Splitting the Data

We will now discuss the train/val/test splits and perplexity.

Train/Val/Test splits

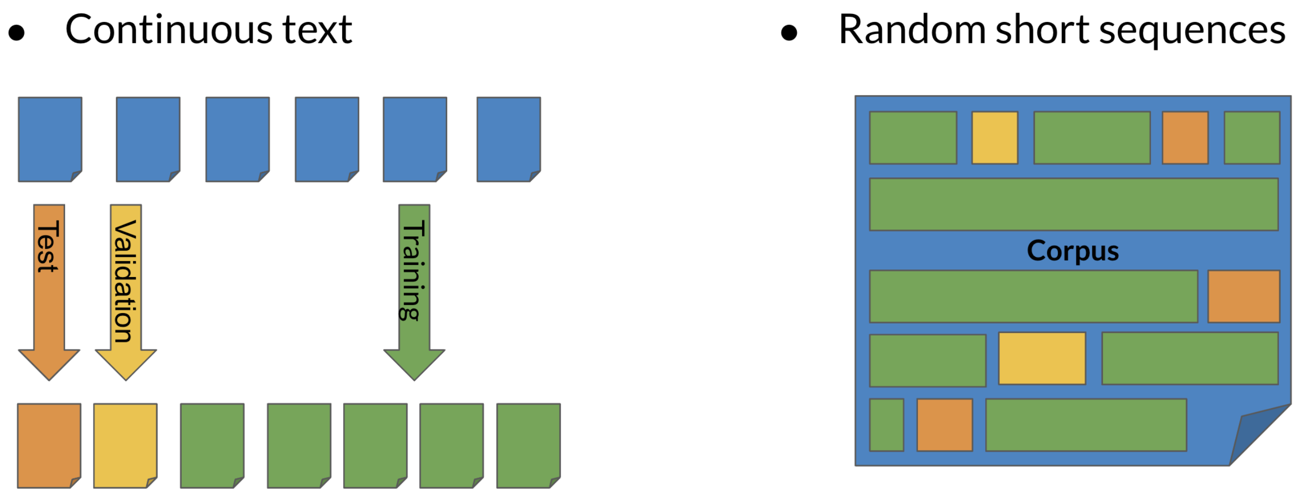
Smaller Corpora:

- 80% train
- 10% val
- 10% test

Larger Corpora:

- 98% train
- 1% val
- 1% test

There are two main methods for splitting the data:



- Continuous text
- Random short sequences