2_0_Dataset_Preparation_with_Original_Split_Balance_FINAL

April 13, 2025

This notebook performs a thorough EDA of the dataset, as well as engineers features, and performs additional QA on these modifications.

We: - Standardize columns

- Search for duplicates and other flaws
- Understand and QA (pre and post-modifications) the distributions of our spans across subcorpora
- Understand and QA (pre and post-modifications) the quartiles of the continous outcome variable 'complexity', the counts across datasets and subsets, the counts of subcorpora, the quantity of more complex vs. less complex spans—coming to the conclusion that we have a rigorously balanced and consistent dataset
- Eliminate contractions across the data, creating 2 versions of the original X variable, per single and multi set (4 total)
- Create two binarized outcome variables (derived from the continous 'complexity' outcome variable), such that we split on the median of the single set's train set and apply it to the validation and test set, and repeat the procedure for the multi set's train set. Then, in order to test for excessive neutrality (given that the continuous label was derived from a 5-increment likert scale, which itself was derived from an average of continuous annotator ratings), we split on the 75th percentile and trained models on that in order to rule out intrinsic issues with the dataset. The resulting datset balances can be seen in the binomial distribution plots in this notebook. Thus, we create 2 versions of the original Y variable, per single and multi set (4 total), excluding the original continous variable from consideration for training.
- Enrich and augment the dataset with features derived from SpaCy, creating 4 new purely derived features per set (8 total). We then systematically leverage the derived features to generate 7 new versions of our X variable per dataset (14 total)—such that 3 features per set (6 total) contain concatenations of engineered features and our expanded raw X variable, 3 features per set (6 total) interleave each token of both the SpaCy features and the original X variable, and 1 feature per set was generated concatenating the raw X variable with a SpaCy-derived complexity score (2 total)—thus injecting an alternative continous value into the input sequence.

At each step, modifications were tested for quality control, and later used systematically in training experiments.

[ ]:

```
#@title Install Packages
```

```
[ ]: !pip install -q transformers
     !pip install -q torchinfo
     !pip install -q datasets
     !pip install -q evaluate
     !pip install -q nltk
     !pip install -q contractions
```

```
[ ]: !sudo apt-get update
     ! sudo apt-get install tree
```

```
Hit:1 https://developer.download.nvidia.com/compute/cuda/repos/ubuntu2204/x86_64
InRelease
Hit:2 https://cloud.r-project.org/bin/linux/ubuntu jammy-cran40/ InRelease
Hit:3 http://security.ubuntu.com/ubuntu jammy-security InRelease
Hit:4 http://archive.ubuntu.com/ubuntu jammy InRelease
Hit:5 http://archive.ubuntu.com/ubuntu jammy-updates InRelease
Hit:6 https://r2u.stat.illinois.edu/ubuntu jammy InRelease
Hit:7 https://ppa.launchpadcontent.net/deadsnakes/ppa/ubuntu jammy InRelease
Hit:8 http://archive.ubuntu.com/ubuntu jammy-backports InRelease
Hit:9 https://ppa.launchpadcontent.net/graphics-drivers/ppa/ubuntu jammy
InRelease
Hit:10 https://ppa.launchpadcontent.net/ubuntugis/ppa/ubuntu jammy InRelease
Reading package lists… Done
W: Skipping acquire of configured file 'main/source/Sources' as repository
'https://r2u.stat.illinois.edu/ubuntu jammy InRelease' does not seem to provide
it (sources.list entry misspelt?)
Reading package lists… Done
Building dependency tree… Done
Reading state information… Done
tree is already the newest version (2.0.2-1).
0 upgraded, 0 newly installed, 0 to remove and 32 not upgraded.
```

```
[ ]: #@title Imports
     import nltk
     from nltk.tokenize import RegexpTokenizer

     import evaluate
     import transformers

     import contractions

     from torchinfo import summary
     from datasets import load_dataset

     from transformers import AutoTokenizer, AutoModel,␣
       ↪AutoModelForSequenceClassification
     from transformers import TrainingArguments, Trainer
```

```
import os
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

import sklearn

import spacy
```

[ ]:

```
# @title Mount Google Drive
```

[ ]:

```
from google.colab import drive
drive.mount('/content/drive')
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call
drive.mount("/content/drive", force_remount=True).
```

[ ]:

```
dir_root = '/content/drive/MyDrive/266-final/'
# dir_data = '/content/drive/MyDrive/266-final/data/'
# dir_data = '/content/drive/MyDrive/266-final/data/se21-t1-comp-lex-master/'
dir_data = '/content/drive/MyDrive/266-final/data/266-comp-lex-master'
dir_models = '/content/drive/MyDrive/266-final/models/'
dir_results = '/content/drive/MyDrive/266-final/results/'
```

[ ]:

```
!tree /content/drive/MyDrive/266-final/data/266-comp-lex-master/
```

```
/content/drive/MyDrive/266-final/data/266-comp-lex-master/
  fe-test-labels
      test_multi_df.csv
      test_single_df.csv
  fe-train
      train_multi_df.csv
      train_single_df.csv
  fe-trial-val
      trial_val_multi_df.csv
      trial_val_single_df.csv
  test-labels
      lcp_multi_test.tsv
      lcp_single_test.tsv
  train
      lcp_multi_train.tsv
      lcp_single_train.tsv
  trial
      lcp_multi_trial.tsv
      lcp_single_trial.tsv
```

```
6 directories, 12 files
```

[ ]:

```
!ls -R /content/drive/MyDrive/266-final/data/266-comp-lex-master/
```

```
/content/drive/MyDrive/266-final/data/266-comp-lex-master/:
fe-test-labels  fe-train  fe-trial-val  test-labels  train  trial

/content/drive/MyDrive/266-final/data/266-comp-lex-master/fe-test-labels:
test_multi_df.csv  test_single_df.csv

/content/drive/MyDrive/266-final/data/266-comp-lex-master/fe-train:
train_multi_df.csv  train_single_df.csv

/content/drive/MyDrive/266-final/data/266-comp-lex-master/fe-trial-val:
trial_val_multi_df.csv  trial_val_single_df.csv

/content/drive/MyDrive/266-final/data/266-comp-lex-master/test-labels:
lcp_multi_test.tsv  lcp_single_test.tsv

/content/drive/MyDrive/266-final/data/266-comp-lex-master/train:
lcp_multi_train.tsv  lcp_single_train.tsv

/content/drive/MyDrive/266-final/data/266-comp-lex-master/trial:
lcp_multi_trial.tsv  lcp_single_trial.tsv
```

[ ]:

```
#@title Import Data
```

[ ]:

```
# Load train data into train_*_df
train_single_df = pd.read_csv(
    os.path.join(dir_data, "train", "lcp_single_train.tsv"),
    sep = "\t",
    engine = "python",
    quoting = 3
)
train_multi_df = pd.read_csv(
    os.path.join(dir_data, "train", "lcp_multi_train.tsv"),
    sep = "\t",
    engine = "python",
    quoting = 3
)

# Load trial data into trial_val_*_df
trial_val_single_df = pd.read_csv(
    os.path.join(dir_data, "trial", "lcp_single_trial.tsv"),
    sep = "\t",
    engine = "python",
    quoting = 3
)
```

```
trial_val_multi_df = pd.read_csv(
    os.path.join(dir_data, "trial", "lcp_multi_trial.tsv"),
    sep = "\t",
    engine = "python",
    quoting = 3
)

# Load test data (with labels) into test_*_df
test_single_df = pd.read_csv(
    os.path.join(dir_data, "test-labels", "lcp_single_test.tsv"),
    sep = "\t",
    engine = "python",
    quoting = 3
)
test_multi_df = pd.read_csv(
    os.path.join(dir_data, "test-labels", "lcp_multi_test.tsv"),
    sep = "\t",
    engine = "python",
    quoting = 3
)

print("Data successfully loaded into train, trial-val, and test variables")
```

```
Data successfully loaded into train, trial-val, and test variables
```

[ ]:

```
#@title EDA
```

[ ]:

```
def print_dataframe_summary(df_name, df):
    # Print section header
    print(f"========== {df_name} ==========")

    # Shape and Columns
    print(f"Shape: {df.shape}")
    print(f"Columns: {list(df.columns)}\n")

    # Data Types
    print("Data Types:")
    print(df.dtypes)
    print()

    # Missing Values
    print("Missing Values (by column):")
    print(df.isna().sum())
    print()

    # 'complexity' column stats
    desc = df['complexity'].describe()  # count, mean, std, min, 25%, 50%, 75%,␣
  ↪max
```

```
    print("'complexity' Column Stats (incl. quartiles and median):")
    print(desc)

    # Calculate frequency counts for each quartile range
    q1 = desc['25%']
    q2 = desc['50%']  # This is the median
    q3 = desc['75%']
    q_max = desc['max']

    # Note: We'll define the ranges as:
    #   <= Q1
    #   > Q1 and <= Q2
    #   > Q2 and <= Q3
    #   > Q3

    freq_q1 = np.sum(df['complexity'] <= q1)
    freq_q2 = np.sum((df['complexity'] > q1) & (df['complexity'] <= q2))
    freq_q3 = np.sum((df['complexity'] > q2) & (df['complexity'] <= q3))
    freq_q4 = np.sum(df['complexity'] > q3)

    print()
    print("Quartile Frequency Counts (tab-separated next to each quartile):")
    print(f"25%: {q1}\tCount (<= Q1): {freq_q1}")
    print(f"50% (Median): {q2}\tCount (Q1 < x <= Q2): {freq_q2}")
    print(f"75%: {q3}\tCount (Q2 < x <= Q3): {freq_q3}")
    print(f"100% (Max): {q_max}\tCount (Q3 < x <= Max): {freq_q4}")

    print("======================================\n")

# Now we call this for each of our dataframes
print_dataframe_summary("train_single_df", train_single_df)
print_dataframe_summary("train_multi_df", train_multi_df)
print_dataframe_summary("trial_val_single_df", trial_val_single_df)
print_dataframe_summary("trial_val_multi_df", trial_val_multi_df)
print_dataframe_summary("test_single_df", test_single_df)
print_dataframe_summary("test_multi_df", test_multi_df)
```

```
========== train_single_df ==========
Shape: (7662, 5)
Columns: ['id', 'corpus', 'sentence', 'token', 'complexity']

Data Types:
id               object
corpus           object
sentence         object
token            object
complexity      float64
dtype: object
```

```
Missing Values (by column):
id            0
corpus        0
sentence      0
token         7
complexity    0
dtype: int64

'complexity' Column Stats (incl. quartiles and median):
count    7662.000000
mean        0.302288
std         0.132977
min         0.000000
25%         0.211538
50%         0.279412
75%         0.375000
max         0.861111
Name: complexity, dtype: float64

Quartile Frequency Counts (tab-separated next to each quartile):
25%: 0.2115384615384615 Count (<= Q1): 1928
50% (Median): 0.2794117647058823        Count (Q1 < x <= Q2): 1937
75%: 0.375      Count (Q2 < x <= Q3): 1984
100% (Max): 0.8611111111111112  Count (Q3 < x <= Max): 1813
======================================

========== train_multi_df ==========
Shape: (1517, 5)
Columns: ['id', 'corpus', 'sentence', 'token', 'complexity']

Data Types:
id             object
corpus         object
sentence       object
token          object
complexity    float64
dtype: object

Missing Values (by column):
id            0
corpus        0
sentence      0
token         0
complexity    0
dtype: int64

'complexity' Column Stats (incl. quartiles and median):
```

```
count    1517.000000
mean        0.418362
std         0.155536
min         0.027778
25%         0.302632
50%         0.409091
75%         0.529412
max         0.975000
Name: complexity, dtype: float64

Quartile Frequency Counts (tab-separated next to each quartile):
25%: 0.3026315789473685 Count (<= Q1): 382
50% (Median): 0.409090909090909 Count (Q1 < x <= Q2): 377
75%: 0.5294117647058824 Count (Q2 < x <= Q3): 380
100% (Max): 0.975       Count (Q3 < x <= Max): 378
======================================

========== trial_val_single_df ==========
Shape: (421, 5)
Columns: ['id', 'subcorpus', 'sentence', 'token', 'complexity']

Data Types:
id               object
subcorpus        object
sentence         object
token            object
complexity      float64
dtype: object

Missing Values (by column):
id            0
subcorpus     0
sentence      0
token         0
complexity    0
dtype: int64

'complexity' Column Stats (incl. quartiles and median):
count    421.000000
mean       0.298631
std        0.137619
min        0.000000
25%        0.214286
50%        0.266667
75%        0.359375
max        0.875000
Name: complexity, dtype: float64
```

```
Quartile Frequency Counts (tab-separated next to each quartile):
25%: 0.2142857142857143 Count (<= Q1): 106
50% (Median): 0.2666666666666667        Count (Q1 < x <= Q2): 107
75%: 0.359375   Count (Q2 < x <= Q3): 103
100% (Max): 0.875       Count (Q3 < x <= Max): 105
======================================

========== trial_val_multi_df ==========
Shape: (99, 5)
Columns: ['id', 'subcorpus', 'sentence', 'token', 'complexity']

Data Types:
id             object
subcorpus      object
sentence       object
token          object
complexity    float64
dtype: object

Missing Values (by column):
id            0
subcorpus     0
sentence      0
token         0
complexity    0
dtype: int64

'complexity' Column Stats (incl. quartiles and median):
count    99.000000
mean      0.417961
std       0.153752
min       0.000000
25%       0.309028
50%       0.421875
75%       0.513932
max       0.825000
Name: complexity, dtype: float64

Quartile Frequency Counts (tab-separated next to each quartile):
25%: 0.3090277777777778 Count (<= Q1): 25
50% (Median): 0.421875  Count (Q1 < x <= Q2): 25
75%: 0.5139318885448916 Count (Q2 < x <= Q3): 24
100% (Max): 0.825       Count (Q3 < x <= Max): 25
======================================

========== test_single_df ==========
Shape: (917, 5)
Columns: ['id', 'corpus', 'sentence', 'token', 'complexity']
```

```
Data Types:
id             object
corpus         object
sentence       object
token          object
complexity    float64
dtype: object

Missing Values (by column):
id            0
corpus        0
sentence      0
token         0
complexity    0
dtype: int64

'complexity' Column Stats (incl. quartiles and median):
count    917.000000
mean       0.296362
std        0.127290
min        0.000000
25%        0.214286
50%        0.276316
75%        0.357143
max        0.777778
Name: complexity, dtype: float64

Quartile Frequency Counts (tab-separated next to each quartile):
25%: 0.2142857142857143 Count (<= Q1): 237
50% (Median): 0.2763157894736842        Count (Q1 < x <= Q2): 224
75%: 0.3571428571428571 Count (Q2 < x <= Q3): 229
100% (Max): 0.7777777777777777  Count (Q3 < x <= Max): 227
======================================

========== test_multi_df ==========
Shape: (184, 5)
Columns: ['id', 'corpus', 'sentence', 'token', 'complexity']

Data Types:
id             object
corpus         object
sentence       object
token          object
complexity    float64
dtype: object

Missing Values (by column):
```

```
id            0
corpus        0
sentence      0
token         0
complexity    0
dtype: int64

'complexity' Column Stats (incl. quartiles and median):
count    184.000000
mean       0.422312
std        0.155785
min        0.000000
25%        0.316667
50%        0.428571
75%        0.527778
max        0.800000
Name: complexity, dtype: float64

Quartile Frequency Counts (tab-separated next to each quartile):
25%: 0.31666666666666665 Count (<= Q1): 47
50% (Median): 0.4285714285714286        Count (Q1 < x <= Q2): 46
75%: 0.5277777777777778 Count (Q2 < x <= Q3): 46
100% (Max): 0.8 Count (Q3 < x <= Max): 45
======================================
```

[ ]:
```python
print(train_single_df.head())
```

```
                               id corpus
sentence     token  complexity
0  3ZLW647WALVGE8EBR50EGUBPU4P32A  bible  Behold, there came up out of the river
seven c…     river    0.000000
1  34R0BODSP1ZBN3DVY8J8XSIY551E5C  bible  I am a fellow bondservant with you and
with yo…  brothers    0.000000
2  3S1WOPCJFGTJU2SGNAN2Y213N6WJE3  bible  The man, the lord of the land, said to
us, 'By…  brothers    0.050000
3  3BFNCI9LYKQN09BHXHH9CLSX5KP738  bible  Shimei had sixteen sons and six
daughters; but…  brothers    0.150000
4  3G5RUKN2EC3YIWSKUXZ8ZVH95R49N2  bible               "He has put my brothers
far from me.  brothers    0.263889
```

[ ]:
```python
print(train_multi_df.head())
```

```
                               id corpus
sentence            token  complexity
0  3S37Y8CWI80N8KVM53U4E6JKCDC4WE  bible  but the seventh day is a Sabbath to
Yahweh you…      seventh day    0.027778
1  3WGCNLZJKF877FYC1Q6COKNWTDWD11  bible  But let each man test his own work,
and then h…         own work    0.050000
```

```
2  3UOMW19E6D6WQ5TH2HDD74IVKTP5CB  bible  To him who by understanding made the
heavens; …  loving kindness    0.050000
3  36JW4WBR06KF9AXMUL4N476OMF8FHD  bible  Remember to me, my God, this also, and
spare m…  loving kindness    0.050000
4  3HRWUH63QU2FH9Q8R7MRNFC7JX2N5A  bible  Because your loving kindness is better
than li…  loving kindness    0.050000
```

[ ]:

```
#@title Data Engineering
```

[ ]:

```
# Assuming you have already loaded the DataFrames:
# train_single_df, train_multi_df, trial_val_single_df, trial_val_multi_df,␣
 ↪test_single_df, test_multi_df

def print_distinct_values(df, column_name):
    """Prints the distinct values of a specified column in a DataFrame."""
    distinct_values = df[column_name].unique()
    print(f"Distinct values in '{column_name}' column:")
    for value in distinct_values:
        print(value)
    print("-" * 30)  # Separator

# Print distinct values for each DataFrame
print_distinct_values(train_single_df, "corpus")
print_distinct_values(train_multi_df, "corpus")
print_distinct_values(trial_val_single_df, "subcorpus")
print_distinct_values(trial_val_multi_df, "subcorpus")
print_distinct_values(test_single_df, "corpus")
print_distinct_values(test_multi_df, "corpus")
```

```
Distinct values in 'corpus' column:
bible
biomed
europarl
------------------------------
Distinct values in 'corpus' column:
bible
biomed
europarl
------------------------------
Distinct values in 'subcorpus' column:
bible
biomed
europarl
------------------------------
Distinct values in 'subcorpus' column:
bible
biomed
europarl
```

```
------------------------------
Distinct values in 'corpus' column:
bible
biomed
europarl
------------------------------
Distinct values in 'corpus' column:
bible
biomed
europarl
------------------------------
```

0.1 standardize column headers: convert trial_val header from ‘subcorpus’ to ‘corpus’

[ ]:

```
# Rename the 'subcorpus' column to 'corpus'
trial_val_single_df = trial_val_single_df.rename(columns={'subcorpus':␣
 ↪'corpus'})
trial_val_multi_df = trial_val_multi_df.rename(columns={'subcorpus': 'corpus'})

# Verify the change (optional)
print(trial_val_single_df.columns)
print(trial_val_multi_df.columns)
```

```
Index(['id', 'corpus', 'sentence', 'token', 'complexity'], dtype='object')
Index(['id', 'corpus', 'sentence', 'token', 'complexity'], dtype='object')
```

[ ]:

```
dataframes = [train_single_df, train_multi_df, trial_val_single_df,␣
 ↪trial_val_multi_df, test_single_df, test_multi_df]

# Get the headers (column names) of the first DataFrame as a reference
reference_headers = list(dataframes[0].columns)

# Loop through the remaining DataFrames and compare headers
all_headers_match = True
for df in dataframes[1:]:
    if list(df.columns) != reference_headers:
        all_headers_match = False
        print(f"Headers do not match for DataFrame: {df.head(0)}")  # Print␣
 ↪which DataFrame has different headers
        break  # Exit the loop if a mismatch is found

# Print the result
if all_headers_match:
    print("All DataFrames have matching headers.")
else:
    print("Headers do not match for all DataFrames.")
```

```
All DataFrames have matching headers.
```

0.2 Interrogate Span Length by Corpus Value by Data Split

[ ]:

```
tokenizer = RegexpTokenizer(r'\w+')

def analyze_sentence_spans_by_corpus_and_quartile(dfs_dict):
    """
    Analyze sentence spans (length metrics) grouped by corpus and complexity␣
 ↪quartile
    for multiple dataframes.
    """
    results = []

    for df_name, df in dfs_dict.items():
        print(f"Processing {df_name}...")

        q1 = df['complexity'].quantile(0.25)
        q2 = df['complexity'].quantile(0.50)
        q3 = df['complexity'].quantile(0.75)

        def get_quartile(x):
            if x <= q1:
                return 'Q1'
            elif x <= q2:
                return 'Q2'
            elif x <= q3:
                return 'Q3'
            else:
                return 'Q4'

        df = df.copy()
        df['quartile'] = df['complexity'].apply(get_quartile)

        def compute_span_metrics(sentence):
            if pd.isna(sentence):
                return pd.Series({'word_count': 0, 'char_count': 0,␣
 ↪'avg_word_len': 0})

            words = tokenizer.tokenize(sentence)
            word_count = len(words)
            char_count = len(sentence)
            avg_word_len = np.mean([len(word) for word in words]) if word_count␣
 ↪> 0 else 0
            return pd.Series({'word_count': word_count, 'char_count':␣
 ↪char_count, 'avg_word_len': avg_word_len})
```

```
        span_metrics = df['sentence'].apply(compute_span_metrics)
        df = pd.concat([df, span_metrics], axis=1)

        corpus_col = 'corpus' if 'corpus' in df.columns else 'subcorpus'

        for corpus_name, corpus_df in df.groupby(corpus_col):
            for quartile, quartile_df in corpus_df.groupby('quartile'):
                complexity_range = f"{quartile_df['complexity'].min():.
 ↪3f}-{quartile_df['complexity'].max():.3f}"
                stats = {
                    'Dataframe': df_name,
                    'Corpus': corpus_name,
                    'Quartile': quartile,
                    'Complexity Range': complexity_range,
                    'Count': len(quartile_df),
                    'Avg Words': quartile_df['word_count'].mean(),
                    'Median Words': quartile_df['word_count'].median(),
                    'Min Words': quartile_df['word_count'].min(),
                    'Max Words': quartile_df['word_count'].max(),
                    'Std Words': quartile_df['word_count'].std(),
                    'Avg Chars': quartile_df['char_count'].mean(),
                    'Avg Word Len': quartile_df['avg_word_len'].mean()
                }
                results.append(stats)

    results_df = pd.DataFrame(results)
    results_df = results_df.sort_values(['Dataframe', 'Corpus', 'Quartile'])

    return results_df

dfs = {
    'train_single_df': train_single_df,
    'train_multi_df': train_multi_df,
    'trial_val_single_df': trial_val_single_df,
    'trial_val_multi_df': trial_val_multi_df,
    'test_single_df': test_single_df,
    'test_multi_df': test_multi_df
}

span_analysis = analyze_sentence_spans_by_corpus_and_quartile(dfs)

pd.set_option('display.max_rows', None)
pd.set_option('display.max_columns', None)
pd.set_option('display.width', 1000)
display(span_analysis)

results_path = os.path.join(dir_results, 'sentence_span_analysis.csv')
```

```python
span_analysis.to_csv(results_path, index=False)
print(f"Analysis saved to: {results_path}")
```

```
Processing train_single_df…
Processing train_multi_df…
Processing trial_val_single_df…
Processing trial_val_multi_df…
Processing test_single_df…
Processing test_multi_df…

              Dataframe    Corpus Quartile Complexity Range  Count   Avg Words ␣
 ↪Median Words  Min Words  Max Words  Std Words   Avg Chars  Avg Word Len
60        test_multi_df     bible       Q1      0.025-0.317     26  23.076923   ␣
 ↪      22.0         4.0       48.0  11.831900  118.653846      4.128898
61        test_multi_df     bible       Q2      0.325-0.417     11  20.545455   ␣
 ↪      17.0         7.0       47.0  12.917923  109.545455      4.209752
62        test_multi_df     bible       Q3      0.432-0.528     18  21.111111   ␣
 ↪      21.5         4.0       43.0  10.889222  112.777778      4.474206
63        test_multi_df     bible       Q4      0.542-0.694     11  22.363636   ␣
 ↪      20.0         7.0       51.0  11.935432  126.181818      4.605062
64        test_multi_df    biomed       Q1      0.000-0.312     11  29.818182   ␣
 ↪      29.0        17.0       47.0   8.388304  195.727273      5.491145
65        test_multi_df    biomed       Q2      0.324-0.417     11  27.090909   ␣
 ↪      24.0         9.0       47.0  11.449494  171.818182      5.436237
66        test_multi_df    biomed       Q3      0.456-0.528     10  26.900000   ␣
 ↪      26.5        10.0       49.0  10.712921  177.500000      5.497409
67        test_multi_df    biomed       Q4      0.562-0.800     21  32.285714   ␣
 ↪      34.0        14.0       56.0  13.598319  209.285714      5.460101
68        test_multi_df  europarl       Q1      0.214-0.303     10  24.700000   ␣
 ↪      24.5         7.0       56.0  14.189589  146.900000      5.049688
69        test_multi_df  europarl       Q2      0.321-0.429     24  27.833333   ␣
 ↪      27.0         9.0       73.0  15.352855  172.291667      5.269610
70        test_multi_df  europarl       Q3      0.432-0.516     18  32.944444   ␣
 ↪      32.0         6.0       68.0  19.129504  209.888889      5.512245
71        test_multi_df  europarl       Q4      0.531-0.562     13  39.000000   ␣
 ↪      36.0         6.0       95.0  29.631065  237.076923      5.100616
48       test_single_df     bible       Q1      0.000-0.214     79  22.835443   ␣
 ↪      22.0         7.0       49.0  10.602891  116.797468      4.031532
49       test_single_df     bible       Q2      0.217-0.276     68  24.176471   ␣
 ↪      21.0         2.0       77.0  14.393138  125.955882      4.167352
50       test_single_df     bible       Q3      0.278-0.353     67  22.388060   ␣
 ↪      20.0         4.0       63.0  11.306950  119.731343      4.254090
51       test_single_df     bible       Q4      0.359-0.732     69  20.579710   ␣
 ↪      19.0         1.0       55.0  11.264736  110.550725      4.337010
52       test_single_df    biomed       Q1      0.000-0.214     75  27.080000   ␣
 ↪      25.0        10.0       84.0  12.025603  172.893333      5.271985
```

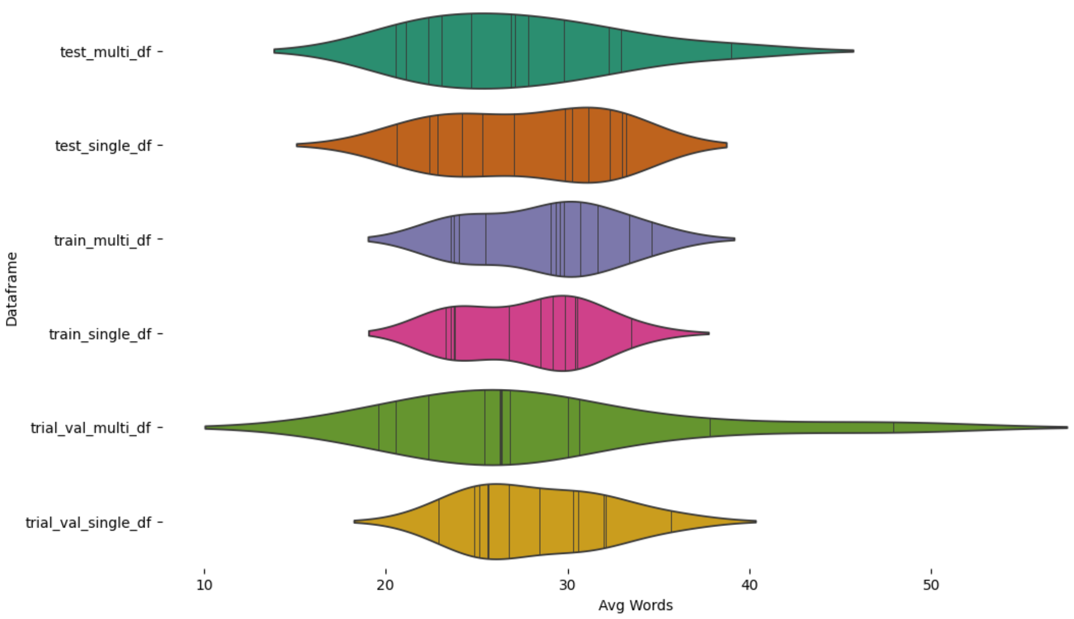
```
53        test_single_df    biomed       Q2      0.217-0.275      58  30.275862    ␣
 ↪       26.0         10.0        83.0  15.856587  197.775862      5.434573
54        test_single_df    biomed       Q3      0.278-0.357      66  29.833333    ␣
 ↪       29.0         13.0        85.0  11.754650  191.863636      5.334048
55        test_single_df    biomed       Q4      0.359-0.778      90  31.144444    ␣
 ↪       30.0         14.0        83.0  12.089146  203.055556      5.393138
56        test_single_df  europarl       Q1      0.000-0.214      83  25.337349    ␣
 ↪       21.0          3.0        82.0  16.032191  151.891566      5.044222
57        test_single_df  europarl       Q2      0.217-0.276      98  32.326531    ␣
 ↪       30.0          1.0        97.0  18.707061  195.653061      5.062296
58        test_single_df  europarl       Q3      0.278-0.357      96  33.000000    ␣
 ↪       30.0          3.0       141.0  21.404377  201.760417      5.124551
59        test_single_df  europarl       Q4      0.361-0.583      68  33.235294    ␣
 ↪       29.0          1.0       130.0  20.440023  206.514706      5.164123
12        train_multi_df     bible       Q1      0.028-0.300     163  23.588957    ␣
 ↪       22.0          3.0        67.0  12.429421  124.834356      4.232989
13        train_multi_df     bible       Q2      0.304-0.409     132  24.053030    ␣
 ↪       22.0          6.0        65.0  11.738444  129.575758      4.302615
14        train_multi_df     bible       Q3      0.411-0.529     131  23.770992    ␣
 ↪       23.0          4.0        50.0  11.158691  127.389313      4.324088
15        train_multi_df     bible       Q4      0.533-0.778      79  25.481013    ␣
 ↪       24.0          3.0        81.0  13.490605  139.240506      4.486716
16        train_multi_df    biomed       Q1      0.028-0.303      87  29.091954    ␣
 ↪       28.0          9.0        77.0  11.882792  185.954023      5.276290
17        train_multi_df    biomed       Q2      0.304-0.408      74  30.716216    ␣
 ↪       28.0         11.0        85.0  13.521693  195.864865      5.370313
18        train_multi_df    biomed       Q3      0.411-0.529     111  29.783784    ␣
 ↪       29.0          8.0        61.0  10.912383  193.855856      5.430133
19        train_multi_df    biomed       Q4      0.531-0.975     242  29.595041    ␣
 ↪       28.0         10.0        75.0  12.040443  194.995868      5.534629
20        train_multi_df  europarl       Q1      0.118-0.303     132  29.363636    ␣
 ↪       27.0          3.0       101.0  17.874146  176.553030      5.002618
21        train_multi_df  europarl       Q2      0.304-0.409     171  31.654971    ␣
 ↪       28.0          3.0       108.0  19.099221  195.152047      5.176834
22        train_multi_df  europarl       Q3      0.411-0.529     138  33.398551    ␣
 ↪       30.0          7.0       101.0  18.992715  208.304348      5.286607
23        train_multi_df  europarl       Q4      0.533-0.750      57  34.596491    ␣
 ↪       31.0          6.0        96.0  20.318763  218.350877      5.345891
0        train_single_df     bible       Q1      0.000-0.212     701  23.275321    ␣
 ↪       22.0          4.0        61.0  11.760701  121.607703      4.126789
1        train_single_df     bible       Q2      0.212-0.279     640  23.753125    ␣
 ↪       22.0          3.0        60.0  11.577932  124.576562      4.148961
2        train_single_df     bible       Q3      0.281-0.375     624  23.823718    ␣
 ↪       22.0          3.0        70.0  11.958906  126.230769      4.208102
3        train_single_df     bible       Q4      0.380-0.861     609  23.577997    ␣
 ↪       21.0          3.0        69.0  12.461688  126.518883      4.295608
```

```
4       train_single_df    biomed       Q1      0.000-0.212    586  28.534130   ␣
 ↪      27.0          2.0        85.0  12.115387  182.011945     5.319754
5       train_single_df    biomed       Q2      0.212-0.279    583  30.435678   ␣
 ↪      29.0          7.0        92.0  11.872558  193.789022     5.285758
6       train_single_df    biomed       Q3      0.281-0.375    659  29.860395   ␣
 ↪      28.0          4.0        77.0  11.591263  191.050076     5.328161
7       train_single_df    biomed       Q4      0.381-0.861    748  29.176471   ␣
 ↪      28.0          3.0        85.0  12.246613  186.909091     5.298112
8       train_single_df  europarl       Q1      0.025-0.212    641  26.761310   ␣
 ↪      24.0          2.0       107.0  15.230853  159.180967     4.942557
9       train_single_df  europarl       Q2      0.212-0.279    714  30.420168   ␣
 ↪      27.0          1.0       129.0  18.383783  183.093838     4.995672
10      train_single_df  europarl       Q3      0.281-0.375    701  30.523538   ␣
 ↪      28.0          1.0       122.0  18.163026  185.840228     5.114587
11      train_single_df  europarl       Q4      0.381-0.775    456  33.528509   ␣
 ↪      31.0          2.0       235.0  21.704693  203.592105     5.054701
36   trial_val_multi_df     bible       Q1      0.000-0.292     11  26.272727   ␣
 ↪      21.0         13.0        64.0  13.950562  141.363636     4.282457
37   trial_val_multi_df     bible       Q2      0.333-0.400      7  20.571429   ␣
 ↪      23.0          5.0        28.0   7.412987  110.857143     4.279406
38   trial_val_multi_df     bible       Q3      0.425-0.500      5  19.600000   ␣
 ↪      19.0          9.0        32.0   8.905055  109.200000     4.431391
39   trial_val_multi_df     bible       Q4      0.525-0.661      6  22.333333   ␣
 ↪      20.5          9.0        44.0  12.242004  117.833333     4.178525
40   trial_val_multi_df    biomed       Q1      0.083-0.303      6  26.833333   ␣
 ↪      25.0         15.0        49.0  11.771434  159.166667     4.899969
41   trial_val_multi_df    biomed       Q2      0.317-0.422      7  25.428571   ␣
 ↪      21.0         15.0        48.0  11.588171  156.000000     5.194383
42   trial_val_multi_df    biomed       Q3      0.438-0.513      6  37.833333   ␣
 ↪      39.5         26.0        44.0   6.675827  247.500000     5.438593
43   trial_val_multi_df    biomed       Q4      0.537-0.825     14  30.642857   ␣
 ↪      29.5         17.0        43.0   9.849695  211.428571     5.730623
44   trial_val_multi_df  europarl       Q1      0.176-0.306      8  30.000000   ␣
 ↪      25.5          4.0        64.0  20.361027  186.750000     5.306837
45   trial_val_multi_df  europarl       Q2      0.312-0.412     11  47.909091   ␣
 ↪      46.0         24.0        78.0  18.651834  296.909091     5.058375
46   trial_val_multi_df  europarl       Q3      0.432-0.500     13  26.307692   ␣
 ↪      26.0          5.0        66.0  18.167666  166.153846     5.263847
47   trial_val_multi_df  europarl       Q4      0.515-0.714      5  26.400000   ␣
 ↪      15.0          6.0        66.0  24.316661  164.600000     4.998182
24  trial_val_single_df     bible       Q1      0.000-0.214     52  26.750000   ␣
 ↪      26.0          5.0        73.0  15.530962  137.230769     4.071006
25  trial_val_single_df     bible       Q2      0.217-0.266     38  24.868421   ␣
 ↪      23.0          7.0        50.0  10.768249  131.236842     4.195550
26  trial_val_single_df     bible       Q3      0.268-0.355     26  22.884615   ␣
 ↪      20.5          5.0        44.0   9.961233  121.269231     4.312026
```

```
27  trial_val_single_df     bible       Q4      0.361-0.633     27  25.666667    ␣
 ↪       23.0        6.0       49.0  12.554497  137.555556      4.212685
28  trial_val_single_df    biomed       Q1      0.028-0.214     21  25.571429    ␣
 ↪       21.0       13.0       65.0  11.543706  163.904762      5.305404
29  trial_val_single_df    biomed       Q2      0.217-0.267     28  30.571429    ␣
 ↪       27.5       11.0       57.0  12.099674  198.142857      5.315287
30  trial_val_single_df    biomed       Q3      0.268-0.359     38  32.105263    ␣
 ↪       29.0       11.0       61.0  12.710476  206.947368      5.364934
31  trial_val_single_df    biomed       Q4      0.364-0.875     48  25.145833    ␣
 ↪       25.5        6.0       56.0  11.721937  163.979167      5.439709
32  trial_val_single_df  europarl       Q1      0.050-0.214     33  31.969697    ␣
 ↪       28.0        5.0       81.0  20.356947  185.969697      4.799024
33  trial_val_single_df  europarl       Q2      0.217-0.267     41  28.463415    ␣
 ↪       28.0        4.0       71.0  15.386841  172.780488      4.997706
34  trial_val_single_df  europarl       Q3      0.268-0.359     39  30.282051    ␣
 ↪       28.0        3.0       99.0  20.040681  184.358974      5.086945
35  trial_val_single_df  europarl       Q4      0.367-0.605     30  35.700000    ␣
 ↪       30.5        5.0       77.0  20.142852  215.400000      4.910759

Analysis saved to:
/content/drive/MyDrive/266-final/results/sentence_span_analysis.csv
```

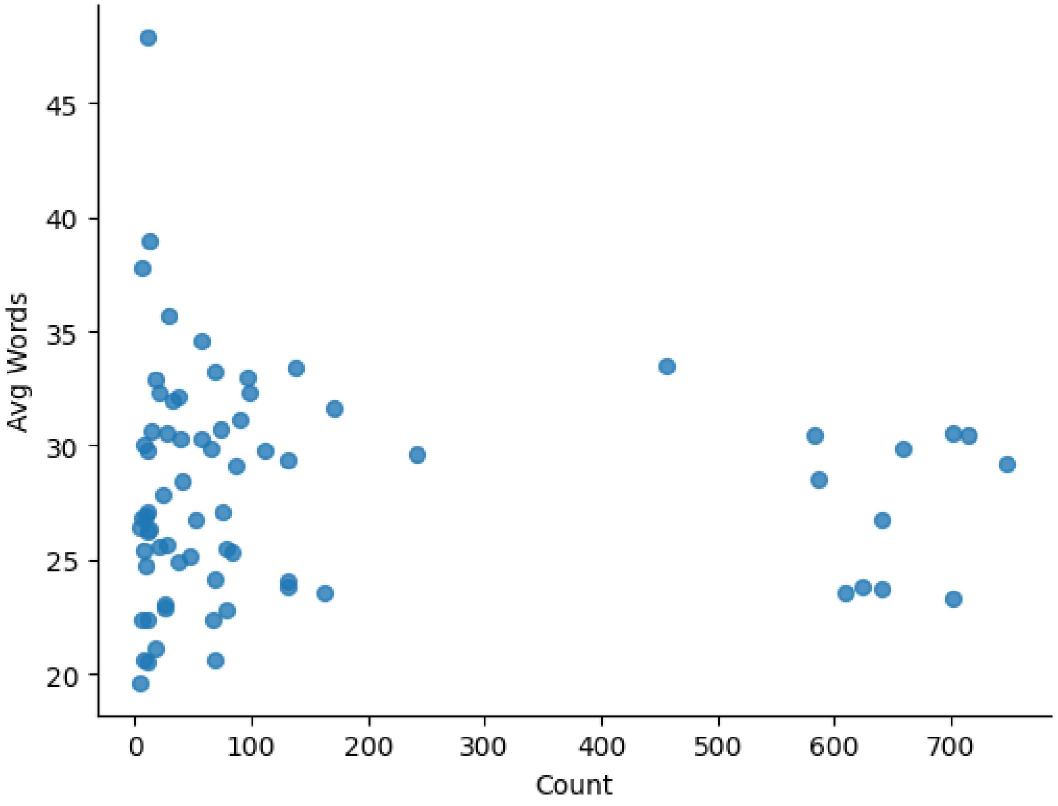
```
[ ]: from matplotlib import pyplot as plt
     import seaborn as sns
     figsize = (12, 1.2 * len(span_analysis['Dataframe'].unique()))
     plt.figure(figsize=figsize)
     sns.violinplot(span_analysis, x='Avg Words', y='Dataframe', inner='stick',␣
       ↪palette='Dark2')
     sns.despine(top=True, right=True, bottom=True, left=True)
```

```
<ipython-input-56-00a8ad5642c1>:5: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in
v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same
effect.

  sns.violinplot(span_analysis, x='Avg Words', y='Dataframe', inner='stick',
palette='Dark2')
```



[ ]:

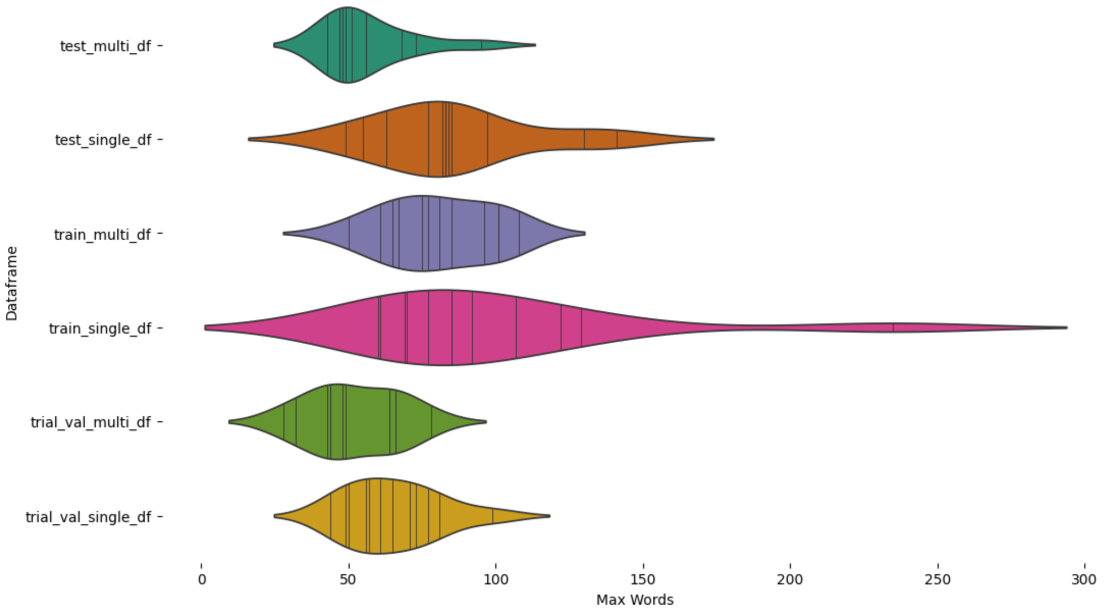
```
from matplotlib import pyplot as plt
span_analysis.plot(kind='scatter', x='Count', y='Avg Words', s=32, alpha=.8)
plt.gca().spines[['top', 'right',]].set_visible(False)
```



[ ]:

```python
from matplotlib import pyplot as plt
import seaborn as sns
figsize = (12, 1.2 * len(span_analysis['Dataframe'].unique()))
plt.figure(figsize=figsize)
sns.violinplot(span_analysis, x='Max Words', y='Dataframe', inner='stick',␣
 ↪palette='Dark2')
sns.despine(top=True, right=True, bottom=True, left=True)
```

```
<ipython-input-58-01bf0c89d620>:5: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in
v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same
effect.

  sns.violinplot(span_analysis, x='Max Words', y='Dataframe', inner='stick',
palette='Dark2')
```

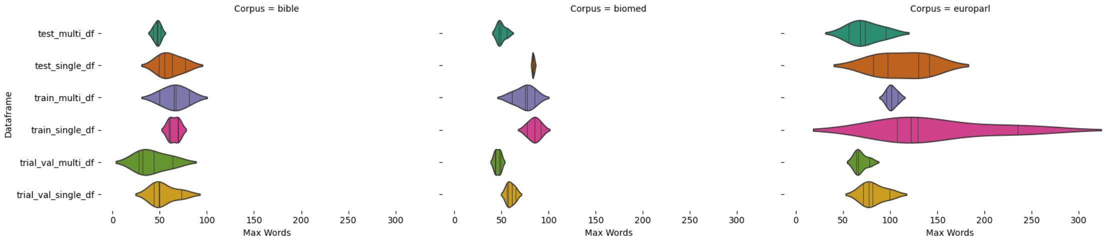


[ ]:

```python
g = sns.FacetGrid(span_analysis, col="Corpus", col_wrap=3, height=4, aspect=1.5)
g.map(sns.violinplot, "Max Words", "Dataframe", inner='stick', palette='Dark2')
g.despine(top=True, right=True, bottom=True, left=True)
plt.tight_layout()
plt.show()
```

```
/usr/local/lib/python3.11/dist-packages/seaborn/axisgrid.py:718: UserWarning:
Using the violinplot function without specifying `order` is likely to produce an
```

```
incorrect plot.
  warnings.warn(warning)
/usr/local/lib/python3.11/dist-packages/seaborn/axisgrid.py:854: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in
v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same
effect.

  func(*plot_args, **plot_kwargs)
/usr/local/lib/python3.11/dist-packages/seaborn/axisgrid.py:854: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in
v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same
effect.

  func(*plot_args, **plot_kwargs)
/usr/local/lib/python3.11/dist-packages/seaborn/axisgrid.py:854: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in
v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same
effect.

  func(*plot_args, **plot_kwargs)
```



- decision: no modifications to sentence spans will be applied, except for Contraction standardization

0.3 Normalize / Eliminate Contractions

```
[ ]: def expand_contractions_in_df(df):
         """
         1) Creates a new column 'sentence_no_contractions' by expanding any␣
     ↪contractions.
         2) Identifies rows where a contraction was actually expanded (the text␣
     ↪changed).
         3) Returns the updated DataFrame and a grouped subset of rows for printing␣
     ↪examples.
```

```
    """
    df = df.copy()
    df['sentence_no_contractions'] = df['sentence'].apply(
        lambda s: contractions.fix(s) if pd.notna(s) else s
    )

    df['contraction_expanded'] = df.apply(
        lambda row: row['sentence'] != row['sentence_no_contractions'], axis=1
    )

    results_by_corpus = {}
    for corpus_val, group in df.groupby('corpus'):
        changed_rows = group[group['contraction_expanded']]
        first_three = changed_rows.head(3)
        results_by_corpus[corpus_val] = first_three
    return df, results_by_corpus


dataframes_info = [
    ("train_single_df", train_single_df),
    ("train_multi_df", train_multi_df),
    ("trial_val_single_df", trial_val_single_df),
    ("trial_val_multi_df", trial_val_multi_df),
    ("test_single_df", test_single_df),
    ("test_multi_df", test_multi_df),
]

for df_name, df in dataframes_info:
    updated_df, corpus_examples = expand_contractions_in_df(df)
    globals()[df_name] = updated_df

    print(f"\n{'='*60}")
    print(f"DataFrame: {df_name}")
    print(f"{'='*60}")

    for corpus_val in sorted(corpus_examples.keys()):
        subset = corpus_examples[corpus_val]
        if len(subset) == 0:
            continue
        print(f"\n  Corpus: {corpus_val}")
        print("    -- BEFORE --")
        for _, row in subset.iterrows():
            print(f"       {row['sentence']}")
        print("    -- AFTER  --")
        for _, row in subset.iterrows():
            print(f"       {row['sentence_no_contractions']}")
```

```
============================================================
DataFrame: train_single_df
============================================================

  Corpus: bible
    -- BEFORE --
       Shimei had sixteen sons and six daughters; but his brothers didn't have
many children, neither did all their family multiply like the children of Judah.
       When his speech is charming, don't believe him; for there are seven
abominations in his heart.
       Jesus said, "Father, forgive them, for they don't know what they are
doing."
    -- AFTER  --
       Shimei had sixteen sons and six daughters; but his brothers did not have
many children, neither did all their family multiply like the children of Judah.
       When his speech is charming, do not believe him; for there are seven
abominations in his heart.
       Jesus said, "Father, forgive them, for they do not know what they are
doing."

  Corpus: biomed
    -- BEFORE --
       Although missense mutation of ITPR1 had previously been ruled out [2] and
the mode of inheritance was inconsistent with that seen in the Itpr1Δ18 and
Itpr1opt mice, the phenotypic presence of ataxia in the mice led us to reexamine
this candidate gene as a possible cause of SCA15.
       Human germline mutations in APC cause FAP [4,5], which is characterized
by hundreds of adenomatous colorectal polyps, with an almost inevitable
progression to colorectal cancer in the third and fourth decades of life.
       Null mutations in Bmpr1a cause early embryonic lethality, with defects in
gastrulation similar to those seen in mice with mutations in Bmp4 (Mishina et
al. 1995; Winnier et al. 1995).
    -- AFTER  --
       Although missense mutation of ITPR1 had previously been ruled out [2] and
the mode of inheritance was inconsistent with that seen in the Itpr1Δ18 and
Itpr1opt mice, the phenotypic presence of ataxia in the mice led us to reexamine
this candidate gene as a possible because of SCA15.
       Human germline mutations in APC because FAP [4,5], which is characterized
by hundreds of adenomatous colorectal polyps, with an almost inevitable
progression to colorectal cancer in the third and fourth decades of life.
       Null mutations in Bmpr1a because early embryonic lethality, with defects
in gastrulation similar to those seen in mice with mutations in Bmp4 (Mishina et
al. 1995; Winnier et al. 1995).

  Corpus: europarl
    -- BEFORE --
       At the same time, you will also have an important role in winning over
the general public of the Member States to the cause of enlargement, of
```

```
enlargement based on conditionality.
       the recommendation for second reading from the Committee on Transport and
Tourism on the common position adopted by the Council with a view to the
adoption of a Regulation of the European Parliament and of the Council
establishing common rules concerning the conditions to be complied with to
pursue the occupation of road transport operator and repealing Council Directive
96/26/EC (11783/1/2008 - C6-0015/2009 - (Rapporteur: Silvia-Adriana Ţicău), and
       Yet, although credit rating agencies were not the main cause of the
recent financial crisis, they did have a harmful influence.
    -- AFTER  --
       At the same time, you will also have an important role in winning over
the general public of the Member States to the because of enlargement, of
enlargement based on conditionality.
       the recommendation for second reading from the Committee on Transport and
Tourism on the common position adopted by the Council with a view to the
adoption of a Regulation of the European Parliament and of the Council
establishing common rules concerning the conditions to be complied with to
pursue the occupation of road transport operator and repealing Council Directive
96/26/EC (11783/1/2008 - C6-0015/2009 - (Rapporteur: Silvia-Adriana Ţicăyou),
and
       Yet, although credit rating agencies were not the main because of the
recent financial crisis, they did have a harmful influence.

============================================================
DataFrame: train_multi_df
============================================================

  Corpus: bible
    -- BEFORE --
       Jahath was the chief, and Zizah the second: but Jeush and Beriah didn't
have many sons; therefore they became a fathers' house in one reckoning.
       But Yahweh said to Samuel, "Don't look on his face, or on the height of
his stature; because I have rejected him: for I see not as man sees; for man
looks at the outward appearance, but Yahweh looks at the heart."
       Because indeed a notable miracle has been done through them, as can be
plainly seen by all who dwell in Jerusalem, and we can't deny it.
    -- AFTER  --
       Jahath was the chief, and Zizah the second: but Jeush and Beriah did not
have many sons; therefore they became a fathers' house in one reckoning.
       But Yahweh said to Samuel, "Do not look on his face, or on the height of
his stature; because I have rejected him: for I see not as man sees; for man
looks at the outward appearance, but Yahweh looks at the heart."
       Because indeed a notable miracle has been done through them, as can be
plainly seen by all who dwell in Jerusalem, and we cannot deny it.

  Corpus: biomed
    -- BEFORE --
       The aim in the present study was to determine the location of pendrin and
```

```
the cause of deafness in Slc26a4-/- mice.
       These characteristics should make RMCE-ASAP a robust and general
technology for analysis of mammalian genes under conditions that preserve normal
control mechanisms in different tissues.
       It was also demonstrated that mutations leading to abolishment of the
enzymatic activity of CLN2 were the direct cause of a fatal inherited
neurodegenerative disease, classical late-infantile neuronal ceroid
lipofuscinosis [2].
    -- AFTER  --
       The aim in the present study was to determine the location of pendrin and
the because of deafness in Slc26a4-/- mice.
       These characteristics should make RMCE-AS SOON AS POSSIBLE a robust and
general technology for analysis of mammalian genes under conditions that
preserve normal control mechanisms in different tissues.
       It was also demonstrated that mutations leading to abolishment of the
enzymatic activity of CLN2 were the direct because of a fatal inherited
neurodegenerative disease, classical late-infantile neuronal ceroid
lipofuscinosis [2].

  Corpus: europarl
    -- BEFORE --
       Account must also be taken of the costs to health, the environment and
the climate of the fact that vehicles emit different types of particles and
that, in burning fossil fuels, they cause increased pollution and thus more
global warming.
       However, this unequal trade relationship is not the only cause for
concern; another is the case of unsafe products coming from China.
       (IT) Madam President, ladies and gentlemen, the oral amendment that our
Group is proposing involves replacing the words 'all forms of glorifying' by the
word 'apology'.
    -- AFTER  --
       Account must also be taken of the costs to health, the environment and
the climate of the fact that vehicles emit different types of particles and
that, in burning fossil fuels, they because increased pollution and thus more
global warming.
       However, this unequal trade relationship is not the only because for
concern; another is the case of unsafe products coming from China.
       (IT) Madam President, ladies and gentlemen, the oral amendment that our
Group is proposing involves replacing the words  forms of glorifying' by the
word 'apology'.

============================================================
DataFrame: trial_val_single_df
============================================================

  Corpus: bible
    -- BEFORE --
       Don't curse the king, no, not in your thoughts; and don't curse the rich
```

```
in your bedroom: for a bird of the sky may carry your voice, and that which has
wings may tell the matter.
       The young man didn't wait to do this thing, because he had delight in
Jacob's daughter, and he was honored above all the house of his father.
       If the axe is blunt, and one doesn't sharpen the edge, then he must use
more strength; but skill brings success.
    -- AFTER  --
       Do not curse the king, no, not in your thoughts; and do not curse the
rich in your bedroom: for a bird of the sky may carry your voice, and that which
has wings may tell the matter.
       The young man did not wait to do this thing, because he had delight in
Jacob's daughter, and he was honored above all the house of his father.
       If the axe is blunt, and one does not sharpen the edge, then he must use
more strength; but skill brings success.

  Corpus: biomed
    -- BEFORE --
       For example, the non-BC individual and BC individual groups are not
perfectly matched with respect to age, gender or smoking history (Table 1) and
each of these factors could contribute to the observed difference in correlation
between groups.
       EM and ER conducted transmission electron microscopy.
    -- AFTER  --
       For example, the non-BECAUSE individual and BECAUSE individual groups are
not perfectly matched with respect to age, gender or smoking history (Table 1)
and each of these factors could contribute to the observed difference in
correlation between groups.
       THEM and ER conducted transmission electron microscopy.

  Corpus: europarl
    -- BEFORE --
       With their help, John has sought to shed light on what has been a very
murky area, and to bring clarity where uncertainty prevailed before, based
consistently on the twin principles that the patient must always come first and
that patient choice should be determined by needs and not by means.
    -- AFTER  --
       With their help, John has sought to she would light on what has been a
very murky area, and to bring clarity where uncertainty prevailed before, based
consistently on the twin principles that the patient must always come first and
that patient choice should be determined by needs and not by means.

============================================================
DataFrame: trial_val_multi_df
============================================================

============================================================
DataFrame: test_single_df
============================================================
  Corpus: bible
    -- BEFORE --
       the ten sons of Haman the son of Hammedatha, the Jew's enemy, but they
didn't lay their hand on the plunder.
       Hezekiah listened to them, and showed them all the house of his precious
things, the silver, and the gold, and the spices, and the precious oil, and the
house of his armor, and all that was found in his treasures: there was nothing
in his house, nor in all his dominion, that Hezekiah didn't show them.
       Of Manasseh also there fell away some to David, when he came with the
Philistines against Saul to battle; but they didn't help them; for the lords of
the Philistines sent him away after consultation, saying, "He will fall away to
his master Saul to the jeopardy of our heads."
    -- AFTER  --
       the ten sons of Haman the son of Hammedatha, the Jew's enemy, but they
did not lay their hand on the plunder.
       Hezekiah listened to them, and showed them all the house of his precious
things, the silver, and the gold, and the spices, and the precious oil, and the
house of his armor, and all that was found in his treasures: there was nothing
in his house, nor in all his dominion, that Hezekiah did not show them.
       Of Manasseh also there fell away some to David, when he came with the
Philistines against Saul to battle; but they did not help them; for the lords of
the Philistines sent him away after consultation, saying, "He will fall away to
his master Saul to the jeopardy of our heads."

  Corpus: biomed
    -- BEFORE --
       In that study, there was a tendency towards correlation in transcript
abundance between several pairs of antioxidant or DNA repair genes in non-BC
individuals, but not in BC individuals.
       This, in turn, leads to increased representation among BC individuals of
individuals with lack of correlation between CEBPG and each of the affected
antioxidant and/or DNA repair genes.
       The 'pregnancy rate' in mice is defined as successful pregnancies per
detected vaginal plug, a phenotype associated with early pregnancy failure,
which in turn possibly could have an inflammatory cause.
    -- AFTER  --
       In that study, there was a tendency towards correlation in transcript
abundance between several pairs of antioxidant or DNA repair genes in non-
BECAUSE individuals, but not in BECAUSE individuals.
       This, in turn, leads to increased representation among BECAUSE
individuals of individuals with lack of correlation between CEBPG and each of
the affected antioxidant and/or DNA repair genes.
       The 'pregnancy rate' in mice is defined as successful pregnancies per
detected vaginal plug, a phenotype associated with early pregnancy failure,
which in turn possibly could have an inflammatory because.

  Corpus: europarl
```

```
    -- BEFORE --
       The next item is the oral question to the Commission (B7-0240/2009) by
Silvia-Adriana Ţicău, Brian Simpson, János Áder, Hannes Swoboda, Eva
Lichtenberger, Michael Cramer, Saïd El Khadraoui, Mathieu Grosch, Iuliu Winkler,
Victor Boştinaru, Ioan Mircea Paşcu, Marian-Jean Marinescu, Ivailo Kalfin,
Norica Nicolai, Dirk Sterckx, Csaba Sándor Tabajdi, Michael Theurer, Ismail
Ertug, Inés Ayala Sender, Jiří Havel, Edit Herczog, Stanimir Ilchev, Iliana
Malinova Iotova, Jelko Kacin, Evgeni Kirilov, Ádám Kósa, Ioan Enciu, Eduard
Kukan, Gesine Meissner, Alajos Mészáros, Nadezhda Neynsky, Katarína Neveďalová,
Daciana Octavia Sârbu, Vilja Savisaar, Olga Sehnalová, Catherine Stihler, Peter
van Dalen, Louis Grech, Corina Creţu, George Sabin Cutaş, Vasilica Viorica
Dăncilă, Cătălin Sorin Ivan, Tanja Fajon, Kinga Göncz, Antonyia Parvanova,
Adina-Ioana Vălean and Rovana Plumb, on the European Strategy for the Danube
Region.
    -- AFTER  --
       The next item is the oral question to the Commission (B7-0240/2009) by
Silvia-Adriana Ţicăyou, Brian Simpson, János Áder, Hannes Swoboda, Eva
Lichtenberger, Michael Cramer, Saïd El Khadraoui, Mathieu Grosch, Iuliu Winkler,
Victor Boştinaru, Ioan Mircea Paşcu, Marian-Jean Marinescu, Ivailo Kalfin,
Norica Nicolai, Dirk Sterckx, Csaba Sándor Tabajdi, Michael Theurer, Ismail
Ertug, Inés Ayala Sender, Jiří Havel, Edit Herczog, Stanimir Ilchev, Iliana
Malinova Iotova, Jelko Kacin, Evgeni Kirilov, Ádám Kósa, Ioan Enciu, Eduard
Kukan, Gesine Meissner, Alajos Mészáros, Nadezhda Neynsky, Katarína Neveďalová,
Daciana Octavia Sârbu, Vilja Savisaar, Olga Sehnalová, Catherine Stihler, Peter
van Dalen, Louis Grech, Corina Creţyou, George Sabin Cutaş, Vasilica Viorica
Dăncilă, Cătălin Sorin Ivan, Tanja Fajon, Kinga Göncz, Antonyia Parvanova,
Adina-Ioana Vălean and Rovana Plumb, on the European Strategy for the Danube
Region.

============================================================
DataFrame: test_multi_df
============================================================

  Corpus: bible
    -- BEFORE --
       Yet he didn't leave himself without witness, in that he did good and gave
you rains from the sky and fruitful seasons, filling our hearts with food and
gladness."
       When he has leveled its surface, doesn't he plant the dill, and scatter
the cumin seed, and put in the wheat in rows, the barley in the appointed place,
and the spelt in its place?
       Don't count your handmaid for a wicked woman; for I have been speaking
out of the abundance of my complaint and my provocation."
    -- AFTER  --
       Yet he did not leave himself without witness, in that he did good and
gave you rains from the sky and fruitful seasons, filling our hearts with food
and gladness."
       When he has leveled its surface, does not he plant the dill, and scatter
```

```
the cumin seed, and put in the wheat in rows, the barley in the appointed place,
and the spelt in its place?
       Do not count your handmaid for a wicked woman; for I have been speaking
out of the abundance of my complaint and my provocation."
```

[ ]:

```
# check for null values

dataframes = [train_single_df, train_multi_df, trial_val_single_df,␣
 ↪trial_val_multi_df, test_single_df, test_multi_df]
for df in dataframes:
  print(df['sentence_no_contractions'].isnull().values.any())
```

```
False
False
False
False
False
False
```

[ ]:

```
dataframes = {
    "train_single_df": train_single_df,
    "train_multi_df": train_multi_df,
    "trial_val_single_df": trial_val_single_df,
    "trial_val_multi_df": trial_val_multi_df,
    "test_single_df": test_single_df,
    "test_multi_df": test_multi_df
}

total_true_counts = 0
for df_name, df in dataframes.items():
    true_count = df['contraction_expanded'].sum()
    print(f"{df_name}: {true_count} True values in 'contraction_expanded'")
    total_true_counts += true_count

print(f"\nTotal True values across all dataframes: {total_true_counts}")
```

```
train_single_df: 254 True values in 'contraction_expanded'
train_multi_df: 54 True values in 'contraction_expanded'
trial_val_single_df: 16 True values in 'contraction_expanded'
trial_val_multi_df: 0 True values in 'contraction_expanded'
test_single_df: 31 True values in 'contraction_expanded'
test_multi_df: 7 True values in 'contraction_expanded'

Total True values across all dataframes: 362
```

[ ]:

```
# verify column headers
```

```
dataframes = [train_single_df, train_multi_df, trial_val_single_df,␣
 ↪trial_val_multi_df, test_single_df, test_multi_df]
for df in dataframes:
  print(df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7662 entries, 0 to 7661
Data columns (total 7 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   id                       7662 non-null   object
 1   corpus                   7662 non-null   object
 2   sentence                 7662 non-null   object
 3   token                    7655 non-null   object
 4   complexity               7662 non-null   float64
 5   sentence_no_contractions  7662 non-null   object
 6   contraction_expanded     7662 non-null   bool
dtypes: bool(1), float64(1), object(5)
memory usage: 366.8+ KB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1517 entries, 0 to 1516
Data columns (total 7 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   id                       1517 non-null   object
 1   corpus                   1517 non-null   object
 2   sentence                 1517 non-null   object
 3   token                    1517 non-null   object
 4   complexity               1517 non-null   float64
 5   sentence_no_contractions  1517 non-null   object
 6   contraction_expanded     1517 non-null   bool
dtypes: bool(1), float64(1), object(5)
memory usage: 72.7+ KB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 421 entries, 0 to 420
Data columns (total 7 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   id                       421 non-null    object
 1   corpus                   421 non-null    object
 2   sentence                 421 non-null    object
 3   token                    421 non-null    object
 4   complexity               421 non-null    float64
 5   sentence_no_contractions  421 non-null    object
 6   contraction_expanded     421 non-null    bool
dtypes: bool(1), float64(1), object(5)
```

```
memory usage: 20.3+ KB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 99 entries, 0 to 98
Data columns (total 7 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   id                        99 non-null     object
 1   corpus                    99 non-null     object
 2   sentence                  99 non-null     object
 3   token                     99 non-null     object
 4   complexity                99 non-null     float64
 5   sentence_no_contractions  99 non-null     object
 6   contraction_expanded      99 non-null     bool
dtypes: bool(1), float64(1), object(5)
memory usage: 4.9+ KB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 917 entries, 0 to 916
Data columns (total 7 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   id                        917 non-null    object
 1   corpus                    917 non-null    object
 2   sentence                  917 non-null    object
 3   token                     917 non-null    object
 4   complexity                917 non-null    float64
 5   sentence_no_contractions  917 non-null    object
 6   contraction_expanded      917 non-null    bool
dtypes: bool(1), float64(1), object(5)
memory usage: 44.0+ KB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 184 entries, 0 to 183
Data columns (total 7 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   id                        184 non-null    object
 1   corpus                    184 non-null    object
 2   sentence                  184 non-null    object
 3   token                     184 non-null    object
 4   complexity                184 non-null    float64
 5   sentence_no_contractions  184 non-null    object
 6   contraction_expanded      184 non-null    bool
dtypes: bool(1), float64(1), object(5)
memory usage: 8.9+ KB
None
```

[ ]:

```
# inspect each df

dataframes = [train_single_df, train_multi_df, trial_val_single_df,␣
 ↪trial_val_multi_df, test_single_df, test_multi_df]
for df in dataframes:
  print(df.head())
```

```
                               id corpus
sentence     token  complexity
sentence_no_contractions  contraction_expanded
0  3ZLW647WALVGE8EBR50EGUBPU4P32A  bible  Behold, there came up out of the river
seven c…     river    0.000000  Behold, there came up out of the river seven
c…                 False
1  34R0B0DSP1ZBN3DVY8J8XSIY551E5C  bible  I am a fellow bondservant with you and
with yo…  brothers    0.000000  I am a fellow bondservant with you and with
yo…                 False
2  3S1WOPCJFGTJU2SGNAN2Y213N6WJE3  bible  The man, the lord of the land, said to
us, 'By…  brothers    0.050000  The man, the lord of the land, said to us,
'By…                 False
3  3BFNCI9LYKQN09BHXHH9CLSX5KP738  bible  Shimei had sixteen sons and six
daughters; but…  brothers    0.150000  Shimei had sixteen sons and six
daughters; but…                  True
4  3G5RUKN2EC3YIWSKUXZ8ZVH95R49N2  bible               "He has put my brothers
far from me.  brothers    0.263889               "He has put my brothers far
from me.                 False
                               id corpus
sentence            token  complexity
sentence_no_contractions  contraction_expanded
0  3S37Y8CWI80N8KVM53U4E6JKCDC4WE  bible  but the seventh day is a Sabbath to
Yahweh you…      seventh day    0.027778  but the seventh day is a Sabbath to
Yahweh you…                 False
1  3WGCNLZJKF877FYC1Q6C0KNWTDWD11  bible  But let each man test his own work,
and then h…         own work    0.050000  But let each man test his own work,
and then h…                 False
2  3U0MW19E6D6WQ5TH2HDD74IVKTP5CB  bible  To him who by understanding made the
heavens; …  loving kindness    0.050000  To him who by understanding made the
heavens; …                 False
3  36JW4WBR06KF9AXMUL4N476OMF8FHD  bible  Remember to me, my God, this also, and
spare m…  loving kindness    0.050000  Remember to me, my God, this also, and
spare m…                 False
4  3HRWUH63QU2FH9Q8R7MRNFC7JX2N5A  bible  Because your loving kindness is better
than li…  loving kindness    0.075000  Because your loving kindness is better
than li…                 False
                               id corpus
sentence token  complexity                           sentence_no_contractions
contraction_expanded
0  3QI9WAYOGQB8GQIR4MDIEF0D2RLS67  bible  They will not hurt nor destroy in all
my holy …   sea    0.000000  They will not hurt nor destroy in all my holy …
```

```
False
1  3T8DUCXY0N6WD9X4RTLK8UN1U929TF  bible  that sends ambassadors by the sea,
even in ves…   sea    0.102941  that sends ambassadors by the sea, even in
ves…                 False
2  3I7KR83SNADXAQ7HXK7S7305BYB9KD  bible  and they entered into the boat, and
were going…   sea    0.109375  and they entered into the boat, and were
going…                 False
3  3BO3NEOQMOHK9ERYPN0GQIWCPC4IAQ  bible  Joseph laid up grain as the sand of
the sea, v…   sea    0.160714  Joseph laid up grain as the sand of the sea,
v…                 False
4  3Y3CZJSZ9KT0W7I0KE38WZHHKSW5RH  bible  There will be a highway for the
remnant that i…  land    0.000000  There will be a highway for the remnant
that i…                 False
                               id corpus
sentence          token  complexity
sentence_no_contractions  contraction_expanded
0  31HLTCK4BLVQ5BO1AUR91TX9V9IVGH  bible  The name of one son was Gershom, for
Moses sai…   foreign land    0.000000  The name of one son was Gershom, for
Moses sai…                 False
1  389A2A304OIXVY7G5B71Q9M43LE0CL  bible  unleavened bread, unleavened cakes
mixed with …    wheat flour    0.157895  unleavened bread, unleavened cakes
mixed with …                 False
2  31N9JPQXIPIRX2A3S9N0CCFXO6TNHR  bible  However the high places were not taken
away; t…  burnt incense    0.200000  However the high places were not taken
away; t…                 False
3  3JVP4ZJHDPSO81TGXL3N1CKZGQY0IN  bible  and he burnt incense of sweet spices
on it, as…  burnt incense    0.250000  and he burnt incense of sweet spices on
it, as…                 False
4  3JAOYN9IHL25ZQAUV5EJZ4GH0KL33L  bible  The same day the king made the middle
of the c…   bronze altar    0.214286  The same day the king made the middle of
the c…                 False
                               id corpus
sentence     token  complexity
sentence_no_contractions  contraction_expanded
0  3K8CQCU3KE19US5SN890DFPK3SANWR  bible  But he, beckoning to them with his
hand to be …      hand    0.000000  But he, beckoning to them with his hand to
be …                 False
1  3Q2T3FD0ON86LCI41NJYV3PN0BW3MV  bible  If I forget you, Jerusalem, let my
right hand …      hand    0.197368  If I forget you, Jerusalem, let my right
hand …                 False
2  3ULIZ0H1VA5C32JJMKOTQ8Z4GUS51B  bible  the ten sons of Haman the son of
Hammedatha, t…      hand    0.200000  the ten sons of Haman the son of
Hammedatha, t…                  True
3  3BFF0DJK8XCEIOT30ZLBPPSRMZQTSD  bible  Let your hand be lifted up above your
adversar…      hand    0.267857  Let your hand be lifted up above your
adversar…                 False
4  3QREJ3J433XSBS8QMHAICCR0BQ1LKR  bible  Abimelech chased him, and he fled
before him, …  entrance    0.000000  Abimelech chased him, and he fled before
```

```
him, …                 False
                                id corpus
sentence           token  complexity
sentence_no_contractions  contraction_expanded
0  3UXQ63NLAAMRIP4WG4XPD98AOYOBLX  bible  for he had an only daughter, about
twelve year…  only daughter    0.025000  for he had an only daughter, about
twelve year…                False
1  3FJ2RVH25Z62TA3R8E1O77EBUYU92W  bible  All these were cities fortified with
high wall…      high walls    0.100000  All these were cities fortified with
high wall…                False
2  3YO4AH2FPDK1PZHZAT8WAEBL7OEQ0F  bible  In the morning, 'It will be foul
weather today…  weather today    0.125000  In the morning, 'It will be foul
weather today…                False
3  3X52SWXE0X5Q3O81YI0MX4V84QTCWZ  bible  Her young children also were dashed in
pieces …  young children    0.160714  Her young children also were dashed in
pieces …                False
4  32K26U12DNONTREA84Q1V8UCIH2VD7  bible  All king Solomon's drinking vessels
were of go…       pure gold    0.178571  All king Solomon's drinking vessels
were of go…                False
```

[ ]:

```python
tokenizer = RegexpTokenizer(r'\w+')

def analyze_sentence_spans_by_corpus_and_quartile_no_contracts(dfs_dict):
    """
    Analyze sentence spans (length metrics) grouped by corpus and complexity␣
 ↪quartile
    for multiple dataframes, but this time using the 'sentence_no_contractions'␣
 ↪column
    instead of the original 'sentence'.
    """
    results = []

    for df_name, df in dfs_dict.items():
        print(f"Processing {df_name} on 'sentence_no_contractions'...")
        df = df.copy()

        q1 = df['complexity'].quantile(0.25)
        q2 = df['complexity'].quantile(0.50)
        q3 = df['complexity'].quantile(0.75)

        def get_quartile(x):
            if x <= q1:
                return 'Q1'
            elif x <= q2:
                return 'Q2'
            elif x <= q3:
                return 'Q3'
            else:
                return 'Q4'

        df['quartile'] = df['complexity'].apply(get_quartile)

        def compute_span_metrics_no_contracts(sentence):
            if pd.isna(sentence):
                return pd.Series({'word_count': 0, 'char_count': 0,␣
↪'avg_word_len': 0})

            words = tokenizer.tokenize(sentence)
            word_count = len(words)
            char_count = len(sentence)
            avg_word_len = np.mean([len(w) for w in words]) if word_count > 0␣
↪else 0

            return pd.Series({
                'word_count': word_count,
                'char_count': char_count,
                'avg_word_len': avg_word_len
            })

        span_metrics_nc = df['sentence_no_contractions'].
↪apply(compute_span_metrics_no_contracts)
        df = pd.concat([df, span_metrics_nc], axis=1)

        corpus_col = 'corpus'
        for corpus_name, corpus_df in df.groupby(corpus_col):
            for quartile, quartile_df in corpus_df.groupby('quartile'):
                complexity_range = f"{quartile_df['complexity'].min():.
↪3f}-{quartile_df['complexity'].max():.3f}"
                stats = {
                    'Dataframe': df_name,
                    'Corpus': corpus_name,
                    'Quartile': quartile,
                    'Complexity Range': complexity_range,
                    'Count': len(quartile_df),
                    'Avg Words': quartile_df['word_count'].mean(),
                    'Median Words': quartile_df['word_count'].median(),
                    'Min Words': quartile_df['word_count'].min(),
                    'Max Words': quartile_df['word_count'].max(),
                    'Std Words': quartile_df['word_count'].std(),
                    'Avg Chars': quartile_df['char_count'].mean(),
                    'Avg Word Len': quartile_df['avg_word_len'].mean()
                }
                results.append(stats)
```

```
    results_df = pd.DataFrame(results)
    results_df = results_df.sort_values(['Dataframe', 'Corpus', 'Quartile'])
    return results_df


dfs = {
    'train_single_df': train_single_df,
    'train_multi_df': train_multi_df,
    'trial_val_single_df': trial_val_single_df,
    'trial_val_multi_df': trial_val_multi_df,
    'test_single_df': test_single_df,
    'test_multi_df': test_multi_df
}

span_analysis_nc =␣
 ↪analyze_sentence_spans_by_corpus_and_quartile_no_contracts(dfs)

pd.set_option('display.max_rows', None)
pd.set_option('display.max_columns', None)
pd.set_option('display.width', 1000)
display(span_analysis_nc)
```

```
Processing train_single_df on 'sentence_no_contractions'…
Processing train_multi_df on 'sentence_no_contractions'…
Processing trial_val_single_df on 'sentence_no_contractions'…
Processing trial_val_multi_df on 'sentence_no_contractions'…
Processing test_single_df on 'sentence_no_contractions'…
Processing test_multi_df on 'sentence_no_contractions'…

              Dataframe    Corpus Quartile Complexity Range  Count  Avg Words ␣
 ↪Median Words  Min Words  Max Words  Std Words   Avg Chars  Avg Word Len
60        test_multi_df     bible       Q1      0.025-0.317     26  23.076923   ␣
 ↪      22.0         4.0       48.0  11.831900  118.730769      4.131249
61        test_multi_df     bible       Q2      0.325-0.417     11  20.545455   ␣
 ↪      17.0         7.0       47.0  12.917923  109.636364      4.213539
62        test_multi_df     bible       Q3      0.432-0.528     18  21.055556   ␣
 ↪      21.5         4.0       43.0  10.843660  113.166667      4.498610
63        test_multi_df     bible       Q4      0.542-0.694     11  22.363636   ␣
 ↪      20.0         7.0       51.0  11.935432  126.181818      4.605062
64        test_multi_df    biomed       Q1      0.000-0.312     11  29.818182   ␣
 ↪      29.0        17.0       47.0   8.388304  195.727273      5.491145
65        test_multi_df    biomed       Q2      0.324-0.417     11  27.090909   ␣
 ↪      24.0         9.0       47.0  11.449494  171.818182      5.436237
66        test_multi_df    biomed       Q3      0.456-0.528     10  26.900000   ␣
 ↪      26.5        10.0       49.0  10.712921  177.500000      5.497409
67        test_multi_df    biomed       Q4      0.562-0.800     21  32.285714   ␣
 ↪      34.0        14.0       56.0  13.598319  209.285714      5.460101
```

```
68         test_multi_df  europarl       Q1      0.214-0.303      10  24.700000   ␣
 ↪      24.5         7.0         56.0  14.189589  146.900000      5.049688
69         test_multi_df  europarl       Q2      0.321-0.429      24  27.833333   ␣
 ↪      27.0         9.0         73.0  15.352855  172.291667      5.269610
70         test_multi_df  europarl       Q3      0.432-0.516      18  32.944444   ␣
 ↪      32.0         6.0         68.0  19.129504  209.888889      5.512245
71         test_multi_df  europarl       Q4      0.531-0.562      13  39.000000   ␣
 ↪      36.0         6.0         95.0  29.631065  237.076923      5.100616
48        test_single_df     bible       Q1      0.000-0.214      79  22.822785   ␣
 ↪      22.0         7.0         49.0  10.585137  116.924051      4.040893
49        test_single_df     bible       Q2      0.217-0.276      68  24.176471   ␣
 ↪      21.0         2.0         77.0  14.393138  126.088235      4.172273
50        test_single_df     bible       Q3      0.278-0.353      67  22.388060   ␣
 ↪      20.0         4.0         63.0  11.306950  119.776119      4.256042
51        test_single_df     bible       Q4      0.359-0.732      69  20.579710   ␣
 ↪      19.0         1.0         55.0  11.264736  110.637681      4.341070
52        test_single_df    biomed       Q1      0.000-0.214      75  27.080000   ␣
 ↪      25.0        10.0         84.0  12.025603  172.986667      5.277318
53        test_single_df    biomed       Q2      0.217-0.275      58  30.275862   ␣
 ↪      26.0        10.0         83.0  15.856587  198.293103      5.446788
54        test_single_df    biomed       Q3      0.278-0.357      66  29.833333   ␣
 ↪      29.0        13.0         85.0  11.754650  191.863636      5.334048
55        test_single_df    biomed       Q4      0.359-0.778      90  31.144444   ␣
 ↪      30.0        14.0         83.0  12.089146  203.077778      5.393791
56        test_single_df  europarl       Q1      0.000-0.214      83  25.337349   ␣
 ↪      21.0         3.0         82.0  16.032191  151.891566      5.044222
57        test_single_df  europarl       Q2      0.217-0.276      98  32.326531   ␣
 ↪      30.0         1.0         97.0  18.707061  195.653061      5.062296
58        test_single_df  europarl       Q3      0.278-0.357      96  33.000000   ␣
 ↪      30.0         3.0        141.0  21.404377  201.760417      5.124551
59        test_single_df  europarl       Q4      0.361-0.583      68  33.235294   ␣
 ↪      29.0         1.0        130.0  20.440023  206.573529      5.164576
12        train_multi_df     bible       Q1      0.028-0.300     163  23.570552   ␣
 ↪      22.0         3.0         67.0  12.429043  124.871166      4.237932
13        train_multi_df     bible       Q2      0.304-0.409     132  24.053030   ␣
 ↪      22.0         6.0         65.0  11.738444  129.659091      4.305703
14        train_multi_df     bible       Q3      0.411-0.529     131  23.778626   ␣
 ↪      23.0         4.0         50.0  11.179163  127.564885      4.331458
15        train_multi_df     bible       Q4      0.533-0.778      79  25.481013   ␣
 ↪      24.0         3.0         81.0  13.490605  139.405063      4.491816
16        train_multi_df    biomed       Q1      0.028-0.303      87  29.091954   ␣
 ↪      28.0         9.0         77.0  11.882792  185.977011      5.277384
17        train_multi_df    biomed       Q2      0.304-0.408      74  30.756757   ␣
 ↪      28.0        11.0         85.0  13.511853  196.067568      5.367302
18        train_multi_df    biomed       Q3      0.411-0.529     111  29.783784   ␣
 ↪      29.0         8.0         61.0  10.912383  193.873874      5.430754
```

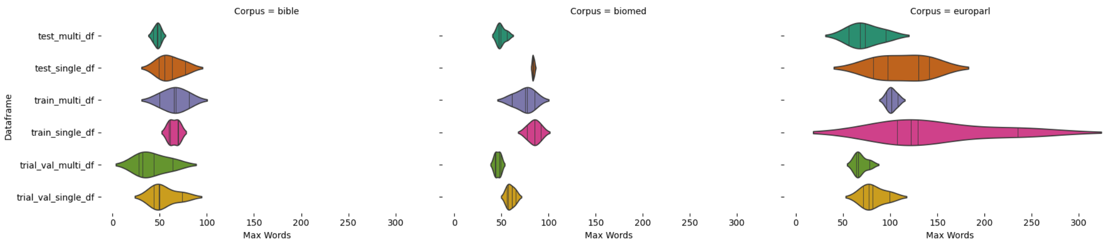
```
19       train_multi_df    biomed       Q4     0.531-0.975    242  29.607438   ␣
 ↪      28.0        10.0        75.0  12.029995  195.107438     5.535387
20       train_multi_df  europarl       Q1     0.118-0.303    132  29.363636   ␣
 ↪      27.0         3.0       101.0  17.874146  176.583333     5.003685
21       train_multi_df  europarl       Q2     0.304-0.409    171  31.666667   ␣
 ↪      28.0         3.0       108.0  19.112977  195.198830     5.176456
22       train_multi_df  europarl       Q3     0.411-0.529    138  33.398551   ␣
 ↪      30.0         7.0       101.0  18.992715  208.304348     5.286607
23       train_multi_df  europarl       Q4     0.533-0.750     57  34.596491   ␣
 ↪      31.0         6.0        96.0  20.318763  218.350877     5.345891
0       train_single_df     bible       Q1     0.000-0.212    701  23.269615   ␣
 ↪      22.0         4.0        61.0  11.764113  121.714693     4.135685
1       train_single_df     bible       Q2     0.212-0.279    640  23.750000   ␣
 ↪      22.0         3.0        60.0  11.579622  124.671875     4.153925
2       train_single_df     bible       Q3     0.281-0.375    624  23.825321   ␣
 ↪      22.0         3.0        70.0  11.963291  126.338141     4.213931
3       train_single_df     bible       Q4     0.380-0.861    609  23.586207   ␣
 ↪      21.0         3.0        69.0  12.460182  126.602627     4.298065
4       train_single_df    biomed       Q1     0.000-0.212    586  28.534130   ␣
 ↪      27.0         2.0        85.0  12.115387  182.076792     5.322266
5       train_single_df    biomed       Q2     0.212-0.279    583  30.442539   ␣
 ↪      29.0         7.0        92.0  11.863182  193.921098     5.289166
6       train_single_df    biomed       Q3     0.281-0.375    659  29.860395   ␣
 ↪      28.0         4.0        77.0  11.591263  191.098634     5.329940
7       train_single_df    biomed       Q4     0.381-0.861    748  29.181818   ␣
 ↪      28.0         3.0        85.0  12.249267  186.978610     5.299963
8       train_single_df  europarl       Q1     0.025-0.212    641  26.761310   ␣
 ↪      24.0         2.0       107.0  15.230853  159.190328     4.942926
9       train_single_df  europarl       Q2     0.212-0.279    714  30.420168   ␣
 ↪      27.0         1.0       129.0  18.383783  183.105042     4.995897
10      train_single_df  europarl       Q3     0.281-0.375    701  30.523538   ␣
 ↪      28.0         1.0       122.0  18.163026  185.843081     5.114626
11      train_single_df  europarl       Q4     0.381-0.775    456  33.543860   ␣
 ↪      31.0         2.0       235.0  21.708515  203.664474     5.054387
36   trial_val_multi_df     bible       Q1     0.000-0.292     11  26.272727   ␣
 ↪      21.0        13.0        64.0  13.950562  141.363636     4.282457
37   trial_val_multi_df     bible       Q2     0.333-0.400      7  20.571429   ␣
 ↪      23.0         5.0        28.0   7.412987  110.857143     4.279406
38   trial_val_multi_df     bible       Q3     0.425-0.500      5  19.600000   ␣
 ↪      19.0         9.0        32.0   8.905055  109.200000     4.431391
39   trial_val_multi_df     bible       Q4     0.525-0.661      6  22.333333   ␣
 ↪      20.5         9.0        44.0  12.242004  117.833333     4.178525
40   trial_val_multi_df    biomed       Q1     0.083-0.303      6  26.833333   ␣
 ↪      25.0        15.0        49.0  11.771434  159.166667     4.899969
41   trial_val_multi_df    biomed       Q2     0.317-0.422      7  25.428571   ␣
 ↪      21.0        15.0        48.0  11.588171  156.000000     5.194383
```

```
42   trial_val_multi_df    biomed       Q3     0.438-0.513      6  37.833333   ␣
 ↪      39.5         26.0        44.0   6.675827  247.500000      5.438593
43   trial_val_multi_df    biomed       Q4     0.537-0.825     14  30.642857   ␣
 ↪      29.5         17.0        43.0   9.849695  211.428571      5.730623
44   trial_val_multi_df  europarl       Q1     0.176-0.306      8  30.000000   ␣
 ↪      25.5          4.0        64.0  20.361027  186.750000      5.306837
45   trial_val_multi_df  europarl       Q2     0.312-0.412     11  47.909091   ␣
 ↪      46.0         24.0        78.0  18.651834  296.909091      5.058375
46   trial_val_multi_df  europarl       Q3     0.432-0.500     13  26.307692   ␣
 ↪      26.0          5.0        66.0  18.167666  166.153846      5.263847
47   trial_val_multi_df  europarl       Q4     0.515-0.714      5  26.400000   ␣
 ↪      15.0          6.0        66.0  24.316661  164.600000      4.998182
24  trial_val_single_df     bible       Q1     0.000-0.214     52  26.769231   ␣
 ↪      26.0          5.0        74.0  15.589860  137.423077      4.074456
25  trial_val_single_df     bible       Q2     0.217-0.266     38  24.868421   ␣
 ↪      23.0          7.0        50.0  10.768249  131.342105      4.200230
26  trial_val_single_df     bible       Q3     0.268-0.355     26  22.884615   ␣
 ↪      20.5          5.0        44.0   9.961233  121.423077      4.316593
27  trial_val_single_df     bible       Q4     0.361-0.633     27  25.666667   ␣
 ↪      23.0          6.0        49.0  12.554497  137.592593      4.213842
28  trial_val_single_df    biomed       Q1     0.028-0.214     21  25.571429   ␣
 ↪      21.0         13.0        65.0  11.543706  164.380952      5.317614
29  trial_val_single_df    biomed       Q2     0.217-0.267     28  30.571429   ␣
 ↪      27.5         11.0        57.0  12.099674  198.142857      5.315287
30  trial_val_single_df    biomed       Q3     0.268-0.359     38  32.105263   ␣
 ↪      29.0         11.0        61.0  12.710476  206.947368      5.364934
31  trial_val_single_df    biomed       Q4     0.364-0.875     48  25.145833   ␣
 ↪      25.5          6.0        56.0  11.721937  164.020833      5.445661
32  trial_val_single_df  europarl       Q1     0.050-0.214     33  31.969697   ␣
 ↪      28.0          5.0        81.0  20.356947  185.969697      4.799024
33  trial_val_single_df  europarl       Q2     0.217-0.267     41  28.487805   ␣
 ↪      28.0          4.0        71.0  15.424205  172.902439      4.997384
34  trial_val_single_df  europarl       Q3     0.268-0.359     39  30.282051   ␣
 ↪      28.0          3.0        99.0  20.040681  184.358974      5.086945
35  trial_val_single_df  europarl       Q4     0.367-0.605     30  35.700000   ␣
 ↪      30.5          5.0        77.0  20.142852  215.400000      4.910759
```

[ ]:
```
g = sns.FacetGrid(span_analysis_nc, col="Corpus", col_wrap=3, height=4,␣
 ↪aspect=1.5)
g.map(sns.violinplot, "Max Words", "Dataframe", inner='stick', palette='Dark2')
g.despine(top=True, right=True, bottom=True, left=True)
plt.tight_layout()
plt.show()
```

```
/usr/local/lib/python3.11/dist-packages/seaborn/axisgrid.py:718: UserWarning:
Using the violinplot function without specifying `order` is likely to produce an
incorrect plot.
```

```
  warnings.warn(warning)
/usr/local/lib/python3.11/dist-packages/seaborn/axisgrid.py:854: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in
v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same
effect.

  func(*plot_args, **plot_kwargs)
/usr/local/lib/python3.11/dist-packages/seaborn/axisgrid.py:854: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in
v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same
effect.

  func(*plot_args, **plot_kwargs)
/usr/local/lib/python3.11/dist-packages/seaborn/axisgrid.py:854: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in
v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same
effect.

  func(*plot_args, **plot_kwargs)
```



- contraction processing successfuly, confirmed with Avg Word deltas between ‘sentence’ and ‘sentence_no_contractions’

0.4 Enrich Datset with PoS Tags, Dependency Parsing, and Morphological Complexity

[ ]:
```
# !pip install -q spacy
# !python -m spacy download en_core_web_trf
!python -m spacy download en_core_web_lg
```

```
Collecting en-core-web-lg==3.8.0
  Using cached https://github.com/explosion/spacy-
models/releases/download/en_core_web_lg-3.8.0/en_core_web_lg-3.8.0-py3-none-
any.whl (400.7 MB)
  Download and installation successful
```

```
You can now load the package via spacy.load('en_core_web_lg')
 Restart to reload dependencies
If you are in a Jupyter or Colab notebook, you may need to restart Python in
order to load all the package's dependencies. You can do this by selecting the
'Restart kernel' or 'Restart runtime' option.
```

[ ]:

```python
nlp = spacy.load("en_core_web_lg")
```

[ ]:

```python
text = "This is a sample sentence for testing spaCy."

doc = nlp(text)

for token in doc:
    print(f"Token: {token.text}, POS: {token.pos_}, Dependency: {token.dep_}")
```

```
Token: This, POS: PRON, Dependency: nsubj
Token: is, POS: AUX, Dependency: ROOT
Token: a, POS: DET, Dependency: det
Token: sample, POS: NOUN, Dependency: compound
Token: sentence, POS: NOUN, Dependency: attr
Token: for, POS: ADP, Dependency: prep
Token: testing, POS: VERB, Dependency: pcomp
Token: spaCy, POS: PROPN, Dependency: dobj
Token: ., POS: PUNCT, Dependency: punct
```

[ ]:

```python
def enrich_with_spacy(df, text_col='sentence_no_contractions'):
    """
    Processes the 'text_col' with spaCy and appends:
      pos_sequence, dep_sequence, morph_sequence,
      and morph_complexity (float) per row.
    """
    df = df.copy()

    pos_tags = []
    dep_tags = []
    morph_tags = []
    morph_complexities = []

    for text in df[text_col]:
        if pd.isna(text) or not text.strip():
            pos_tags.append([])
            dep_tags.append([])
            morph_tags.append([])
            morph_complexities.append(0.0)
            continue

        doc = nlp(text)
        pos_seq = [token.pos_ for token in doc]
        dep_seq = [token.dep_ for token in doc]
        morph_seq = [token.morph for token in doc]

        total_features = 0
        for token in doc:
            features_dict = token.morph.to_dict()
            total_features += len(features_dict)

        avg_morph = total_features / len(doc)

        pos_tags.append(pos_seq)
        dep_tags.append(dep_seq)
        morph_tags.append(morph_seq)
        morph_complexities.append(avg_morph)

    df['pos_sequence'] = pos_tags
    df['dep_sequence'] = dep_tags
    df['morph_sequence'] = morph_tags
    df['morph_complexity'] = morph_complexities

    return df
```

[ ]:

```
dataframes_info = [
    ("train_single_df", train_single_df),
    ("train_multi_df", train_multi_df),
    ("trial_val_single_df", trial_val_single_df),
    ("trial_val_multi_df", trial_val_multi_df),
    ("test_single_df", test_single_df),
    ("test_multi_df", test_multi_df),
]

for df_name, df in dataframes_info:
    print(f"Enriching {df_name} with spaCy features...")
    enriched_df = enrich_with_spacy(df, text_col='sentence_no_contractions')
    globals()[df_name] = enriched_df
    print(f"Done! Now '{df_name}' has columns: pos_sequence, dep_sequence,␣
 ↪morph_sequence, morph_complexity.\n")
```

```
Enriching train_single_df with spaCy features…
Done! Now 'train_single_df' has columns: pos_sequence, dep_sequence,
morph_sequence, morph_complexity.

Enriching train_multi_df with spaCy features…
Done! Now 'train_multi_df' has columns: pos_sequence, dep_sequence,
morph_sequence, morph_complexity.

Enriching trial_val_single_df with spaCy features…
```

```
Done! Now 'trial_val_single_df' has columns: pos_sequence, dep_sequence,
morph_sequence, morph_complexity.

Enriching trial_val_multi_df with spaCy features…
Done! Now 'trial_val_multi_df' has columns: pos_sequence, dep_sequence,
morph_sequence, morph_complexity.

Enriching test_single_df with spaCy features…
Done! Now 'test_single_df' has columns: pos_sequence, dep_sequence,
morph_sequence, morph_complexity.

Enriching test_multi_df with spaCy features…
Done! Now 'test_multi_df' has columns: pos_sequence, dep_sequence,
morph_sequence, morph_complexity.
```

```
[ ]: for df_name, df in dataframes_info:
         print(f"\n{'='*50}")
         print(f"DataFrame: {df_name}")
         print(f"{'='*50}\n")
         sample_df = globals()[df_name].sample(3, random_state=42)
         display(sample_df[['sentence_no_contractions', 'pos_sequence',␣
      ↪'dep_sequence', 'morph_sequence', 'morph_complexity']])
```

```
==================================================
DataFrame: train_single_df
==================================================


                              sentence_no_contractions                          ␣
 ↪               pos_sequence                                       dep_sequence  ␣
 ↪                                    morph_sequence  morph_complexity
5061  The transgenic approach that was used to creat…  [DET, ADJ, NOUN, PRON,␣
 ↪AUX, VERB, PART, VERB, …  [det, amod, nsubjpass, nsubjpass, auxpass, rel… ␣
 ↪[(Definite=Def, PronType=Art), (Degree=Pos), (…          1.500000
2471  When the report comes to Egypt, they will be i…  [SCONJ, DET, NOUN, VERB,␣
 ↪ADP, PROPN, PUNCT, PR…  [advmod, det, nsubj, advcl, prep, pobj, punct,…  [(),␣
 ↪(Definite=Def, PronType=Art), (Number=Sin…          1.166667
800   Saul asked counsel of God, "Shall I go down af…  [PROPN, VERB, NOUN, ADP,␣
 ↪PROPN, PUNCT, PUNCT, …  [nsubj, ROOT, dobj, prep, pobj, punct, punct, … ␣
 ↪[(Number=Sing), (Tense=Past, VerbForm=Fin), (N…          1.200000


==================================================
DataFrame: train_multi_df
==================================================
                              sentence_no_contractions                              ␣
↪             pos_sequence                                         dep_sequence    ␣
↪                                  morph_sequence  morph_complexity
724   BRCA2 may thus promote RAD51 assembly into rec…  [PROPN, AUX, ADV, VERB,␣
↪PROPN, NOUN, ADP, ADJ,…  [nsubj, aux, advmod, ROOT, compound, dobj, pre… ␣
↪[(Number=Sing), (VerbForm=Fin), (), (VerbForm=…          1.222222
812   Therefore, BMPR1A appears to maintain articula…  [ADV, PUNCT, PROPN, VERB,␣
↪PART, VERB, ADJ, NOU…  [advmod, punct, nsubj, ROOT, aux, xcomp, amod,…  [(),␣
↪(PunctType=Comm), (Number=Sing), (Number=…          1.000000
1466  Continued support for the renewal and modernis…  [VERB, NOUN, ADP, DET,␣
↪NOUN, CCONJ, NOUN, ADP,…  [amod, nsubj, prep, det, pobj, cc, conj, prep,… ␣
↪[(Aspect=Perf, Tense=Past, VerbForm=Part), (Nu…          1.205882


==================================================
DataFrame: trial_val_single_df
==================================================


                              sentence_no_contractions                              ␣
↪             pos_sequence                                         dep_sequence    ␣
↪                                  morph_sequence  morph_complexity
145  However, this reduction in bone resorption occ…  [ADV, PUNCT, DET, NOUN,␣
↪ADP, NOUN, NOUN, VERB,…  [advmod, punct, det, nsubj, prep, compound, po…  [(),␣
↪(PunctType=Comm), (Number=Sing, PronType=…          1.0000
335  A word of thanks is also due to many non-gover…  [DET, NOUN, ADP, NOUN,␣
↪AUX, ADV, ADJ, ADP, ADJ…  [det, nsubj, prep, pobj, ROOT, advmod, prep, p… ␣
↪[(Definite=Ind, PronType=Art), (Number=Sing), …          1.0625
175  To test the hypothesis that a temporal delay i…  [PART, VERB, DET, NOUN,␣
↪SCONJ, DET, ADJ, NOUN,…  [aux, advcl, det, dobj, mark, det, amod, nsubj…  [(),␣
↪(VerbForm=Inf), (Definite=Def, PronType=A…          1.2000


==================================================
DataFrame: trial_val_multi_df
==================================================


                              sentence_no_contractions                              ␣
↪             pos_sequence                                         dep_sequence    ␣
↪                                  morph_sequence  morph_complexity
62  by Mr Virrankoski, on behalf of the Committee …  [ADP, PROPN, PROPN, PUNCT,␣
↪ADP, NOUN, ADP, DET…  [prep, compound, pobj, punct, prep, pobj, prep…  [(),␣
↪(Number=Sing), (Number=Sing), (PunctType=…          0.892857
40  Indeed, we recently showed that neural crest c…  [ADV, PUNCT, PRON, ADV,␣
↪VERB, SCONJ, ADJ, PROP…  [advmod, punct, nsubj, advmod, ROOT, mark, com…  [(),␣
↪(PunctType=Comm), (Case=Nom, Number=Plur,…          1.108696
95  It is not an easy task, particularly for the c…  [PRON, AUX, PART, DET, ADJ,␣
↪NOUN, PUNCT, ADV, …  [nsubj, ROOT, neg, det, amod, attr, punct, adv… ␣
↪[(Case=Nom, Gender=Neut, Number=Sing, Person=3…          1.180328


==================================================
DataFrame: test_single_df
==================================================


                              sentence_no_contractions                          ␣
↪            pos_sequence                                       dep_sequence    ␣
↪                                    morph_sequence  morph_complexity
668  It is therefore not a matter of indifference h…  [PRON, AUX, ADV, PART,␣
↪DET, NOUN, ADP, NOUN, S…  [nsubj, ROOT, advmod, neg, det, attr, prep, po… ␣
↪[(Case=Nom, Gender=Neut, Number=Sing, Person=3…          1.200000
30   then shall he offer with the bull a meal offer…  [ADV, AUX, PRON, VERB,␣
↪ADP, DET, NOUN, DET, NO…  [advmod, aux, nsubj, ROOT, prep, det, pobj, de… ␣
↪[(PronType=Dem), (VerbType=Mod), (Case=Nom, Ge…          1.071429
377  While they do have their limitations (e.g. dev…  [SCONJ, PRON, AUX, VERB,␣
↪PRON, NOUN, PUNCT, AD…  [mark, nsubj, aux, advcl, poss, dobj, punct, a…  [(),␣
↪(Case=Nom, Number=Plur, Person=3, PronTyp…          1.157895


==================================================
DataFrame: test_multi_df
==================================================


                              sentence_no_contractions                          ␣
↪            pos_sequence                                       dep_sequence    ␣
↪                                    morph_sequence  morph_complexity
19   God said, "Let the earth yield grass, herbs yi…  [PROPN, VERB, PUNCT,␣
↪PUNCT, VERB, DET, NOUN, V…  [nsubj, ROOT, punct, punct, xcomp, det, nsubj,… ␣
↪[(Number=Sing), (Tense=Past, VerbForm=Fin), (P…          1.564103
42   Moreover I will make a covenant of peace with …  [ADV, PRON, AUX, VERB,␣
↪DET, NOUN, ADP, NOUN, A…  [advmod, nsubj, aux, ccomp, det, dobj, prep, p… ␣
↪[(), (Case=Nom, Number=Sing, Person=1, PronTyp…          1.550000
156  Developing innovation policy is crucial to EU …  [VERB, NOUN, NOUN, AUX,␣
↪ADJ, ADP, PROPN, NOUN,…  [csubj, compound, dobj, ROOT, acomp, prep, com… ␣
↪[(Aspect=Prog, Tense=Pres, VerbForm=Part), (Nu…          1.333333
```

[ ]:

```python
# verify column headers

dataframes = [train_single_df, train_multi_df, trial_val_single_df,␣
 ↪trial_val_multi_df, test_single_df, test_multi_df]
for df in dataframes:
  print(df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7662 entries, 0 to 7661
Data columns (total 11 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   id                       7662 non-null   object
 1   corpus                   7662 non-null   object
 2   sentence                 7662 non-null   object
 3   token                    7655 non-null   object
 4   complexity               7662 non-null   float64
 5   sentence_no_contractions  7662 non-null   object
 6   contraction_expanded     7662 non-null   bool
 7   pos_sequence             7662 non-null   object
 8   dep_sequence             7662 non-null   object
 9   morph_sequence           7662 non-null   object
 10  morph_complexity         7662 non-null   float64
dtypes: bool(1), float64(2), object(8)
memory usage: 606.2+ KB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1517 entries, 0 to 1516
Data columns (total 11 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   id                       1517 non-null   object
 1   corpus                   1517 non-null   object
 2   sentence                 1517 non-null   object
 3   token                    1517 non-null   object
 4   complexity               1517 non-null   float64
 5   sentence_no_contractions  1517 non-null   object
 6   contraction_expanded     1517 non-null   bool
 7   pos_sequence             1517 non-null   object
 8   dep_sequence             1517 non-null   object
 9   morph_sequence           1517 non-null   object
 10  morph_complexity         1517 non-null   float64
dtypes: bool(1), float64(2), object(8)
memory usage: 120.1+ KB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 421 entries, 0 to 420
Data columns (total 11 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   id                       421 non-null    object
 1   corpus                   421 non-null    object
 2   sentence                 421 non-null    object
 3   token                    421 non-null    object
 4   complexity               421 non-null    float64
 5   sentence_no_contractions  421 non-null    object
 6   contraction_expanded      421 non-null    bool
 7   pos_sequence              421 non-null    object
 8   dep_sequence              421 non-null    object
 9   morph_sequence            421 non-null    object
 10  morph_complexity          421 non-null    float64
dtypes: bool(1), float64(2), object(8)
memory usage: 33.4+ KB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 99 entries, 0 to 98
Data columns (total 11 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   id                        99 non-null     object
 1   corpus                    99 non-null     object
 2   sentence                  99 non-null     object
 3   token                     99 non-null     object
 4   complexity                99 non-null     float64
 5   sentence_no_contractions  99 non-null     object
 6   contraction_expanded      99 non-null     bool
 7   pos_sequence              99 non-null     object
 8   dep_sequence              99 non-null     object
 9   morph_sequence            99 non-null     object
 10  morph_complexity          99 non-null     float64
dtypes: bool(1), float64(2), object(8)
memory usage: 8.0+ KB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 917 entries, 0 to 916
Data columns (total 11 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   id                        917 non-null    object
 1   corpus                    917 non-null    object
 2   sentence                  917 non-null    object
 3   token                     917 non-null    object
 4   complexity                917 non-null    float64
 5   sentence_no_contractions  917 non-null    object
 6   contraction_expanded      917 non-null    bool
 7   pos_sequence              917 non-null    object
 8   dep_sequence              917 non-null    object
 9   morph_sequence            917 non-null    object
 10  morph_complexity          917 non-null    float64
dtypes: bool(1), float64(2), object(8)
memory usage: 72.7+ KB
None
<class 'pandas.core.frame.DataFrame'>
```

```
RangeIndex: 184 entries, 0 to 183
Data columns (total 11 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   id                        184 non-null    object
 1   corpus                    184 non-null    object
 2   sentence                  184 non-null    object
 3   token                     184 non-null    object
 4   complexity                184 non-null    float64
 5   sentence_no_contractions  184 non-null    object
 6   contraction_expanded      184 non-null    bool
 7   pos_sequence              184 non-null    object
 8   dep_sequence              184 non-null    object
 9   morph_sequence            184 non-null    object
 10  morph_complexity          184 non-null    float64
dtypes: bool(1), float64(2), object(8)
memory usage: 14.7+ KB
None
```

0.5 Create Binarized Outcome Variable, based on train_single_df median and train_multi_df median, applied to trial-val and test

[ ]:

```
train_single_median = train_single_df['complexity'].median()

def binarize_complexity(value, threshold):
    """
    If value <= threshold, return 0, else return 1.
    """
    if value <= threshold:
        return 0
    else:
        return 1

train_single_df['binary_complexity'] = train_single_df['complexity'].apply(
    lambda x: binarize_complexity(x, train_single_median)
)
trial_val_single_df['binary_complexity'] = trial_val_single_df['complexity'].
 ↪apply(
    lambda x: binarize_complexity(x, train_single_median)
)
test_single_df['binary_complexity'] = test_single_df['complexity'].apply(
    lambda x: binarize_complexity(x, train_single_median)
)

train_multi_median = train_multi_df['complexity'].median()

train_multi_df['binary_complexity'] = train_multi_df['complexity'].apply(
    lambda x: binarize_complexity(x, train_multi_median)
)
trial_val_multi_df['binary_complexity'] = trial_val_multi_df['complexity'].
  ↪apply(
    lambda x: binarize_complexity(x, train_multi_median)
)
test_multi_df['binary_complexity'] = test_multi_df['complexity'].apply(
    lambda x: binarize_complexity(x, train_multi_median)
)

print(f"Median complexity (single): {train_single_median}")
print(f"Median complexity (multi): {train_multi_median}")

print("\nSample rows from train_single_df:")
print(train_single_df[['id', 'complexity', 'binary_complexity']].head())

print("\nSample rows from train_multi_df:")
print(train_multi_df[['id', 'complexity', 'binary_complexity']].head())
```

```
Median complexity (single): 0.2794117647058823
Median complexity (multi): 0.409090909090909

Sample rows from train_single_df:
                               id  complexity  binary_complexity
0  3ZLW647WALVGE8EBR50EGUBPU4P32A    0.000000                  0
1  34R0B0DSP1ZBN3DVY8J8XSIY551E5C    0.000000                  0
2  3S1W0PCJFGTJU2SGNAN2Y213N6WJE3    0.050000                  0
3  3BFNCI9LYKQN09BHXHH9CLSX5KP738    0.150000                  0
4  3G5RUKN2EC3YIWSKUXZ8ZVH95R49N2    0.263889                  0

Sample rows from train_multi_df:
                               id  complexity  binary_complexity
0  3S37Y8CWI80N8KVM53U4E6JKCDC4WE    0.027778                  0
1  3WGCNLZJKF877FYC1Q6C0KNWTDWD11    0.050000                  0
2  3U0MW19E6D6WQ5TH2HDD74IVKTP5CB    0.050000                  0
3  36JW4WBR06KF9AXMUL4N476OMF8FHD    0.050000                  0
4  3HRWUH63QU2FH9Q8R7MRNFC7JX2N5A    0.075000                  0
```

[ ]:
```
# verify column headers

dataframes = [train_single_df, train_multi_df, trial_val_single_df,␣
  ↪trial_val_multi_df, test_single_df, test_multi_df]
for df in dataframes:
  print(df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7662 entries, 0 to 7661
Data columns (total 12 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   id                        7662 non-null   object 
 1   corpus                    7662 non-null   object 
 2   sentence                  7662 non-null   object 
 3   token                     7655 non-null   object 
 4   complexity                7662 non-null   float64
 5   sentence_no_contractions  7662 non-null   object 
 6   contraction_expanded      7662 non-null   bool   
 7   pos_sequence              7662 non-null   object 
 8   dep_sequence              7662 non-null   object 
 9   morph_sequence            7662 non-null   object 
 10  morph_complexity          7662 non-null   float64
 11  binary_complexity         7662 non-null   int64  
dtypes: bool(1), float64(2), int64(1), object(8)
memory usage: 666.1+ KB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1517 entries, 0 to 1516
Data columns (total 12 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   id                        1517 non-null   object 
 1   corpus                    1517 non-null   object 
 2   sentence                  1517 non-null   object 
 3   token                     1517 non-null   object 
 4   complexity                1517 non-null   float64
 5   sentence_no_contractions  1517 non-null   object 
 6   contraction_expanded      1517 non-null   bool   
 7   pos_sequence              1517 non-null   object 
 8   dep_sequence              1517 non-null   object 
 9   morph_sequence            1517 non-null   object 
 10  morph_complexity          1517 non-null   float64
 11  binary_complexity         1517 non-null   int64  
dtypes: bool(1), float64(2), int64(1), object(8)
memory usage: 132.0+ KB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 421 entries, 0 to 420
Data columns (total 12 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   id                        421 non-null    object 
 1   corpus                    421 non-null    object 
 2   sentence                  421 non-null    object 
 3   token                     421 non-null    object 
 4   complexity                421 non-null    float64
 5   sentence_no_contractions  421 non-null    object 
 6   contraction_expanded      421 non-null    bool
 7   pos_sequence              421 non-null    object
 8   dep_sequence              421 non-null    object
 9   morph_sequence            421 non-null    object
 10  morph_complexity          421 non-null    float64
 11  binary_complexity         421 non-null    int64
dtypes: bool(1), float64(2), int64(1), object(8)
memory usage: 36.7+ KB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 99 entries, 0 to 98
Data columns (total 12 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   id                        99 non-null     object
 1   corpus                    99 non-null     object
 2   sentence                  99 non-null     object
 3   token                     99 non-null     object
 4   complexity                99 non-null     float64
 5   sentence_no_contractions  99 non-null     object
 6   contraction_expanded      99 non-null     bool
 7   pos_sequence              99 non-null     object
 8   dep_sequence              99 non-null     object
 9   morph_sequence            99 non-null     object
 10  morph_complexity          99 non-null     float64
 11  binary_complexity         99 non-null     int64
dtypes: bool(1), float64(2), int64(1), object(8)
memory usage: 8.7+ KB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 917 entries, 0 to 916
Data columns (total 12 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   id                        917 non-null    object
 1   corpus                    917 non-null    object
 2   sentence                  917 non-null    object
 3   token                     917 non-null    object
 4   complexity                917 non-null    float64
 5   sentence_no_contractions  917 non-null    object
 6   contraction_expanded      917 non-null    bool
 7   pos_sequence              917 non-null    object
 8   dep_sequence              917 non-null    object
 9   morph_sequence            917 non-null    object
 10  morph_complexity          917 non-null    float64
 11  binary_complexity         917 non-null    int64
dtypes: bool(1), float64(2), int64(1), object(8)
memory usage: 79.8+ KB
```

```
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 184 entries, 0 to 183
Data columns (total 12 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   id                         184 non-null    object
 1   corpus                     184 non-null    object
 2   sentence                   184 non-null    object
 3   token                      184 non-null    object
 4   complexity                 184 non-null    float64
 5   sentence_no_contractions   184 non-null    object
 6   contraction_expanded       184 non-null    bool
 7   pos_sequence               184 non-null    object
 8   dep_sequence               184 non-null    object
 9   morph_sequence             184 non-null    object
 10  morph_complexity           184 non-null    float64
 11  binary_complexity          184 non-null    int64
dtypes: bool(1), float64(2), int64(1), object(8)
memory usage: 16.1+ KB
None
```

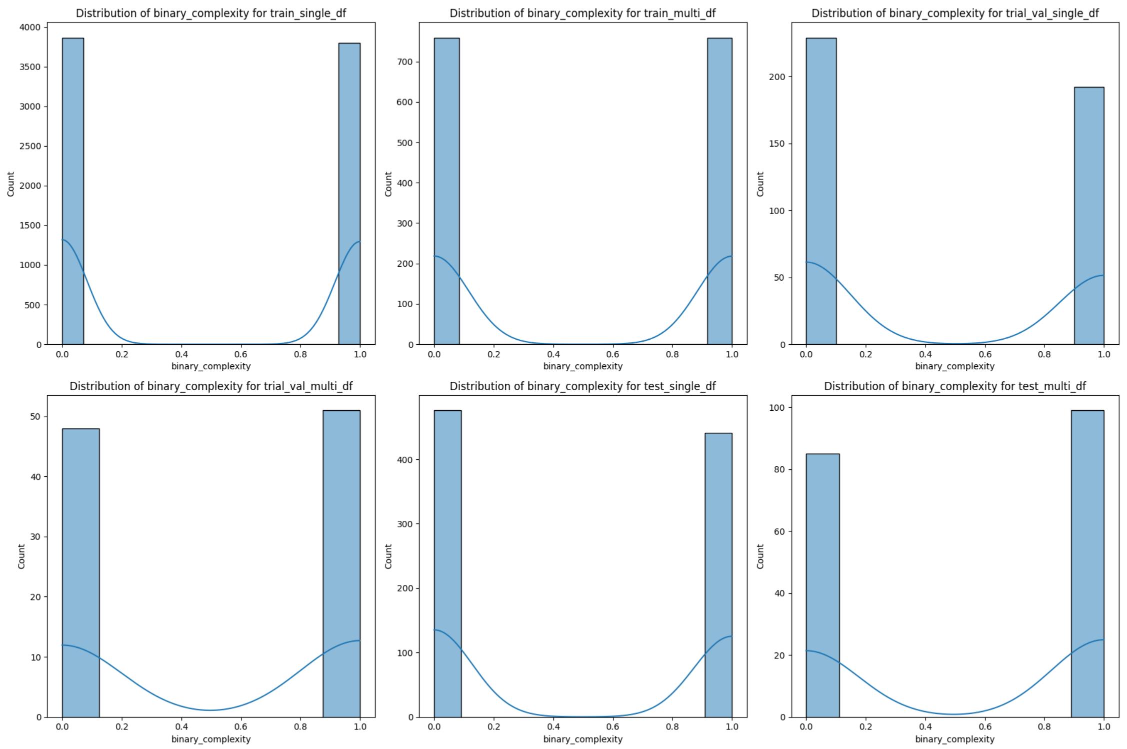
[ ]:
```
# inspect each df

dataframes = [train_single_df, train_multi_df, trial_val_single_df,␣
 ↪trial_val_multi_df, test_single_df, test_multi_df]
for df in dataframes:
  print(df.head())
```

```
                               id corpus
sentence     token  complexity
sentence_no_contractions  contraction_expanded
pos_sequence                                        dep_sequence
morph_sequence  morph_complexity  binary_complexity
0  3ZLW647WALVGE8EBR50EGUBPU4P32A  bible  Behold, there came up out of the river
seven c…     river    0.000000  Behold, there came up out of the river seven
c…                 False  [ADV, PUNCT, PRON, VERB, ADP, ADP, ADP, DET, N…
[advmod, punct, expl, ROOT, prt, prep, prep, d…  [(), (PunctType=Comm), (),
(Tense=Past, VerbFo…          1.041667                  0
1  34R0BODSP1ZBN3DVY8J8XSIY551E5C  bible  I am a fellow bondservant with you and
with yo…  brothers    0.000000  I am a fellow bondservant with you and with
yo…                 False  [PRON, AUX, DET, ADJ, NOUN, ADP, PRON, CCONJ, …
[nsubj, ROOT, det, amod, attr, prep, pobj, cc,…  [(Case=Nom, Number=Sing,
Person=1, PronType=Pr…          1.461538                  0
2  3S1WOPCJFGTJU2SGNAN2Y213N6WJE3  bible  The man, the lord of the land, said to
us, 'By…  brothers    0.050000  The man, the lord of the land, said to us,
'By…                 False  [DET, NOUN, PUNCT, DET, PROPN, ADP, DET, NOUN,…
[det, nsubj, punct, det, appos, prep, det, pob…  [(Definite=Def,
PronType=Art), (Number=Sing), …          1.354167                  0
3  3BFNCI9LYKQN09BHXHH9CLSX5KP738  bible  Shimei had sixteen sons and six
daughters; but…  brothers    0.150000  Shimei had sixteen sons and six
daughters; but…                  True  [PROPN, VERB, NUM, NOUN, CCONJ, NUM,
NOUN, PUN…  [nsubj, ROOT, nummod, dobj, cc, nummod, conj, …  [(Number=Sing),
(Tense=Past, VerbForm=Fin), (N…          1.275862                  0
4  3G5RUKN2EC3YIWSKUXZ8ZVH95R49N2  bible               "He has put my brothers
far from me.  brothers    0.263889               "He has put my brothers far
from me.                 False  [PUNCT, PRON, AUX, VERB, PRON, NOUN, ADV,
ADP,…  [punct, nsubj, aux, ROOT, poss, dobj, advmod, …  [(PunctSide=Ini,
PunctType=Quot), (Case=Nom, G…         2.500000                  0
                               id corpus
sentence            token  complexity
sentence_no_contractions  contraction_expanded
pos_sequence                                      dep_sequence
morph_sequence  morph_complexity  binary_complexity
0  3S37Y8CWI80N8KVM53U4E6JKCDC4WE  bible  but the seventh day is a Sabbath to
Yahweh you…      seventh day    0.027778  but the seventh day is a Sabbath to
Yahweh you…                 False  [CCONJ, DET, ADJ, NOUN, AUX, DET, PROPN,
ADP, …  [cc, det, amod, nsubj, ccomp, det, attr, prep,…  [(ConjType=Cmp),
(Definite=Def, PronType=Art),…         1.341772                  0
1  3WGCNLZJKF877FYC1Q6COKNWTDWD11  bible  But let each man test his own work,
and then h…         own work    0.050000  But let each man test his own work,
and then h…                 False  [CCONJ, VERB, DET, NOUN, VERB, PRON, ADJ,
NOUN…  [cc, ROOT, det, nsubj, ccomp, poss, amod, dobj…  [(ConjType=Cmp),
(VerbForm=Inf), (), (Number=S…         1.608696                  0
2  3UOMW19E6D6WQ5TH2HDD74IVKTP5CB  bible  To him who by understanding made the
heavens; …  loving kindness    0.050000  To him who by understanding made the
heavens; …                 False  [ADP, PRON, PRON, ADP, VERB, VERB, DET,
NOUN, …  [prep, pobj, nsubj, prep, pcomp, advcl, det, d…  [(), (Case=Acc,
Gender=Masc, Number=Sing, Pers…          1.562500                  0
3  36JW4WBR06KF9AXMUL4N476OMF8FHD  bible  Remember to me, my God, this also, and
spare m…  loving kindness    0.050000  Remember to me, my God, this also, and
spare m…                 False  [VERB, ADP, PRON, PUNCT, PRON, PROPN, PUNCT,
P…  [ROOT, prep, pobj, punct, poss, npadvmod, punc…  [(VerbForm=Inf), (),
(Case=Acc, Number=Sing, P…          1.590909                  0
4  3HRWUH63QU2FH9Q8R7MRNFC7JX2N5A  bible  Because your loving kindness is better
than li…  loving kindness    0.075000  Because your loving kindness is better
than li…                 False  [SCONJ, PRON, ADJ, NOUN, AUX, ADJ, ADP, NOUN,
…  [mark, poss, amod, nsubj, advcl, acomp, prep, …  [(), (Person=2,
Poss=Yes, PronType=Prs), (Degr…          1.600000                  0
                               id corpus
sentence token  complexity                          sentence_no_contractions
contraction_expanded                                        pos_sequence
dep_sequence                                    morph_sequence
morph_complexity  binary_complexity
0  3QI9WAYOGQB8GQIR4MDIEF0D2RLS67  bible  They will not hurt nor destroy in all
my holy …   sea    0.000000  They will not hurt nor destroy in all my holy …
```

```
False  [PRON, AUX, PART, VERB, CCONJ, VERB, ADP, PRON…  [nsubj, aux, neg,
ccomp, cc, conj, prep, prede…  [(Case=Nom, Number=Plur, Person=3,
PronType=Pr…          1.129032                  0
1  3T8DUCXY0N6WD9X4RTLK8UN1U929TF  bible  that sends ambassadors by the sea,
even in ves…   sea    0.102941  that sends ambassadors by the sea, even in
ves…                 False  [PRON, VERB, NOUN, ADP, DET, NOUN, PUNCT, ADV,…
[nsubj, ROOT, dobj, prep, det, pobj, punct, ad…  [(PronType=Rel),
(Number=Sing, Person=3, Tense…         1.263158                  0
2  3I7KR83SNADXAQ7HXK7S7305BYB9KD  bible  and they entered into the boat, and
were going…   sea    0.109375  and they entered into the boat, and were
going…                 False  [CCONJ, PRON, VERB, ADP, DET, NOUN, PUNCT,
CCO…  [cc, nsubj, ROOT, prep, det, pobj, punct, cc, …  [(ConjType=Cmp),
(Case=Nom, Number=Plur, Perso…         1.437500                  0
3  3BO3NEOQM0HK9ERYPN0GQIWCPC4IAQ  bible  Joseph laid up grain as the sand of
the sea, v…   sea    0.160714  Joseph laid up grain as the sand of the sea,
v…                 False  [PROPN, VERB, ADP, NOUN, ADP, DET, NOUN, ADP, …
[nsubj, ROOT, prt, dobj, prep, det, pobj, prep…  [(Number=Sing), (Tense=Past,
VerbForm=Fin), ()…          1.400000                  0
4  3Y3CZJSZ9KT0W7I0KE38WZHHKSW5RH  bible  There will be a highway for the
remnant that i…  land    0.000000  There will be a highway for the remnant
that i…                 False  [PRON, AUX, AUX, DET, NOUN, ADP, DET, NOUN,
PR…  [expl, aux, ROOT, det, attr, prep, det, pobj, …  [(), (VerbForm=Fin),
(VerbForm=Inf), (Definite…         1.277778                  0
                               id corpus
sentence          token  complexity
sentence_no_contractions  contraction_expanded
pos_sequence                                       dep_sequence
morph_sequence  morph_complexity  binary_complexity
0  31HLTCK4BLVQ5BO1AUR91TX9V9IVGH  bible  The name of one son was Gershom, for
Moses sai…   foreign land    0.000000  The name of one son was Gershom, for
Moses sai…                 False  [DET, NOUN, ADP, NUM, NOUN, AUX, PROPN,
PUNCT,…  [det, nsubj, prep, nummod, pobj, ROOT, attr, p…  [(Definite=Def,
PronType=Art), (Number=Sing), …          1.520000                  0
1  389A2A304OIXVY7G5B71Q9M43LE0CL  bible  unleavened bread, unleavened cakes
mixed with …    wheat flour    0.157895  unleavened bread, unleavened cakes
mixed with …                 False  [ADJ, NOUN, PUNCT, ADJ, NOUN, VERB, ADP,
NOUN,…  [amod, dep, punct, amod, appos, acl, prep, pob…  [(Degree=Pos),
(Number=Sing), (PunctType=Comm)…          1.200000                  0
2  31N9JPQXIPIRX2A3S9N0CCFXO6TNHR  bible  However the high places were not taken
away; t…  burnt incense    0.200000  However the high places were not taken
away; t…                 False  [ADV, DET, ADJ, NOUN, AUX, PART, VERB, ADV,
PU…  [advmod, det, amod, nsubjpass, auxpass, neg, c…  [(), (Definite=Def,
PronType=Art), (Degree=Pos…          1.190476                  0
3  3JVP4ZJHDPSO81TGXL3N1CKZGQY0IN  bible  and he burnt incense of sweet spices
on it, as…  burnt incense    0.250000  and he burnt incense of sweet spices on
it, as…                 False  [CCONJ, PRON, VERB, NOUN, ADP, ADJ, NOUN,
ADP,…  [cc, nsubj, ROOT, dobj, prep, amod, pobj, prep…  [(ConjType=Cmp),
(Case=Nom, Gender=Masc, Numbe…         1.466667                  0
4  3JAOYN9IHL25ZQAUV5EJZ4GH0KL33L  bible  The same day the king made the middle
of the c…   bronze altar    0.214286  The same day the king made the middle of
the c…                 False  [DET, ADJ, NOUN, DET, NOUN, VERB, DET, NOUN,
A…  [det, amod, npadvmod, det, nsubj, ccomp, det, …  [(Definite=Def,
PronType=Art), (Degree=Pos), (…         1.352113                  0
                               id corpus
sentence     token  complexity
sentence_no_contractions  contraction_expanded
pos_sequence                                       dep_sequence
morph_sequence  morph_complexity  binary_complexity
0  3K8CQCU3KE19US5SN890DFPK3SANWR  bible  But he, beckoning to them with his
hand to be …      hand    0.000000  But he, beckoning to them with his hand to
be …                 False  [CCONJ, PRON, PUNCT, VERB, ADP, PRON, ADP, PRO…
[cc, nsubj, punct, advcl, prep, pobj, prep, po…  [(ConjType=Cmp), (Case=Nom,
Gender=Masc, Numbe…         1.703704                  0
1  3Q2T3FD0ON86LCI41NJYV3PN0BW3MV  bible  If I forget you, Jerusalem, let my
right hand …      hand    0.197368  If I forget you, Jerusalem, let my right
hand …                 False  [SCONJ, PRON, VERB, PRON, PUNCT, PROPN,
PUNCT,…  [mark, nsubj, advcl, dobj, punct, npadvmod, pu…  [(), (Case=Nom,
Number=Sing, Person=1, PronTyp…         1.800000                  0
2  3ULIZ0H1VA5C32JJMK0TQ8Z4GUS51B  bible  the ten sons of Haman the son of
Hammedatha, t…      hand    0.200000  the ten sons of Haman the son of
Hammedatha, t…                 True  [DET, NUM, NOUN, ADP, PROPN, DET, NOUN,
ADP, P…  [det, nummod, ROOT, prep, pobj, det, appos, pr…  [(Definite=Def,
PronType=Art), (NumType=Card),…         1.269231                  0
3  3BFF0DJK8XCEIOT30ZLBPPSRMZQTSD  bible  Let your hand be lifted up above your
adversar…      hand    0.267857  Let your hand be lifted up above your
adversar…                 False  [VERB, PRON, NOUN, AUX, VERB, ADP, ADP, PRON,
…  [ROOT, poss, nsubjpass, auxpass, ccomp, prt, p…  [(VerbForm=Inf),
(Person=2, Poss=Yes, PronType…         1.250000                  0
4  3QREJ3J433XSBS8QMHAICCR0BQ1LKR  bible  Abimelech chased him, and he fled
before him, …  entrance    0.000000  Abimelech chased him, and he fled before
him, …                 False  [PROPN, VERB, PRON, PUNCT, CCONJ, PRON, VERB,
…  [nsubj, ROOT, dobj, punct, cc, nsubj, conj, pr…  [(Number=Sing),
(Tense=Past, VerbForm=Fin), (C…         1.652174                  0
                               id corpus
sentence           token  complexity
sentence_no_contractions  contraction_expanded
pos_sequence                                       dep_sequence
morph_sequence  morph_complexity  binary_complexity
0  3UXQ63NLAAMRIP4WG4XPD98AOYOBLX  bible  for he had an only daughter, about
twelve year…  only daughter    0.025000  for he had an only daughter, about
twelve year…                 False  [SCONJ, PRON, VERB, DET, ADJ, NOUN, PUNCT,
ADV…  [mark, nsubj, ROOT, det, amod, dobj, punct, ad…  [(), (Case=Nom,
Gender=Masc, Number=Sing, Pers…         1.722222                  0
1  3FJ2RVH25Z62TA3R8E1O77EBUYU92W  bible  All these were cities fortified with
high wall…      high walls    0.100000  All these were cities fortified with
high wall…                 False  [DET, PRON, AUX, NOUN, VERB, ADP, ADJ, NOUN,
P…  [predet, nsubj, ROOT, attr, acl, prep, amod, p…  [(), (Number=Plur,
PronType=Dem), (Mood=Ind, T…          1.136364                  0
2  3YO4AH2FPDK1PZHZAT8WAEBL7OEQOF  bible  In the morning, 'It will be foul
weather today…   weather today    0.125000  In the morning, 'It will be foul
weather today…                 False  [ADP, DET, NOUN, PUNCT, PUNCT, PRON,
AUX, AUX,…  [prep, det, pobj, punct, punct, nsubj, aux, RO…  [(),
(Definite=Def, PronType=Art), (Number=Sin…          1.476190
0
3  3X52SWXE0X5Q3O81YIOMX4V84QTCWZ  bible  Her young children also were dashed in
pieces …  young children    0.160714  Her young children also were dashed in
pieces …                 False  [PRON, ADJ, NOUN, ADV, AUX, VERB, ADP, NOUN,
A…  [poss, amod, nsubjpass, advmod, auxpass, ROOT,…  [(Gender=Fem,
Number=Sing, Person=3, Poss=Yes,…          1.514286                  0
4  32K26U12DNONTREA84Q1V8UCIH2VD7  bible  All king Solomon's drinking vessels
were of go…       pure gold    0.178571  All king Solomon's drinking vessels
were of go…                 False  [DET, NOUN, PROPN, PART, NOUN, NOUN, AUX,
ADP,…  [det, compound, poss, case, compound, nsubj, c…  [(), (Number=Sing),
(Number=Sing), (), (Number…          1.162791                  0
```

[ ]:

```python
dataframes = {
    "train_single_df": train_single_df,
    "train_multi_df": train_multi_df,
    "trial_val_single_df": trial_val_single_df,
    "trial_val_multi_df": trial_val_multi_df,
    "test_single_df": test_single_df,
    "test_multi_df": test_multi_df
}

fig, axes = plt.subplots(2, 3, figsize=(18, 12))

for i, (df_name, df) in enumerate(dataframes.items()):
  row = i // 3
  col = i % 3
  ax = axes[row, col]
  sns.histplot(df['binary_complexity'], kde=True, ax=ax)
  ax.set_title(f'Distribution of binary_complexity for {df_name}')
  ax.set_xlabel('binary_complexity')

plt.tight_layout()
plt.show()
```



```
[ ]: train_single_75th = train_single_df['complexity'].quantile(0.75)
     train_multi_75th = train_multi_df['complexity'].quantile(0.75)

     print("75th percentile (single-track):", train_single_75th)
     print("75th percentile (multi-track):", train_multi_75th)

     def binarize_complexity_75th(value, threshold):
         """
         Returns 0 if 'value' <= threshold, else 1.
         """
         if value <= threshold:
             return 0
         else:
             return 1

     train_single_df['binary_complexity_75th_split'] = train_single_df['complexity'].
      ↪apply(
         lambda x: binarize_complexity_75th(x, train_single_75th)
     )
     trial_val_single_df['binary_complexity_75th_split'] =␣
      ↪trial_val_single_df['complexity'].apply(
         lambda x: binarize_complexity_75th(x, train_single_75th)
     )
```

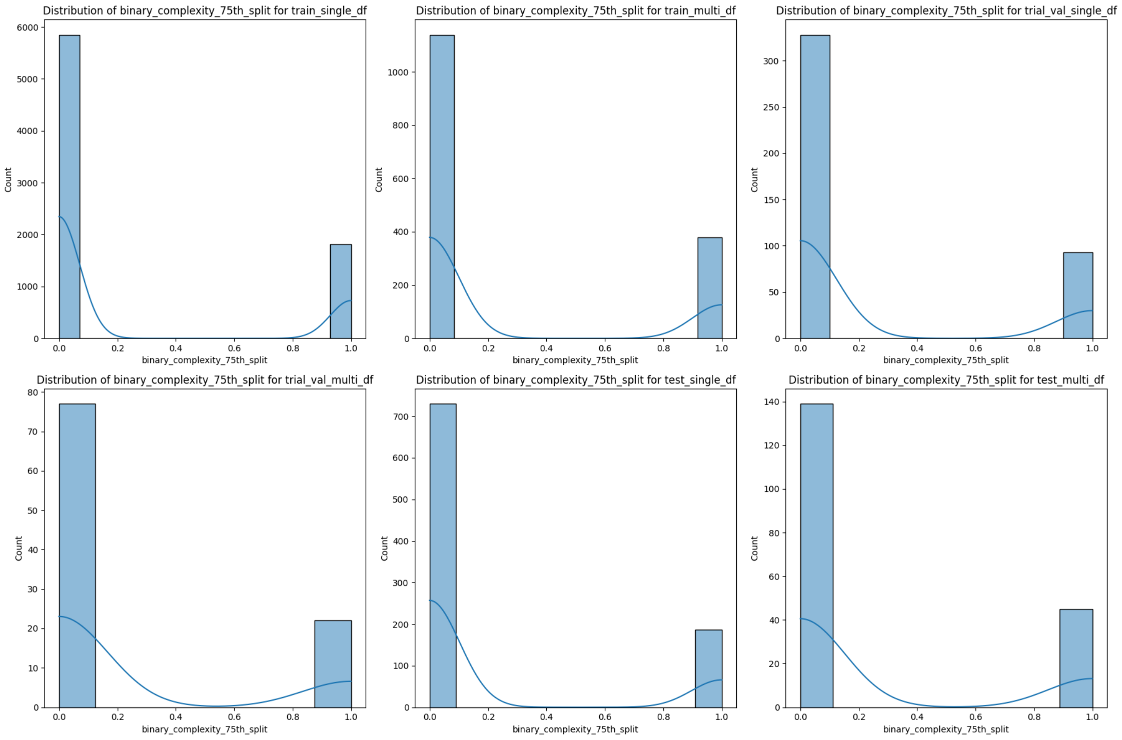
```
test_single_df['binary_complexity_75th_split'] = test_single_df['complexity'].
 ↪apply(
    lambda x: binarize_complexity_75th(x, train_single_75th)
)

train_multi_df['binary_complexity_75th_split'] = train_multi_df['complexity'].
 ↪apply(
    lambda x: binarize_complexity_75th(x, train_multi_75th)
)
trial_val_multi_df['binary_complexity_75th_split'] =␣
 ↪trial_val_multi_df['complexity'].apply(
    lambda x: binarize_complexity_75th(x, train_multi_75th)
)
test_multi_df['binary_complexity_75th_split'] = test_multi_df['complexity'].
 ↪apply(
    lambda x: binarize_complexity_75th(x, train_multi_75th)
)

print("\nDistribution of 'binary_complexity_75th_split' in train_single_df:")
print(train_single_df['binary_complexity_75th_split'].value_counts())

print("\nDistribution of 'binary_complexity_75th_split' in train_multi_df:")
print(train_multi_df['binary_complexity_75th_split'].value_counts())
```

```
75th percentile (single-track): 0.375
75th percentile (multi-track): 0.5294117647058824

Distribution of 'binary_complexity_75th_split' in train_single_df:
binary_complexity_75th_split
0    5849
1    1813
Name: count, dtype: int64

Distribution of 'binary_complexity_75th_split' in train_multi_df:
binary_complexity_75th_split
0    1139
1     378
Name: count, dtype: int64
```

[ ]:

```
dataframes = {
    "train_single_df": train_single_df,
    "train_multi_df": train_multi_df,
    "trial_val_single_df": trial_val_single_df,
    "trial_val_multi_df": trial_val_multi_df,
    "test_single_df": test_single_df,
    "test_multi_df": test_multi_df
}
```

```
fig, axes = plt.subplots(2, 3, figsize=(18, 12))

for i, (df_name, df) in enumerate(dataframes.items()):
  row = i // 3
  col = i % 3
  ax = axes[row, col]
  sns.histplot(df['binary_complexity_75th_split'], kde=True, ax=ax)
  ax.set_title(f'Distribution of binary_complexity_75th_split for {df_name}')
  ax.set_xlabel('binary_complexity_75th_split')

plt.tight_layout()
plt.show()
```



```
[ ]: !ls -R /content/drive/MyDrive/266-final/data/266-comp-lex-master/
```

```
/content/drive/MyDrive/266-final/data/266-comp-lex-master/:
fe-test-labels  fe-train  fe-trial-val  test-labels  train  trial

/content/drive/MyDrive/266-final/data/266-comp-lex-master/fe-test-labels:
test_multi_df.csv  test_single_df.csv

/content/drive/MyDrive/266-final/data/266-comp-lex-master/fe-train:
train_multi_df.csv  train_single_df.csv
```

```
/content/drive/MyDrive/266-final/data/266-comp-lex-master/fe-trial-val:
trial_val_multi_df.csv  trial_val_single_df.csv

/content/drive/MyDrive/266-final/data/266-comp-lex-master/test-labels:
lcp_multi_test.tsv  lcp_single_test.tsv

/content/drive/MyDrive/266-final/data/266-comp-lex-master/train:
lcp_multi_train.tsv  lcp_single_train.tsv

/content/drive/MyDrive/266-final/data/266-comp-lex-master/trial:
lcp_multi_trial.tsv  lcp_single_trial.tsv
```

[ ]:

```
# verify column headers

dataframes = [train_single_df, train_multi_df, trial_val_single_df,␣
 ↪trial_val_multi_df, test_single_df, test_multi_df]
for df in dataframes:
  print(df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7662 entries, 0 to 7661
Data columns (total 13 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   id                            7662 non-null   object
 1   corpus                        7662 non-null   object
 2   sentence                      7662 non-null   object
 3   token                         7655 non-null   object
 4   complexity                    7662 non-null   float64
 5   sentence_no_contractions      7662 non-null   object
 6   contraction_expanded          7662 non-null   bool
 7   pos_sequence                  7662 non-null   object
 8   dep_sequence                  7662 non-null   object
 9   morph_sequence                7662 non-null   object
 10  morph_complexity              7662 non-null   float64
 11  binary_complexity             7662 non-null   int64
 12  binary_complexity_75th_split  7662 non-null   int64
dtypes: bool(1), float64(2), int64(2), object(8)
memory usage: 725.9+ KB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1517 entries, 0 to 1516
Data columns (total 13 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   id                            1517 non-null   object
 1   corpus                        1517 non-null   object
 2   sentence                      1517 non-null   object
 3   token                         1517 non-null   object
 4   complexity                    1517 non-null   float64
 5   sentence_no_contractions      1517 non-null   object
 6   contraction_expanded          1517 non-null   bool
 7   pos_sequence                  1517 non-null   object
 8   dep_sequence                  1517 non-null   object
 9   morph_sequence                1517 non-null   object
 10  morph_complexity              1517 non-null   float64
 11  binary_complexity             1517 non-null   int64
 12  binary_complexity_75th_split  1517 non-null   int64
dtypes: bool(1), float64(2), int64(2), object(8)
memory usage: 143.8+ KB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 421 entries, 0 to 420
Data columns (total 13 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   id                            421 non-null    object
 1   corpus                        421 non-null    object
 2   sentence                      421 non-null    object
 3   token                         421 non-null    object
 4   complexity                    421 non-null    float64
 5   sentence_no_contractions      421 non-null    object
 6   contraction_expanded          421 non-null    bool
 7   pos_sequence                  421 non-null    object
 8   dep_sequence                  421 non-null    object
 9   morph_sequence                421 non-null    object
 10  morph_complexity              421 non-null    float64
 11  binary_complexity             421 non-null    int64
 12  binary_complexity_75th_split  421 non-null    int64
dtypes: bool(1), float64(2), int64(2), object(8)
memory usage: 40.0+ KB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 99 entries, 0 to 98
Data columns (total 13 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   id                            99 non-null     object
 1   corpus                        99 non-null     object
 2   sentence                      99 non-null     object
 3   token                         99 non-null     object
 4   complexity                    99 non-null     float64
 5   sentence_no_contractions      99 non-null     object
 6   contraction_expanded          99 non-null     bool
 7   pos_sequence                  99 non-null     object
 8   dep_sequence                  99 non-null     object 
 9   morph_sequence                99 non-null     object 
 10  morph_complexity              99 non-null     float64
 11  binary_complexity             99 non-null     int64  
 12  binary_complexity_75th_split  99 non-null     int64  
dtypes: bool(1), float64(2), int64(2), object(8)
memory usage: 9.5+ KB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 917 entries, 0 to 916
Data columns (total 13 columns):
 #   Column                        Non-Null Count  Dtype  
---  ------                        --------------  -----  
 0   id                            917 non-null    object 
 1   corpus                        917 non-null    object 
 2   sentence                      917 non-null    object 
 3   token                         917 non-null    object 
 4   complexity                    917 non-null    float64
 5   sentence_no_contractions      917 non-null    object 
 6   contraction_expanded          917 non-null    bool   
 7   pos_sequence                  917 non-null    object 
 8   dep_sequence                  917 non-null    object 
 9   morph_sequence                917 non-null    object 
 10  morph_complexity              917 non-null    float64
 11  binary_complexity             917 non-null    int64  
 12  binary_complexity_75th_split  917 non-null    int64  
dtypes: bool(1), float64(2), int64(2), object(8)
memory usage: 87.0+ KB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 184 entries, 0 to 183
Data columns (total 13 columns):
 #   Column                        Non-Null Count  Dtype  
---  ------                        --------------  -----  
 0   id                            184 non-null    object 
 1   corpus                        184 non-null    object 
 2   sentence                      184 non-null    object 
 3   token                         184 non-null    object 
 4   complexity                    184 non-null    float64
 5   sentence_no_contractions      184 non-null    object 
 6   contraction_expanded          184 non-null    bool   
 7   pos_sequence                  184 non-null    object 
 8   dep_sequence                  184 non-null    object 
 9   morph_sequence                184 non-null    object 
 10  morph_complexity              184 non-null    float64
 11  binary_complexity             184 non-null    int64  
 12  binary_complexity_75th_split  184 non-null    int64  
dtypes: bool(1), float64(2), int64(2), object(8)
```

```
memory usage: 17.6+ KB
None
```

```
[ ]: # inspect each df

     dataframes = [train_single_df, train_multi_df, trial_val_single_df,␣
      ↪trial_val_multi_df, test_single_df, test_multi_df]
     for df in dataframes:
       print(df.head())
```

```
                               id corpus
sentence     token  complexity
sentence_no_contractions  contraction_expanded
pos_sequence                                           dep_sequence
morph_sequence  morph_complexity  binary_complexity
binary_complexity_75th_split
0  3ZLW647WALVGE8EBR50EGUBPU4P32A  bible  Behold, there came up out of the river
seven c…     river    0.000000  Behold, there came up out of the river seven
c…                   False  [ADV, PUNCT, PRON, VERB, ADP, ADP, ADP, DET, N…
[advmod, punct, expl, ROOT, prt, prep, prep, d…  [(), (PunctType=Comm), (),
(Tense=Past, VerbFo…          1.041667                  0
0
1  34R0BODSP1ZBN3DVY8J8XSIY551E5C  bible  I am a fellow bondservant with you and
with yo…  brothers    0.000000  I am a fellow bondservant with you and with
yo…                   False  [PRON, AUX, DET, ADJ, NOUN, ADP, PRON, CCONJ, …
[nsubj, ROOT, det, amod, attr, prep, pobj, cc,…  [(Case=Nom, Number=Sing,
Person=1, PronType=Pr…          1.461538                  0
0
2  3S1WOPCJFGTJU2SGNAN2Y213N6WJE3  bible  The man, the lord of the land, said to
us, 'By…  brothers    0.050000  The man, the lord of the land, said to us,
'By…                   False  [DET, NOUN, PUNCT, DET, PROPN, ADP, DET, NOUN,…
[det, nsubj, punct, det, appos, prep, det, pob…  [(Definite=Def,
PronType=Art), (Number=Sing), …          1.354167                  0
0
3  3BFNCI9LYKQNO9BHXHH9CLSX5KP738  bible  Shimei had sixteen sons and six
daughters; but…  brothers    0.150000  Shimei had sixteen sons and six
daughters; but…                   True  [PROPN, VERB, NUM, NOUN, CCONJ, NUM,
NOUN, PUN…  [nsubj, ROOT, nummod, dobj, cc, nummod, conj, …  [(Number=Sing),
(Tense=Past, VerbForm=Fin), (N…          1.275862                  0
0
4  3G5RUKN2EC3YIWSKUXZ8ZVH95R49N2  bible               "He has put my brothers
far from me.  brothers    0.263889               "He has put my brothers far
from me.                 False  [PUNCT, PRON, AUX, VERB, PRON, NOUN, ADV,
ADP,…  [punct, nsubj, aux, ROOT, poss, dobj, advmod, …  [(PunctSide=Ini,
PunctType=Quot), (Case=Nom, G…          2.500000                  0
0
                               id corpus
sentence            token  complexity
```

```
sentence_no_contractions  contraction_expanded
pos_sequence                                       dep_sequence
morph_sequence  morph_complexity  binary_complexity
binary_complexity_75th_split
0  3S37Y8CWI80N8KVM53U4E6JKCDC4WE  bible  but the seventh day is a Sabbath to
Yahweh you…      seventh day    0.027778  but the seventh day is a Sabbath to
Yahweh you…                 False  [CCONJ, DET, ADJ, NOUN, AUX, DET, PROPN,
ADP, …  [cc, det, amod, nsubj, ccomp, det, attr, prep,…  [(ConjType=Cmp),
(Definite=Def, PronType=Art),…          1.341772                  0
0
1  3WGCNLZJKF877FYC1Q6COKNWTDWD11  bible  But let each man test his own work,
and then h…         own work    0.050000  But let each man test his own work,
and then h…                 False  [CCONJ, VERB, DET, NOUN, VERB, PRON, ADJ,
NOUN…  [cc, ROOT, det, nsubj, ccomp, poss, amod, dobj…  [(ConjType=Cmp),
(VerbForm=Inf), (), (Number=S…          1.608696                  0
0
2  3UOMW19E6D6WQ5TH2HDD74IVKTP5CB  bible  To him who by understanding made the
heavens; …  loving kindness    0.050000  To him who by understanding made the
heavens; …                 False  [ADP, PRON, PRON, ADP, VERB, VERB, DET,
NOUN, …  [prep, pobj, nsubj, prep, pcomp, advcl, det, d…  [(), (Case=Acc,
Gender=Masc, Number=Sing, Pers…          1.562500                  0
0
3  36JW4WBR06KF9AXMUL4N476OMF8FHD  bible  Remember to me, my God, this also, and
spare m…  loving kindness    0.050000  Remember to me, my God, this also, and
spare m…                 False  [VERB, ADP, PRON, PUNCT, PRON, PROPN, PUNCT,
P…  [ROOT, prep, pobj, punct, poss, npadvmod, punc…  [(VerbForm=Inf), (),
(Case=Acc, Number=Sing, P…          1.590909                  0
0
4  3HRWUH63QU2FH9Q8R7MRNFC7JX2N5A  bible  Because your loving kindness is better
than li…  loving kindness    0.075000  Because your loving kindness is better
than li…                 False  [SCONJ, PRON, ADJ, NOUN, AUX, ADJ, ADP, NOUN,
…  [mark, poss, amod, nsubj, advcl, acomp, prep, …  [(), (Person=2,
Poss=Yes, PronType=Prs), (Degr…          1.600000                  0
0
                               id corpus
sentence token  complexity                           sentence_no_contractions
contraction_expanded                                       pos_sequence
dep_sequence                                     morph_sequence
morph_complexity  binary_complexity  binary_complexity_75th_split
0  3QI9WAYOGQB8GQIR4MDIEF0D2RLS67  bible  They will not hurt nor destroy in all
my holy …   sea    0.000000  They will not hurt nor destroy in all my holy …
False  [PRON, AUX, PART, VERB, CCONJ, VERB, ADP, PRON…  [nsubj, aux, neg,
ccomp, cc, conj, prep, prede…  [(Case=Nom, Number=Plur, Person=3,
PronType=Pr…          1.129032                  0
0
1  3T8DUCXY0N6WD9X4RTLK8UN1U929TF  bible  that sends ambassadors by the sea,
even in ves…   sea    0.102941  that sends ambassadors by the sea, even in
ves…                 False  [PRON, VERB, NOUN, ADP, DET, NOUN, PUNCT, ADV,…
```

```
[nsubj, ROOT, dobj, prep, det, pobj, punct, ad…  [(PronType=Rel),
(Number=Sing, Person=3, Tense…          1.263158                  0
0
2  3I7KR83SNADXAQ7HXK7S7305BYB9KD  bible  and they entered into the boat, and
were going…   sea    0.109375  and they entered into the boat, and were
going…                 False  [CCONJ, PRON, VERB, ADP, DET, NOUN, PUNCT,
CCO…  [cc, nsubj, ROOT, prep, det, pobj, punct, cc, …  [(ConjType=Cmp),
(Case=Nom, Number=Plur, Perso…          1.437500                  0
0
3  3BO3NEOQM0HK9ERYPN0GQIWCPC4IAQ  bible  Joseph laid up grain as the sand of
the sea, v…   sea    0.160714  Joseph laid up grain as the sand of the sea,
v…                 False  [PROPN, VERB, ADP, NOUN, ADP, DET, NOUN, ADP, …
[nsubj, ROOT, prt, dobj, prep, det, pobj, prep…  [(Number=Sing), (Tense=Past,
VerbForm=Fin), ()…          1.400000                  0
0
4  3Y3CZJSZ9KT0W7I0KE38WZHHKSW5RH  bible  There will be a highway for the
remnant that i…  land    0.000000  There will be a highway for the remnant
that i…                 False  [PRON, AUX, AUX, DET, NOUN, ADP, DET, NOUN,
PR…  [expl, aux, ROOT, det, attr, prep, det, pobj, …  [(), (VerbForm=Fin),
(VerbForm=Inf), (Definite…          1.277778                  0
0
                              id corpus
sentence          token  complexity
sentence_no_contractions  contraction_expanded
pos_sequence                                       dep_sequence
morph_sequence  morph_complexity  binary_complexity
binary_complexity_75th_split
0  31HLTCK4BLVQ5BO1AUR91TX9V9IVGH  bible  The name of one son was Gershom, for
Moses sai…   foreign land    0.000000  The name of one son was Gershom, for
Moses sai…                 False  [DET, NOUN, ADP, NUM, NOUN, AUX, PROPN,
PUNCT,…  [det, nsubj, prep, nummod, pobj, ROOT, attr, p…  [(Definite=Def,
PronType=Art), (Number=Sing), …          1.520000                  0
0
1  389A2A304OIXVY7G5B71Q9M43LE0CL  bible  unleavened bread, unleavened cakes
mixed with …    wheat flour    0.157895  unleavened bread, unleavened cakes
mixed with …                 False  [ADJ, NOUN, PUNCT, ADJ, NOUN, VERB, ADP,
NOUN,…  [amod, dep, punct, amod, appos, acl, prep, pob…  [(Degree=Pos),
(Number=Sing), (PunctType=Comm)…          1.200000                  0
0
2  31N9JPQXIPIRX2A3S9N0CCFXO6TNHR  bible  However the high places were not taken
away; t…  burnt incense    0.200000  However the high places were not taken
away; t…                 False  [ADV, DET, ADJ, NOUN, AUX, PART, VERB, ADV,
PU…  [advmod, det, amod, nsubjpass, auxpass, neg, c…  [(), (Definite=Def,
PronType=Art), (Degree=Pos…          1.190476                  0
0
3  3JVP4ZJHDPSO81TGXL3N1CKZGQY0IN  bible  and he burnt incense of sweet spices
on it, as…  burnt incense    0.250000  and he burnt incense of sweet spices on
it, as…                 False  [CCONJ, PRON, VERB, NOUN, ADP, ADJ, NOUN,
ADP,…  [cc, nsubj, ROOT, dobj, prep, amod, pobj, prep…  [(ConjType=Cmp),
(Case=Nom, Gender=Masc, Numbe…         1.466667                  0
0
4  3JAOYN9IHL25ZQAUV5EJZ4GHOKL33L  bible  The same day the king made the middle
of the c…   bronze altar    0.214286  The same day the king made the middle of
the c…                 False  [DET, ADJ, NOUN, DET, NOUN, VERB, DET, NOUN,
A…  [det, amod, npadvmod, det, nsubj, ccomp, det, …  [(Definite=Def,
PronType=Art), (Degree=Pos), (…         1.352113                  0
0
                               id corpus
sentence     token  complexity
sentence_no_contractions  contraction_expanded
pos_sequence                                       dep_sequence
morph_sequence  morph_complexity  binary_complexity
binary_complexity_75th_split
0  3K8CQCU3KE19US5SN890DFPK3SANWR  bible  But he, beckoning to them with his
hand to be …      hand    0.000000  But he, beckoning to them with his hand to
be …                 False  [CCONJ, PRON, PUNCT, VERB, ADP, PRON, ADP, PRO…
[cc, nsubj, punct, advcl, prep, pobj, prep, po…  [(ConjType=Cmp), (Case=Nom,
Gender=Masc, Numbe…          1.703704                  0
0
1  3Q2T3FD0ON86LCI41NJYV3PN0BW3MV  bible  If I forget you, Jerusalem, let my
right hand …      hand    0.197368  If I forget you, Jerusalem, let my right
hand …                 False  [SCONJ, PRON, VERB, PRON, PUNCT, PROPN,
PUNCT,…  [mark, nsubj, advcl, dobj, punct, npadvmod, pu…  [(), (Case=Nom,
Number=Sing, Person=1, PronTyp…         1.800000                  0
0
2  3ULIZ0H1VA5C32JJMKOTQ8Z4GUS51B  bible  the ten sons of Haman the son of
Hammedatha, t…      hand    0.200000  the ten sons of Haman the son of
Hammedatha, t…                  True  [DET, NUM, NOUN, ADP, PROPN, DET, NOUN,
ADP, P…  [det, nummod, ROOT, prep, pobj, det, appos, pr…  [(Definite=Def,
PronType=Art), (NumType=Card),…         1.269231                  0
0
3  3BFF0DJK8XCEIOT30ZLBPPSRMZQTSD  bible  Let your hand be lifted up above your
adversar…      hand    0.267857  Let your hand be lifted up above your
adversar…                  False  [VERB, PRON, NOUN, AUX, VERB, ADP, ADP, PRON,
…  [ROOT, poss, nsubjpass, auxpass, ccomp, prt, p…  [(VerbForm=Inf),
(Person=2, Poss=Yes, PronType…         1.250000                  0
0
4  3QREJ3J433XSBS8QMHAICCR0BQ1LKR  bible  Abimelech chased him, and he fled
before him, …  entrance    0.000000  Abimelech chased him, and he fled before
him, …                 False  [PROPN, VERB, PRON, PUNCT, CCONJ, PRON, VERB,
…  [nsubj, ROOT, dobj, punct, cc, nsubj, conj, pr…  [(Number=Sing),
(Tense=Past, VerbForm=Fin), (C…         1.652174                  0
0
                               id corpus
sentence           token  complexity
sentence_no_contractions  contraction_expanded
```

```
pos_sequence                                       dep_sequence
morph_sequence  morph_complexity  binary_complexity
binary_complexity_75th_split
0  3UXQ63NLAAMRIP4WG4XPD98AOYOBLX  bible  for he had an only daughter, about
twelve year…   only daughter    0.025000  for he had an only daughter, about
twelve year…                 False  [SCONJ, PRON, VERB, DET, ADJ, NOUN, PUNCT,
ADV…  [mark, nsubj, ROOT, det, amod, dobj, punct, ad…  [(), (Case=Nom,
Gender=Masc, Number=Sing, Pers…          1.722222                  0
0
1  3FJ2RVH25Z62TA3R8E1077EBUYU92W  bible  All these were cities fortified with
high wall…      high walls    0.100000  All these were cities fortified with
high wall…                 False  [DET, PRON, AUX, NOUN, VERB, ADP, ADJ, NOUN,
P…  [predet, nsubj, ROOT, attr, acl, prep, amod, p…  [(), (Number=Plur,
PronType=Dem), (Mood=Ind, T…          1.136364                  0
0
2  3YO4AH2FPDK1PZHZAT8WAEBL7OEQOF  bible  In the morning, 'It will be foul
weather today…   weather today    0.125000  In the morning, 'It will be foul
weather today…                 False  [ADP, DET, NOUN, PUNCT, PUNCT, PRON,
AUX, AUX,…  [prep, det, pobj, punct, punct, nsubj, aux, RO…  [(),
(Definite=Def, PronType=Art), (Number=Sin…          1.476190
0                             0
3  3X52SWXE0X5Q3O81YIOMX4V84QTCWZ  bible  Her young children also were dashed in
pieces …  young children    0.160714  Her young children also were dashed in
pieces …                 False  [PRON, ADJ, NOUN, ADV, AUX, VERB, ADP, NOUN,
A…  [poss, amod, nsubjpass, advmod, auxpass, ROOT,…  [(Gender=Fem,
Number=Sing, Person=3, Poss=Yes,…          1.514286                  0
0
4  32K26U12DNONTREA84Q1V8UCIH2VD7  bible  All king Solomon's drinking vessels
were of go…       pure gold    0.178571  All king Solomon's drinking vessels
were of go…                 False  [DET, NOUN, PROPN, PART, NOUN, NOUN, AUX,
ADP,…  [det, compound, poss, case, compound, nsubj, c…  [(), (Number=Sing),
(Number=Sing), (), (Number…          1.162791                  0
0
```

[ ]:

```python
dataframes = {
    "train_single_df": train_single_df,
    "train_multi_df": train_multi_df,
    "trial_val_single_df": trial_val_single_df,
    "trial_val_multi_df": trial_val_multi_df,
    "test_single_df": test_single_df,
    "test_multi_df": test_multi_df
}

for df_name, df in dataframes.items():
    print(f"\n=== {df_name} ===")
    print(df['binary_complexity'].value_counts())
```

```
=== train_single_df ===
binary_complexity
0    3865
1    3797
Name: count, dtype: int64

=== train_multi_df ===
binary_complexity
0    759
1    758
Name: count, dtype: int64

=== trial_val_single_df ===
binary_complexity
0    229
1    192
Name: count, dtype: int64

=== trial_val_multi_df ===
binary_complexity
1    51
0    48
Name: count, dtype: int64

=== test_single_df ===
binary_complexity
0    476
1    441
Name: count, dtype: int64

=== test_multi_df ===
binary_complexity
1    99
0    85
Name: count, dtype: int64
```

0.5.1 Create Concatenated and Alternating Features

[ ]:

```python
def pos_method1_concat(row):
    """
    Row-level function for Method 1 (POS):
    sentence_no_contractions + " [" + comma-separated pos_sequence + "]"
    """
    sentence = row['sentence_no_contractions']
    tags = row['pos_sequence']  # list of POS
    if not isinstance(tags, list):
        return sentence  # gracefully handle missing or non-list
    joined_tags = ", ".join(tags)
    return f"{sentence} [{joined_tags}]"

def pos_method2_concat(row):
    """
    Row-level function for Method 2 (POS):
    Interleave tokens with [POS_TAG].
    """
    sentence = row['sentence_no_contractions']
    tags = row['pos_sequence']
    if not isinstance(tags, list):
        return sentence
    tokens = sentence.split()
    interleaved = []
    for tok, pos in zip(tokens, tags):
        interleaved.append(f"{tok} [{pos}]")
    leftover_tokens = tokens[len(tags):]
    interleaved.extend(leftover_tokens)
    return " ".join(interleaved)

def create_pos_method1(df):
    """Creates column snc_pos_seq using pos_method1_concat."""
    df['snc_pos_seq'] = df.apply(pos_method1_concat, axis=1)

def create_pos_method2(df):
    """Creates column snc_pos_alt using pos_method2_concat."""
    df['snc_pos_alt'] = df.apply(pos_method2_concat, axis=1)

for df_name, df in dataframes.items():
    create_pos_method1(df)   # => snc_pos_seq
    create_pos_method2(df)   # => snc_pos_alt
```

[ ]:

```
def morph_method1_concat(row):
    """
    Row-level function for Method 1 (Morph):
    sentence_no_contractions + " [" + comma-separated morph_sequence + "]"
    Where each morph is parenthesized like (Number=Sing), etc.
    """
    sentence = row['sentence_no_contractions']
    morphs = row['morph_sequence']  # list of morph feature strings
    if not isinstance(morphs, list):
        return sentence
    joined_morphs = ", ".join(f"({m})" for m in morphs)
    return f"{sentence} [{joined_morphs}]"

def morph_method2_concat(row):
    """
    Row-level function for Method 2 (Morph):
    Interleave tokens with [({morph})].
    Example:  "bread [(Number=Sing)] dough [(Degree=Pos)] ..."
    """
    sentence = row['sentence_no_contractions']
    morphs = row['morph_sequence']
    if not isinstance(morphs, list):
        return sentence

    tokens = sentence.split()
    interleaved = []
    for tok, morph in zip(tokens, morphs):
        interleaved.append(f"{tok} [({morph})]")
    leftover_tokens = tokens[len(morphs):]
    interleaved.extend(leftover_tokens)
    return " ".join(interleaved)

def create_morph_method1(df):
    """Creates column snc_morph_seq using morph_method1_concat."""
    df['snc_morph_seq'] = df.apply(morph_method1_concat, axis=1)

def create_morph_method2(df):
    """Creates column snc_morph_alt using morph_method2_concat."""
    df['snc_morph_alt'] = df.apply(morph_method2_concat, axis=1)

for df_name, df in dataframes.items():
    create_morph_method1(df)  # => snc_morph_seq
    create_morph_method2(df)  # => snc_morph_alt
```

```
[ ]: def dep_method1_concat(row):
         """
         Row-level function for Method 1 (Dependency):
         sentence_no_contractions + " [" + comma-separated dep_sequence + "]"
         """
         sentence = row['sentence_no_contractions']
         deps = row['dep_sequence']  # list of dependency tags
         if not isinstance(deps, list):
             return sentence
         joined_deps = ", ".join(deps)
         return f"{sentence} [{joined_deps}]"

     def dep_method2_concat(row):
         """
         Row-level function for Method 2 (Dependency):
         Interleave tokens with [DEP_TAG].
         """
         sentence = row['sentence_no_contractions']
```

```
    deps = row['dep_sequence']
    if not isinstance(deps, list):
        return sentence

    tokens = sentence.split()
    interleaved = []
    for tok, dep in zip(tokens, deps):
        interleaved.append(f"{tok} [{dep}]")
    leftover_tokens = tokens[len(deps):]
    interleaved.extend(leftover_tokens)
    return " ".join(interleaved)

def create_dep_method1(df):
    """Creates column snc_dep_seq using dep_method1_concat."""
    df['snc_dep_seq'] = df.apply(dep_method1_concat, axis=1)

def create_dep_method2(df):
    """Creates column snc_dep_alt using dep_method2_concat."""
    df['snc_dep_alt'] = df.apply(dep_method2_concat, axis=1)

for df_name, df in dataframes.items():
    create_dep_method1(df)   # => snc_dep_seq
    create_dep_method2(df)   # => snc_dep_alt (optional if needed)
```

[ ]:

```
def morph_complexity_concat(row):
    """
    Row-level function for appending the numeric 'morph_complexity'
    to the end of sentence_no_contractions.
    """
    sentence = row['sentence_no_contractions']
    mc = row['morph_complexity']
    if pd.isna(mc):
        return sentence  # handle missing
    return f"{sentence} {mc}"

def create_morph_complexity_value(df):
    """
    - For each row, produce:
         sentence_no_contractions + " " + str(morph_complexity)
    - Store result in 'snc_morph_complexity_value'.
    """
    df['snc_morph_complexity_value'] = df.apply(morph_complexity_concat, axis=1)

for df_name, df in dataframes.items():
    create_morph_complexity_value(df)   # => snc_morph_complexity_value
```

[ ]:

```python
# verify column headers

dataframes = [train_single_df, train_multi_df, trial_val_single_df,␣
 ↪trial_val_multi_df, test_single_df, test_multi_df]
for df in dataframes:
  print(df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7662 entries, 0 to 7661
Data columns (total 20 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   id                           7662 non-null   object
 1   corpus                       7662 non-null   object
 2   sentence                     7662 non-null   object
 3   token                        7655 non-null   object
 4   complexity                   7662 non-null   float64
 5   sentence_no_contractions     7662 non-null   object
 6   contraction_expanded         7662 non-null   bool
 7   pos_sequence                 7662 non-null   object
 8   dep_sequence                 7662 non-null   object
 9   morph_sequence               7662 non-null   object
 10  morph_complexity             7662 non-null   float64
 11  binary_complexity            7662 non-null   int64
 12  binary_complexity_75th_split  7662 non-null   int64
 13  snc_pos_seq                  7662 non-null   object
 14  snc_pos_alt                  7662 non-null   object
 15  snc_morph_seq                7662 non-null   object
 16  snc_morph_alt                7662 non-null   object
 17  snc_dep_seq                  7662 non-null   object
 18  snc_dep_alt                  7662 non-null   object
 19  snc_morph_complexity_value   7662 non-null   object
dtypes: bool(1), float64(2), int64(2), object(15)
memory usage: 1.1+ MB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1517 entries, 0 to 1516
Data columns (total 20 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   id                           1517 non-null   object
 1   corpus                       1517 non-null   object
 2   sentence                     1517 non-null   object
 3   token                        1517 non-null   object
 4   complexity                   1517 non-null   float64
 5   sentence_no_contractions     1517 non-null   object
 6   contraction_expanded         1517 non-null   bool
 7   pos_sequence                 1517 non-null   object
 8   dep_sequence                   1517 non-null   object
 9   morph_sequence                 1517 non-null   object
 10  morph_complexity               1517 non-null   float64
 11  binary_complexity              1517 non-null   int64
 12  binary_complexity_75th_split  1517 non-null   int64
 13  snc_pos_seq                    1517 non-null   object
 14  snc_pos_alt                    1517 non-null   object
 15  snc_morph_seq                  1517 non-null   object
 16  snc_morph_alt                  1517 non-null   object
 17  snc_dep_seq                    1517 non-null   object
 18  snc_dep_alt                    1517 non-null   object
 19  snc_morph_complexity_value    1517 non-null   object
dtypes: bool(1), float64(2), int64(2), object(15)
memory usage: 226.8+ KB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 421 entries, 0 to 420
Data columns (total 20 columns):
 #   Column                         Non-Null Count  Dtype
---  ------                         --------------  -----
 0   id                             421 non-null    object
 1   corpus                         421 non-null    object
 2   sentence                       421 non-null    object
 3   token                          421 non-null    object
 4   complexity                     421 non-null    float64
 5   sentence_no_contractions       421 non-null    object
 6   contraction_expanded           421 non-null    bool
 7   pos_sequence                   421 non-null    object
 8   dep_sequence                   421 non-null    object
 9   morph_sequence                 421 non-null    object
 10  morph_complexity               421 non-null    float64
 11  binary_complexity              421 non-null    int64
 12  binary_complexity_75th_split  421 non-null    int64
 13  snc_pos_seq                    421 non-null    object
 14  snc_pos_alt                    421 non-null    object
 15  snc_morph_seq                  421 non-null    object
 16  snc_morph_alt                  421 non-null    object
 17  snc_dep_seq                    421 non-null    object
 18  snc_dep_alt                    421 non-null    object
 19  snc_morph_complexity_value    421 non-null    object
dtypes: bool(1), float64(2), int64(2), object(15)
memory usage: 63.0+ KB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 99 entries, 0 to 98
Data columns (total 20 columns):
 #   Column                         Non-Null Count  Dtype
---  ------                         --------------  -----
```

```
 0   id                            99 non-null     object 
 1   corpus                        99 non-null     object 
 2   sentence                      99 non-null     object 
 3   token                         99 non-null     object 
 4   complexity                    99 non-null     float64
 5   sentence_no_contractions      99 non-null     object 
 6   contraction_expanded          99 non-null     bool   
 7   pos_sequence                  99 non-null     object 
 8   dep_sequence                  99 non-null     object 
 9   morph_sequence                99 non-null     object 
 10  morph_complexity              99 non-null     float64
 11  binary_complexity             99 non-null     int64  
 12  binary_complexity_75th_split  99 non-null     int64  
 13  snc_pos_seq                   99 non-null     object 
 14  snc_pos_alt                   99 non-null     object 
 15  snc_morph_seq                 99 non-null     object 
 16  snc_morph_alt                 99 non-null     object 
 17  snc_dep_seq                   99 non-null     object 
 18  snc_dep_alt                   99 non-null     object 
 19  snc_morph_complexity_value    99 non-null     object 
dtypes: bool(1), float64(2), int64(2), object(15)
memory usage: 14.9+ KB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 917 entries, 0 to 916
Data columns (total 20 columns):
 #   Column                        Non-Null Count  Dtype  
---  ------                        --------------  -----  
 0   id                            917 non-null    object 
 1   corpus                        917 non-null    object 
 2   sentence                      917 non-null    object 
 3   token                         917 non-null    object 
 4   complexity                    917 non-null    float64
 5   sentence_no_contractions      917 non-null    object 
 6   contraction_expanded          917 non-null    bool   
 7   pos_sequence                  917 non-null    object 
 8   dep_sequence                  917 non-null    object 
 9   morph_sequence                917 non-null    object 
 10  morph_complexity              917 non-null    float64
 11  binary_complexity             917 non-null    int64  
 12  binary_complexity_75th_split  917 non-null    int64  
 13  snc_pos_seq                   917 non-null    object 
 14  snc_pos_alt                   917 non-null    object 
 15  snc_morph_seq                 917 non-null    object 
 16  snc_morph_alt                 917 non-null    object 
 17  snc_dep_seq                   917 non-null    object 
 18  snc_dep_alt                   917 non-null    object 
 19  snc_morph_complexity_value    917 non-null    object 
```

```
dtypes: bool(1), float64(2), int64(2), object(15)
memory usage: 137.1+ KB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 184 entries, 0 to 183
Data columns (total 20 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   id                            184 non-null    object
 1   corpus                        184 non-null    object
 2   sentence                      184 non-null    object
 3   token                         184 non-null    object
 4   complexity                    184 non-null    float64
 5   sentence_no_contractions      184 non-null    object
 6   contraction_expanded          184 non-null    bool
 7   pos_sequence                  184 non-null    object
 8   dep_sequence                  184 non-null    object
 9   morph_sequence                184 non-null    object
 10  morph_complexity              184 non-null    float64
 11  binary_complexity             184 non-null    int64
 12  binary_complexity_75th_split  184 non-null    int64
 13  snc_pos_seq                   184 non-null    object
 14  snc_pos_alt                   184 non-null    object
 15  snc_morph_seq                 184 non-null    object
 16  snc_morph_alt                 184 non-null    object
 17  snc_dep_seq                   184 non-null    object
 18  snc_dep_alt                   184 non-null    object
 19  snc_morph_complexity_value    184 non-null    object
dtypes: bool(1), float64(2), int64(2), object(15)
memory usage: 27.6+ KB
None
```

[ ]:

```python
# inspect each df

dataframes = [train_single_df, train_multi_df, trial_val_single_df,␣
 ↪trial_val_multi_df, test_single_df, test_multi_df]
for df in dataframes:
  print(df.head())
```

```
                               id corpus
sentence     token  complexity
sentence_no_contractions  contraction_expanded
pos_sequence                                      dep_sequence
morph_sequence  morph_complexity  binary_complexity
binary_complexity_75th_split                                       snc_pos_seq
snc_pos_alt                                     snc_morph_seq
snc_morph_alt                                       snc_dep_seq
snc_dep_alt                         snc_morph_complexity_value
```

```
0  3ZLW647WALVGE8EBR50EGUBPU4P32A  bible  Behold, there came up out of the river
seven c…     river    0.000000  Behold, there came up out of the river seven
c…                 False  [ADV, PUNCT, PRON, VERB, ADP, ADP, ADP, DET, N…
[advmod, punct, expl, ROOT, prt, prep, prep, d…  [(), (PunctType=Comm), (),
(Tense=Past, VerbFo…          1.041667                  0
0  Behold, there came up out of the river seven c…  Behold, [ADV] there
[PUNCT] came [PRON] up [VE…  Behold, there came up out of the river seven c…
Behold, [()] there [(PunctType=Comm)] came [()…  Behold, there came up out of
the river seven c…  Behold, [advmod] there [punct] came [expl] up …  Behold,
there came up out of the river seven c…
1  34R0BODSP1ZBN3DVY8J8XSIY551E5C  bible  I am a fellow bondservant with you and
with yo…  brothers    0.000000  I am a fellow bondservant with you and with
yo…                 False  [PRON, AUX, DET, ADJ, NOUN, ADP, PRON, CCONJ, …
[nsubj, ROOT, det, amod, attr, prep, pobj, cc,…  [(Case=Nom, Number=Sing,
Person=1, PronType=Pr…          1.461538                  0
0  I am a fellow bondservant with you and with yo…  I [PRON] am [AUX] a [DET]
fellow [ADJ] bondser…  I am a fellow bondservant with you and with yo…  I
[(Case=Nom|Number=Sing|Person=1|PronType=Prs…  I am a fellow bondservant with
you and with yo…  I [nsubj] am [ROOT] a [det] fellow [amod] bond…  I am a
fellow bondservant with you and with yo…
2  3S1WOPCJFGTJU2SGNAN2Y213N6WJE3  bible  The man, the lord of the land, said to
us, 'By…  brothers    0.050000  The man, the lord of the land, said to us,
'By…                 False  [DET, NOUN, PUNCT, DET, PROPN, ADP, DET, NOUN,…
[det, nsubj, punct, det, appos, prep, det, pob…  [(Definite=Def,
PronType=Art), (Number=Sing), …          1.354167                  0
0  The man, the lord of the land, said to us, 'By…  The [DET] man, [NOUN] the
[PUNCT] lord [DET] o…  The man, the lord of the land, said to us, 'By…  The
[(Definite=Def|PronType=Art)] man, [(Numbe…  The man, the lord of the land,
said to us, 'By…  The [det] man, [nsubj] the [punct] lord [det] …  The man,
the lord of the land, said to us, 'By…
3  3BFNCI9LYKQNO9BHXHH9CLSX5KP738  bible  Shimei had sixteen sons and six
daughters; but…  brothers    0.150000  Shimei had sixteen sons and six
daughters; but…                 True  [PROPN, VERB, NUM, NOUN, CCONJ, NUM,
NOUN, PUN…  [nsubj, ROOT, nummod, dobj, cc, nummod, conj, …  [(Number=Sing),
(Tense=Past, VerbForm=Fin), (N…          1.275862                  0
0  Shimei had sixteen sons and six daughters; but…  Shimei [PROPN] had [VERB]
sixteen [NUM] sons […  Shimei had sixteen sons and six daughters; but…
Shimei [(Number=Sing)] had [(Tense=Past|VerbFo…  Shimei had sixteen sons and
six daughters; but…  Shimei [nsubj] had [ROOT] sixteen [nummod] son…  Shimei
had sixteen sons and six daughters; but…
4  3G5RUKN2EC3YIWSKUXZ8ZVH95R49N2  bible               "He has put my brothers
far from me.  brothers    0.263889               "He has put my brothers far
from me.                 False  [PUNCT, PRON, AUX, VERB, PRON, NOUN, ADV,
ADP,…  [punct, nsubj, aux, ROOT, poss, dobj, advmod, …  [(PunctSide=Ini,
PunctType=Quot), (Case=Nom, G…          2.500000                  0
0  "He has put my brothers far from me. [PUNCT, P…  "He [PUNCT] has [PRON] put
[AUX] my [VERB] bro…  "He has put my brothers far from me. [(PunctSi…  "He
[(PunctSide=Ini|PunctType=Quot)] has [(Cas…  "He has put my brothers far from
me. [punct, n…  "He [punct] has [nsubj] put [aux] my [ROOT] br…
"He has put my brothers far from me. 2.5
                               id corpus
sentence            token  complexity
sentence_no_contractions  contraction_expanded
pos_sequence                                       dep_sequence
morph_sequence  morph_complexity  binary_complexity
binary_complexity_75th_split                                        snc_pos_seq
snc_pos_alt                                      snc_morph_seq
snc_morph_alt                                        snc_dep_seq
snc_dep_alt                         snc_morph_complexity_value
0  3S37Y8CWI80N8KVM53U4E6JKCDC4WE  bible  but the seventh day is a Sabbath to
Yahweh you…      seventh day    0.027778  but the seventh day is a Sabbath to
Yahweh you…                False  [CCONJ, DET, ADJ, NOUN, AUX, DET, PROPN,
ADP, …  [cc, det, amod, nsubj, ccomp, det, attr, prep,…  [(ConjType=Cmp),
(Definite=Def, PronType=Art),…          1.341772                 0
0  but the seventh day is a Sabbath to Yahweh you…  but [CCONJ] the [DET]
seventh [ADJ] day [NOUN]…  but the seventh day is a Sabbath to Yahweh you…
but [(ConjType=Cmp)] the [(Definite=Def|PronTy…  but the seventh day is a
Sabbath to Yahweh you…  but [cc] the [det] seventh [amod] day [nsubj] …  but
the seventh day is a Sabbath to Yahweh you…
1  3WGCNLZJKF877FYC1Q6COKNWTDWD11  bible  But let each man test his own work,
and then h…         own work    0.050000  But let each man test his own work,
and then h…                False  [CCONJ, VERB, DET, NOUN, VERB, PRON, ADJ,
NOUN…  [cc, ROOT, det, nsubj, ccomp, poss, amod, dobj…  [(ConjType=Cmp),
(VerbForm=Inf), (), (Number=S…          1.608696                 0
0  But let each man test his own work, and then h…  But [CCONJ] let [VERB]
each [DET] man [NOUN] t…  But let each man test his own work, and then h…
But [(ConjType=Cmp)] let [(VerbForm=Inf)] each…  But let each man test his own
work, and then h…  But [cc] let [ROOT] each [det] man [nsubj] tes…  But let
each man test his own work, and then h…
2  3UOMW19E6D6WQ5TH2HDD74IVKTP5CB  bible  To him who by understanding made the
heavens; …  loving kindness    0.050000  To him who by understanding made the
heavens; …                False  [ADP, PRON, PRON, ADP, VERB, VERB, DET,
NOUN, …  [prep, pobj, nsubj, prep, pcomp, advcl, det, d…  [(), (Case=Acc,
Gender=Masc, Number=Sing, Pers…          1.562500                 0
0  To him who by understanding made the heavens; …  To [ADP] him [PRON] who
[PRON] by [ADP] unders…  To him who by understanding made the heavens; …  To
[()] him [(Case=Acc|Gender=Masc|Number=Sing…  To him who by understanding made
the heavens; …  To [prep] him [pobj] who [nsubj] by [prep] und…  To him who
by understanding made the heavens; …
3  36JW4WBR06KF9AXMUL4N476OMF8FHD  bible  Remember to me, my God, this also, and
spare m…  loving kindness    0.050000  Remember to me, my God, this also, and
spare m…                False  [VERB, ADP, PRON, PUNCT, PRON, PROPN, PUNCT,
P…  [ROOT, prep, pobj, punct, poss, npadvmod, punc…  [(VerbForm=Inf), (),
(Case=Acc, Number=Sing, P…          1.590909                 0
0  Remember to me, my God, this also, and spare m…  Remember [VERB] to [ADP]
me, [PRON] my [PUNCT]…  Remember to me, my God, this also, and spare m…
```

```
Remember [(VerbForm=Inf)] to [()] me, [(Case=A…  Remember to me, my God, this
also, and spare m…  Remember [ROOT] to [prep] me, [pobj] my [punct…
Remember to me, my God, this also, and spare m…
4  3HRWUH63QU2FH9Q8R7MRNFC7JX2N5A  bible  Because your loving kindness is better
than li…  loving kindness    0.075000  Because your loving kindness is better
than li…                 False  [SCONJ, PRON, ADJ, NOUN, AUX, ADJ, ADP, NOUN,
…  [mark, poss, amod, nsubj, advcl, acomp, prep, …  [(), (Person=2,
Poss=Yes, PronType=Prs), (Degr…          1.600000                  0
0  Because your loving kindness is better than li…  Because [SCONJ] your
[PRON] loving [ADJ] kindn…  Because your loving kindness is better than li…
Because [()] your [(Person=2|Poss=Yes|PronType…  Because your loving kindness
is better than li…  Because [mark] your [poss] loving [amod] kindn…  Because
your loving kindness is better than li…
                               id corpus
sentence token  complexity                           sentence_no_contractions
contraction_expanded                                       pos_sequence
dep_sequence                                     morph_sequence
morph_complexity  binary_complexity  binary_complexity_75th_split
snc_pos_seq                                        snc_pos_alt
snc_morph_seq                                      snc_morph_alt
snc_dep_seq                                        snc_dep_alt
snc_morph_complexity_value
0  3QI9WAYOGQB8GQIR4MDIEF0D2RLS67  bible  They will not hurt nor destroy in all
my holy …   sea    0.000000  They will not hurt nor destroy in all my holy …
False  [PRON, AUX, PART, VERB, CCONJ, VERB, ADP, PRON…  [nsubj, aux, neg,
ccomp, cc, conj, prep, prede…  [(Case=Nom, Number=Plur, Person=3,
PronType=Pr…          1.129032                  0
0  They will not hurt nor destroy in all my holy …  They [PRON] will [AUX] not
[PART] hurt [VERB] …  They will not hurt nor destroy in all my holy …  They
[(Case=Nom|Number=Plur|Person=3|PronType=…  They will not hurt nor destroy in
all my holy …  They [nsubj] will [aux] not [neg] hurt [ccomp]…  They will
not hurt nor destroy in all my holy …
1  3T8DUCXY0N6WD9X4RTLK8UN1U929TF  bible  that sends ambassadors by the sea,
even in ves…   sea    0.102941  that sends ambassadors by the sea, even in
ves…                 False  [PRON, VERB, NOUN, ADP, DET, NOUN, PUNCT, ADV,…
[nsubj, ROOT, dobj, prep, det, pobj, punct, ad…  [(PronType=Rel),
(Number=Sing, Person=3, Tense…          1.263158                  0
0  that sends ambassadors by the sea, even in ves…  that [PRON] sends [VERB]
ambassadors [NOUN] by…  that sends ambassadors by the sea, even in ves…
that [(PronType=Rel)] sends [(Number=Sing|Pers…  that sends ambassadors by the
sea, even in ves…  that [nsubj] sends [ROOT] ambassadors [dobj] b…  that
sends ambassadors by the sea, even in ves…
2  3I7KR83SNADXAQ7HXK7S7305BYB9KD  bible  and they entered into the boat, and
were going…   sea    0.109375  and they entered into the boat, and were
going…                 False  [CCONJ, PRON, VERB, ADP, DET, NOUN, PUNCT,
CCO…  [cc, nsubj, ROOT, prep, det, pobj, punct, cc, …  [(ConjType=Cmp),
(Case=Nom, Number=Plur, Perso…          1.437500                  0
0  and they entered into the boat, and were going…  and [CCONJ] they [PRON]
```

```
entered [VERB] into [A…  and they entered into the boat, and were going…
and [(ConjType=Cmp)] they [(Case=Nom|Number=Pl…  and they entered into the
boat, and were going…  and [cc] they [nsubj] entered [ROOT] into [pre…  and
they entered into the boat, and were going…
3  3BO3NEOQM0HK9ERYPN0GQIWCPC4IAQ  bible  Joseph laid up grain as the sand of
the sea, v…   sea    0.160714  Joseph laid up grain as the sand of the sea,
v…                   False  [PROPN, VERB, ADP, NOUN, ADP, DET, NOUN, ADP, …
[nsubj, ROOT, prt, dobj, prep, det, pobj, prep…  [(Number=Sing), (Tense=Past,
VerbForm=Fin), ()…          1.400000                 0
0  Joseph laid up grain as the sand of the sea, v…  Joseph [PROPN] laid [VERB]
up [ADP] grain [NOU…  Joseph laid up grain as the sand of the sea, v…
Joseph [(Number=Sing)] laid [(Tense=Past|VerbF…  Joseph laid up grain as the
sand of the sea, v…  Joseph [nsubj] laid [ROOT] up [prt] grain [dob…  Joseph
laid up grain as the sand of the sea, v…
4  3Y3CZJSZ9KT0W7I0KE38WZHHKSW5RH  bible  There will be a highway for the
remnant that i…  land    0.000000  There will be a highway for the remnant
that i…                False  [PRON, AUX, AUX, DET, NOUN, ADP, DET, NOUN,
PR…  [expl, aux, ROOT, det, attr, prep, det, pobj, …  [(), (VerbForm=Fin),
(VerbForm=Inf), (Definite…         1.277778                 0
0  There will be a highway for the remnant that i…  There [PRON] will [AUX] be
[AUX] a [DET] highw…  There will be a highway for the remnant that i…  There
[()] will [(VerbForm=Fin)] be [(VerbForm…  There will be a highway for the
remnant that i…  There [expl] will [aux] be [ROOT] a [det] high…  There will
be a highway for the remnant that i…
                               id corpus
sentence          token  complexity
sentence_no_contractions  contraction_expanded
pos_sequence                                        dep_sequence
morph_sequence  morph_complexity  binary_complexity
binary_complexity_75th_split                                        snc_pos_seq
snc_pos_alt                                     snc_morph_seq
snc_morph_alt                                       snc_dep_seq
snc_dep_alt                         snc_morph_complexity_value
0  31HLTCK4BLVQ5BO1AUR91TX9V9IVGH  bible  The name of one son was Gershom, for
Moses sai…   foreign land    0.000000  The name of one son was Gershom, for
Moses sai…                 False  [DET, NOUN, ADP, NUM, NOUN, AUX, PROPN,
PUNCT,…  [det, nsubj, prep, nummod, pobj, ROOT, attr, p…  [(Definite=Def,
PronType=Art), (Number=Sing), …          1.520000                 0
0  The name of one son was Gershom, for Moses sai…  The [DET] name [NOUN] of
[ADP] one [NUM] son […  The name of one son was Gershom, for Moses sai…  The
[(Definite=Def|PronType=Art)] name [(Numbe…  The name of one son was Gershom,
for Moses sai…  The [det] name [nsubj] of [prep] one [nummod] …  The name of
one son was Gershom, for Moses sai…
1  389A2A304OIXVY7G5B71Q9M43LE0CL  bible  unleavened bread, unleavened cakes
mixed with …    wheat flour    0.157895  unleavened bread, unleavened cakes
mixed with …                 False  [ADJ, NOUN, PUNCT, ADJ, NOUN, VERB, ADP,
NOUN,…  [amod, dep, punct, amod, appos, acl, prep, pob…  [(Degree=Pos),
(Number=Sing), (PunctType=Comm)…          1.200000                 0
```

```
0  unleavened bread, unleavened cakes mixed with …  unleavened [ADJ] bread,
[NOUN] unleavened [PUN…  unleavened bread, unleavened cakes mixed with …
unleavened [(Degree=Pos)] bread, [(Number=Sing…  unleavened bread, unleavened
cakes mixed with …  unleavened [amod] bread, [dep] unleavened [pun…
unleavened bread, unleavened cakes mixed with …
2  31N9JPQXIPIRX2A3S9N0CCFXO6TNHR  bible  However the high places were not taken
away; t…  burnt incense    0.200000  However the high places were not taken
away; t…                 False  [ADV, DET, ADJ, NOUN, AUX, PART, VERB, ADV,
PU…  [advmod, det, amod, nsubjpass, auxpass, neg, c…  [(), (Definite=Def,
PronType=Art), (Degree=Pos…          1.190476                0
0  However the high places were not taken away; t…  However [ADV] the [DET]
high [ADJ] places [NOU…  However the high places were not taken away; t…
However [()] the [(Definite=Def|PronType=Art)]…  However the high places were
not taken away; t…  However [advmod] the [det] high [amod] places …  However
the high places were not taken away; t…
3  3JVP4ZJHDPSO81TGXL3N1CKZGQY0IN  bible  and he burnt incense of sweet spices
on it, as…  burnt incense    0.250000  and he burnt incense of sweet spices on
it, as…                 False  [CCONJ, PRON, VERB, NOUN, ADP, ADJ, NOUN,
ADP,…  [cc, nsubj, ROOT, dobj, prep, amod, pobj, prep…  [(ConjType=Cmp),
(Case=Nom, Gender=Masc, Numbe…         1.466667                0
0  and he burnt incense of sweet spices on it, as…  and [CCONJ] he [PRON]
burnt [VERB] incense [NO…  and he burnt incense of sweet spices on it, as…
and [(ConjType=Cmp)] he [(Case=Nom|Gender=Masc…  and he burnt incense of sweet
spices on it, as…  and [cc] he [nsubj] burnt [ROOT] incense [dobj…  and he
burnt incense of sweet spices on it, as…
4  3JAOYN9IHL25ZQAUV5EJZ4GH0KL33L  bible  The same day the king made the middle
of the c…   bronze altar    0.214286  The same day the king made the middle of
the c…                 False  [DET, ADJ, NOUN, DET, NOUN, VERB, DET, NOUN,
A…  [det, amod, npadvmod, det, nsubj, ccomp, det, …  [(Definite=Def,
PronType=Art), (Degree=Pos), (…          1.352113                0
0  The same day the king made the middle of the c…  The [DET] same [ADJ] day
[NOUN] the [DET] king…  The same day the king made the middle of the c…  The
[(Definite=Def|PronType=Art)] same [(Degre…  The same day the king made the
middle of the c…  The [det] same [amod] day [npadvmod] the [det]…  The same
day the king made the middle of the c…
                              id corpus
sentence     token  complexity
sentence_no_contractions  contraction_expanded
pos_sequence                                      dep_sequence
morph_sequence  morph_complexity  binary_complexity
binary_complexity_75th_split                                        snc_pos_seq
snc_pos_alt                                      snc_morph_seq
snc_morph_alt                                        snc_dep_seq
snc_dep_alt                         snc_morph_complexity_value
0  3K8CQCU3KE19US5SN890DFPK3SANWR  bible  But he, beckoning to them with his
hand to be …      hand    0.000000  But he, beckoning to them with his hand to
be …                 False  [CCONJ, PRON, PUNCT, VERB, ADP, PRON, ADP, PRO…
[cc, nsubj, punct, advcl, prep, pobj, prep, po…  [(ConjType=Cmp), (Case=Nom,
Gender=Masc, Numbe…          1.703704                  0
0  But he, beckoning to them with his hand to be …  But [CCONJ] he, [PRON]
beckoning [PUNCT] to [V…  But he, beckoning to them with his hand to be …
But [(ConjType=Cmp)] he, [(Case=Nom|Gender=Mas…  But he, beckoning to them
with his hand to be …  But [cc] he, [nsubj] beckoning [punct] to [adv…  But
he, beckoning to them with his hand to be …
1  3Q2T3FD0ON86LCI41NJYV3PN0BW3MV  bible  If I forget you, Jerusalem, let my
right hand …      hand    0.197368  If I forget you, Jerusalem, let my right
hand …                 False  [SCONJ, PRON, VERB, PRON, PUNCT, PROPN,
PUNCT,…  [mark, nsubj, advcl, dobj, punct, npadvmod, pu…  [(), (Case=Nom,
Number=Sing, Person=1, PronTyp…          1.800000                  0
0  If I forget you, Jerusalem, let my right hand …  If [SCONJ] I [PRON] forget
[VERB] you, [PRON] …  If I forget you, Jerusalem, let my right hand …  If
[()] I [(Case=Nom|Number=Sing|Person=1|Pron…  If I forget you, Jerusalem, let
my right hand …  If [mark] I [nsubj] forget [advcl] you, [dobj]…  If I
forget you, Jerusalem, let my right hand …
2  3ULIZ0H1VA5C32JJMK0TQ8Z4GUS51B  bible  the ten sons of Haman the son of
Hammedatha, t…      hand    0.200000  the ten sons of Haman the son of
Hammedatha, t…                  True  [DET, NUM, NOUN, ADP, PROPN, DET, NOUN,
ADP, P…  [det, nummod, ROOT, prep, pobj, det, appos, pr…  [(Definite=Def,
PronType=Art), (NumType=Card),…          1.269231                  0
0  the ten sons of Haman the son of Hammedatha, t…  the [DET] ten [NUM] sons
[NOUN] of [ADP] Haman…  the ten sons of Haman the son of Hammedatha, t…  the
[(Definite=Def|PronType=Art)] ten [(NumTyp…  the ten sons of Haman the son of
Hammedatha, t…  the [det] ten [nummod] sons [ROOT] of [prep] H…  the ten
sons of Haman the son of Hammedatha, t…
3  3BFF0DJK8XCEIOT30ZLBPPSRMZQTSD  bible  Let your hand be lifted up above your
adversar…      hand    0.267857  Let your hand be lifted up above your
adversar…                 False  [VERB, PRON, NOUN, AUX, VERB, ADP, ADP, PRON,
…  [ROOT, poss, nsubjpass, auxpass, ccomp, prt, p…  [(VerbForm=Inf),
(Person=2, Poss=Yes, PronType…          1.250000                  0
0  Let your hand be lifted up above your adversar…  Let [VERB] your [PRON]
hand [NOUN] be [AUX] li…  Let your hand be lifted up above your adversar…
Let [(VerbForm=Inf)] your [(Person=2|Poss=Yes|…  Let your hand be lifted up
above your adversar…  Let [ROOT] your [poss] hand [nsubjpass] be [au…  Let
your hand be lifted up above your adversar…
4  3QREJ3J433XSBS8QMHAICCR0BQ1LKR  bible  Abimelech chased him, and he fled
before him, …  entrance    0.000000  Abimelech chased him, and he fled before
him, …                 False  [PROPN, VERB, PRON, PUNCT, CCONJ, PRON, VERB,
…  [nsubj, ROOT, dobj, punct, cc, nsubj, conj, pr…  [(Number=Sing),
(Tense=Past, VerbForm=Fin), (C…          1.652174                  0
0  Abimelech chased him, and he fled before him, …  Abimelech [PROPN] chased
[VERB] him, [PRON] an…  Abimelech chased him, and he fled before him, …
Abimelech [(Number=Sing)] chased [(Tense=Past|…  Abimelech chased him, and he
fled before him, …  Abimelech [nsubj] chased [ROOT] him, [dobj] an…
Abimelech chased him, and he fled before him, …
                               id corpus
sentence           token  complexity
```

```
sentence_no_contractions  contraction_expanded
pos_sequence                                       dep_sequence
morph_sequence  morph_complexity  binary_complexity
binary_complexity_75th_split                                        snc_pos_seq
snc_pos_alt                                      snc_morph_seq
snc_morph_alt                                        snc_dep_seq
snc_dep_alt                         snc_morph_complexity_value
0  3UXQ63NLAAMRIP4WG4XPD98AOYOBLX  bible  for he had an only daughter, about
twelve year…   only daughter    0.025000  for he had an only daughter, about
twelve year…                  False  [SCONJ, PRON, VERB, DET, ADJ, NOUN, PUNCT,
ADV…  [mark, nsubj, ROOT, det, amod, dobj, punct, ad…  [(), (Case=Nom,
Gender=Masc, Number=Sing, Pers…         1.722222                 0
0  for he had an only daughter, about twelve year…  for [SCONJ] he [PRON] had
[VERB] an [DET] only…  for he had an only daughter, about twelve year…  for
[()] he [(Case=Nom|Gender=Masc|Number=Sing…  for he had an only daughter,
about twelve year…  for [mark] he [nsubj] had [ROOT] an [det] only…  for he
had an only daughter, about twelve year…
1  3FJ2RVH25Z62TA3R8E1O77EBUYU92W  bible  All these were cities fortified with
high wall…      high walls    0.100000  All these were cities fortified with
high wall…                 False  [DET, PRON, AUX, NOUN, VERB, ADP, ADJ, NOUN,
P…  [predet, nsubj, ROOT, attr, acl, prep, amod, p…  [(), (Number=Plur,
PronType=Dem), (Mood=Ind, T…         1.136364                 0
0  All these were cities fortified with high wall…  All [DET] these [PRON]
were [AUX] cities [NOUN…  All these were cities fortified with high wall…
All [()] these [(Number=Plur|PronType=Dem)] we…  All these were cities
fortified with high wall…  All [predet] these [nsubj] were [ROOT] cities …
All these were cities fortified with high wall…
2  3YO4AH2FPDK1PZHZAT8WAEBL7OEQOF  bible  In the morning, 'It will be foul
weather today…   weather today    0.125000  In the morning, 'It will be foul
weather today…                 False  [ADP, DET, NOUN, PUNCT, PUNCT, PRON,
AUX, AUX,…  [prep, det, pobj, punct, punct, nsubj, aux, RO…  [(),
(Definite=Def, PronType=Art), (Number=Sin…         1.476190
0                             0  In the morning, 'It will be foul weather
today…  In [ADP] the [DET] morning, [NOUN] 'It [PUNCT]…  In the morning, 'It
will be foul weather today…  In [()] the [(Definite=Def|PronType=Art)] morn…
In the morning, 'It will be foul weather today…  In [prep] the [det] morning,
[pobj] 'It [punct…  In the morning, 'It will be foul weather today…
3  3X52SWXE0X5Q3O81YI0MX4V84QTCWZ  bible  Her young children also were dashed in
pieces …  young children    0.160714  Her young children also were dashed in
pieces …                 False  [PRON, ADJ, NOUN, ADV, AUX, VERB, ADP, NOUN,
A…  [poss, amod, nsubjpass, advmod, auxpass, ROOT,…  [(Gender=Fem,
Number=Sing, Person=3, Poss=Yes,…         1.514286                 0
0  Her young children also were dashed in pieces …  Her [PRON] young [ADJ]
children [NOUN] also [A…  Her young children also were dashed in pieces …
Her [(Gender=Fem|Number=Sing|Person=3|Poss=Yes…  Her young children also were
dashed in pieces …  Her [poss] young [amod] children [nsubjpass] a…  Her
young children also were dashed in pieces …
4  32K26U12DN0NTREA84Q1V8UCIH2VD7  bible  All king Solomon's drinking vessels
```

```
were of go…       pure gold    0.178571  All king Solomon's drinking vessels
were of go…                 False  [DET, NOUN, PROPN, PART, NOUN, NOUN, AUX,
ADP,…  [det, compound, poss, case, compound, nsubj, c…  [(), (Number=Sing),
(Number=Sing), (), (Number…          1.162791                   0
0  All king Solomon's drinking vessels were of go…  All [DET] king [NOUN]
Solomon's [PROPN] drinki…  All king Solomon's drinking vessels were of go…
All [()] king [(Number=Sing)] Solomon's [(Numb…  All king Solomon's drinking
vessels were of go…  All [det] king [compound] Solomon's [poss] dri…  All
king Solomon's drinking vessels were of go…
```

```
[ ]: dataframes = [train_single_df, train_multi_df, trial_val_single_df,␣
      ↪trial_val_multi_df, test_single_df, test_multi_df]

     for df in dataframes:
         if hasattr(df, 'columns') and 'corpus' in df.columns:
             print(df[df['corpus'] == 'biomed'].head())
         else:
             pass
```

```
                                  id  corpus
sentence    token  complexity                           sentence_no_contractions
contraction_expanded                                       pos_sequence
dep_sequence                                    morph_sequence
morph_complexity  binary_complexity  binary_complexity_75th_split
snc_pos_seq                                         snc_pos_alt
snc_morph_seq                                       snc_morph_alt
snc_dep_seq                                         snc_dep_alt
snc_morph_complexity_value
2574  37ZQELHEQ0YDPGBEJ63D4HNT5SBNMJ  biomed  In fact, this situation gave an
opportunity to…     fact    0.000000  In fact, this situation gave an
opportunity to…                 False  [ADP, NOUN, PUNCT, DET, NOUN, VERB,
DET, NOUN,…  [prep, pobj, punct, det, nsubj, ROOT, det, dob…  [(),
(Number=Sing), (PunctType=Comm), (Number=…          1.000000
0                              0  In fact, this situation gave an opportunity
to…  In [ADP] fact, [NOUN] this [PUNCT] situation […  In fact, this
situation gave an opportunity to…  In [()] fact, [(Number=Sing)] this
[(PunctType…  In fact, this situation gave an opportunity to…  In [prep]
fact, [pobj] this [punct] situation …  In fact, this situation gave an
opportunity to…
2575  3XUSYT70IT170QDU572CAF4MOM1D0B  biomed  It can be inferred from this fact
that Nrl is …     fact    0.183333  It can be inferred from this fact that Nrl
is …                 False  [PRON, AUX, AUX, VERB, ADP, DET, NOUN, SCONJ, …
[nsubjpass, aux, auxpass, ROOT, prep, det, pob…  [(Gender=Neut, Number=Sing,
Person=3, PronType…          1.291667                  0
0  It can be inferred from this fact that Nrl is …  It [PRON] can [AUX] be
[AUX] inferred [VERB] f…  It can be inferred from this fact that Nrl is …
It [(Gender=Neut|Number=Sing|Person=3|PronType…  It can be inferred from this
fact that Nrl is …  It [nsubjpass] can [aux] be [auxpass] inferred…  It can
```

```
be inferred from this fact that Nrl is …
2576  34R3P23QHS1HKWJHKAEN8VSOHJ9WH5  biomed  The site of mutation is of
interest, particula…     fact    0.300000  The site of mutation is of
interest, particula…                 False  [DET, NOUN, ADP, NOUN, AUX, ADP,
NOUN, PUNCT, …  [det, nsubj, prep, pobj, ccomp, prep, pobj, pu…
[(Definite=Def, PronType=Art), (Number=Sing), …          1.083333
1                             0  The site of mutation is of interest,
particula…  The [DET] site [NOUN] of [ADP] mutation [NOUN]…  The site of
mutation is of interest, particula…  The [(Definite=Def|PronType=Art)] site
[(Numbe…  The site of mutation is of interest, particula…  The [det] site
[nsubj] of [prep] mutation [pob…  The site of mutation is of interest,
particula…
2577  3L21G7IH47WA5QT3XMTQ15XXB1L1YG  biomed  This model reflects many other
observed change…  studies    0.000000  This model reflects many other observed
change…                 False  [DET, NOUN, VERB, ADJ, ADJ, VERB, NOUN, VERB,
…  [det, nsubj, ROOT, amod, amod, amod, dobj, acl…  [(Number=Sing,
PronType=Dem), (Number=Sing), (…          1.428571                 0
0  This model reflects many other observed change…  This [DET] model [NOUN]
reflects [VERB] many […  This model reflects many other observed change…
This [(Number=Sing|PronType=Dem)] model [(Numb…  This model reflects many
other observed change…  This [det] model [nsubj] reflects [ROOT] many …
This model reflects many other observed change…
2578  3ZXNP4Z39RL4GD163NL987ME58H7LR  biomed  Several studies have been carried
out to detec…  studies    0.125000  Several studies have been carried out to
detec…                 False  [ADJ, NOUN, AUX, AUX, VERB, ADP, PART, VERB,
N…  [amod, nsubjpass, aux, auxpass, ROOT, prt, aux…  [(Degree=Pos),
(Number=Plur), (Mood=Ind, Tense…          1.000000                 0
0  Several studies have been carried out to detec…  Several [ADJ] studies
[NOUN] have [AUX] been […  Several studies have been carried out to detec…
Several [(Degree=Pos)] studies [(Number=Plur)]…  Several studies have been
carried out to detec…  Several [amod] studies [nsubjpass] have [aux] …
Several studies have been carried out to detec…
                                 id  corpus
sentence              token  complexity
sentence_no_contractions  contraction_expanded
pos_sequence                                      dep_sequence
morph_sequence  morph_complexity  binary_complexity
binary_complexity_75th_split                                        snc_pos_seq
snc_pos_alt                                      snc_morph_seq
snc_morph_alt                                        snc_dep_seq
snc_dep_alt                       snc_morph_complexity_value
505  3D7VY91L65XB07MHGGY4DMNZO4QMBO  biomed  We have found similar values for
Plg-/- mice i…     similar values    0.027778  We have found similar values
for Plg-/- mice i…                False  [PRON, AUX, VERB, ADJ, NOUN, ADP,
NOUN, NOUN, …  [nsubj, aux, ROOT, amod, dobj, prep, compound,…  [(Case=Nom,
Number=Plur, Person=1, PronType=Pr…          1.275862                 0
0  We have found similar values for Plg-/- mice i…  We [PRON] have [AUX] found
[VERB] similar [ADJ…  We have found similar values for Plg-/- mice i…  We
```

```
[(Case=Nom|Number=Plur|Person=1|PronType=Pr…  We have found similar values for
Plg-/- mice i…  We [nsubj] have [aux] found [ROOT] similar [am…  We have
found similar values for Plg-/- mice i…
506  3NZ1E5QA6Z1DG01BOHHIWKCD290B5N  biomed  Our results and the sequences we
provide will …     global studies    0.075000  Our results and the sequences
we provide will …                 False  [PRON, NOUN, CCONJ, DET, NOUN, PRON,
VERB, AUX…  [poss, nsubj, cc, det, conj, nsubj, relcl, aux…  [(Number=Plur,
Person=1, Poss=Yes, PronType=Pr…          1.304348                 0
0  Our results and the sequences we provide will …  Our [PRON] results [NOUN]
and [CCONJ] the [DET…  Our results and the sequences we provide will …  Our
[(Number=Plur|Person=1|Poss=Yes|PronType=P…  Our results and the sequences we
provide will …  Our [poss] results [nsubj] and [cc] the [det] …  Our results
and the sequences we provide will …
507  3XUSYT70IT170QDU572CAF4MOM10DY  biomed  Although great effort was put forth
to elimina…      other factors    0.075000  Although great effort was put
forth to elimina…                 False  [SCONJ, ADJ, NOUN, AUX, VERB, ADV,
PART, VERB,…  [mark, amod, nsubjpass, auxpass, advcl, advmod…  [(),
(Degree=Pos), (Number=Sing), (Mood=Ind, N…          1.121951
0                            0  Although great effort was put forth to
elimina…  Although [SCONJ] great [ADJ] effort [NOUN] was…  Although great
effort was put forth to elimina…  Although [()] great [(Degree=Pos)] effort
[(Nu…  Although great effort was put forth to elimina…  Although [mark]
great [amod] effort [nsubjpass…  Although great effort was put forth to
elimina…
508  3S1WOPCJFGTJU2SGNAN2Y213N78JEH  biomed  Complex traits, such as polygenic
growth and o…     direct effects    0.083333  Complex traits, such as
polygenic growth and o…                 False  [ADJ, NOUN, PUNCT, ADJ, ADP,
ADJ, NOUN, CCONJ,…  [amod, nsubjpass, punct, amod, prep, amod, pob…
[(Degree=Pos), (Number=Plur), (PunctType=Comm)…          1.051282
0                            0  Complex traits, such as polygenic growth and
o…  Complex [ADJ] traits, [NOUN] such [PUNCT] as [… Complex traits, such as
polygenic growth and o…  Complex [(Degree=Pos)] traits, [(Number=Plur)]…
Complex traits, such as polygenic growth and o…  Complex [amod] traits,
[nsubjpass] such [punct…  Complex traits, such as polygenic growth and o…
509  3RBI0I35XE36FT7IKQ79PYCU9MQY3Y  biomed  As known from the frequent human
vWF-syndrome …  normal conditions    0.100000  As known from the frequent
human vWF-syndrome …                 False  [SCONJ, VERB, ADP, DET, ADJ, ADJ,
NOUN, PUNCT,…  [mark, advcl, prep, det, amod, amod, compound,…  [(),
(Aspect=Perf, Tense=Past, VerbForm=Part),…          1.032258
0                            0  As known from the frequent human vWF-syndrome
…  As [SCONJ] known [VERB] from [ADP] the [DET] f…  As known from the
frequent human vWF-syndrome …  As [()] known
[(Aspect=Perf|Tense=Past|VerbFor…  As known from the frequent human vWF-
syndrome …  As [mark] known [advcl] from [prep] the [det] …  As known from
the frequent human vWF-syndrome …
                                 id  corpus
sentence token  complexity                          sentence_no_contractions
contraction_expanded                                      pos_sequence
```

```
dep_sequence                                    morph_sequence
morph_complexity  binary_complexity  binary_complexity_75th_split
snc_pos_seq                                       snc_pos_alt
snc_morph_seq                                     snc_morph_alt
snc_dep_seq                                       snc_dep_alt
snc_morph_complexity_value
143  3BAKUKE49HC18PHHJR1WT9408E0R1Y  biomed  The expression of Sam68 was not
altered in agi…  bone    0.027778  The expression of Sam68 was not altered in
agi…                 False  [DET, NOUN, ADP, PROPN, AUX, PART, VERB, ADP, …
[det, nsubjpass, prep, pobj, auxpass, neg, ROO…  [(Definite=Def,
PronType=Art), (Number=Sing), …         1.434783                  0
0  The expression of Sam68 was not altered in agi…  The [DET] expression
[NOUN] of [ADP] Sam68 [PR…  The expression of Sam68 was not altered in agi…
The [(Definite=Def|PronType=Art)] expression […  The expression of Sam68 was
not altered in agi…  The [det] expression [nsubjpass] of [prep] Sam…  The
expression of Sam68 was not altered in agi…
144  3900SQZVJN7FJBWJ87I5UJVDB007RB  biomed  At skeletal maturity, bone mass is
maintained …  bone    0.107143  At skeletal maturity, bone mass is maintained
…                 False  [ADP, ADJ, NOUN, PUNCT, NOUN, NOUN, AUX, VERB,…
[prep, amod, pobj, punct, compound, nsubjpass,…  [(), (Degree=Pos),
(Number=Sing), (PunctType=C…         1.190476                  0
0  At skeletal maturity, bone mass is maintained …  At [ADP] skeletal [ADJ]
maturity, [NOUN] bone …  At skeletal maturity, bone mass is maintained …  At
[()] skeletal [(Degree=Pos)] maturity, [(Nu…  At skeletal maturity, bone mass
is maintained …  At [prep] skeletal [amod] maturity, [pobj] bon…  At
skeletal maturity, bone mass is maintained …
145  3SR6AEG6W5TL91EHZBWBTSD4SQHHYV  biomed  However, this reduction in bone
resorption occ…  bone    0.156250  However, this reduction in bone resorption
occ…                 False  [ADV, PUNCT, DET, NOUN, ADP, NOUN, NOUN, VERB,…
[advmod, punct, det, nsubj, prep, compound, po…  [(), (PunctType=Comm),
(Number=Sing, PronType=…         1.000000                  0
0  However, this reduction in bone resorption occ…  However, [ADV] this
[PUNCT] reduction [DET] in…  However, this reduction in bone resorption occ…
However, [()] this [(PunctType=Comm)] reductio…  However, this reduction in
bone resorption occ…  However, [advmod] this [punct] reduction [det]…
However, this reduction in bone resorption occ…
146  3MIVREZQVHY32PO3EMIETYQULYYKQT  biomed  In contrast, our analysis of Bmpr1a
mutant art…  bone    0.218750  In contrast, our analysis of Bmpr1a mutant
art…                 False  [ADP, NOUN, PUNCT, PRON, NOUN, ADP, NOUN, ADJ,…
[prep, pobj, punct, poss, nsubj, prep, pobj, a…  [(), (Number=Sing),
(PunctType=Comm), (Number=…         1.000000                  0
0  In contrast, our analysis of Bmpr1a mutant art…  In [ADP] contrast, [NOUN]
our [PUNCT] analysis…  In contrast, our analysis of Bmpr1a mutant art…  In
[()] contrast, [(Number=Sing)] our [(PunctT…  In contrast, our analysis of
Bmpr1a mutant art…  In [prep] contrast, [pobj] our [punct] analysi…  In
contrast, our analysis of Bmpr1a mutant art…
147  3G9UA71JVVUYLND6029WSS9M1JBJ70  biomed  The skeletal phenotyping of cohorts
of Sam68+/…  bone    0.228261  The skeletal phenotyping of cohorts of
```

```
Sam68+/…                False  [DET, ADJ, NOUN, ADP, NOUN, ADP, PROPN,
CCONJ,…  [det, amod, nsubj, prep, pobj, prep, pobj, cc,…  [(Definite=Def,
PronType=Art), (Degree=Pos), (…         1.190476                 0
0  The skeletal phenotyping of cohorts of Sam68+/…  The [DET] skeletal [ADJ]
phenotyping [NOUN] of…  The skeletal phenotyping of cohorts of Sam68+/…  The
[(Definite=Def|PronType=Art)] skeletal [(D…  The skeletal phenotyping of
cohorts of Sam68+/…  The [det] skeletal [amod] phenotyping [nsubj] …  The
skeletal phenotyping of cohorts of Sam68+/…
                                id  corpus
sentence              token  complexity
sentence_no_contractions  contraction_expanded
pos_sequence                                       dep_sequence
morph_sequence  morph_complexity  binary_complexity
binary_complexity_75th_split                                          snc_pos_seq
snc_pos_alt                                      snc_morph_seq
snc_morph_alt                                        snc_dep_seq
snc_dep_alt                        snc_morph_complexity_value
29  3NSM4HLQNRUPDSMYRR2BPK23K5OQQR  biomed  While it is possible that dox acts
in some oth…  multiple studies    0.083333  While it is possible that dox
acts in some oth…                False  [SCONJ, PRON, AUX, ADJ, SCONJ, NOUN,
VERB, ADP…  [mark, nsubj, advcl, acomp, mark, nsubj, ccomp…  [(), (Case=Nom,
Gender=Neut, Number=Sing, Pers…         1.339286                 0
0  While it is possible that dox acts in some oth…  While [SCONJ] it [PRON] is
[AUX] possible [ADJ…  While it is possible that dox acts in some oth…  While
[()] it [(Case=Nom|Gender=Neut|Number=Si…  While it is possible that dox acts
in some oth…  While [mark] it [nsubj] is [advcl] possible [a…  While it is
possible that dox acts in some oth…
30  3GL25Y6843UI1APILCQM2JER77EMXX  biomed  Detailed reports on appearance and
distributio…  brain development    0.125000  Detailed reports on appearance
and distributio…                False  [ADJ, NOUN, ADP, NOUN, CCONJ, NOUN,
ADP, PROPN…  [amod, nsubj, prep, pobj, cc, conj, prep, comp…  [(Degree=Pos),
(Number=Plur), (), (Number=Sing…         0.937500                 0
0  Detailed reports on appearance and distributio…  Detailed [ADJ] reports
[NOUN] on [ADP] appeara…  Detailed reports on appearance and distributio…
Detailed [(Degree=Pos)] reports [(Number=Plur)…  Detailed reports on
appearance and distributio…  Detailed [amod] reports [nsubj] on [prep] appe…
Detailed reports on appearance and distributio…
31  33CLA8O0MIBSY4BPQQGHIB8U9PHRFZ  biomed  The discovery of multiple and
diverse roles fo…  brain development    0.279412  The discovery of multiple
and diverse roles fo…                False  [DET, NOUN, ADP, ADJ, CCONJ, ADJ,
NOUN, ADP, P…  [det, nsubj, prep, amod, cc, conj, pobj, prep,…
[(Definite=Def, PronType=Art), (Number=Sing), …         1.068966
0                             0  The discovery of multiple and diverse roles
fo…  The [DET] discovery [NOUN] of [ADP] multiple [… The discovery of
multiple and diverse roles fo…  The [(Definite=Def|PronType=Art)] discovery
[(…  The discovery of multiple and diverse roles fo…  The [det] discovery
[nsubj] of [prep] multiple…  The discovery of multiple and diverse roles fo…
32  3HEM8MA6H9C4DGLJRENMPFCTF92QP0  biomed  In the development of the mammalian
```

```
retina, a …      diverse range    0.194444  In the development of the
mammalian retina, a …                False  [ADP, DET, NOUN, ADP, DET, ADJ,
NOUN, PUNCT, D…  [prep, det, pobj, prep, det, amod, pobj, punct…  [(),
(Definite=Def, PronType=Art), (Number=Sin…          1.200000
0                             0  In the development of the mammalian retina, a
…  In [ADP] the [DET] development [NOUN] of [ADP]…  In the development of
the mammalian retina, a …  In [()] the [(Definite=Def|PronType=Art)] deve…
In the development of the mammalian retina, a …  In [prep] the [det]
development [pobj] of [pre…  In the development of the mammalian retina, a …
33  3M67TQBQQHORYDYVLTU3DPX96NY9AH  biomed  A total of 200 female mice (ten
months of age)…        female mice    0.234375  A total of 200 female mice
(ten months of age)…                 False  [DET, NOUN, ADP, NUM, ADJ, NOUN,
PUNCT, NUM, N…  [det, ROOT, prep, nummod, amod, pobj, punct, n…
[(Definite=Ind, PronType=Art), (Number=Sing), …          0.920000
0                             0  A total of 200 female mice (ten months of
age)…  A [DET] total [NOUN] of [ADP] 200 [NUM] female…  A total of 200
female mice (ten months of age)…  A [(Definite=Ind|PronType=Art)] total
[(Number…  A total of 200 female mice (ten months of age)…  A [det] total
[ROOT] of [prep] 200 [nummod] fe…  A total of 200 female mice (ten months of
age)…
                                 id  corpus
sentence token  complexity                          sentence_no_contractions
contraction_expanded                                      pos_sequence
dep_sequence                                    morph_sequence
morph_complexity  binary_complexity  binary_complexity_75th_split
snc_pos_seq                                       snc_pos_alt
snc_morph_seq                                     snc_morph_alt
snc_dep_seq                                       snc_dep_alt
snc_morph_complexity_value
283  31KSVEGZ34SU9QXKGFQHMZUU4REWR7  biomed  The role of CAF-1 in the nuclear
organization …  role    0.000000  The role of CAF-1 in the nuclear
organization …                False  [DET, NOUN, ADP, PROPN, ADP, DET, ADJ,
NOUN, A…  [det, nsubj, prep, pobj, prep, det, amod, pobj…  [(Definite=Def,
PronType=Art), (Number=Sing), …          1.045455                  0
0  The role of CAF-1 in the nuclear organization …  The [DET] role [NOUN] of
[ADP] CAF-1 [PROPN] i…  The role of CAF-1 in the nuclear organization …  The
[(Definite=Def|PronType=Art)] role [(Numbe…  The role of CAF-1 in the nuclear
organization …  The [det] role [nsubj] of [prep] CAF-1 [pobj] …  The role of
CAF-1 in the nuclear organization …
284  3K1H3NEY7LZ4BUOFJ9RFV7R2V2XGDM  biomed  These studies might clarify whether
ADAM11 pla…  role    0.203125  These studies might clarify whether ADAM11
pla…                False  [DET, NOUN, AUX, VERB, SCONJ, PROPN, VERB, DET…
[det, nsubj, aux, ROOT, mark, nsubj, ccomp, de…  [(Number=Plur, PronType=Dem),
(Number=Plur), (…          1.360000                  0
0  These studies might clarify whether ADAM11 pla…  These [DET] studies [NOUN]
might [AUX] clarify…  These studies might clarify whether ADAM11 pla…  These
[(Number=Plur|PronType=Dem)] studies [(N…  These studies might clarify whether
ADAM11 pla…  These [det] studies [nsubj] might [aux] clarif…  These studies
```

```
might clarify whether ADAM11 pla…
285  34OWYT6U3WH64VHTXHMGUNLSJUB9IR  biomed  These findings led us to
hypothesize that ADAM…  role    0.205882  These findings led us to hypothesize
that ADAM…                False  [DET, NOUN, VERB, PRON, PART, VERB, SCONJ,
PRO…  [det, nsubj, ROOT, dobj, aux, xcomp, mark, nsu…  [(Number=Plur,
PronType=Dem), (Number=Plur), (…         1.576923                 0
0  These findings led us to hypothesize that ADAM…  These [DET] findings
[NOUN] led [VERB] us [PRO…  These findings led us to hypothesize that ADAM…
These [(Number=Plur|PronType=Dem)] findings [(…  These findings led us to
hypothesize that ADAM…  These [det] findings [nsubj] led [ROOT] us [do…
These findings led us to hypothesize that ADAM…
286  37SQU136V7ODFKIOLXMHNIMN4G711D  biomed  An important role for annexins in
mediating th…  role    0.233333  An important role for annexins in mediating
th…                False  [DET, ADJ, NOUN, ADP, NOUN, ADP, VERB, DET, NO…
[det, amod, nsubjpass, prep, pobj, prep, pcomp…  [(Definite=Ind,
PronType=Art), (Degree=Pos), (…         1.400000                 0
0  An important role for annexins in mediating th…  An [DET] important [ADJ]
role [NOUN] for [ADP]…  An important role for annexins in mediating th…  An
[(Definite=Ind|PronType=Art)] important [(D…  An important role for annexins
in mediating th…  An [det] important [amod] role [nsubjpass] for…  An
important role for annexins in mediating th…
287  30EMX9PEVKJFF53G6Q7JOY5V3MJKSI  biomed  Positional cloning was used to
identify the mo…  role    0.234375  Positional cloning was used to identify
the mo…                False  [ADJ, NOUN, AUX, VERB, PART, VERB, DET,
PROPN,…  [amod, nsubjpass, auxpass, ROOT, aux, xcomp, d…  [(Degree=Pos),
(Number=Sing), (Mood=Ind, Numbe…         1.307692                 0
0  Positional cloning was used to identify the mo…  Positional [ADJ] cloning
[NOUN] was [AUX] used…  Positional cloning was used to identify the mo…
Positional [(Degree=Pos)] cloning [(Number=Sin…  Positional cloning was used
to identify the mo…  Positional [amod] cloning [nsubjpass] was [aux…
Positional cloning was used to identify the mo…
                                id  corpus
sentence                token  complexity
sentence_no_contractions  contraction_expanded
pos_sequence                                      dep_sequence
morph_sequence  morph_complexity  binary_complexity
binary_complexity_75th_split                                        snc_pos_seq
snc_pos_alt                                      snc_morph_seq
snc_morph_alt                                        snc_dep_seq
snc_dep_alt                          snc_morph_complexity_value
66  3HXCEECSQMT7OMEB5X2ITZH9OHQYZW  biomed  The work presented here has
clarified two impo…  important questions    0.000000  The work presented here
has clarified two impo…                False  [DET, NOUN, VERB, ADV, AUX,
VERB, NUM, ADJ, NO…  [det, nsubj, acl, advmod, aux, ROOT, nummod, a…
[(Definite=Def, PronType=Art), (Number=Sing), …          1.423077
0                             0  The work presented here has clarified two
impo…  The [DET] work [NOUN] presented [VERB] here [A…  The work presented
here has clarified two impo…  The [(Definite=Def|PronType=Art)] work
```

```
[(Numbe…  The work presented here has clarified two impo…  The [det] work
[nsubj] presented [acl] here [a…  The work presented here has clarified two
impo…
67  306W7JMRYYYW3IKDMFOL84M44Z1B8P  biomed  These findings are in complete
agreement with …   complete agreement    0.100000  These findings are in
complete agreement with …                  False  [DET, NOUN, AUX, ADP, ADJ,
NOUN, ADP, ADJ, NOU…  [det, nsubj, ROOT, prep, amod, pobj, prep, amo…
[(Number=Plur, PronType=Dem), (Number=Plur), (…          1.000000
0                              0  These findings are in complete agreement with
…  These [DET] findings [NOUN] are [AUX] in [ADP]…  These findings are in
complete agreement with …  These [(Number=Plur|PronType=Dem)] findings [(…
These findings are in complete agreement with …  These [det] findings [nsubj]
are [ROOT] in [pr…  These findings are in complete agreement with …
68  3CMIQF80GNQW3A3ECIODJFLCK4CQ6U  biomed  Recent human genetic studies have
also demonst…           many cases    0.132353  Recent human genetic studies
have also demonst…                 False  [ADJ, ADJ, ADJ, NOUN, AUX, ADV,
VERB, NOUN, AD…  [amod, amod, amod, nsubj, aux, advmod, ROOT, d…
[(Degree=Pos), (Degree=Pos), (Degree=Pos), (Nu…          1.035714
0                              0  Recent human genetic studies have also
demonst…  Recent [ADJ] human [ADJ] genetic [ADJ] studies…  Recent human
genetic studies have also demonst…  Recent [(Degree=Pos)] human [(Degree=Pos)]
gen…  Recent human genetic studies have also demonst…  Recent [amod] human
[amod] genetic [amod] stud…  Recent human genetic studies have also demonst…
69  3P7RGTLO6EDBF9HMPQLS3YBPHHAAKC  biomed  This technology should provide new
possibiliti…    new possibilities    0.160714  This technology should provide
new possibiliti…                 False  [DET, NOUN, AUX, VERB, ADJ, NOUN, ADP,
VERB, D…  [det, nsubj, aux, ROOT, amod, dobj, prep, pcom…  [(Number=Sing,
PronType=Dem), (Number=Sing), (…          1.250000                0
0  This technology should provide new possibiliti…  This [DET] technology
[NOUN] should [AUX] prov…  This technology should provide new possibiliti…
This [(Number=Sing|PronType=Dem)] technology […  This technology should
provide new possibiliti…  This [det] technology [nsubj] should [aux] pro…
This technology should provide new possibiliti…
70  3XJOUITW8UR258EQ8VW6UPDQ4CNQT5  biomed  Detailed genetic studies in
Drosophila and Cae…      genetic studies    0.194444  Detailed genetic studies
in Drosophila and Cae…                 False  [ADJ, ADJ, NOUN, ADP, PROPN,
CCONJ, PROPN, NOU…  [amod, amod, nsubj, prep, nmod, cc, conj, pobj…
[(Degree=Pos), (Degree=Pos), (Number=Plur), ()…          1.057143
0                              0  Detailed genetic studies in Drosophila and
Cae…  Detailed [ADJ] genetic [ADJ] studies [NOUN] in…  Detailed genetic
studies in Drosophila and Cae…  Detailed [(Degree=Pos)] genetic
[(Degree=Pos)]…  Detailed genetic studies in Drosophila and Cae…  Detailed
[amod] genetic [amod] studies [nsubj]…  Detailed genetic studies in Drosophila
and Cae…
```

[ ]:

```
dataframes = [train_single_df, train_multi_df, trial_val_single_df,␣
 ↪trial_val_multi_df, test_single_df, test_multi_df]
```

```python
for df in dataframes:
    if hasattr(df, 'columns') and 'corpus' in df.columns:
        print(df[df['corpus'] == 'europarl'].head())
    else:
        pass
```

```
                                  id    corpus
sentence    token  complexity                           sentence_no_contractions
contraction_expanded                                       pos_sequence
dep_sequence                                     morph_sequence
morph_complexity  binary_complexity  binary_complexity_75th_split
snc_pos_seq                                        snc_pos_alt
snc_morph_seq                                      snc_morph_alt
snc_dep_seq                                        snc_dep_alt
snc_morph_complexity_value
5150  3Y40HMYLL1I1EIURUEH8TTVLKTKUX0  europarl  Despite the fact that the Treaty
does not requ…     fact    0.156250  Despite the fact that the Treaty does not
requ…                 False  [SCONJ, DET, NOUN, SCONJ, DET, PROPN, AUX, PAR…
[prep, det, pobj, mark, det, nsubj, aux, neg, …  [(), (Definite=Def,
PronType=Art), (Number=Sin…          1.666667                  0
0  Despite the fact that the Treaty does not requ…  Despite [SCONJ] the [DET]
fact [NOUN] that [SC…  Despite the fact that the Treaty does not requ…
Despite [()] the [(Definite=Def|PronType=Art)]…  Despite the fact that the
Treaty does not requ…  Despite [prep] the [det] fact [pobj] that [mar…
Despite the fact that the Treaty does not requ…
5151  3OZ4VAIBEXF0WDE2I0CCY6PPN3VVJL  europarl  The average consumption in the
EU fluctuates b…     fact    0.236842  The average consumption in the EU
fluctuates b…                 False  [DET, ADJ, NOUN, ADP, DET, PROPN, VERB,
ADP, N…  [det, amod, nsubj, prep, det, pobj, ROOT, quan…  [(Definite=Def,
PronType=Art), (Degree=Pos), (…          0.937500                  0
0  The average consumption in the EU fluctuates b…  The [DET] average [ADJ]
consumption [NOUN] in …  The average consumption in the EU fluctuates b…
The [(Definite=Def|PronType=Art)] average [(De…  The average consumption in
the EU fluctuates b…  The [det] average [amod] consumption [nsubj] i…  The
average consumption in the EU fluctuates b…
5152  3NFWQRSHVEE19E2BAFM5J7UN7HQFGD  europarl  The main Charlemagne Prize was
presented on 13…     days    0.111111  The main Charlemagne Prize was
presented on 13…                 False  [DET, ADJ, PROPN, PROPN, AUX, VERB,
ADP, NUM, …  [det, amod, compound, nsubjpass, auxpass, ROOT…
[(Definite=Def, PronType=Art), (Degree=Pos), (…          1.100000
0                             0  The main Charlemagne Prize was presented on
13…  The [DET] main [ADJ] Charlemagne [PROPN] Prize…  The main Charlemagne
Prize was presented on 13…  The [(Definite=Def|PronType=Art)] main [(Degre…
The main Charlemagne Prize was presented on 13…  The [det] main [amod]
Charlemagne [compound] P…  The main Charlemagne Prize was presented on 13…
5153  3TZ0XG8CBUKDFP5G0VAPHYREGZ298H  europarl  Commissioner, ladies and
gentlemen, we have al…     days    0.116667  Commissioner, ladies and
```

```
gentlemen, we have al…                 False  [PROPN, PUNCT, NOUN, CCONJ,
NOUN, PUNCT, PRON,…  [npadvmod, punct, conj, cc, conj, punct, nsubj…
[(Number=Sing), (PunctType=Comm), (Number=Plur…          1.258065
0                             0  Commissioner, ladies and gentlemen, we have
al…  Commissioner, [PROPN] ladies [PUNCT] and [NOUN…  Commissioner, ladies
and gentlemen, we have al…  Commissioner, [(Number=Sing)] ladies [(PunctTy…
Commissioner, ladies and gentlemen, we have al…  Commissioner, [npadvmod]
ladies [punct] and [c…  Commissioner, ladies and gentlemen, we have al…
5154  3M7OI89LVYOS99TV7ONIZAWVGPFC6F  europarl  (For the outcome and other
details of the vote…  details    0.075000  (For the outcome and other details
of the vote…                 False  [PUNCT, ADP, DET, NOUN, CCONJ, ADJ, NOUN,
ADP,…  [punct, prep, det, pobj, cc, amod, conj, prep,…  [(PunctSide=Ini,
PunctType=Brck), (), (Definit…         1.071429                 0
0  (For the outcome and other details of the vote…  (For [PUNCT] the [ADP]
outcome [DET] and [NOUN…  (For the outcome and other details of the vote…
(For [(PunctSide=Ini|PunctType=Brck)] the [()]…  (For the outcome and other
details of the vote…  (For [punct] the [prep] outcome [det] and [pob…  (For
the outcome and other details of the vote…
                                  id    corpus
sentence             token  complexity
sentence_no_contractions  contraction_expanded
pos_sequence                                      dep_sequence
morph_sequence  morph_complexity  binary_complexity
binary_complexity_75th_split                                        snc_pos_seq
snc_pos_alt                                      snc_morph_seq
snc_morph_alt                                        snc_dep_seq
snc_dep_alt                         snc_morph_complexity_value
1019  37M4O367VJI9ZR58F67RA0N7E9RM5C  europarl  We do not know how many people
are affected, b…       many people    0.222222  We do not know how many people
are affected, b…                 False  [PRON, AUX, PART, VERB, SCONJ, ADJ,
NOUN, AUX,…  [nsubj, aux, neg, ROOT, advmod, amod, nsubjpas…  [(Case=Nom,
Number=Plur, Person=1, PronType=Pr…          1.480000                  0
0  We do not know how many people are affected, b…  We [PRON] do [AUX] not
[PART] know [VERB] how …  We do not know how many people are affected, b…
We [(Case=Nom|Number=Plur|Person=1|PronType=Pr…  We do not know how many
people are affected, b…  We [nsubj] do [aux] not [neg] know [ROOT] how …  We
do not know how many people are affected, b…
1020  3W1K7D6QSBHBNEL0V5OYL0J839VBZJ  europarl  The issue we were discussing
comes within this…       major issue    0.117647  The issue we were discussing
comes within this…                 False  [DET, NOUN, PRON, AUX, VERB, VERB,
ADP, DET, A…  [det, nsubj, nsubj, aux, relcl, ccomp, prep, d…
[(Definite=Def, PronType=Art), (Number=Sing), …          1.621622
0                             0  The issue we were discussing comes within
this…  The [DET] issue [NOUN] we [PRON] were [AUX] di…  The issue we were
discussing comes within this…  The [(Definite=Def|PronType=Art)] issue
[(Numb…  The issue we were discussing comes within this…  The [det] issue
[nsubj] we [nsubj] were [aux] …  The issue we were discussing comes within
this…
```

```
1021  37SQU136V7ODFKI0LXMHNIMN4IS112  europarl  A renewed EU tourism policy:
towards a stronge…  European tourism    0.142857  A renewed EU tourism policy:
towards a stronge…                 False  [DET, VERB, PROPN, NOUN, NOUN,
PUNCT, ADP, DET…  [det, amod, compound, compound, ROOT, punct, p…
[(Definite=Ind, PronType=Art), (Aspect=Perf, T…          1.187500
0                             0  A renewed EU tourism policy: towards a
stronge…  A [DET] renewed [VERB] EU [PROPN] tourism [NOU…  A renewed EU
tourism policy: towards a stronge…  A [(Definite=Ind|PronType=Art)] renewed
[(Aspe…  A renewed EU tourism policy: towards a stronge…  A [det] renewed
[amod] EU [compound] tourism […  A renewed EU tourism policy: towards a
stronge…
1022  3XBYQ44Z6P47P5ACK4VCMEVCSERTW1  europarl  In fact, I can tell you that
there was an exce…   other occasions    0.156250  In fact, I can tell you that
there was an exce…                 False  [ADP, NOUN, PUNCT, PRON, AUX, VERB,
PRON, SCON…  [prep, pobj, punct, nsubj, aux, ROOT, dobj, ma…  [(),
(Number=Sing), (PunctType=Comm), (Case=No…          1.222222
0                             0  In fact, I can tell you that there was an
exce…  In [ADP] fact, [NOUN] I [PUNCT] can [PRON] tel…  In fact, I can tell
you that there was an exce…  In [()] fact, [(Number=Sing)] I [(PunctType=Co…
In fact, I can tell you that there was an exce…  In [prep] fact, [pobj] I
[punct] can [nsubj] t…  In fact, I can tell you that there was an exce…
1023  3MZ3TAMYTLNC8VDFRYM2L8LMPIWIR6  europarl  He did not pursue the pressing
imperative of r…    land ownership    0.160714  He did not pursue the pressing
imperative of r…                 False  [PRON, AUX, PART, VERB, DET, VERB,
NOUN, ADP, …  [nsubj, aux, neg, ROOT, det, amod, dobj, prep,…  [(Case=Nom,
Gender=Masc, Number=Sing, Person=3…          1.750000                  0
0  He did not pursue the pressing imperative of r…  He [PRON] did [AUX] not
[PART] pursue [VERB] t…  He did not pursue the pressing imperative of r…  He
[(Case=Nom|Gender=Masc|Number=Sing|Person=3…  He did not pursue the pressing
imperative of r…  He [nsubj] did [aux] not [neg] pursue [ROOT] t…  He did
not pursue the pressing imperative of r…
                                 id    corpus
sentence   token  complexity                           sentence_no_contractions
contraction_expanded                                       pos_sequence
dep_sequence                                    morph_sequence
morph_complexity  binary_complexity  binary_complexity_75th_split
snc_pos_seq                                         snc_pos_alt
snc_morph_seq                                       snc_morph_alt
snc_dep_seq                                         snc_dep_alt
snc_morph_complexity_value
278  3H6W48L9F4P9XDH53NMSH4UF3B5WPY  europarl  It is estimated that a staggering
10 000 conta…     sea    0.220588  It is estimated that a staggering 10 000
conta…                 False  [PRON, AUX, VERB, SCONJ, DET, ADJ, NUM, NUM,
N…  [nsubjpass, auxpass, ROOT, mark, det, amod, co…  [(Gender=Neut,
Number=Sing, Person=3, PronType…          1.687500                  0
0  It is estimated that a staggering 10 000 conta…  It [PRON] is [AUX]
estimated [VERB] that [SCON…  It is estimated that a staggering 10 000
conta…  It [(Gender=Neut|Number=Sing|Person=3|PronType…  It is estimated
that a staggering 10 000 conta…  It [nsubjpass] is [auxpass] estimated [ROOT]
t…  It is estimated that a staggering 10 000 conta…
279  32W3UF2EZOLEUMPHOCU32CCHKY9C4U  europarl  I would remind you that the
election of the Pr…   Rules    0.050000  I would remind you that the election
of the Pr…                 False  [PRON, AUX, VERB, PRON, SCONJ, DET, NOUN,
ADP,…  [nsubj, aux, ROOT, dobj, mark, det, nsubj, pre…  [(Case=Nom,
Number=Sing, Person=1, PronType=Pr…          1.257576                  0
0  I would remind you that the election of the Pr…  I [PRON] would [AUX]
remind [VERB] you [PRON] …  I would remind you that the election of the Pr…
I [(Case=Nom|Number=Sing|Person=1|PronType=Prs…  I would remind you that the
election of the Pr…  I [nsubj] would [aux] remind [ROOT] you [dobj]…  I
would remind you that the election of the Pr…
280  3P0I4CQYVY7RCD54ON9DS4PPT5QOWO  europarl  We have simply confirmed, in
accordance with o…   Rules    0.178571  We have simply confirmed, in
accordance with o…                 False  [PRON, AUX, ADV, VERB, PUNCT, ADP,
NOUN, ADP, …  [nsubj, aux, advmod, ROOT, punct, prep, pobj, …  [(Case=Nom,
Number=Plur, Person=1, PronType=Pr…          1.187500                  0
0  We have simply confirmed, in accordance with o…  We [PRON] have [AUX]
simply [ADV] confirmed, [… We have simply confirmed, in accordance with o…
We [(Case=Nom|Number=Plur|Person=1|PronType=Pr…  We have simply confirmed, in
accordance with o…  We [nsubj] have [aux] simply [advmod] confirme…  We have
simply confirmed, in accordance with o…
281  3PZDSVZ3J5HXLQM8D23HIN6TJ2N4N4  europarl  What further measures is the
Commission now ta…  prices    0.066667  What further measures is the
Commission now ta…                 False  [PRON, ADJ, NOUN, AUX, DET, PROPN,
ADV, VERB, …  [det, amod, nsubj, ROOT, det, nsubj, advmod, c…  [(),
(Degree=Pos), (Number=Plur), (Mood=Ind, N…          1.142857
0                             0  What further measures is the Commission now
ta…  What [PRON] further [ADJ] measures [NOUN] is [… What further measures
is the Commission now ta…  What [()] further [(Degree=Pos)] measures [(Nu…
What further measures is the Commission now ta…  What [det] further [amod]
measures [nsubj] is …  What further measures is the Commission now ta…
282  3GITHABACYLNIC7L9OKTP89VZOR2N6  europarl  Many economic operators are in an
even more se…  prices    0.115385  Many economic operators are in an even more
se…                 False  [ADJ, ADJ, NOUN, AUX, ADP, DET, ADV, ADV, ADJ,…
[amod, amod, nsubj, ROOT, prep, det, advmod, a…  [(Degree=Pos), (Degree=Pos),
(Number=Plur), (M…          1.027778                  0
0  Many economic operators are in an even more se…  Many [ADJ] economic [ADJ]
operators [NOUN] are…  Many economic operators are in an even more se…  Many
[(Degree=Pos)] economic [(Degree=Pos)] op…  Many economic operators are in an
even more se…  Many [amod] economic [amod] operators [nsubj] …  Many
economic operators are in an even more se…
                                id    corpus
sentence              token  complexity
sentence_no_contractions  contraction_expanded
pos_sequence                                       dep_sequence
morph_sequence  morph_complexity  binary_complexity
binary_complexity_75th_split                                        snc_pos_seq
```

```
snc_pos_alt                                       snc_morph_seq
snc_morph_alt                                          snc_dep_seq
snc_dep_alt                            snc_morph_complexity_value
62  3BA7SXOG1JQJJP12ICAB8JR8MMRR87  europarl  by Mr Virrankoski, on behalf of
the Committee …    management tool    0.176471  by Mr Virrankoski, on behalf
of the Committee …                 False  [ADP, PROPN, PROPN, PUNCT, ADP,
NOUN, ADP, DET…  [prep, compound, pobj, punct, prep, pobj, prep…  [(),
(Number=Sing), (Number=Sing), (PunctType=…          0.892857
0                             0  by Mr Virrankoski, on behalf of the Committee
…  by [ADP] Mr [PROPN] Virrankoski, [PROPN] on [P…  by Mr Virrankoski, on
behalf of the Committee …  by [()] Mr [(Number=Sing)] Virrankoski, [(Numb…
by Mr Virrankoski, on behalf of the Committee …  by [prep] Mr [compound]
Virrankoski, [pobj] on…  by Mr Virrankoski, on behalf of the Committee …
63  3Z8UJEJOCZDRESZACEFTQHJ3OET93A  europarl  'Considers it appropriate,
therefore, to explo…  debt cancellation    0.250000  'Considers it
appropriate, therefore, to explo…                False  [PUNCT, VERB, PRON,
ADJ, PUNCT, ADV, PUNCT, PA…  [punct, ccomp, nsubj, ccomp, punct, advmod, pu…
[(PunctSide=Ini, PunctType=Quot), (Number=Sing…          1.150000
0                             0  'Considers it appropriate, therefore, to
explo…  'Considers [PUNCT] it [VERB] appropriate, [PRO…  'Considers it
appropriate, therefore, to explo…  'Considers [(PunctSide=Ini|PunctType=Quot)]
it…  'Considers it appropriate, therefore, to explo…  'Considers [punct] it
[ccomp] appropriate, [ns…  'Considers it appropriate, therefore, to explo…
64  31ANT7FQN82N7D4XO9REIVFBXNSH5Y  europarl  Mobilisation of the European
Globalisation Adj…   textile industry    0.250000  Mobilisation of the
European Globalisation Adj…                 False  [NOUN, ADP, DET, PROPN,
PROPN, PROPN, PROPN, P…  [ROOT, prep, det, compound, compound, compound…
[(Number=Sing), (), (Definite=Def, PronType=Ar…          0.941176
0                             0  Mobilisation of the European Globalisation
Adj…  Mobilisation [NOUN] of [ADP] the [DET] Europea…  Mobilisation of the
European Globalisation Adj…  Mobilisation [(Number=Sing)] of [()] the [(Def…
Mobilisation of the European Globalisation Adj…  Mobilisation [ROOT] of [prep]
the [det] Europe…  Mobilisation of the European Globalisation Adj…
65  3D06DR5225J65XHPA2Y8IB3T6NSMAI  europarl  At the time, we sent messages to
the President…  Russian elections    0.264706  At the time, we sent messages
to the President…                False  [ADP, DET, NOUN, PUNCT, PRON, VERB,
NOUN, ADP,…  [prep, det, pobj, punct, nsubj, ROOT, dobj, da…  [(),
(Definite=Def, PronType=Art), (Number=Sin…          1.290323
0                             0  At the time, we sent messages to the
President…  At [ADP] the [DET] time, [NOUN] we [PUNCT] sen…  At the time, we
sent messages to the President…  At [()] the [(Definite=Def|PronType=Art)]
time…  At the time, we sent messages to the President…  At [prep] the [det]
time, [pobj] we [punct] se…  At the time, we sent messages to the President…
66  3FI3OCQHVKJ9Z41PT0RNOQQDY5Y6BN  europarl                         Both are
workable options.   workable options    0.281250                        Both
are workable options.                False                          [PRON, AUX,
ADJ, NOUN, PUNCT]                  [nsubj, ROOT, amod, attr, punct]  [(),
(Mood=Ind, Tense=Pres, VerbForm=Fin), (De…          1.200000
0                             0  Both are workable options. [PRON, AUX, ADJ,
NO…  Both [PRON] are [AUX] workable [ADJ] options. …  Both are workable
options. [(), (Mood=Ind|Tens…  Both [()] are
[(Mood=Ind|Tense=Pres|VerbForm=F…  Both are workable options. [nsubj, ROOT,
amod,…  Both [nsubj] are [ROOT] workable [amod] option…
Both are workable options. 1.2
                                 id    corpus
sentence    token  complexity                           sentence_no_contractions
contraction_expanded                                       pos_sequence
dep_sequence                                     morph_sequence
morph_complexity  binary_complexity  binary_complexity_75th_split
snc_pos_seq                                        snc_pos_alt
snc_morph_seq                                      snc_morph_alt
snc_dep_seq                                        snc_dep_alt
snc_morph_complexity_value
572  3X2LT8FDHWIORLIOH6KHVIZPE138WO  europarl  Europe, on the other hand,
unfortunately too o…     hand     0.15625  Europe, on the other hand,
unfortunately too o…                 False  [PROPN, PUNCT, ADP, DET, ADJ,
NOUN, PUNCT, ADV…  [nsubj, punct, prep, det, amod, pobj, punct, a…
[(Number=Sing), (PunctType=Comm), (), (Definit…          1.250000
0                             0  Europe, on the other hand, unfortunately too
o…  Europe, [PROPN] on [PUNCT] the [ADP] other [DE…  Europe, on the other
hand, unfortunately too o…  Europe, [(Number=Sing)] on [(PunctType=Comm)] …
Europe, on the other hand, unfortunately too o…  Europe, [nsubj] on [punct]
the [prep] other [d…  Europe, on the other hand, unfortunately too o…
573  3QX22DUVOOHQXLKNLXP4EYH6RZBVME  europarl  That is why we want to introduce
the role of m…     role     0.05000  That is why we want to introduce the role
of m…                 False  [PRON, AUX, SCONJ, PRON, VERB, PART, VERB, DET…
[nsubj, ROOT, advmod, nsubj, advcl, aux, xcomp…  [(Number=Sing, PronType=Dem),
(Mood=Ind, Numbe…          1.500000                  0
0  That is why we want to introduce the role of m…  That [PRON] is [AUX] why
[SCONJ] we [PRON] wan…  That is why we want to introduce the role of m…
That [(Number=Sing|PronType=Dem)] is [(Mood=In…  That is why we want to
introduce the role of m…  That [nsubj] is [ROOT] why [advmod] we [nsubj]…
That is why we want to introduce the role of m…
574  3NBFJK3IOHIVFRF49I5V6131ZH1GOI  europarl  The Union also has the aim of
encouraging deve…     size     0.00000  The Union also has the aim of
encouraging deve…                 False  [DET, PROPN, ADV, VERB, DET, NOUN,
ADP, VERB, …  [det, nsubj, advmod, ccomp, det, dobj, prep, p…
[(Definite=Def, PronType=Art), (Number=Sing), …          1.366667
0                             0  The Union also has the aim of encouraging
deve…  The [DET] Union [PROPN] also [ADV] has [VERB] …  The Union also has
the aim of encouraging deve…  The [(Definite=Def|PronType=Art)] Union
[(Numb…  The Union also has the aim of encouraging deve…  The [det] Union
[nsubj] also [advmod] has [cco…  The Union also has the aim of encouraging
deve…
575  3LN50BUKPVBTMJ56Z9FQ8TDZ56KLPH  europarl  We are taking note of your
comment and it will…  comment     0.05000  We are taking note of your comment
```

```
and it will…                 False  [PRON, AUX, VERB, NOUN, ADP, PRON, NOUN,
CCONJ…  [nsubj, aux, ROOT, dobj, prep, poss, pobj, cc,…  [(Case=Nom,
Number=Plur, Person=1, PronType=Pr…         1.857143                  0
0  We are taking note of your comment and it will…  We [PRON] are [AUX] taking
[VERB] note [NOUN] …  We are taking note of your comment and it will…  We
[(Case=Nom|Number=Plur|Person=1|PronType=Pr…  We are taking note of your
comment and it will…  We [nsubj] are [aux] taking [ROOT] note [dobj]…  We
are taking note of your comment and it will…
576  3CZH926SICETRK9VK30YS0CK5AME4P  europarl     We have taken note of your
comment, Mr Helmer.  comment     0.05000     We have taken note of your comment,
Mr Helmer.                 False  [PRON, AUX, VERB, NOUN, ADP, PRON, NOUN,
PUNCT…  [nsubj, aux, ROOT, dobj, prep, poss, pobj, pun…  [(Case=Nom,
Number=Plur, Person=1, PronType=Pr…         1.727273                  0
0  We have taken note of your comment, Mr Helmer…  We [PRON] have [AUX] taken
[VERB] note [NOUN] …  We have taken note of your comment, Mr Helmer…  We
[(Case=Nom|Number=Plur|Person=1|PronType=Pr…  We have taken note of your
comment, Mr Helmer…  We [nsubj] have [aux] taken [ROOT] note [dobj]…  We
have taken note of your comment, Mr Helmer…
                                id    corpus
sentence               token  complexity
sentence_no_contractions  contraction_expanded
pos_sequence                                      dep_sequence
morph_sequence  morph_complexity  binary_complexity
binary_complexity_75th_split                                        snc_pos_seq
snc_pos_alt                                      snc_morph_seq
snc_morph_alt                                        snc_dep_seq
snc_dep_alt                         snc_morph_complexity_value
119  3VGET1QSZ0ZKR7D571SBHI3U3H0W7S  europarl  I have been assured by our
technical services …  technical services    0.214286  I have been assured by
our technical services …                 False  [PRON, AUX, AUX, VERB, ADP,
PRON, ADJ, NOUN, S…  [nsubjpass, aux, auxpass, ROOT, agent, poss, a…
[(Case=Nom, Number=Sing, Person=1, PronType=Pr…         1.384615
0                             0  I have been assured by our technical services
…  I [PRON] have [AUX] been [AUX] assured [VERB] …  I have been assured by
our technical services …  I [(Case=Nom|Number=Sing|Person=1|PronType=Prs…  I
have been assured by our technical services …  I [nsubjpass] have [aux] been
[auxpass] assure…  I have been assured by our technical services …
120  3L7SUC0TTUUA4KJ7I01FT5RGX1GM0R  europarl  You understand the importance of
free peoples …        free peoples    0.234375  You understand the importance
of free peoples …                False  [PRON, VERB, DET, NOUN, ADP, ADJ,
NOUN, CCONJ,…  [nsubj, ROOT, det, dobj, prep, amod, pobj, cc,…  [(Case=Nom,
Person=2, PronType=Prs), (Tense=Pr…         1.294118                  0
0  You understand the importance of free peoples …  You [PRON] understand
[VERB] the [DET] importa…  You understand the importance of free peoples …
You [(Case=Nom|Person=2|PronType=Prs)] underst…  You understand the importance
of free peoples …  You [nsubj] understand [ROOT] the [det] import…  You
understand the importance of free peoples …
121  3JTPR5MTZSCE9355UUUBVNV3P4WK55  europarl  We launched the debate on 24
```

```
January 2007 and …      valuable input    0.234375  We launched the debate on
24 January 2007 and …                  False  [PRON, VERB, DET, NOUN, ADP, NUM,
PROPN, NUM, …  [nsubj, ROOT, det, dobj, prep, nummod, pobj, n…  [(Case=Nom,
Number=Plur, Person=1, PronType=Pr…          1.235294                  0
0  We launched the debate on 24 January 2007 and …  We [PRON] launched [VERB]
the [DET] debate [NO…  We launched the debate on 24 January 2007 and …  We
[(Case=Nom|Number=Plur|Person=1|PronType=Pr…  We launched the debate on 24
January 2007 and …  We [nsubj] launched [ROOT] the [det] debate [d…  We
launched the debate on 24 January 2007 and …
122  3MGHRFQY2LPAY18L13PQNOEN6BTYOS  europarl  In subsequent budgetary policy I
think that Pa…          own choice    0.250000  In subsequent budgetary policy
I think that Pa…                  False  [ADP, ADJ, ADJ, NOUN, PRON, VERB,
SCONJ, PROPN…  [prep, amod, amod, pobj, nsubj, ROOT, mark, ns…  [(),
(Degree=Pos), (Degree=Pos), (Number=Sing)…          1.338983
0                              0  In subsequent budgetary policy I think that
Pa…  In [ADP] subsequent [ADJ] budgetary [ADJ] poli…  In subsequent
budgetary policy I think that Pa…  In [()] subsequent [(Degree=Pos)] budgetary
[(…  In subsequent budgetary policy I think that Pa…  In [prep] subsequent
[amod] budgetary [amod] p…  In subsequent budgetary policy I think that Pa…
123  302Y2UIUCQU6B0YU067KHZMGEYAFKJ  europarl     Council position at first
reading: see Minutes       first reading    0.272727     Council position at
first reading: see Minutes                 False   [NOUN, NOUN, ADP, ADJ, NOUN,
PUNCT, VERB, PROPN]  [compound, nsubj, prep, amod, pobj, punct, ROO…
[(Number=Sing), (Number=Sing), (), (Degree=Pos…          0.750000
0                              0  Council position at first reading: see
Minutes…  Council [NOUN] position [NOUN] at [ADP] first …  Council position
at first reading: see Minutes…  Council [(Number=Sing)] position
[(Number=Sing…  Council position at first reading: see Minutes…  Council
[compound] position [nsubj] at [prep] …  Council position at first reading:
see Minutes…
```

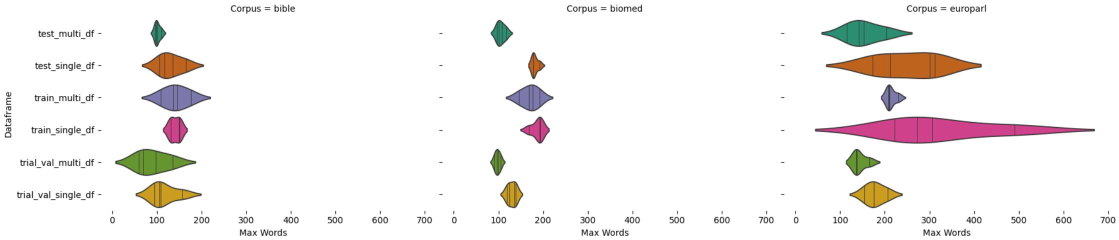
[ ]:

```python
tokenizer = RegexpTokenizer(r'\w+')

def analyze_sentence_spans_by_corpus_and_quartile_no_contracts(dfs_dict):
    results = []

    for df_name, df in dfs_dict.items():
        print(f"Processing {df_name} on 'newly created columns'...")
        df = df.copy()

        q1 = df['complexity'].quantile(0.25)
        q2 = df['complexity'].quantile(0.50)
        q3 = df['complexity'].quantile(0.75)

        def get_quartile(x):
            if x <= q1:
                return 'Q1'
            elif x <= q2:
                return 'Q2'
            elif x <= q3:
                return 'Q3'
            else:
                return 'Q4'

        df['quartile'] = df['complexity'].apply(get_quartile)

        def compute_span_metrics_no_contracts(sentence):
            if pd.isna(sentence):
                return pd.Series({'word_count': 0, 'char_count': 0,␣
↪'avg_word_len': 0})

            words = tokenizer.tokenize(sentence)
            word_count = len(words)
            char_count = len(sentence)
            avg_word_len = np.mean([len(w) for w in words]) if word_count > 0␣
↪else 0

            return pd.Series({
                'word_count': word_count,
                'char_count': char_count,
                'avg_word_len': avg_word_len
            })

        span_metrics_nc = df['snc_pos_seq'].
↪apply(compute_span_metrics_no_contracts)
        df = pd.concat([df, span_metrics_nc], axis=1)

        corpus_col = 'corpus'
        for corpus_name, corpus_df in df.groupby(corpus_col):
            for quartile, quartile_df in corpus_df.groupby('quartile'):
                complexity_range = f"{quartile_df['complexity'].min():.
↪3f}-{quartile_df['complexity'].max():.3f}"
                stats = {
                    'Dataframe': df_name,
                    'Corpus': corpus_name,
                    'Quartile': quartile,
                    'Complexity Range': complexity_range,
                    'Count': len(quartile_df),
                    'Avg Words': quartile_df['word_count'].mean(),
                    'Median Words': quartile_df['word_count'].median(),
                    'Min Words': quartile_df['word_count'].min(),
                    'Max Words': quartile_df['word_count'].max(),
                    'Std Words': quartile_df['word_count'].std(),
                    'Avg Chars': quartile_df['char_count'].mean(),
                    'Avg Word Len': quartile_df['avg_word_len'].mean()
                }
                results.append(stats)

    results_df = pd.DataFrame(results)
    results_df = results_df.sort_values(['Dataframe', 'Corpus', 'Quartile'])
    return results_df


dfs = {
    'train_single_df': train_single_df,
    'train_multi_df': train_multi_df,
    'trial_val_single_df': trial_val_single_df,
    'trial_val_multi_df': trial_val_multi_df,
    'test_single_df': test_single_df,
    'test_multi_df': test_multi_df
}

span_analysis_nc =␣
 ↪analyze_sentence_spans_by_corpus_and_quartile_no_contracts(dfs)

pd.set_option('display.max_rows', None)
pd.set_option('display.max_columns', None)
pd.set_option('display.width', 1000)
# display(span_analysis_nc)

results_path_nc = os.path.join(dir_results,␣
 ↪'sentence_span_analysis_no_contractions.csv')
span_analysis_nc.to_csv(results_path_nc, index=False)
print(f"Analysis (NO CONTRACTIONS) saved to: {results_path_nc}")

g = sns.FacetGrid(span_analysis_nc, col="Corpus", col_wrap=3, height=4,␣
 ↪aspect=1.5)
g.map(sns.violinplot, "Max Words", "Dataframe", inner='stick', palette='Dark2')
g.despine(top=True, right=True, bottom=True, left=True)
plt.tight_layout()
plt.show()
```

```
Processing train_single_df on 'newly created columns'…
Processing train_multi_df on 'newly created columns'…
Processing trial_val_single_df on 'newly created columns'…
Processing trial_val_multi_df on 'newly created columns'…
Processing test_single_df on 'newly created columns'…
Processing test_multi_df on 'newly created columns'…
Analysis (NO CONTRACTIONS) saved to: /content/drive/MyDrive/266-
final/results/sentence_span_analysis_no_contractions.csv
```

```
/usr/local/lib/python3.11/dist-packages/seaborn/axisgrid.py:718: UserWarning:
```

```
Using the violinplot function without specifying `order` is likely to produce an
incorrect plot.
  warnings.warn(warning)
/usr/local/lib/python3.11/dist-packages/seaborn/axisgrid.py:854: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in
v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same
effect.

  func(*plot_args, **plot_kwargs)
/usr/local/lib/python3.11/dist-packages/seaborn/axisgrid.py:854: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in
v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same
effect.

  func(*plot_args, **plot_kwargs)
/usr/local/lib/python3.11/dist-packages/seaborn/axisgrid.py:854: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in
v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same
effect.

  func(*plot_args, **plot_kwargs)
```

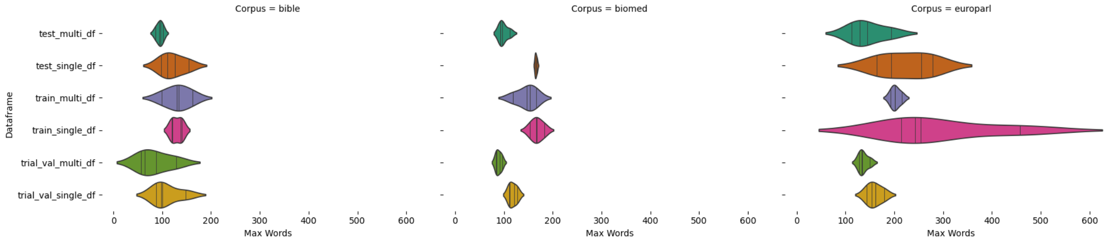


```
[ ]: tokenizer = RegexpTokenizer(r'\w+')

     def analyze_sentence_spans_by_corpus_and_quartile_no_contracts(dfs_dict):
         results = []

         for df_name, df in dfs_dict.items():
             print(f"Processing {df_name} on 'newly created columns'...")
             df = df.copy()

             q1 = df['complexity'].quantile(0.25)
             q2 = df['complexity'].quantile(0.50)
             q3 = df['complexity'].quantile(0.75)
```

```
        def get_quartile(x):
            if x <= q1:
                return 'Q1'
            elif x <= q2:
                return 'Q2'
            elif x <= q3:
                return 'Q3'
            else:
                return 'Q4'

        df['quartile'] = df['complexity'].apply(get_quartile)

        def compute_span_metrics_no_contracts(sentence):
            if pd.isna(sentence):
                return pd.Series({'word_count': 0, 'char_count': 0,␣
↪'avg_word_len': 0})

            words = tokenizer.tokenize(sentence)
            word_count = len(words)
            char_count = len(sentence)
            avg_word_len = np.mean([len(w) for w in words]) if word_count > 0␣
↪else 0

            return pd.Series({
                'word_count': word_count,
                'char_count': char_count,
                'avg_word_len': avg_word_len
            })

        span_metrics_nc = df['snc_pos_alt'].
↪apply(compute_span_metrics_no_contracts)
        df = pd.concat([df, span_metrics_nc], axis=1)

        corpus_col = 'corpus'
        for corpus_name, corpus_df in df.groupby(corpus_col):
            for quartile, quartile_df in corpus_df.groupby('quartile'):
                complexity_range = f"{quartile_df['complexity'].min():.
↪3f}-{quartile_df['complexity'].max():.3f}"
                stats = {
                    'Dataframe': df_name,
                    'Corpus': corpus_name,
                    'Quartile': quartile,
                    'Complexity Range': complexity_range,
                    'Count': len(quartile_df),
                    'Avg Words': quartile_df['word_count'].mean(),
                    'Median Words': quartile_df['word_count'].median(),
                    'Min Words': quartile_df['word_count'].min(),
                    'Max Words': quartile_df['word_count'].max(),
                    'Std Words': quartile_df['word_count'].std(),
                    'Avg Chars': quartile_df['char_count'].mean(),
                    'Avg Word Len': quartile_df['avg_word_len'].mean()
                }
                results.append(stats)

    results_df = pd.DataFrame(results)
    results_df = results_df.sort_values(['Dataframe', 'Corpus', 'Quartile'])
    return results_df


dfs = {
    'train_single_df': train_single_df,
    'train_multi_df': train_multi_df,
    'trial_val_single_df': trial_val_single_df,
    'trial_val_multi_df': trial_val_multi_df,
    'test_single_df': test_single_df,
    'test_multi_df': test_multi_df
}

span_analysis_nc =␣
  ↪analyze_sentence_spans_by_corpus_and_quartile_no_contracts(dfs)

pd.set_option('display.max_rows', None)
pd.set_option('display.max_columns', None)
pd.set_option('display.width', 1000)
# display(span_analysis_nc)

results_path_nc = os.path.join(dir_results,␣
  ↪'sentence_span_analysis_no_contractions.csv')
span_analysis_nc.to_csv(results_path_nc, index=False)
print(f"Analysis (NO CONTRACTIONS) saved to: {results_path_nc}")

g = sns.FacetGrid(span_analysis_nc, col="Corpus", col_wrap=3, height=4,␣
  ↪aspect=1.5)
g.map(sns.violinplot, "Max Words", "Dataframe", inner='stick', palette='Dark2')
g.despine(top=True, right=True, bottom=True, left=True)
plt.tight_layout()
plt.show()
```

```
Processing train_single_df on 'newly created columns'…
Processing train_multi_df on 'newly created columns'…
Processing trial_val_single_df on 'newly created columns'…
Processing trial_val_multi_df on 'newly created columns'…
Processing test_single_df on 'newly created columns'…
Processing test_multi_df on 'newly created columns'…
Analysis (NO CONTRACTIONS) saved to: /content/drive/MyDrive/266-
final/results/sentence_span_analysis_no_contractions.csv
```

```
/usr/local/lib/python3.11/dist-packages/seaborn/axisgrid.py:718: UserWarning:
Using the violinplot function without specifying `order` is likely to produce an
incorrect plot.
  warnings.warn(warning)
/usr/local/lib/python3.11/dist-packages/seaborn/axisgrid.py:854: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in
v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same
effect.

  func(*plot_args, **plot_kwargs)
/usr/local/lib/python3.11/dist-packages/seaborn/axisgrid.py:854: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in
v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same
effect.

  func(*plot_args, **plot_kwargs)
/usr/local/lib/python3.11/dist-packages/seaborn/axisgrid.py:854: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in
v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same
effect.

  func(*plot_args, **plot_kwargs)
```

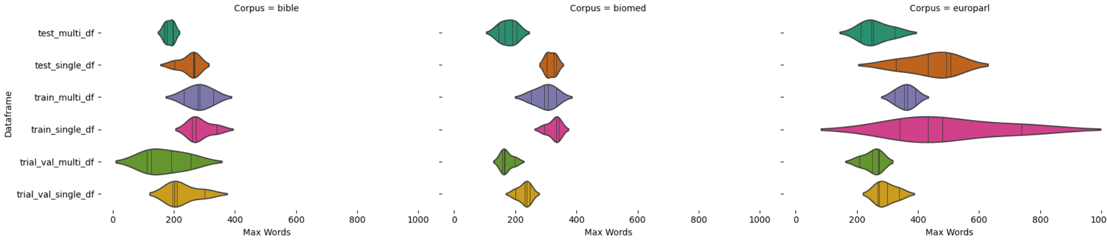


[ ]:

```
tokenizer = RegexpTokenizer(r'\w+')

def analyze_sentence_spans_by_corpus_and_quartile_no_contracts(dfs_dict):
    results = []

    for df_name, df in dfs_dict.items():
        print(f"Processing {df_name} on 'newly created columns'...")
        df = df.copy()
        q1 = df['complexity'].quantile(0.25)
        q2 = df['complexity'].quantile(0.50)
        q3 = df['complexity'].quantile(0.75)

        def get_quartile(x):
            if x <= q1:
                return 'Q1'
            elif x <= q2:
                return 'Q2'
            elif x <= q3:
                return 'Q3'
            else:
                return 'Q4'

        df['quartile'] = df['complexity'].apply(get_quartile)

        def compute_span_metrics_no_contracts(sentence):
            if pd.isna(sentence):
                return pd.Series({'word_count': 0, 'char_count': 0,␣
↪'avg_word_len': 0})

            words = tokenizer.tokenize(sentence)
            word_count = len(words)
            char_count = len(sentence)
            avg_word_len = np.mean([len(w) for w in words]) if word_count > 0␣
↪else 0

            return pd.Series({
                'word_count': word_count,
                'char_count': char_count,
                'avg_word_len': avg_word_len
            })

        span_metrics_nc = df['snc_morph_seq'].
↪apply(compute_span_metrics_no_contracts)
        df = pd.concat([df, span_metrics_nc], axis=1)

        corpus_col = 'corpus'
        for corpus_name, corpus_df in df.groupby(corpus_col):
            for quartile, quartile_df in corpus_df.groupby('quartile'):
                complexity_range = f"{quartile_df['complexity'].min():.
↪3f}-{quartile_df['complexity'].max():.3f}"
                stats = {
                    'Dataframe': df_name,
                    'Corpus': corpus_name,
                    'Quartile': quartile,
                    'Complexity Range': complexity_range,
                    'Count': len(quartile_df),
                    'Avg Words': quartile_df['word_count'].mean(),
                    'Median Words': quartile_df['word_count'].median(),
                    'Min Words': quartile_df['word_count'].min(),
                    'Max Words': quartile_df['word_count'].max(),
                    'Std Words': quartile_df['word_count'].std(),
                    'Avg Chars': quartile_df['char_count'].mean(),
                    'Avg Word Len': quartile_df['avg_word_len'].mean()
                }
                results.append(stats)

    results_df = pd.DataFrame(results)
    results_df = results_df.sort_values(['Dataframe', 'Corpus', 'Quartile'])
    return results_df


dfs = {
    'train_single_df': train_single_df,
    'train_multi_df': train_multi_df,
    'trial_val_single_df': trial_val_single_df,
    'trial_val_multi_df': trial_val_multi_df,
    'test_single_df': test_single_df,
    'test_multi_df': test_multi_df
}

span_analysis_nc =␣
 ↪analyze_sentence_spans_by_corpus_and_quartile_no_contracts(dfs)

pd.set_option('display.max_rows', None)
pd.set_option('display.max_columns', None)
pd.set_option('display.width', 1000)
# display(span_analysis_nc)

results_path_nc = os.path.join(dir_results,␣
 ↪'sentence_span_analysis_no_contractions.csv')
span_analysis_nc.to_csv(results_path_nc, index=False)
print(f"Analysis (NO CONTRACTIONS) saved to: {results_path_nc}")

g = sns.FacetGrid(span_analysis_nc, col="Corpus", col_wrap=3, height=4,␣
 ↪aspect=1.5)
g.map(sns.violinplot, "Max Words", "Dataframe", inner='stick', palette='Dark2')
g.despine(top=True, right=True, bottom=True, left=True)
plt.tight_layout()
plt.show()
```

```
Processing train_single_df on 'newly created columns'…
Processing train_multi_df on 'newly created columns'…
Processing trial_val_single_df on 'newly created columns'…
Processing trial_val_multi_df on 'newly created columns'…
Processing test_single_df on 'newly created columns'…
Processing test_multi_df on 'newly created columns'…
Analysis (NO CONTRACTIONS) saved to: /content/drive/MyDrive/266-
final/results/sentence_span_analysis_no_contractions.csv
```

```
/usr/local/lib/python3.11/dist-packages/seaborn/axisgrid.py:718: UserWarning:
Using the violinplot function without specifying `order` is likely to produce an
incorrect plot.
  warnings.warn(warning)
/usr/local/lib/python3.11/dist-packages/seaborn/axisgrid.py:854: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in
v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same
effect.

  func(*plot_args, **plot_kwargs)
/usr/local/lib/python3.11/dist-packages/seaborn/axisgrid.py:854: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in
v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same
effect.

  func(*plot_args, **plot_kwargs)
/usr/local/lib/python3.11/dist-packages/seaborn/axisgrid.py:854: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in
v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same
effect.

  func(*plot_args, **plot_kwargs)
```



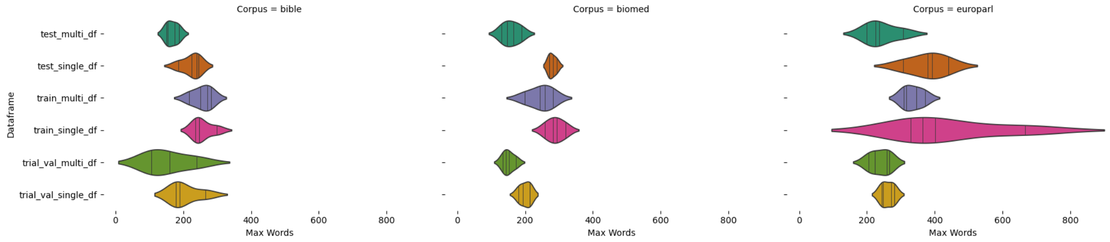
[ ]:
```
tokenizer = RegexpTokenizer(r'\w+')

def analyze_sentence_spans_by_corpus_and_quartile_no_contracts(dfs_dict):
    results = []
    for df_name, df in dfs_dict.items():
        print(f"Processing {df_name} on 'newly created columns'...")
        df = df.copy()

        q1 = df['complexity'].quantile(0.25)
        q2 = df['complexity'].quantile(0.50)
        q3 = df['complexity'].quantile(0.75)

        def get_quartile(x):
            if x <= q1:
                return 'Q1'
            elif x <= q2:
                return 'Q2'
            elif x <= q3:
                return 'Q3'
            else:
                return 'Q4'

        df['quartile'] = df['complexity'].apply(get_quartile)

        def compute_span_metrics_no_contracts(sentence):
            if pd.isna(sentence):
                return pd.Series({'word_count': 0, 'char_count': 0,␣
↪'avg_word_len': 0})

            words = tokenizer.tokenize(sentence)
            word_count = len(words)
            char_count = len(sentence)
            avg_word_len = np.mean([len(w) for w in words]) if word_count > 0␣
↪else 0

            return pd.Series({
                'word_count': word_count,
                'char_count': char_count,
                'avg_word_len': avg_word_len
            })

        span_metrics_nc = df['snc_morph_alt'].
↪apply(compute_span_metrics_no_contracts)
        df = pd.concat([df, span_metrics_nc], axis=1)

        corpus_col = 'corpus'
        for corpus_name, corpus_df in df.groupby(corpus_col):
            for quartile, quartile_df in corpus_df.groupby('quartile'):
                complexity_range = f"{quartile_df['complexity'].min():.
↪3f}-{quartile_df['complexity'].max():.3f}"
```

```
                stats = {
                    'Dataframe': df_name,
                    'Corpus': corpus_name,
                    'Quartile': quartile,
                    'Complexity Range': complexity_range,
                    'Count': len(quartile_df),
                    'Avg Words': quartile_df['word_count'].mean(),
                    'Median Words': quartile_df['word_count'].median(),
                    'Min Words': quartile_df['word_count'].min(),
                    'Max Words': quartile_df['word_count'].max(),
                    'Std Words': quartile_df['word_count'].std(),
                    'Avg Chars': quartile_df['char_count'].mean(),
                    'Avg Word Len': quartile_df['avg_word_len'].mean()
                }
                results.append(stats)

    results_df = pd.DataFrame(results)
    results_df = results_df.sort_values(['Dataframe', 'Corpus', 'Quartile'])
    return results_df


dfs = {
    'train_single_df': train_single_df,
    'train_multi_df': train_multi_df,
    'trial_val_single_df': trial_val_single_df,
    'trial_val_multi_df': trial_val_multi_df,
    'test_single_df': test_single_df,
    'test_multi_df': test_multi_df
}

span_analysis_nc =␣
 ↪analyze_sentence_spans_by_corpus_and_quartile_no_contracts(dfs)

pd.set_option('display.max_rows', None)
pd.set_option('display.max_columns', None)
pd.set_option('display.width', 1000)
# display(span_analysis_nc)

results_path_nc = os.path.join(dir_results,␣
 ↪'sentence_span_analysis_no_contractions.csv')
span_analysis_nc.to_csv(results_path_nc, index=False)
print(f"Analysis (NO CONTRACTIONS) saved to: {results_path_nc}")

g = sns.FacetGrid(span_analysis_nc, col="Corpus", col_wrap=3, height=4,␣
 ↪aspect=1.5)
g.map(sns.violinplot, "Max Words", "Dataframe", inner='stick', palette='Dark2')
g.despine(top=True, right=True, bottom=True, left=True)
```

```python
plt.tight_layout()
plt.show()
```

```
Processing train_single_df on 'newly created columns'…
Processing train_multi_df on 'newly created columns'…
Processing trial_val_single_df on 'newly created columns'…
Processing trial_val_multi_df on 'newly created columns'…
Processing test_single_df on 'newly created columns'…
Processing test_multi_df on 'newly created columns'…
Analysis (NO CONTRACTIONS) saved to: /content/drive/MyDrive/266-
final/results/sentence_span_analysis_no_contractions.csv
```

```
/usr/local/lib/python3.11/dist-packages/seaborn/axisgrid.py:718: UserWarning:
Using the violinplot function without specifying `order` is likely to produce an
incorrect plot.
  warnings.warn(warning)
/usr/local/lib/python3.11/dist-packages/seaborn/axisgrid.py:854: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in
v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same
effect.

  func(*plot_args, **plot_kwargs)
/usr/local/lib/python3.11/dist-packages/seaborn/axisgrid.py:854: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in
v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same
effect.

  func(*plot_args, **plot_kwargs)
/usr/local/lib/python3.11/dist-packages/seaborn/axisgrid.py:854: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in
v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same
effect.

  func(*plot_args, **plot_kwargs)
```



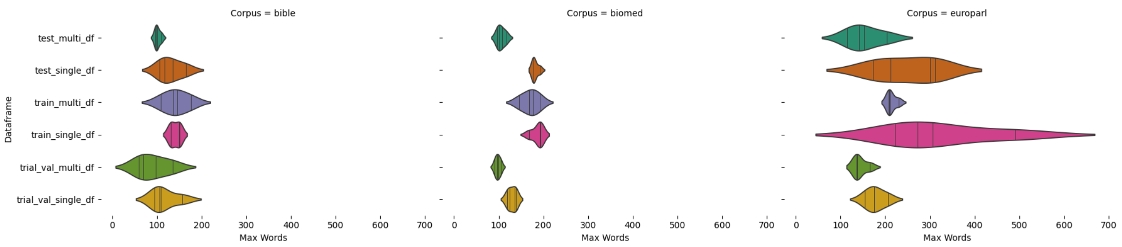
[ ]:

```
tokenizer = RegexpTokenizer(r'\w+')

def analyze_sentence_spans_by_corpus_and_quartile_no_contracts(dfs_dict):
    results = []

    for df_name, df in dfs_dict.items():
        print(f"Processing {df_name} on 'newly created columns'...")
        df = df.copy()

        q1 = df['complexity'].quantile(0.25)
        q2 = df['complexity'].quantile(0.50)
        q3 = df['complexity'].quantile(0.75)

        def get_quartile(x):
            if x <= q1:
                return 'Q1'
            elif x <= q2:
                return 'Q2'
            elif x <= q3:
                return 'Q3'
            else:
                return 'Q4'

        df['quartile'] = df['complexity'].apply(get_quartile)

        def compute_span_metrics_no_contracts(sentence):
            if pd.isna(sentence):
                return pd.Series({'word_count': 0, 'char_count': 0,␣
 ↪'avg_word_len': 0})

            words = tokenizer.tokenize(sentence)
            word_count = len(words)
            char_count = len(sentence)
            avg_word_len = np.mean([len(w) for w in words]) if word_count > 0␣
 ↪else 0

            return pd.Series({
                'word_count': word_count,
                'char_count': char_count,
                'avg_word_len': avg_word_len
            })

        span_metrics_nc = df['snc_dep_seq'].
 ↪apply(compute_span_metrics_no_contracts)
        df = pd.concat([df, span_metrics_nc], axis=1)

        corpus_col = 'corpus'
        for corpus_name, corpus_df in df.groupby(corpus_col):
            for quartile, quartile_df in corpus_df.groupby('quartile'):
                complexity_range = f"{quartile_df['complexity'].min():.
 ↪3f}-{quartile_df['complexity'].max():.3f}"
                stats = {
                    'Dataframe': df_name,
                    'Corpus': corpus_name,
                    'Quartile': quartile,
                    'Complexity Range': complexity_range,
                    'Count': len(quartile_df),
                    'Avg Words': quartile_df['word_count'].mean(),
                    'Median Words': quartile_df['word_count'].median(),
                    'Min Words': quartile_df['word_count'].min(),
                    'Max Words': quartile_df['word_count'].max(),
                    'Std Words': quartile_df['word_count'].std(),
                    'Avg Chars': quartile_df['char_count'].mean(),
                    'Avg Word Len': quartile_df['avg_word_len'].mean()
                }
                results.append(stats)

    results_df = pd.DataFrame(results)
    results_df = results_df.sort_values(['Dataframe', 'Corpus', 'Quartile'])
    return results_df


dfs = {
    'train_single_df': train_single_df,
    'train_multi_df': train_multi_df,
    'trial_val_single_df': trial_val_single_df,
    'trial_val_multi_df': trial_val_multi_df,
    'test_single_df': test_single_df,
    'test_multi_df': test_multi_df
}

span_analysis_nc =␣
 ↪analyze_sentence_spans_by_corpus_and_quartile_no_contracts(dfs)

pd.set_option('display.max_rows', None)
pd.set_option('display.max_columns', None)
pd.set_option('display.width', 1000)
# display(span_analysis_nc)

results_path_nc = os.path.join(dir_results,␣
 ↪'sentence_span_analysis_no_contractions.csv')
span_analysis_nc.to_csv(results_path_nc, index=False)
print(f"Analysis (NO CONTRACTIONS) saved to: {results_path_nc}")
```

```python
g = sns.FacetGrid(span_analysis_nc, col="Corpus", col_wrap=3, height=4, aspect=1.5)
g.map(sns.violinplot, "Max Words", "Dataframe", inner='stick', palette='Dark2')
g.despine(top=True, right=True, bottom=True, left=True)
plt.tight_layout()
plt.show()
```

```
Processing train_single_df on 'newly created columns'…
Processing train_multi_df on 'newly created columns'…
Processing trial_val_single_df on 'newly created columns'…
Processing trial_val_multi_df on 'newly created columns'…
Processing test_single_df on 'newly created columns'…
Processing test_multi_df on 'newly created columns'…
Analysis (NO CONTRACTIONS) saved to: /content/drive/MyDrive/266-
final/results/sentence_span_analysis_no_contractions.csv
```

```
/usr/local/lib/python3.11/dist-packages/seaborn/axisgrid.py:718: UserWarning:
Using the violinplot function without specifying `order` is likely to produce an
incorrect plot.
  warnings.warn(warning)
/usr/local/lib/python3.11/dist-packages/seaborn/axisgrid.py:854: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in
v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same
effect.

  func(*plot_args, **plot_kwargs)
/usr/local/lib/python3.11/dist-packages/seaborn/axisgrid.py:854: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in
v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same
effect.

  func(*plot_args, **plot_kwargs)
/usr/local/lib/python3.11/dist-packages/seaborn/axisgrid.py:854: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in
v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same
effect.

  func(*plot_args, **plot_kwargs)
```



[ ]:

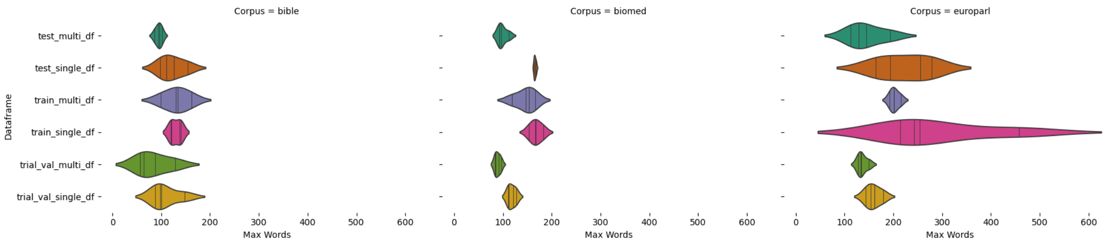
```
tokenizer = RegexpTokenizer(r'\w+')

def analyze_sentence_spans_by_corpus_and_quartile_no_contracts(dfs_dict):
    results = []

    for df_name, df in dfs_dict.items():
        print(f"Processing {df_name} on 'newly created columns'...")
        df = df.copy()

        q1 = df['complexity'].quantile(0.25)
        q2 = df['complexity'].quantile(0.50)
        q3 = df['complexity'].quantile(0.75)

        def get_quartile(x):
            if x <= q1:
                return 'Q1'
            elif x <= q2:
                return 'Q2'
            elif x <= q3:
                return 'Q3'
            else:
                return 'Q4'

        df['quartile'] = df['complexity'].apply(get_quartile)

        def compute_span_metrics_no_contracts(sentence):
            if pd.isna(sentence):
                return pd.Series({'word_count': 0, 'char_count': 0,␣
 ↪'avg_word_len': 0})

            words = tokenizer.tokenize(sentence)
            word_count = len(words)
            char_count = len(sentence)
            avg_word_len = np.mean([len(w) for w in words]) if word_count > 0␣
 ↪else 0

            return pd.Series({
                'word_count': word_count,
                'char_count': char_count,
                'avg_word_len': avg_word_len
            })

        span_metrics_nc = df['snc_dep_alt'].
 ↪apply(compute_span_metrics_no_contracts)
        df = pd.concat([df, span_metrics_nc], axis=1)

        corpus_col = 'corpus'
        for corpus_name, corpus_df in df.groupby(corpus_col):
            for quartile, quartile_df in corpus_df.groupby('quartile'):
                complexity_range = f"{quartile_df['complexity'].min():.
 ↪3f}-{quartile_df['complexity'].max():.3f}"
                stats = {
                    'Dataframe': df_name,
                    'Corpus': corpus_name,
                    'Quartile': quartile,
                    'Complexity Range': complexity_range,
                    'Count': len(quartile_df),
                    'Avg Words': quartile_df['word_count'].mean(),
                    'Median Words': quartile_df['word_count'].median(),
                    'Min Words': quartile_df['word_count'].min(),
                    'Max Words': quartile_df['word_count'].max(),
                    'Std Words': quartile_df['word_count'].std(),
                    'Avg Chars': quartile_df['char_count'].mean(),
                    'Avg Word Len': quartile_df['avg_word_len'].mean()
                }
                results.append(stats)

    results_df = pd.DataFrame(results)
    results_df = results_df.sort_values(['Dataframe', 'Corpus', 'Quartile'])
    return results_df


dfs = {
    'train_single_df': train_single_df,
    'train_multi_df': train_multi_df,
    'trial_val_single_df': trial_val_single_df,
    'trial_val_multi_df': trial_val_multi_df,
    'test_single_df': test_single_df,
    'test_multi_df': test_multi_df
}

span_analysis_nc =␣
 ↪analyze_sentence_spans_by_corpus_and_quartile_no_contracts(dfs)
```

```
pd.set_option('display.max_rows', None)
pd.set_option('display.max_columns', None)
pd.set_option('display.width', 1000)
# display(span_analysis_nc)

results_path_nc = os.path.join(dir_results,␣
 ↪'sentence_span_analysis_no_contractions.csv')
span_analysis_nc.to_csv(results_path_nc, index=False)
print(f"Analysis (NO CONTRACTIONS) saved to: {results_path_nc}")

g = sns.FacetGrid(span_analysis_nc, col="Corpus", col_wrap=3, height=4,␣
 ↪aspect=1.5)
g.map(sns.violinplot, "Max Words", "Dataframe", inner='stick', palette='Dark2')
g.despine(top=True, right=True, bottom=True, left=True)
plt.tight_layout()
plt.show()
```

```
Processing train_single_df on 'newly created columns'…
Processing train_multi_df on 'newly created columns'…
Processing trial_val_single_df on 'newly created columns'…
Processing trial_val_multi_df on 'newly created columns'…
Processing test_single_df on 'newly created columns'…
Processing test_multi_df on 'newly created columns'…
Analysis (NO CONTRACTIONS) saved to: /content/drive/MyDrive/266-
final/results/sentence_span_analysis_no_contractions.csv
```

```
/usr/local/lib/python3.11/dist-packages/seaborn/axisgrid.py:718: UserWarning:
Using the violinplot function without specifying `order` is likely to produce an
incorrect plot.
  warnings.warn(warning)
/usr/local/lib/python3.11/dist-packages/seaborn/axisgrid.py:854: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in
v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same
effect.

  func(*plot_args, **plot_kwargs)
/usr/local/lib/python3.11/dist-packages/seaborn/axisgrid.py:854: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in
v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same
effect.

  func(*plot_args, **plot_kwargs)
/usr/local/lib/python3.11/dist-packages/seaborn/axisgrid.py:854: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in
v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same
```

```
effect.

  func(*plot_args, **plot_kwargs)
```



[ ]:

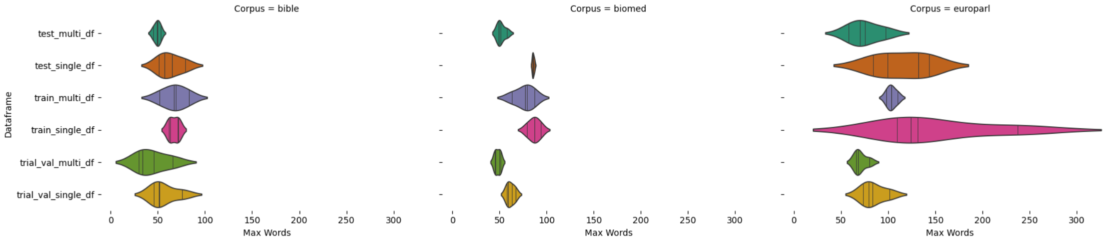
```
tokenizer = RegexpTokenizer(r'\w+')

def analyze_sentence_spans_by_corpus_and_quartile_no_contracts(dfs_dict):
    results = []

    for df_name, df in dfs_dict.items():
        print(f"Processing {df_name} on 'newly created columns'...")
        df = df.copy()

        q1 = df['complexity'].quantile(0.25)
        q2 = df['complexity'].quantile(0.50)
        q3 = df['complexity'].quantile(0.75)

        def get_quartile(x):
            if x <= q1:
                return 'Q1'
            elif x <= q2:
                return 'Q2'
            elif x <= q3:
                return 'Q3'
            else:
                return 'Q4'

        df['quartile'] = df['complexity'].apply(get_quartile)

        def compute_span_metrics_no_contracts(sentence):
            if pd.isna(sentence):
                return pd.Series({'word_count': 0, 'char_count': 0,␣
↪'avg_word_len': 0})

            words = tokenizer.tokenize(sentence)
            word_count = len(words)
            char_count = len(sentence)
            avg_word_len = np.mean([len(w) for w in words]) if word_count > 0␣
 ↪else 0

            return pd.Series({
                'word_count': word_count,
                'char_count': char_count,
                'avg_word_len': avg_word_len
            })

        span_metrics_nc = df['snc_morph_complexity_value'].
 ↪apply(compute_span_metrics_no_contracts)
        df = pd.concat([df, span_metrics_nc], axis=1)

        corpus_col = 'corpus'
        for corpus_name, corpus_df in df.groupby(corpus_col):
            for quartile, quartile_df in corpus_df.groupby('quartile'):
                complexity_range = f"{quartile_df['complexity'].min():.
 ↪3f}-{quartile_df['complexity'].max():.3f}"
                stats = {
                    'Dataframe': df_name,
                    'Corpus': corpus_name,
                    'Quartile': quartile,
                    'Complexity Range': complexity_range,
                    'Count': len(quartile_df),
                    'Avg Words': quartile_df['word_count'].mean(),
                    'Median Words': quartile_df['word_count'].median(),
                    'Min Words': quartile_df['word_count'].min(),
                    'Max Words': quartile_df['word_count'].max(),
                    'Std Words': quartile_df['word_count'].std(),
                    'Avg Chars': quartile_df['char_count'].mean(),
                    'Avg Word Len': quartile_df['avg_word_len'].mean()
                }
                results.append(stats)

    results_df = pd.DataFrame(results)
    results_df = results_df.sort_values(['Dataframe', 'Corpus', 'Quartile'])
    return results_df


dfs = {
    'train_single_df': train_single_df,
    'train_multi_df': train_multi_df,
    'trial_val_single_df': trial_val_single_df,
    'trial_val_multi_df': trial_val_multi_df,
    'test_single_df': test_single_df,
    'test_multi_df': test_multi_df
}
```

```
span_analysis_nc =␣
  ↪analyze_sentence_spans_by_corpus_and_quartile_no_contracts(dfs)

pd.set_option('display.max_rows', None)
pd.set_option('display.max_columns', None)
pd.set_option('display.width', 1000)
# display(span_analysis_nc)

results_path_nc = os.path.join(dir_results,␣
  ↪'sentence_span_analysis_no_contractions.csv')
span_analysis_nc.to_csv(results_path_nc, index=False)
print(f"Analysis (NO CONTRACTIONS) saved to: {results_path_nc}")

g = sns.FacetGrid(span_analysis_nc, col="Corpus", col_wrap=3, height=4,␣
  ↪aspect=1.5)
g.map(sns.violinplot, "Max Words", "Dataframe", inner='stick', palette='Dark2')
g.despine(top=True, right=True, bottom=True, left=True)
plt.tight_layout()
plt.show()
```

```
Processing train_single_df on 'newly created columns'…
Processing train_multi_df on 'newly created columns'…
Processing trial_val_single_df on 'newly created columns'…
Processing trial_val_multi_df on 'newly created columns'…
Processing test_single_df on 'newly created columns'…
Processing test_multi_df on 'newly created columns'…
Analysis (NO CONTRACTIONS) saved to: /content/drive/MyDrive/266-
final/results/sentence_span_analysis_no_contractions.csv
```

```
/usr/local/lib/python3.11/dist-packages/seaborn/axisgrid.py:718: UserWarning:
Using the violinplot function without specifying `order` is likely to produce an
incorrect plot.
  warnings.warn(warning)
/usr/local/lib/python3.11/dist-packages/seaborn/axisgrid.py:854: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in
v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same
effect.

  func(*plot_args, **plot_kwargs)
/usr/local/lib/python3.11/dist-packages/seaborn/axisgrid.py:854: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in
v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same
effect.

  func(*plot_args, **plot_kwargs)
```

```
/usr/local/lib/python3.11/dist-packages/seaborn/axisgrid.py:854: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in
v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same
effect.

  func(*plot_args, **plot_kwargs)
```



0.5.2 Save Dataframes as CSVs

[ ]:
```
### Save Dataframes as CSVs
```

[ ]:
```
!tree /content/drive/MyDrive/266-final/data/266-comp-lex-master/
```

```
/content/drive/MyDrive/266-final/data/266-comp-lex-master/
  fe-test-labels
     test_multi_df.csv
     test_single_df.csv
  fe-train
     train_multi_df.csv
     train_single_df.csv
  fe-trial-val
     trial_val_multi_df.csv
     trial_val_single_df.csv
  test-labels
     lcp_multi_test.tsv
     lcp_single_test.tsv
  train
     lcp_multi_train.tsv
     lcp_single_train.tsv
  trial
      lcp_multi_trial.tsv
      lcp_single_trial.tsv

6 directories, 12 files
```

[ ]:
```
import os
dataframes = {
    "train_single_df": train_single_df,
    "train_multi_df": train_multi_df,
    "trial_val_single_df": trial_val_single_df,
    "trial_val_multi_df": trial_val_multi_df,
    "test_single_df": test_single_df,
    "test_multi_df": test_multi_df
}

base_dir = "/content/drive/MyDrive/266-final/data/266-comp-lex-master/"

for df_name, df in dataframes.items():
    subdir = None
    if "train" in df_name:
      subdir = "fe-train"
    elif "trial_val" in df_name:
      subdir = "fe-trial-val"
    elif "test" in df_name:
      subdir = "fe-test-labels"

    if subdir:
      save_path = os.path.join(base_dir, subdir, f"{df_name}.csv")
      os.makedirs(os.path.dirname(save_path), exist_ok=True)
      df.to_csv(save_path, index=False)
      print(f"Saved {df_name} to {save_path}")
```

```
Saved train_single_df to /content/drive/MyDrive/266-final/data/266-comp-lex-
master/fe-train/train_single_df.csv
Saved train_multi_df to /content/drive/MyDrive/266-final/data/266-comp-lex-
master/fe-train/train_multi_df.csv
Saved trial_val_single_df to /content/drive/MyDrive/266-final/data/266-comp-lex-
master/fe-trial-val/trial_val_single_df.csv
Saved trial_val_multi_df to /content/drive/MyDrive/266-final/data/266-comp-lex-
master/fe-trial-val/trial_val_multi_df.csv
Saved test_single_df to /content/drive/MyDrive/266-final/data/266-comp-lex-
master/fe-test-labels/test_single_df.csv
Saved test_multi_df to /content/drive/MyDrive/266-final/data/266-comp-lex-
master/fe-test-labels/test_multi_df.csv
```

[ ]:

```
df_names = [
    "train_single_df",
    "train_multi_df",
    "trial_val_single_df",
    "trial_val_multi_df",
    "test_single_df",
    "test_multi_df"
]
```

```python
loaded_dataframes = {}

for df_name in df_names:
    if "train" in df_name:
        subdir = "fe-train"
    elif "trial_val" in df_name:
        subdir = "fe-trial-val"
    elif "test" in df_name:
        subdir = "fe-test-labels"
    else:
        subdir = None

    if subdir:
        read_path = os.path.join(dir_data, subdir, f"{df_name}.csv")
        loaded_df = pd.read_csv(read_path)
        loaded_dataframes[df_name] = loaded_df
        print(f"Loaded {df_name} from {read_path}")

for df_name, df in loaded_dataframes.items():
    print(f"\n>>> {df_name} shape: {df.shape}")
    if 'binary_complexity' in df.columns:
        print(df['binary_complexity'].value_counts())
```

```
Loaded train_single_df from /content/drive/MyDrive/266-final/data/266-comp-lex-master/fe-train/train_single_df.csv
Loaded train_multi_df from /content/drive/MyDrive/266-final/data/266-comp-lex-master/fe-train/train_multi_df.csv
Loaded trial_val_single_df from /content/drive/MyDrive/266-final/data/266-comp-lex-master/fe-trial-val/trial_val_single_df.csv
Loaded trial_val_multi_df from /content/drive/MyDrive/266-final/data/266-comp-lex-master/fe-trial-val/trial_val_multi_df.csv
Loaded test_single_df from /content/drive/MyDrive/266-final/data/266-comp-lex-master/fe-test-labels/test_single_df.csv
Loaded test_multi_df from /content/drive/MyDrive/266-final/data/266-comp-lex-master/fe-test-labels/test_multi_df.csv

>>> train_single_df shape: (7662, 20)
binary_complexity
0    3865
1    3797
Name: count, dtype: int64

>>> train_multi_df shape: (1517, 20)
binary_complexity
0    759
1    758
Name: count, dtype: int64
```

```
>>> trial_val_single_df shape: (421, 20)
binary_complexity
0    229
1    192
Name: count, dtype: int64

>>> trial_val_multi_df shape: (99, 20)
binary_complexity
1    51
0    48
Name: count, dtype: int64

>>> test_single_df shape: (917, 20)
binary_complexity
0    476
1    441
Name: count, dtype: int64

>>> test_multi_df shape: (184, 20)
binary_complexity
1    99
0    85
Name: count, dtype: int64
```

```
[ ]: !tree /content/drive/MyDrive/266-final/data/266-comp-lex-master/
```

```
/content/drive/MyDrive/266-final/data/266-comp-lex-master/
  fe-test-labels
     test_multi_df.csv
     test_single_df.csv
  fe-train
     train_multi_df.csv
     train_single_df.csv
  fe-trial-val
     trial_val_multi_df.csv
     trial_val_single_df.csv
  test-labels
     lcp_multi_test.tsv
     lcp_single_test.tsv
  train
     lcp_multi_train.tsv
     lcp_single_train.tsv
  trial
      lcp_multi_trial.tsv
      lcp_single_trial.tsv

6 directories, 12 files
```

- These counts match my offline calculations exactly.The binarized outcome variables have been split on on the median of the TRAIN_SINGLE and TRAIN_MULTI dataset splits ONLY, thus this median is applied to trial_val and test. The first two quartiles (up to the train median) are equal to 0 in ‘binary_complexity’ and the next two quartiles are equal to 1.

- Because the dataset has been excellently balanced by the Task’s annotators, we’re lucky that no further data processing is necessary prior to moving onto the modeling step, and ensuring protection from data leakage by (later) removing necessary columns prior to vectorization.
- Lastly, a note on the balanced nature of the data. It should be noted that (even in the continuous outome representation of ‘complexity’) the medians were 0.28 in train_single, and 0.27 in both trial_single and test_single—for multi, it was 0.41 in train_multi, and 0.42 in trial_multi and 0.43 in test_multi.
- We also find that after Data Engineering, our sanity checks have come out successfully. No records have been lost, shapes are consistent with our expectations, and we have enriched the dataset with SpaCy-derived features to give us flexibility in multi-channel inputs or vectorization ablations. This is a very thorough dataset, and we are now ready for modeling.

[ ]: