—------------- Dataset A

- Netflix
    - Scraping by using Language Rector
    - 284,229 sentences
- Ted
    - 170,155 sentences
    - CC BY-NC-ND
    - https://huggingface.co/datasets/msarmi9/korean-english-multitarget-ted-talks-task
- Kaggle NaverDictionary
    - 5,412 sentences
    - Copyright © NAVER Corp
- Kaggle Korean Single speaker Speech Dataset
    - 25,733 sentences
    - CC BY-NC-SA 4.0
- Manythings
    - 3,853 sentences
    - CC BY.
    - http://www.manythings.org/anki/
- MRL-2021
    - 9,844 sentences
    - (https://github.com/emorynlp/MRL-2021)
- KISS(Korean-english Idioms in Sentences dataSet)
    - 7,499 sentences
    - (https://github.com/Judy-Choi/KISS-Korean-english-Idioms-in-Sentences-dataSet)
- Korean Parallel corpora(Jungyeul)
    - 98,563 sentences
    - CC BY-SA 3.0
    - https://github.com/emorynlp/MRL-2021

—------------- Dataset B

- Tatoeba
    - 4,672,270 sentences
    - CC
    - https://github.com/Helsinki-NLP/Tatoeba-Challenge

—------------- Dataset C

```
(base) ➜  FINAL wc -l *.ko

  294049 konlp.test.ko

 5293998 konlp.train.ko

  294049 konlp.valid.ko

 5,882,096 total
```

- Custom Set scratched by Yoonjung Choi
  - 1,095 sentences
  - https://www.ef.com/wwen/english-resources/english-idioms/
  - https://www.theidioms.com/
- WMT
  - 4,520,346 sentences
  - CC BY-NC-SA 4.0
- Ko NLP
  - total 1,174,045 sentences
  - 3i4k
    - 52,180 sentences
    - CC-BY-SA 4.0
  - Chatbot
    - 19, 436 sentences
    - MIT License
  - KorNLU
    - 1,144,622 sentences
    - CC-BY-SA-4.0
  - paraKQC
    - 9,987 sentences
    - CC-BY-SA-4.0
  - Sae4k
    - 48,922 sentences
    - CC-BY-SA-4.0
  - Stylekqc
    - 29,970 sentences
    - CC-BY-SA 4.0
- Bitext
  - Inference_5000_samples.
    - MIT License.
    - https://github.com/snoop2head/Deep-Encoder-Shallow-Decoder/tree/main/result
  - Blog download.
    - 35,000 & 12,000 samples