Сначала скриптом собирается файл, содержащий все книги из корпуса (все файлы формата .txt в подкаталоге с названием corpus).

Теперь еще одним скриптом убираем некоторые сложные символы (‘(‘ , ‘)’, ‘*’, ‘[’, ‘]’), лишние пробелы, а также кавычки, двоеточия и точки с запятой. Составляем словарь с количеством появления слов в начале предложения, в конце, тех, что идут вторыми, а также для каждой пары последовательных слов ищем появляющиеся тройки. Считаем длины предложений. Все результаты записываем в файлы в виде словарей. Перед записью происходит предподсчет размера каждого множества, из которого будет происходить случайный выбор и результат записывается в словарь.

Сама генерация начинается со считывания всех словарей из файлов, после чего случайным образом начинается текст одним из слов, которые встречались в начале, аналогичным образом выбирается и второе слово, после чего ищется подходящая случайная тройка и т.д. В начале каждого предложения выбирается случайная его длина на основе распределения длин предложений в общем тексте, но при этом оно не может быть короче трех слов и заканчивается только в том случае, если последнее его слово заканчивало какое-нибудь предложение в корпусе.