

## General Information about used dataset:

- The lecture data was split into 80% train, 10% dev and 10% test datasets
- The daily dialogue data was used as provided and is also split into train/dev/test datasets
- The respective datasets were concatenated (e.g. train-train)
- Each set was then split into chunks of 10 successive sentences
- Those chunks were shuffled
- Concatenation of chunks and creation of 64 word sequences

### *Stanford lecture data:*

Number of sentences in train: **61824**

Number of sentences in dev: **7728**

Number of sentences in test: **7729**

### *Daily Dialogue data:*

Number of sentences in train: **133260**

Number of sentences in dev: **12447**

Number of sentences in test: **12064**

### *Tokens concatenated data:*

Train: **2381400**

Dev: **285287**

Test: **266135**