На начальном этапе работы имелся корпус, состоящий из 37 драматургических произведений Шекспира (полный список см. в библиографии, раздел “Данные”), закодированных по правилам TEI и хранящимся в 37 файлах в формате xml, а также данные о разметке NUPOS v2, использующейся в корпусе, взятые с сайта classify.at.northwestern.edu, раздел corpussearch (полную ссылку см. в библиографии, раздел “Источники, используемые в программной части исследования”). Конечный результат представляет собой выдачу всех предложений из корпуса, в которых встречалась интересующая нас конструкция, с вынесением конструкции в отдельный столбик. Eщё до начала основной работы было произведено выкачивание интересующей нас системы разметки с сайта и её преобразование в формат, позволяющий простейшую машинную обработку (.tsv). Следующими шагами были многочисленные попытки анализа корпуса с помощью различных библиотек, рассчитанных на анализ xml и подобных.

По совету старшего коллеги автор начал работу, используя библиотеку xml.etree.ElementTree (далее просто ElementTree). Довольно быстро стало очевидно, что данная библиотека не подходит для такого типа работы. ElementTree позволяет опускаться на один уровень индентации за шаг: если для работы нужно несколько тэгов <a>, которые лежат в тэге <b>, несколько экземпляров которого лежит в тэге <c>, то сначала необходимо обработать структуру, содержащуюся в тэге <c>, потом несколько раз (по количеству тэгов <b>) опуститься на уровень ниже, а потом из каждого тэга <b> достать все тэги <c>, каждый раз спускаясь ещё на уровень ниже. Необходимый исследованию тэг <w> находится на седьмом уровне индентации, что делает использование ElementTree крайне неудобным, а код – громоздким и «уезжающим вправо».