Documentation de mes experimentations sur fairseq

1. Generer les BPE
2. Learn BPE

subword-nmt learn-bpe -s 4000 < new\_testament\_ew\_train.txt > bpe\_codes\_ew

subword-nmt learn-bpe -s 4000 < new\_testament\_fr\_train.txt > bpe\_codes\_fr

1. Apply BPE

subword-nmt apply-bpe -c bpe\_codes\_ew < new\_testament\_ew\_train.txt > train.bpe.ew

subword-nmt apply-bpe -c bpe\_codes\_fr < new\_testament\_fr\_train.txt > train.bpe.fr

subword-nmt apply-bpe -c bpe\_codes\_ew < new\_testament\_ew\_test.txt > test.bpe.ew

subword-nmt apply-bpe -c bpe\_codes\_ew < new\_testament\_ew\_dev.txt > dev.bpe.ew

subword-nmt apply-bpe -c bpe\_codes\_fr < new\_testament\_fr\_test.txt > test.bpe.fr

subword-nmt apply-bpe -c bpe\_codes\_fr < new\_testament\_fr\_dev.txt > dev.bpe.fr

1. Preprocess Data for Fairseq

Fairseq requires data to be in a binary format:

fairseq-preprocess --source-lang src --target-lang tgt \ --trainpref train.bpe --validpref valid.bpe --testpref test.bpe \ --destdir data-bin --thresholdtgt 0 --thresholdsrc 0 \ --workers 4