Секция «Лингвистика – языки стран Азии и Африки»

Принципы составления корпуса экономических медиатекстов (на материале языка суахили)

Научный руководитель – Громова Нелли Владимировна

Грушина Оксана Андреевна
Аспирант
Московский государственный университет имени М.В.Ломоносова, Институт стран Азии и Африки, Кафедра африканистики, Москва, Россия
E-mail: o.a.grushina@gmail.com

Целью работы является описание принципов создания корпуса медиатекстов на языке суахили для выявления как единичных, так и составных экономических терминов путем анализа полученных данных по частотности употребления n-грамм. Выявленные термины могут быть использованы для составления словаря после лексикографической обработки.

Корпусная лингвистика - один из разделов языкознания, занимающийся разработкой и использованием текстовых корпусов. Под корпусом понимается систематизированное собрание текстов обычно в электронном формате, используемое для лингвистического анализа. При составлении корпуса необходимо руководствоваться рядом принципов, как например, размер, сбалансированность, репрезентативность, релевантность для целей исследования, систематизированность с точки зрения структуры и содержания. Характер и тип текстов, включаемых в корпус, определяется целями исследования, например, общий и специализированный корпус, синхронический и диахронический и другие. Большинство корпусов сформированы в электронной форме, поэтому нужно учитывать формат текстов, который был бы пригоден для обработки компьютерными программами, используемыми для лингвистического анализа. Важен выбор программного обеспечения, с помощью которого проводится анализ.

В условиях глобализации информационного пространства интернет является источником большого количество различных текстов, которые могут быть использованы для составления необходимого корпуса. Для выявления современных актуальных экономических терминов на языке суахили релевантно использование медиатекстов интернет-версий периодических изданий экономической тематики.

Медиатекст - это текст массовой информации; термин является производным от понятия текста, определения которого зависят от специфики подхода к анализу данного понятия. При функционировании в сфере массовой коммуникации, текст обрастает медийными надбавками и получает расширенное толкование. Можно определить медиатекст как текст, бытующий в сфере средств массовой коммуникации, относящийся к любому виду и жанру, посредством которого осуществляется процесс речевого общения в данной сфере. Одна из основных задач средств массовой коммуникации - информирование читательской аудитории о событиях и изменениях, происходящих в различных сферах жизнедеятельности. Для выполнения данной функции лексическая база медиатекстов должна быть актуальной, своевременно реагировать на внешние изменения, находя способы передачи информации в корректной и достоверной форме о новых явлениях и событиях, что достигается благодаря расширению и обогащению языка за счет использования как внутренних, так и внешних ресурсов. Таким образом, именно медиатексты достаточно репрезентативны с точки зрения анализа «живого», современного языка, вычленения терминологии и выражений, актуальных на данный момент времени.

Наличие обновляемой базы интернет-версий медиатекстов танзанийских СМИ позволяет использовать методы корпусной лингвистики для анализа языка медиатекстов. В настоящее время существует небольшое число терминологических словарей на суахили по причине отсутствия достаточных технических и экономических ресурсов. Использование данных корпуса медиатекстов может применяться для выявления актуальной, бытующей терминологии той или иной тематики.

При составлении корпуса использованы медиатексты интернет-версий изданий HabariLeo *Сегодняшние новости* и Mtanzania *Танзаниец*, раздел Biashara na Uchiumi *Торговля и Экономика* за 2016-2017 годы. Для обработки медиатекстов использованы программы операционной системы FreeBSD семейства UNIX. Анализ медиатекстов позволил получить список частотности употребления слов, биграмм и триграмм. Последовательная проверка результатов продемонстрировала отсутствие в данных медиатекстах определений экономической терминологии в явной форме, что свидетельствует о том, что данные СМИ нацелены на аудиторию читателей, владеющих данной терминологией и не нуждающихся в дополнительных пояснениях. Путем анализа списка частотности употребления возможно выделить как единичные термины, так и составные, относящиеся к экономической тематике. Так, например, по списку частотности употребления выделяется термин fedha *деньги*, по списку биграмм можно идентифицировать словосочетания fedha taslimu *наличные деньги*, -pata fedha *получать деньги*, -weka fedha *вкладывать/инвестировать деньги* и другие. По списку триграмм можно выделить следующие выражения: mwaka wa fedha *финансовый год*, fedha za kigeni *иностранная валюта*, fedha za ndani *внутренняя валюта* и другие.

Результаты, полученные путем анализа корпуса экономических медиатекстов при помощи системы UNIX, показали возможность использования данного метода для вычленения терминологии определенной тематики, которая может быть в дальнейшем обработана лексикографически и использована для составления корпусного словаря. Неоспоримым преимуществом анализа корпуса медиатекстов является использование в них «живого» языка с лексической базой, актуальной для определенного периода времени. Сами по себе медиатексты не могут предоставить термины в чистом виде, вычленение терминологии требует применения корректно настроенной компьютерной базы, последовательного анализа списков частотности употребления n-грамм и их контекстов употребления.