Секция «Лингвистика (языки Азии и Африки)»

Википедия как источник параллельных текстов на путунхуа и диалекте Юэ

Научный руководитель – Музыченко Юрий Фёдорович

Краснопольская Елизавета Андреевна

Студент (бакалавр)

Московский государственный университет имени М.В.Ломоносова, Институт стран Азии и Африки, Кафедра китайской филологии, Москва, Россия

E-mail: rr.exile@yahoo.de

В современном обществе возрастает интерес к компьютерной лингвистике. Развитие информационных технологий затрагивает все области человеческой жизни: например, не только различные сферы машинной обработки языка (машинный перевод, информационный поиск, распознавание речи, сочетаемость слов), но и сферу образования [1]. Всё это обуславливает необходимость внедрения информационных технологий и электронных ресурсов.

В последнее время Википедия привлекается для технологий мультиязычного перевода, извлечения данных и для других областей машинной обработки языковых данных. Википедия может рассматриваться в качестве потенциального источника для корпуса параллельных текстов, так как является изначально открытым, свободным и структурированным ресурсом. Тексты статей Википедии содержат интерлинки (интерлинк (interlink) - представленные в тексте статьи ссылки на аналогичный по содержанию материал статьи на другом “языке” (по терминологии Википедии). Подразумевает совпадение по смыслу, параллельность).

Мы проанализировали два “языковых” раздела (обозначенных в В. zh (далее - путунхуа) и zh-yue (далее - диалект Юэ), выделили самые объемные статьи, проанализировали их содержание и попробовали “выровнять” эти тексты для получения параллельного корпуса.

Мы также рассмотрели и проанализировали полученные данные, сопоставили совпадающие статьи на языковых разделах путунхуа и диалекта Юэ. В частности, мы выделили наиболее употребительные слова кантонского диалекта, сопоставили их с соответствиями в путунхуа, выявили закономерность или отсутствие закономерности расположения слов в разных диалектах в совпадающих текстах.

В результате нашего исследования были получены следующие данные:

В разделе на путунхуа число страниц с количеством знаков от 30 000 и более - 1472 (что составляет 0,16%), с количеством знаков от 50 000 - 729 (что составляет 0.08%) и с количеством знаков от 100 000 - 175 (что составляет 0,01%).

В разделе на Юэ число страниц с количеством знаков от 30 000 и более - 246 (что составляет 0,54%), с количеством знаков от 50 000 - 73 (что составляет 0,16%) и с количеством знаков от 100 000 - 10 (что составляет 0,02%).

Несмотря на различие в количестве статей, имеются сопоставимые статьи на уровне предложений. Был произведён анализ десяти самых крупных и сопоставимых статей в разделе (от 50 000 и 100 000 знаков). Их число составляет менее процента от общего числа статей раздела на путунхуа (873 861 статей), и на Юэ (44 550 статей), но они могут служить основой для представительного корпуса, особенно в условиях небольшого количества письменных текстов на Юэ.

С помощью СУБД (система управления базами данных) были обнаружены совпадающие лексические единицы в статьях. Во всех десяти статьях на путунхуа и на Юэ преимущественно наибольшее число совпадений приходится на глаголы, существительные, включая имена собственные.

Наиболее заметные лексические различия наблюдаются в служебных словах. Ниже приведены данные о процентном соотношении служебных слов в текстах (относительно общего объёма слов статей раздела на Юэ), их соответствие служебным словам раздела на путунхуа:

[U+5605] (соотв. [U+7684] в путунхуа) составляет 4% от общего числа слов в отобранных статьях

[U+55BA] (соотв. [U+5728] в путунхуа) составляет 1,1% от общего числа слов в отобранных статьях

[U+5497] (соотв. [U+4E86] и [U+8FC7] в путунхуа) составляет 0,5% от общего числа слов в отобранных статьях

[U+4F62] (соотв. местоимению 3 лица в путунхуа) составляет 0,25% от общего числа слов в отобранных статьях

[U+540C][U+57CB] (соотв. [U+548C]/[U+4E0E] в путунхуа) составляет 0,1% от общего числа слов в отобранных статьях

Для подтверждения нашей точки зрения анализа лексического наполнения текста статей было рассмотрено похожее исследование, проведенное китайским лингвистом. [2] Из него мы видим, что в процентном соотношении данные, полученные нами, в целом совпадают с данными, приведёнными в исследовании: лексическая составляющая диалектных иероглифов менее значительная в текстах, а общая с путунхуа лексика может доходить до 90%. Опираясь на полученные нами данные, можно также сказать, что общая лексика превышает 90% в статьях на диалекте Юэ, а число диалектных слов составляет менее 10%.

***

Таким образом, на основании проведённых исследований, можно сделать вывод о том, что Википедия может использоваться как источник параллельных текстов на путунхуа и письменном варианте Юэ, несмотря на несоответствие размеров разделов. Это обусловлено тем, что содержание текстов сопоставимых статей совпадает более чем на 50%. Что касается самых крупных статей из языковых разделов, то содержание 1/3 статей практически полностью совпадает (93% на уровне предложений), содержание оставшихся 2/3 совпадает не полностью детально, это связано с авторскими особенностями текстов. На основании полученного материала и данных мы можем сделать значимые выводы для изучения и исследования диалекта Юэ.

Источники и литература

1) Сысоев П.В. Лингвистический корпус в методике обучения иностранным языкам // Язык и культура. 2010, No. 1.

2) Ouyang Jueya, Comparison between Putonghua and Cantonese dialect. Beijing, 1993.