Our goal here is to simplify wikipedia articles. We have to maintain similar content while also simplifying the style, ie, making sentences and words shorter and more common. This cannot be done by identifying rare words and replacing them using DRG since that doesn't change sentence length. Summarization algorithms don't work either since they usually cut down on content while maintaining the same style.

Here, existing text generators which can control the style of the output are used with some modification to reach our goal. The text generator is trained on already existing simple wiki articles. We use GPT-2 transformer for text generation and then create embeddings for the text generated using google encoder. Then, we calculate the difference between the embeddings of the sample generated and the original lexem from the article. If this difference is more than a certain threshold, it is rejected. So, the pre-trained text generator controls style, while the threshold on the embedding difference ensures the content is retained.

Testing shows that this does produce remarkable improvements in Flesch scores of readability, but there are still some drawbacks, for example, our algo doesn't know how to deal with new or rare words, doesn't make many replacements if we didn't have similar articles in the training set, replaces words with words with similar embeddings but different meanings.