МНОГОЯЗЫЧНЫЕ КОРПУСА

История компьютерной лингвистики (к которой относится и корпусная) как в капле воды отражается в развитии машинного перевода: надежды и разочарования, интерес и скепсис пользователей – все в этой области отражает успехи и провалы компьютерной лингвистики. История машинного перевода берет начало в далеком 1949 году, когда Уоррен Уивер (Warren Weaver) написал свой знаменитый меморандум ”Translation”:

> "I have a text in front of me which is written in Russian but I am going to pretend that it is really written in English and that it has been coded in some strange symbols. All I need to do is strip off the code in order to retrieve the information contained in the text".
>
> «У меня перед глазами текст, написанный по-русски, но я собираюсь сделать вид, что на самом деле он написан по-английски и закодирован при помощи довольно странных знаков. Все что мне нужно — это взломать код, чтобы извлечь информацию, заключенную в тексте»

Задача оказалось не такой простой, как это представлялось в начале, было изобретено множество систем и программ, пока наконец машинный перевод в конце 1980-х годов не обратился к корпусной лингвистике, в рамках которой было подготовлены разные многоязычных корпусов (англ. multilingual *corpora*), содержащих тексты на разных языках и пригодных для автоматизации перевода: сравнительный (англ. *comparable corpus*) и параллельный, или корпус переводов (англ., *parallel или translation corpus*). Не лишне отметить, что многоязычные корпуса используются не только для подготовки профессиональных компьютеров-переводчиков последней модели.

Сравнительные корпуса

Сравнительный корпус – это многоязычный корпус, в котором собраны похожие тексты на разных языках. Слово «похожие» в этом определении надо понимать в самом широком смысле: тексты могут быть похожи тематически, жанрово, хронологически и т.д. В сравнительном корпусе часто используется общий формат хранения данных, но тексты на одном языке не связаны с другим языком. Унифицированный доступ к такому ресурсу позволяет, например, быстро сопоставлять среднюю длину предложения во французской и немецкой статье или определять лексическое разнообразие в Твиттер-сообщениях.

> Французские исследователи из лаборатории LIMSI создали систему автоматического распознавания языка звонящего для быстрого переключения на нужного оператора. С этой целью они собрали большой четырёхъязычный корпус телефонных разговоров, состоящий из нескольких сотен звонков, сделанных людьми разного пола и возраста, проживающими в разных частях света.

Сравнительным корпусом иногда называют и корпус, содержащий тексты на одном языке: например, переводные и оригинальные тексты на одном и том же языке или территориальные варианты английского языка. Конечно, это не многоязычные корпуса в строгом смысле слова, но и их можно использовать для анализа сравнения языков или их разновидностей. Как бы ни понимать сравнительный корпус, его главным и очевидным минусом является несовпадение текстов на разных языках. То, что мы ищем в одном тексте, может просто отсутствовать в другом, и наше сравнение окажется ненадежным. Для сопоставления языков лучше подходит другой тип корпуса.

Параллельные корпуса

В двуязычной Канаде задачи по переводу всегда были актуальными, и исследователи из этой страны предложили использовать переведенные тексты, вычленяя из них фрагменты (англ. *chunks*), совпадающие с теми, которые требуется перевести. Так возникли корпуса, содержащие один и тот же текст на разных языках.

> Параллельный корпус (англ. *parallel corpus*) – это корпус, состоящий из текстов на одном языке вместе с его переводом на другой язык или языки. Такие параллельные пары или ряды текстов называют битекстами (англ. *bitext*).

Для создания параллельного корпуса мало просто иметь оригинал и его перевод. Надо обеспечить нахождение соответствующего оригиналу фрагмента в переводе. С этой целью используют процедуру выравнивания (англ., *alignment*), в результате которой одинаковые фрагменты параллельных тестов оказываются сопоставлены друг с другом.

Розеттский камень – прекрасный образец параллельного корпуса. Созданный во втором веке до нашей эры, он содержит один и тот же текст на двух вариантах древнеегипетского письма и по-древнегречески. Для точного выравнивания текстов понадобилось бы выбить их еще раз. Желающих сделать это пока не нашлось...

Задача создания параллельных текстов сложнее, чем кажется на первый взгляд. До нашей эры это было связано с тяжелой работой камнереза, но и сейчас подводных камней осталось немало. Основная из них – что, собственно, выравнивать. Идеальным было бы пословное выравнивание (англ. *word alignment*), но оно часто оказывается невозможным по естественным причинам: наборы лексем, словоформ и устойчивых выражений в разных языках не совпадают. Поэтому гораздо чаще тексты выравнивают по предложениям (англ. *sentence alignment*) или по абзацам (англ, paragraph alignment). Но оказывается, что количество слов и даже количество предложений тоже могут не совпадать.

> Приведенный ниже фрагмент русско-финского параллельного корпуса ParRus подготовлен в университете финского города Тампере. Создатель корпуса М. Н. Михайлов подсчитал, что число слов в оригинальных текстах больше, чем в переводах, а число предложений меньше. Попробуйте сами подсчитать количество слов (от

пробела до пробела) и предложений (от точки до точки) в исходном тексте и в его переводе на финский язык.

Лара смеялась и с завистью думала: девочка живет в нужде, трудится. Малолетние из народа рано развиваются. А вот поди же ты, сколько в ней еще неиспорченного, детского. Яйца, Джек – откуда что берется? За что же мне такая участь, – думала Лара, – что я все вижу и так о всем болею? (Пастернак Б.Л., Доктор Живаго)	Lara nauroi ja unohtui kateellisena ajattelemaan Oljaa. Tyttö eli puutteessa, sai tehdä kovaa työtä. Rahvaan lapset kehittyivät nopeasti. Katsopas vain, kuinka paljon hänessä on vielä turmeltumatonta. Tekomunat, Jack, mistä tuo kaikki? Entä miksi minun kohtaloni on tällainen, Lara ajatteli, – näen kaiken ja kärsin kaiken takia? (Konkka J., Tohtori Živago)

Но и это еще не все сложности. Отдельной проблемой при создании корпуса становится несоответствие текстов: переводчики по разным причинам, например цензурным, могут сокращать тексты; авторские переводы при наличии вдохновения или идеологических задач могут существенно отходить от оригинала и т.д.

В 80-ые годы XIX века Пантелеймон Кулиш написал роман «Черная рада», вышедший почти одновременно на украинском и русском языках. Перевод он сделал сам, и вы можете посмотреть, что у него получилось.

І дістав із полички жбан, прехимерно з срібла вилитий і що то вже за приукрашений! Не жалували пани грошей для своєї пихи і потіхи. По боках бігли босоніж дівчата - інша і в бубон б'є; а зверху сидів, мов живий, божок грецеський, Бахус.	И он достал с полки большую серебряную кружку с барельефами, представлявшими греческих вакханок. Крышка была украшена литою статуйкой Фауна.

Даже если вам удастся выровнять тексты, получить автоматический ответ на вопрос, как именно переводится, например, «серебряный» или «жбан» на другой язык, будет непросто. Параллельный корпус не отвечает на такие вопросы, как слово X переводится на другой язык. Он лишь находит по заданным лексемам или грамматическим параметрам фрагменты на одном языке и показывает привязанные (т.е. выровненные) фрагменты на другом языке. Важно понимать, что поиск в параллельном корпусе не отличается от поиска в корпусе одноязычном – второй язык «прицепляется» лишь на последнем этапе вывода результатов на экран.

Известная шутка гласит, что у финнов не будущего. Это верно, но касается только грамматики. Задав поиск форм будущего времени в русской части корпуса ParRus, вы получите пестрый набор переводов, в которых придется самому искать финские соответствия (это будут и формы настоящего времени, и лексикализованные способы указания на время, и даже формы перфекта).

Безусловно, интересной задачей было бы создание параллельных, то есть связанных друг с другом разметок, которые позволяли бы автоматически искать грамматические и лексические соответствия в разных языках, но эта задача отдаленного будущего.

Параллельные корпуса естественным образом делятся на двуязычные (таких большинство) и многоязычные. По направлению перевода можно выделить однонаправленные (англ., *unidirectional*), например, переводы «Слова по полку Игореве» и двунаправленные (англ., *bidirectional*) корпуса.

Главным недостатком параллельного корпуса является то, что тексты не лишены влияния языка источника, переводческих ошибок, индивидуального стиля переводчика. Однако схожие проблемы несбалансированности подстерегают и в одноязычном корпусе. Решение в любом случае одно: чем больше и разнообразнее корпус, тем меньше влияние отдельного текста.

> Существует более сотни переводов «Слова о полку Игореве» на русский язык и две сотни – переводов на другие языки. Большинство из них представлено в виде параллельного корпуса (http://nevmenandr.net/slovo). С его помощью можно легко увидеть, что загадочная «мысь», превратившись русском языке в оборот «растекаться мыслью (по древу)», оказывается, переводилась множеством способов: *соловей, мысль, векша, белка*. Трудно сказать, что из этого более верно...

ЗАДАНИЯ

1. Как выравниваются тексты в программе MS Word? Чтобы проверить это, откройте два файла и выберите команду «**Сравнить рядом с**» в меню «**Окно**».

2. На сайте Национального корпуса размещены параллельные русско-английский, -немецкий, -украинский, -белорусский и многоязычный корпуса: http://www.ruscorpora.ru/search-para.html. Выберите знакомый Вам язык и проверьте, какие варианты перевода существуют для лингвоспецифического слова «пошлость» (или любого другого труднопереводимого слова или конструкции)

3. Vincent Vandeghinste, Removing the Distinction Between a Translation Memory, a Bilingual Dictionary and a Parallel Corpus

Наиболее перспективные технологии машинного перевода основывается на совмещении нескольких подходов, в том числе использовании параллельных корпусов. Прочитайте статью об одном из проектов и ответьте на следующие вопросы:

- Как расшифровываются аббревиатуры: RBMT, SMT, EBMT?
- Какие подходы совмещает описанный в статье проект?
- Можно ли сказать, что параллельные корпуса и Translation Memory – это одно и то же?

4. (По желанию) Подумаете, какие мотивы сподвигли Пантелеймона Кулиша создать такие разные тексты для своего романа «Черная рада».