МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Кормачева Дарья Андреевна

«Морфологический анализ текстов на прибалтийско-финских языках при помощи машинного обучения»

программа: «Прикладная, экспериментальная и математическая лингвистика»

Научный руководитель: к.ф.н., доц. Николаев И. С.

Данная работа ставит своей целью разработку алгоритма машинного обучения, лежащего в основе программы морфологического анализа текстов на прибалтийско-финских языках, которая была реализована на языке Python. Результатом работы этой программы является сегментация слов на основу и аффиксы на основании анализа неразмеченных текстовых данных.

Актуальность данного исследования обусловлена тем, что морфологический анализ является важным компонентом систем обработки естественного языка. Преимуществами применения методов обучения без учителя в этой области являются уменьшение временных затрат на разработку приложений, отсутствие потребности в каких-либо ресурсах помимо входного текста, применение одних и тех же технологий для большого количества новых языков, а также возможность анализа словоформ, которые не встретились в обучающей выборке и/или не присутствуют в словаре.

Новизна данной работы заключается в создании модуля морфологического анализа для текстов на прибалтийско-финских языках, в основе которого лежит разработанный нами вероятностный алгоритм. Насколько нам известно, мы впервые применили методы машинного обучения без учителя для морфологического анализа текстов на ижорском языке. Кроме того в процессе работы мы выявили определенные особенности поэтической метрики в текстах эпических песен на ижорском языке, влияющие на процессы словоизменения.

В качестве материала для эксперимента мы использовали тексты на финском и ижорском языках общим объемом 11662 и 5860 словоупотреблений соответственно.

Работа состоит из введения, трех глав, заключения, списка литературы и одного приложения (диска с программой). Список использованной литературы содержит 53 источника, в том числе 36 – на иностранных языках. Общий объем диссертации составляет 55 страниц.