Sophistication

Step 1 Concatenate all corpus files into one txt file in the data folder

bash concatenate_files.sh prompts_n_coherence/data/ --> feature_extraction/data

Step 2 Run the main script

bash sophistication.sh --> feature_extraction/results/sophistication_scores.csv

Lexical richness

bash lxr_scores.sh prompts_n_coherence/data/

Morphology

Step 1 Extract vocabulary of most frequent words

bash create_most_freq_vocs.sh : prompts_n_coherence/data/ --> scripts/freq_voc/, scripts/lemmas/

Step 2 Run diversity analysis

for single file

python3 shannon_pairwise.py -f ~/switchdrive/IMAGINE_files/datasets/wmtnews21/wmtnews_test_de_A.txt -l de -sys A_wmt -v freq_voc/wmtnews_test_de_A.freq_voc > test.txt

for multiple files in a directory if considering the top 1000 most frequent lemmas with more than 1 morphological form

lang = {"en", "de"}

bash shannon_1000_mostfrequent_script.sh ~/switchdrive/IMAGINE_files/chatGPT/project_2/final_files_simple_prompt/{corpus} lang

if chosing all lemmas with more than one morphological form

lang = {"en", "de"}

bash mrph_all.sh ~/switchdrive/IMAGINE_files/chatGPT/project_2/final_files_simple_prompt/{corpus} lang

Extract Features with TextDescriptives

extract features as is

bash run_extract_fetures.sh --> iterates over a list of corpora --> ../results/per_corpus/{corpus}/{i}.csv"

Collects features from TextDescriptives, replaapplying custom formula for German Flesch Reading Ease metric. Additionally counts connectives.

rearrange results first based on feature then on language

python3 combine_results_per_feat_corpus.py :

```
iterates over:  ../results/per_corpus/{corpus}
writes to :     ../results/per_feature/{feature_to_extract}/{corpus}.csv
                ../results/per_language/german/{feature}.csv
                ../results/per_language/english/{feature}.csv
```

create data