУ, 2005. 48 с.

3. *Захаров В.П.* Корпусный менеджер как поисковая система. URL: http:/download.yandex.ru/class/zakharov/CL_L7.ppt (дата обращения: 12.10.2014).

4. *Adolphs S.* et al. Clinical Linguistics and Corpus Linguistics in Health Care Settings // CHLR. 1998. P. 5.

5. *Francis N.W.* Language corpora B.C. // Directions in Corpus Linguistics / Ed. J. Svartvik. Berlin: Mouton de Gruyter, 1992. P. 17–32.

6. *Laviosa S.* Corpora and Translation: the Methods and Theories of Corpus Work in Translation. Manchester, 2003. P. 3.

7. *Zakharov V.* Corpora of the Russian Language // Text, Speech and Dialogue: Proceedings of the 16th International Conference, TSD 2013, Plzen, Czech Republic, September 1–5, 2013. (Lecture Notes in Artificial Intelligence, 8082) / Ivan Habernal, Václav Matoušek (Eds.). Springer-Verlag, Berlin Heidelberg, 2013. P. 1–13.

8. 崔卫/张岚. 俄汉翻译平行语料库及其应用研究 // 解放军外国语学院学报. 2014. 1. P. 81 - 87.

9. 黄昌宁. 语料库语言学 // 北京：商务印书馆. 2002. P. 35.

10. 陆俭明. 词的具体意义对句子意思理解的影响 // 汉语学习. 2004. P. 1 - 5.

11. 陶源. 基于俄汉平行语料库的翻译单位研究 // 外语教学. 2015. 1.

12. 赵敏善. 俄汉语对比研究 // 上海译文出版社. 1994. P. 68.

13. 赵世举. 试论词汇语义对语法的决定作用. 武汉大学学报 // 2008. 2. P. 173 - 179.

References

Literatura

1. *Vladimov N.V.* Korpusny'j podxod k resheniyu perevodcheskix problem: dis. ... kand. filol. nauk: 10.02.19. M., 2005. 198 s.

2. *Zaxarov V.P.* Korpusnaya lingvistika. SPb.: Izd-vo SPbGU, 2005. 48 s.

3. *Zaxarov V.P.* Korpusny'j menedzher kak poiskovaya sistema. URL: http:/download.yandex.ru/class/zakharov/CL_L7.ppt (data obrasheniya: 12.10.2014).

4. *Adolphs S.* et al. Clinical Linguistics and Corpus Linguistics in Health Care Settings // CHLR. 1998. P. 5.

5. *Francis N.W.* Language corpora B.C. // Directions in Corpus Linguistics / Ed. J. Svartvik. Berlin: Mouton de Gruyter, 1992. P. 17–32.

6. *Laviosa S.* Corpora and Translation: the Methods and Theories of Corpus Work in Translation. Manchester. 2003. P. 3.

7. *Zakharov V.* Corpora of the Russian Language // Text, Speech and Dialogue: Proceedings of the 16th International Conference, TSD 2013, Plzen, Czech Republic, September 1–5, 2013. (Lecture Notes in Artificial Intelligence, 8082) / Ivan Habernal, Václav Matoušek (Eds.). Springer-Verlag, Berlin Heidelberg, 2013. P. 1–13.

8. 崔卫/张岚. 俄汉翻译平行语料库及其应用研究 // 解放军外国语学院学报. 2014. 1. P. 81 - 87.

9. 黄昌宁. 语料库语言学 // 北京：商务印书馆. 2002. P. 35.

10. 陆俭明. 词的具体意义对句子意思理解的影响 // 汉语学习. 2004. P. 1 - 5.

11. 陶源. 基于俄汉平行语料库的翻译单位研究 // 外语教学. 2015. 1.

12. 赵敏善. 俄汉语对比研究 // 上海译文出版社. 1994. P. 68.

13. 赵世举. 试论词汇语义对语法的决定作用. 武汉大学学报 // 2008. 2. P. 173 - 179.